Journal of
Data Acquisition and Processing

ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.777708

# MODIFIED LATENT CONTENT APPROACH TO TEXT SENTIMENT ANALYSIS

**Mr. Dhirendra Negi**
Enrollment No – 200155132949

**Dr. Ajay Agarwal**
(Guide) Department of Computer Science, Singhania University, Pacheri Beri, Jhunjhunu (Rajasthan)

**1. ABSTRACT:** Sentiment categorization has been used in a variety of contexts, including reviews of movies and products as well as analyses of customer feedback. The foundation of sentiment analysis is the ability to label a piece of text as positive or negative. Emotions are hard to categorise since the meanings of words and phrases shift with context. As innovative business intelligence options, data mining and machine learning may facilitate the real-time processing of massive volumes of internet data. Another new innovation in online text mining is sentiment analysis, which measures the positive or negative emotional tone of written content. Sentiment analysis is a method for determining an individual's attitude towards a subject. Finding of the study shows that Text Mining Methodology for Business Intelligence is discussed in this chapter, which covers one of the submodules of the hybrid approach (TMABI). In this chapter, the emphasis was placed on the categorization of sentiments using a modified version of the LSA methodology, which, in comparison to other methods already in use, produces more effective results. In the next chapter, the whole operation of the hybrid system, which is an innovative method for gathering business information using text mining, will be presented.

**KEYWORDS:** Modified Latent Content, Sentiment Analysis, Machine learning, etc.

## 2. INTRODUCTION

Movie reviews, product evaluations, and reviews of consumer comments are just some of the areas where sentiment classification has been put to use. Classifying a piece of text as good or negative is the groundwork of sentiment analysis. Emotion classification is difficult since the meanings of words and phrases change depending on the circumstances. For the dynamic processing of vast amounts of internet data, data mining and machine learning provide useful alternatives as cutting-edge business intelligence solutions. Sentiment analysis, which determines the emotional polarity of text, is another recent development in the field of text mining online. Finding out how someone feels about a topic is what sentiment analysis is all about. The authors attitude might be any evaluation or assessment of the work, as well as the authors emotional state. It is well understood that domains or themes influence the effectiveness of sentiment classifiers (Rintyarna et al.,2018).

The goal of Latent Semantic Analysis (LSA), a Natural Language Processing (NLP) method for semantic analysis, is to use statistical methods to glean meaning from a collection of texts.

LSA is a method for retrieving data that searches for and finds connections between pieces of unstructured text.

**Polarity Classification:** The polarity categorization is seen as a two-way decision tree. The task at hand is to ascertain if the overall tone of a given text (such as a customer review or editorial remark) is positive or negative. Evidently, the definition of the extremes of emotion is a vital aspect. Im confused about the difference between a good and bad opinion. This question has no simple solution. Its important to keep in mind that the context of an actual use is crucial to any given definition, and that variations might be negligible. In the context of political arguments, for instance, positive may denote agreement while negative may indicate disagreement. Classification of customer reviews often takes the texts evaluative tone into account. While reviewing a product, does the reviewer recommend it or not? When tackling the sentiment polarity in a computational setting, considering it as a classification job, a clear description is more important than before. (Pang et a1., 2002) looked at methods of identifying polarity in product reviews and film reviews. Opinions stated by a writer towards a target may be categorised in a variety of ways, including valence (positive, negative, or neutral), discrete measurement (excellent, good, satisfactory, bad, or extremely poor), and a variety of emotions (joy, sorrow, rage, surprise, contempt, or fear). As used here, the term sentiment analysis process refers to a set of activities, each of which expresses a certain feeling (Nigam et a1.,2006).

**Subjectivity Classification:** Classifying subjective data is often seen as a simple yes/no task. Its goal is to separate personal opinions from hard facts. Like before, there are several levels of detail that may be applied to this issue. The purpose at the document level may be to differentiate between review-like papers and other documents, or between news pieces and editorial comments in a newspaper (Riloff et a1.,2003). A major component of sentiment retrieval is the categorization of subjective content. Over the course of several years, many researchers have uncovered many machine learning algorithms for subjective annotation jobs. The subjectivity detection job has made use of a wide variety of sources, including dependency parsing, named entity identification, morphological analyzers, stemmer, SentiWordNet, and WordNet. Initiated to create a subjectivity classifier that can function on unannotated text sources (Wiebe et al., 2005). The goal is to develop an automated procedure that can learn language-rich extraction patterns for subjective expressions and provide rich ontological language-specific (rather than domain dependent) knowledge.

**Emotion Classification:** An improvement on the sentiment polarity classification task might be the ability to identify emotional emotions in written language. The goal here is to assign a piece of text to one of many emotional categories. Emotion classification seeks to discover nuanced differences in expression of emotion, as opposed to the binary polarity commonly attributed to positive and negative sentiments. Anger, contempt, fear, pleasure, grief, and surprise are the six basic emotions commonly used as class names for this endeavour. Class labels may be drawn from psychological theories of emotion, or they can be developed ad hoc based on the needs of a specific application. There are several domains in which emotion classification might be useful, such as market research, public opinion polling, and medical records.

**Latent Semantic Analysis:** To find antonyms in a corpus of texts, Latent Semantic Analysis (LSA) may be used as an unsupervised technique. The method uses statistical computations on a vast corpus of text to extract and capture the meaning of words in context. LSA is a method for retrieving data that searches for and finds connections between pieces of unstructured text. An idea set is generated that draws connections between the texts and the concepts they include, which is an approach to natural language processing. According to the LSA, synonyms will be found in adjacent text segments (the distributional hypothesis). LSA has been used extensively in the past to classify texts. LSA uses singular value decomposition to break down a large term-document matrix into a collection of k orthogonal components (SVD). Its a machine-learned strategy for compressing the semantic space occupied by a body of text by use of the inherent higher-order structure of objects representing connections between words. Concepts with a semantic relationship are created as indexing dimensions, and thus has the potential to greatly decrease dimensionality. By the use of LSA, we transform the initial feature space into a new, lower-dimensional space that is more suited for the semantic understanding of emotional valence. A support vector machine (SVM) is then used to classify the texts polarity. Under the standards of Latent Semantic Analysis (LSA), two documents that share a large number of words are deemed to be conceptually near, while those that share fewer words are seen as conceptually distant. After constructing a matrix of terms and documents, LSA calls for a solitary value. In order to express conceptual term-document relationships, this matrix is divided into a semantic vector space.

## 3. LITERATURE REVIEW

**Kan, et al (2019)** In this research work, the authors investigate a closest neighbour approach for predicting the number of calls that will come into call centres. The approach does not need an embraces for the throughput, and it is able to be used to past information instead of requiring the data to all be preprocessed. The researchers demonstrate that this type of procedures produces a more accurate prediction than the traditional technique, which consists only of calculating averages of the data. The closest neighbour approach that uses the Correlation segmentation method is also prepared to take to consideration ratio did, which may often be observed in the data collected from contact centres. Testing with numbers have shown that using this approach results in fewer mistakes in the predictions and improved staffing levels in contact centres. The findings may be put to use in call centres to provide more adaptable management of the available labour.

**Mandal et al (2018)** This article provides an overview of the primary forecasting methods that are used for pv plants. The many methods of forecasting that are available have just been explored, with a particular emphasis placed on the prediction of wind power and the load and price of energy. The difficulties associated with forecasting have really been broken down into several categories according to time period, application particular region, and forecasting method. The activity of the formed cnn architecture (NN) method that uses a similar month approach to forecast hourly power grid and profit margin, respectively, is demonstrated by using textual evidence from data referring to the Tudor electricity real economy in The region and the PJM wind industry in the United States of America. Both of these markets are used to generate electricity. The inconsistency and unpredictability of pv and wind output poses a significant challenge for power system providers, both in terms of their capacity to effectively plan and manage their operations and in terms of the quality of service they can provide to their customers. This research demonstrates the use of an Adjustable Neural Fuzzy Induction System (ANFIS) to absurdly short winds forecasting employing a study from Tasmania, Australia. The purpose of this work is to solve the challenges that are associated with wind power.

**Ahmed, et al (2018)** Researchers give a large-scale comparative evaluation of the primary machine learning algorithms for time-series prediction in this body of work. To be more exact, researchers utilized the systems to the data from the bimonthly M3 standard statistical championship . The researchers anticipate that this study will fill the void left by the absence of large-scale similar studies on classification techniques applied to challenges involving stagnation or time series prediction. Lstm, Bayesian machine learning, rectified functions, generalised regression deep learning (also known as kernel dependent variable), K-nearest neighbour correlation, CART decision tree, svm regression, and Normal distribution processes are the models that have been taken into consideration. The research demonstrates that there are substantial dissimilarities across the various approaches. The clustering algorithm and the Principle component regression were found to be the two approaches that performed the best overall. In addition to comparing the models, research has been conducted to evaluate various preprocessing approaches, and the results have demonstrated that each of these methods has a unique influence on the results.

**Joseph (2017)** It appears to be a clear association between a nation ’s Gdp and the ever-increasing pervasiveness of globalisation as a result of worldwide commerce, which is mostly conducted on the sea (GDP). Historically, in the academic literature, design parameters have been used to make predictions about the association between GDP and the outsource tonnage on a variety of different sectors. In this study, we investigate how machine teaching and information mining methods might be used to the analysis of publicly accessible information about the imported and exported tonnage of commodity at the naval bases of the country in question. The results of the algorithms that provide real GDP predictions for the fiscal year are then taken into consideration. The information that will be used for the experiment is comprised of the daily imported and exported tonnage now at specified port. Following this, many ports located inside the nation of interest are taken into consideration. The question entails a difficult deep learning difficulties to be assessed, with a properly designed data set, that also is intended to generalise, given that it contains data over many years and an auxiliary GDP predicted on a routine basis. In addition, the question contains data for many years.

**Rodrigues, et al (2013)** The assortment of garments that are available for purchase is typically updated twice a year by the majority of clothing businesses. Each new collection has a substantial number of brand-new pieces, each of which has a limited and clearly outlined selling time that corresponds to one of the collections' sales seasons (20-30 weeks). There is a lack of historical sales data, which, together with changes in both taste and production planning, makes it impossible to correctly estimate customer demand. Therefore, the most important aspects to consider in this setting are how difficult it is to generate reliable estimates of future demand and how short the summer season is for items. Because of this, a system for managing inventory that was developed to function in this setting is restricted by the reality that consumption for many different types of commodities is not likely to continue into the foreseeable future. A novel way of load forecasting that makes use of neural nets is described, and the efficiency of demand predictions derived by classic profile approaches is analysed that use the data from one specific firm. The discussion will focus on some of the most important concerns with the use of neural networks models.

**Khorana et al (2013)** Businesses who have effectively implemented the newest applications of nanotechnology, which is a subfield, have generated millions of dollars in profits as a result. Applications for algorithms are being used extensively by business professionals working in areas such as finance, commerce, and logistics, but even marketing. Research in this area is very important to organisations for a wide variety of reasons, including the development of new products and the restructuring of supply chains, among other things. However, there has not been a sufficient amount of analysis done to determine if the ongoing study is valuable and whether it should continue in the same path. The use of strategies that are continually becoming more complicated has elevated natural language processing ( nlp to the level of a high-tech, increased field of study; yet, reliability and generalizability of the findings have not yet been demonstrated. When researchers concentrate on the beauty of the solution, they end up asking questions that are not as significant as others, despite the fact that there is an unmet need for them to investigate topics that are likely to be more pertinent from a commercial perspective.

## 4. EXPERIMENTAL SETUP

### Mathematical Process of LSA

Consider the case when you have access to a term-document (or term-frequency) matrix. There are m papers (Dl, D2,......... Dm) and n relevant words (W1, W2, W3, W4, W5,......... Wn) have been chosen from these documents. The latent semantic analysis is done once the SVD procedure is completed and the findings are interpreted. Using SVD, we must trisect the provided term-document matrix X into three submatrices.

X S V RTC (4.1)

Where (X, S) is an orthogonal matrix of size (m x n) and (V, X) is a diagonal matrix of size (m x n).

Orthogonal matrices of size n by n are what we call R.

The purpose of singular value decomposition (SVD) is to identify the most important details and convey them in a more compact way.

### Steps for Classification using LSA

First, compile a dictionary of adverbs and negative keywords as well as a glossary of directed words.

After the stoppage has occurred, the second step is to adapt each observation to the vector space model representation. Some of the words have been left out.

Third, youll use SVD to create the latent semantic space for every comment.

The fourth step is to look up each phrase in the dictionary and identify the polarity of the nouns, verbs, and adjectives. The polarity may then be calculated by inspecting the prefix of its modifications.

The fifth step is to calculate the overall density and intensity of the oriented words. thereafter learn the passages orientation grade.

Sixth, group similar passages together.

### Description of Data Set

Web scraping has been used to compile a data set that may be used by a restaurant to examine the quality of its meals by means of sentiment categorization. We also utilised the Amazon customer review data for meals to evaluate the effectiveness of the algorithm. There were a total of 12632 reviews collected for this study. Sites scraped included those at the following URLs:

https://www.tripadvisor.com/Restaurant Review

Following python libraries were used in processing and result analysis:

Pandas — DataFrames and Manipulation

Numpy — Numerical Library

Matplotlib — Visualization

Seaborn — Visualization

Natural Language Toolkit — Library for NLP

Scikit-learn— Processing techniques etc.

## 5. RESULTS ANALYSIS AND DISCUSSION

The modified LSA method was used to classify the sentiment. The analysis is used to determine whether or not the sentence has polarity. Using web scraping, consumer comments or feedbacks on food quality are gathered, and modified latent topic modelling is used to categorise the phrases into distinct subjects. The next stage is to determine whether or not the statement is polarised. Modified LSA is used to accomplish this goal. Sentences are sorted according to whether they are favourable, negative, or neutral. As our goal is to give business information, we have only considered positive and negative remarks for analysis.

The following is a study of some common themes:

Taste: - The polarity of all topic-related sentences has been determined for the taste category. Customers were pleased with the flavour of the dish 84.6% of the time, while 15.4% had complaints. If the companys taste threshold is 15%, then there is room for small improvement.

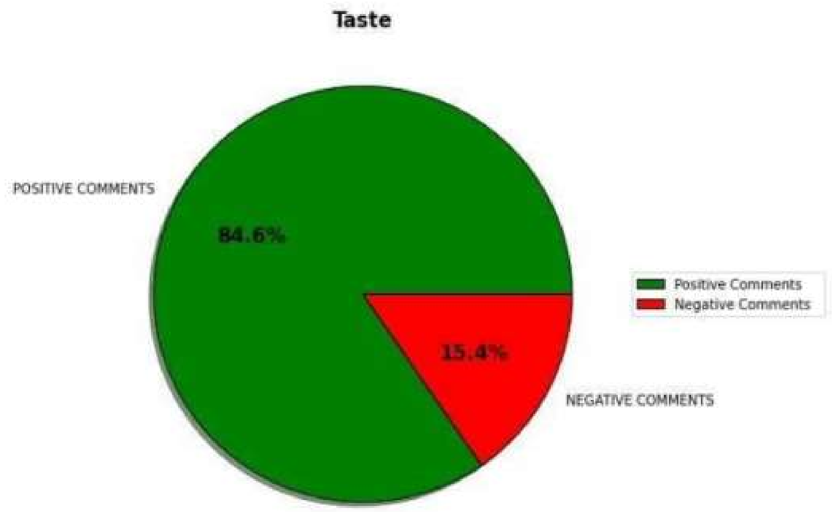

**Figure 5.1: Polarity Analysis of Taste Attribute**

Cost: - Polarity has been determined for all of the sentences that pertain to this issue via analysis, which has been done. 72.8 percent of customers are happy with the price of the meal, whereas 27.2 percent of customers are unhappy with the price. If the corporation has determined that the threshold number is 15%, then the company has to take the right action in this regard.

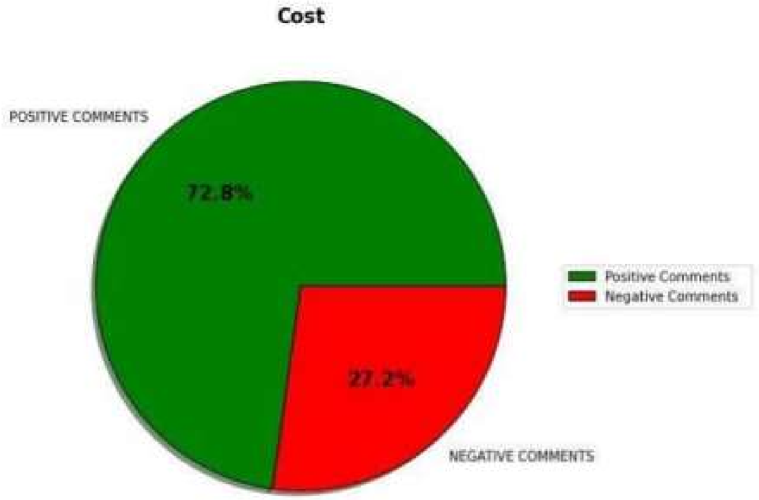

**Figure 5.2: Polarity Analysis of Cost Attribute**

Appearance: - Polarity has been determined for all of the sentences that pertain to this issue via analysis, which has been done. Customers are pleased with the overall appearance of the meal 86.3% of the time, while 13.7% of customers are unhappy with the overall appearance. If the company has established 15% as the threshold value, then this characteristic is considered to be under control.

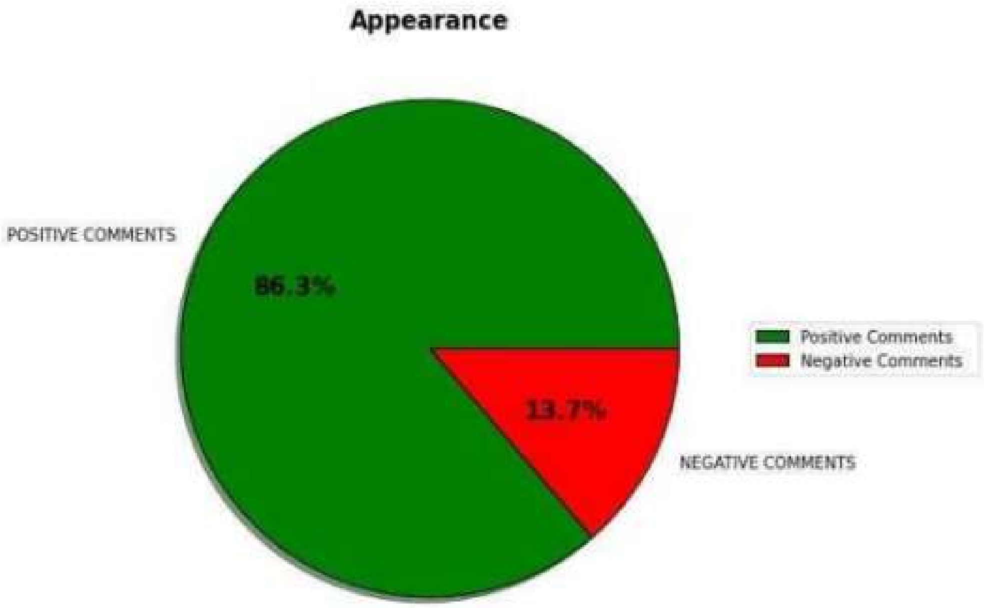

**Figure 5.3: Polarity Analysis of Appearance Attribute**

**Nutrition: -** Polarity has been determined for all of the sentences that pertain to this issue via analysis, which has been done. Customers are pleased with the nutritional content of food 82.7% of the time, whereas 17.3% of customers are unhappy with the nutritional content of the food they purchase. If the organisation has determined that 15% is the appropriate threshold figure, then the nutritional content of the product must be somewhat improved.

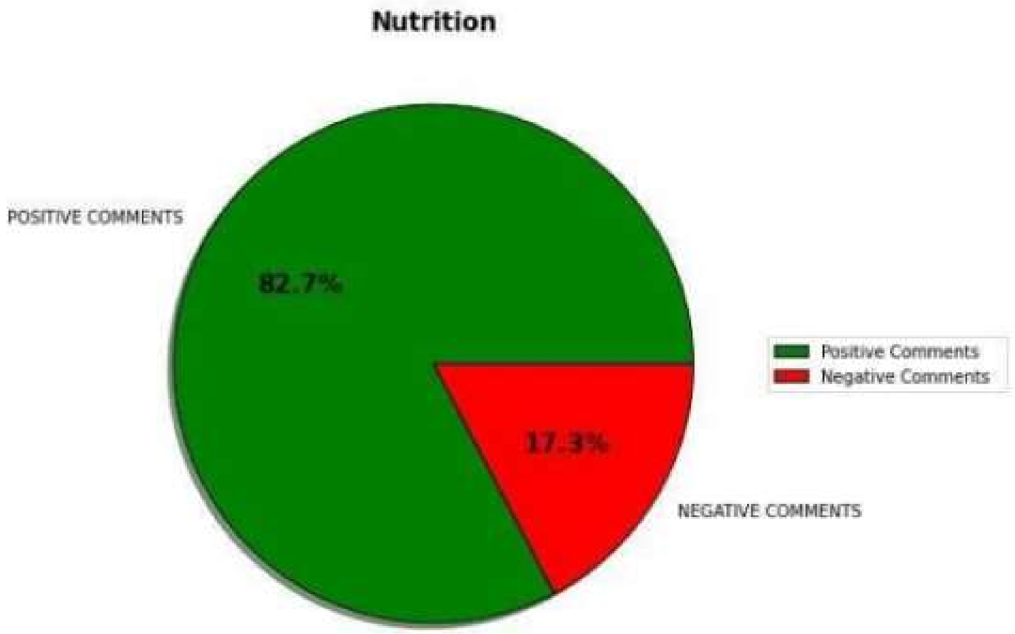

**Figure 5.4: Polarity Analysis of Nutrition Attribute**

**Hygienic: -** Polarity has been determined for all of the sentences that pertain to this issue via analysis, which has been done.

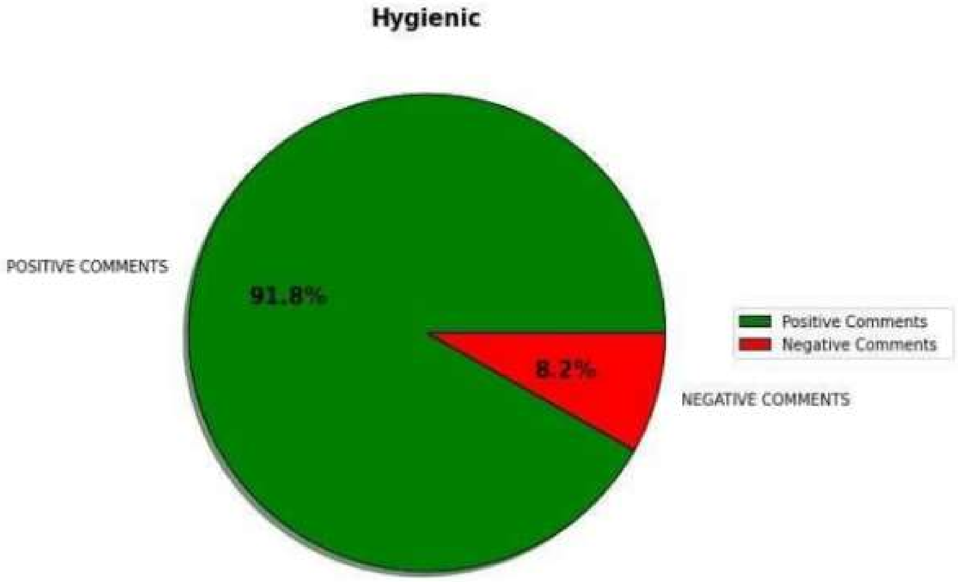

**Figure 5.5: Polarity Analysis of Hygienic Attribute**

Customers are pleased with the sanitary conditions of the meal 91.8% of the time, while just 8.2% of customers are unhappy with the sanitary conditions. If the company has established 15% as the threshold value, then this characteristic is considered to be under control.

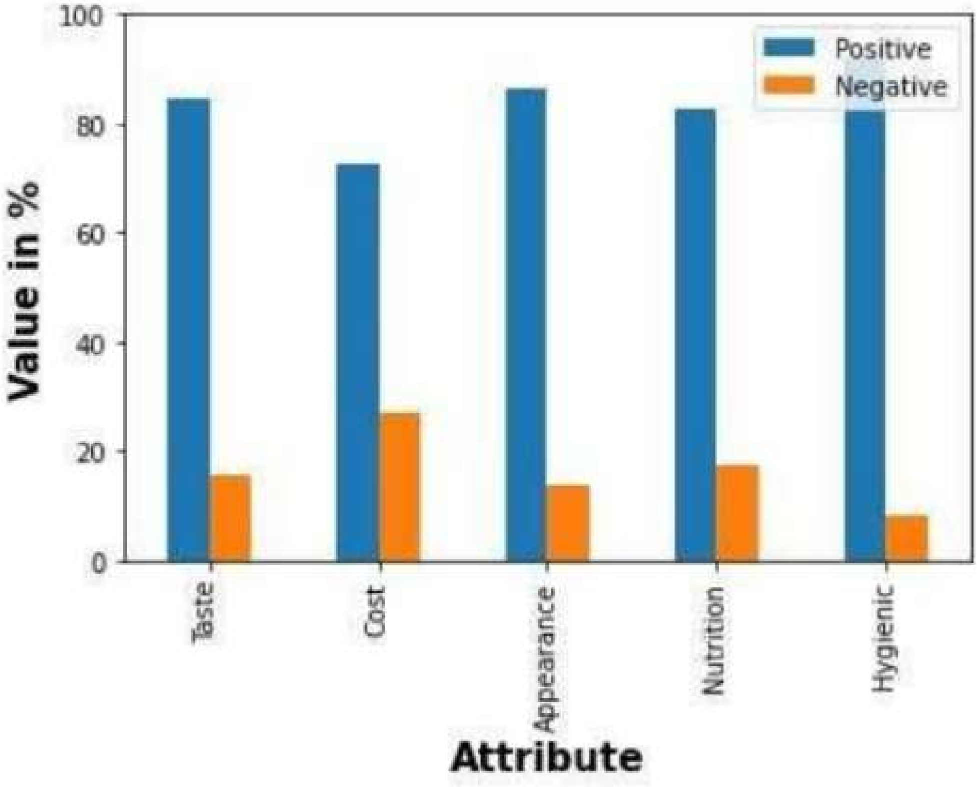

**Figure 5.6: Polarity representation of different attribute of food**

The favourable and negative remarks on various aspects of the dish are compared to one another in Figure 5.7.

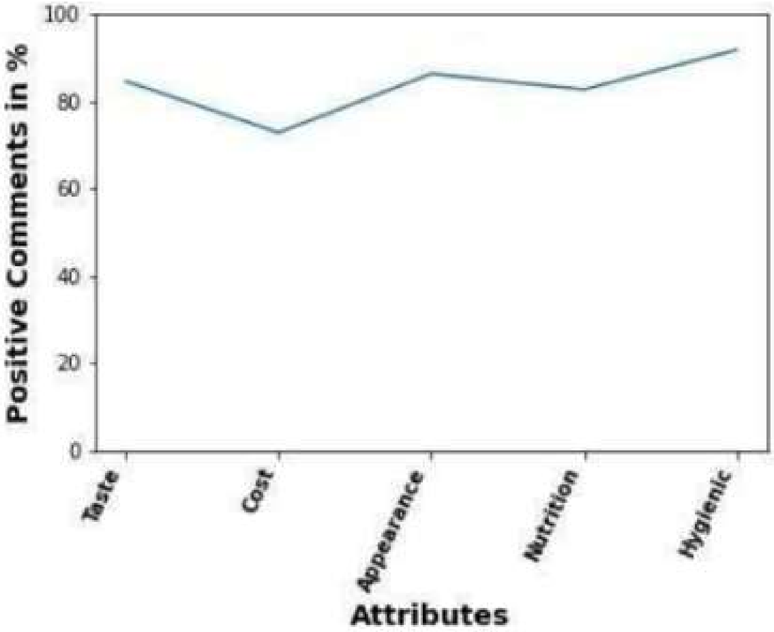

**Figure 5.7: Positive Comments for Different Attributes of Food**

Figure 5.7 displays the good remarks that were received about the various characteristics of food. The graph illustrates quite clearly that good comments on hygiene standards have reached 91.8%, while only 72.8% of remarks have been positive about costs.

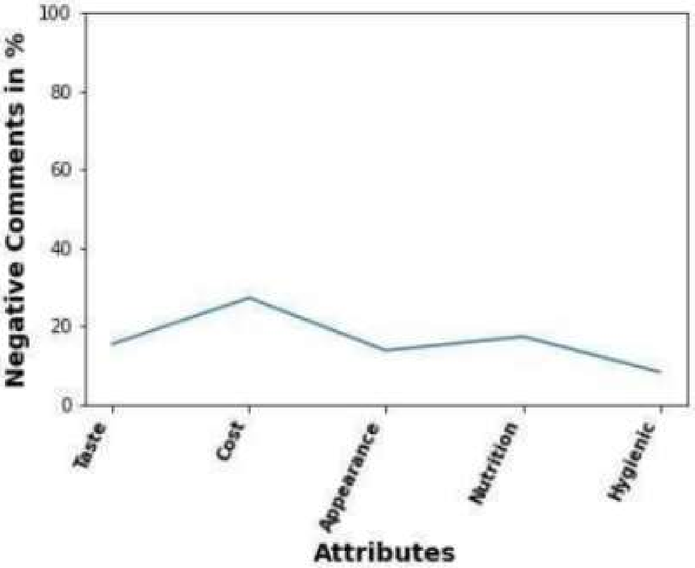

**Figure 5.8: Negative Comments for Different Attributes of Food**

The chart in Figure 5.8 illustrates the unfavourable remarks that were made about the various characteristics of food. The graph demonstrates quite clearly that the percentage of unfavourable comments on the price has reached 27.2%, and that the other criteria that have crossed the threshold value of 15% are taste (15.4%) and nutrition (17.3%). It is needed of the organisation to take the proper actions for regulating these characteristics. It was determined that the modified LSA approach yielded the best results in terms of precision, recall, F-measure, and accuracy when it came to the classification of positive and negative comments regarding the quality of the food. A variety of methods were utilised in order to classify positive and negative comments regarding the quality of the food. The comparison of the various methods is shown in Table 5.1.

**Table 5.1: Comparison of Different Approaches of Sentiment Classification**

| Approach | Precision | Recall | F-Measure | Accuracy |
|---|---|---|---|---|
| Naive Bayes | 95.01% | 93.74% | 94.37% | 96.05% |
| SVM | 95.42% | 93.9% | 94.65% | 96.25% |
| KNN | 95.29% | 93.95% | 94.61% | 96.22% |

| | | | | |
|---|---|---|---|---|
| Linear Regression | 94.61% | 93.64% | 94.12% | 95.87% |
| LSA | 96.21% | 93.95% | 95.07% | 96.54% |
| Modified LSA | 97.01% | 95.71% | 96.36% | 97.46% |

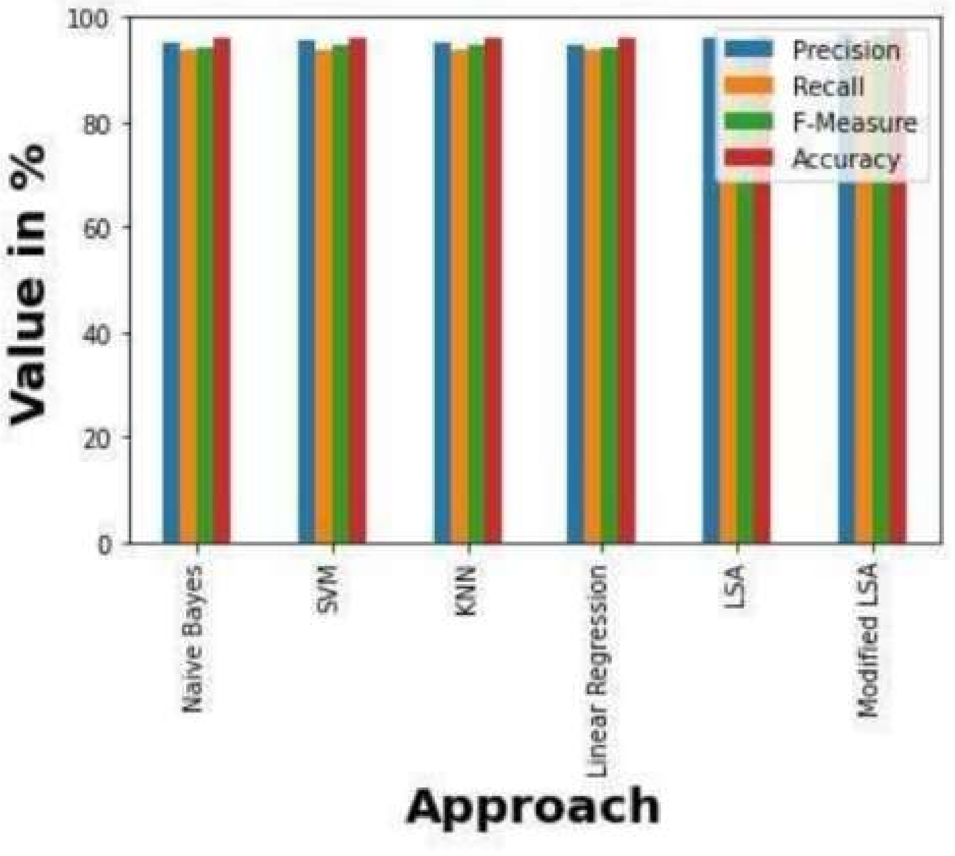

**Figure 5.9: Comparison of Modified LSA with other Approaches**

Figure 5.9 presents a comparison of various methods on the basis of precision, recall, F-measure, and accuracy. It can be seen that the modified LSA method provides the best results in terms of precision, recall, F-Measure, and accuracy when compared to other prominent methods such as Naive Bayes, Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Linear Regression, and LSA.

## 6. CONCLUSION

Text Mining Methodology for Business Intelligence is discussed in this chapter, which covers one of the submodules of the hybrid approach (TMABI). In this chapter, the emphasis was placed on the categorization of sentiments using a modified version of the LSA methodology, which, in comparison to other methods already in use, produces more effective results. In the next chapter, the whole operation of the hybrid system, which is an innovative method for gathering business information using text mining, will be presented.

## 7. REFERENCE

1. *Sandjai Bhulai, Wing Hong Kan, and Elena Marchiori, (2019) Nearest Neighbour Algorithms for Forecasting Call Arrivals in Call Centres..*

2. *Michael Negnevitsky, Paras Mandal, Anurag K. Srivastava, (20180 Machine Learning Applications for Load and Price Forecasting and Wind Power Prediction in Power Systems*

3. *Nesreen K. Ahmed, Amir F. Atiya, Neamat el Gayar, and Hisham el-Shishiny, (20180 An empirical comparison of machine learning models for time series forecasting..*

4. *H Raymond Joseph, (2017) GDP Forecasting through Data Mining of Seaport Export-Import Records .*

5. *Eduardo Miguel Rodrigues, Manuel Carlos Figueiredo, (2013) Forecasting Demand in the Clothing Industry, XI Congreso Galego de Estatística e Investigación de Operacións A Coruña, 24-25-26 de outubro de 2013.*

6. *Anand N. Asthana , Sangeeta Khorana,(2013) Unlearning Machine Learning: The Chal- lenge of Integrating Research in Business Applications, Middle-East Journal of Scientific, 2013.*