ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26feb1207

An Ensemble of Naive Bayes Variants and SentiWordNet with Threshold Adjustment for Aspect Based Sentiment Analysis on Restaurant Reviews

Musa Karatu[1]; Hamza Abdullahi Kwazo[2]

[1,2]Department of Computer Science, Federal University Birnin Kebbi, 860222, Nigeria

Publication Date: 2026/03/02

Abstract: This study proposes an enhanced framework for Aspect-Based Sentiment Analysis (ABSA) applied to restaurant reviews by integrating SentiWordNet with the variants of Naïve Bayes classifier. Existing sentiment analysis techniques often focus on limited aspects such as food quality, service, and price, while overlooking menu variety, which significantly influences customer perceptions. A dataset of 10,000 restaurant reviews was pre-processed using text normalization, tokenization, lemmatization, and stop-word removal. Aspect extraction was conducted through supervised learning with Naïve Bayes, and sentiment polarity scores were assigned using SentiWordNet. To handle mixed feature types, an ensemble combining Multinomial Naïve Bayes for TF-IDF features and Gaussian Naïve Bayes for sentiment polarity features was employed. Experimental results demonstrate that the proposed model achieves 88% accuracy with improved F1-scores for both positive and negative classes compared to baseline approaches. This contribution provides more balanced classification and offers practical insights for restaurant managers to enhance customer satisfaction. The findings highlight the significance of menu variety as a critical aspect of dining experiences. Future work may extend this research by applying deep learning models and multilingual datasets to broaden applicability.

Keywords: *Aspect-Based Sentiment Analysis, Natural Language Processing, SentiWordNet, Naïve Bayes, Restaurant Reviews, Customer Satisfaction.*

How to Cite: Musa Karatu; Hamza Abdullahi Kwazo (2026) An Ensemble of Naive Bayes Variants and SentiWordNet with Threshold Adjustment for Aspect Based Sentiment Analysis on Restaurant Reviews. *International Journal of Innovative Science and Research Technology*, 11(2), 2140-2152. https://doi.org/10.38124/ijisrt/26feb1207

I. INTRODUCTION

Sentiment analysis on restaurant review has become a continued area of research due to the numerous number of attributes that influences the performance of restaurant such as cultural, ambience, price, food, service, menu update, interaction between customers and stakeholders [1]. Many researchers, have applied several SA techniques in restaurant review such as, utilized the J48 decision tree algorithm to classify restaurant reviews collected from TripAdvisor.

Despite progress in sentiment analysis on restaurant review, research is ongoing particularly on asymmetrical attribute performance [2], Customers opinion [3], and ontology [4] which identified six critical aspects like food, service, ambience, cleanliness, location, and price. These ongoing researches focuses on improving model accuracy, addressing bias, and enhancing interpretability by 11.2% as shown in the study of [4].

There are still issue despite these improvements such that large diversification of the attributes on some dataset remain a challenge, menu variety are more important or significance than others like food quality, service quality, ambience, environment, and price fairness [3].

Aspect Based Sentiment Analysis (ABSA) stands as a promising method for classifying reviews into positive or negative categories based on aspect categories [5]. Despite various efforts in sentiment analysis based on multiple aspects, its accuracy remains in development [6]. Nurifan, et al. [7], employed a modified Hybrid ELMo-Wikipedia and Hybrid Expanded Opinion Lexicon-SentiCircle (HEOLS) technique to assess restaurant quality based on four aspects: Physical environment, Food quality, Service quality, and Price fairness. Their results indicates that HEOLS could expand and determine the Opinion Lexicon polarity, thereby increasing the F1 measure of Sentiment Analysis (SA) by 6%. However, this method is limited to only four aspects of restaurant evaluation, overlooking other significant factors such as menu variety, which directly impacts the dining

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26feb1207

experience. Menu variety is not merely about offering more options; it shapes memorable dining experiences, fosters customer loyalty, and drives business growth in a competitive industry.

Incorporating menu variety alongside physical environment, food quality, service quality, and price fairness emphasises its importance, as it directly influences customer perceptions and enjoyment of their dining experience. Nonetheless, the proposed method by [7] lacks efficiency in determining word polarity (sentiment classification), resulting in instances like "noise" being misclassified as positive. This is a critical error since, in most contexts, "noise" retains its negative connotation, indicating disturbance or lack of tranquility. Therefore, there is a need to enhance the accuracy of polarity determination to reduce errors.

Furthermore, the manual labeling of sentences in the dataset by professional annotators may introduce human errors. This paper proposed an ABSA of restaurant reviews considering five aspects: physical environment, food quality, service quality, Price fairness, and menu variety. This will also enhance restaurant quality and better customer satisfaction by integrating SentiWordNet and Naïve Bayes classifier to improve the accuracy of the restaurant review. This was achieved by investigating the impact of attribute importance in sentiment classification by analysing key aspects such as food quality, service quality, physical environment, price fairness, and menu variety; developing a method that will improve the accuracy of aspect based sentiment analysis in restaurant review using hybrid approach and evaluating the performance of SentiWordNet and the Naïve Bayes technique on a newly collected dataset and provide insights for better sentiment classification.

II. RELATED WORKS

Govindarajan [8] made a comparative study on the effectiveness of ensemble techniques use for sentiment classification using a hybrid classification method called arcing classifier. The result show that the hybrid model shows higher percentage of 92.44% classification accuracy than the base classifiers and enhances the testing time due to data dimensions reduction. It is clear from the result that the hybrid classifier shows significant improvement over the single classifiers.

Aye and Aung [9], analysed language-specific challenge by designing a lexicon-based sentiment analysis for reviews of foods and restaurants in Myanmar text using lexicon-based sentiment analysis. The result obtained using lexicon-based sentiment analysis achieved 96% overall accuracy of 500 customers' reviews of food and restaurant domain. This research can also be used to classify the objective and subjective reviews, the aspect and rule-based sentiment analysis for Myanmar text.

Adnan, et al. [10], look for positive or negative judgments and assess the performance of the method and comments in the reviews on the website, especially at restaurants in Surabaya using Decision Tree-J48 algorithm. The result shows the performance of the Decision Tree-J48 method with the average value of Precision 48.7%, Recall 36.8%, F-Measure 41.4% and accuracy 45.6%. The benefit of this classification results is as a recommendation for consumers to choose the best restaurant and it can be done by using other methods for comparison material.

Sharif, et al. [11], proposed Sentiment Analysis of Bengali Texts on Online Restaurant Reviews to classify customer reviews into positive and negative classes based on their sentimental feedback using Multinomial Naive Bayes. The experimental result shows that the proposed system can classify restaurant reviews with 80.48% accuracy.

Hossain, Bhuiyan and Tumpa [3] , present Sentiment Analysis of Restaurant Reviews using Combined Convolutional Neural Network - Long Short-Term Memory (CNN-LSTM) in order to understand the quality of a restaurant by the reviews from other customers. The study combined CNN-LSTM architecture of deep learning techniques used in the dataset and got an accuracy of 94.22%. This model can be reused for others Bangla text perspective. Other pre-trained word vectors were also used to compare the performance of CNN-LSTM.

Hossain, et al. [12], Highlighted a deep Learning approach for sentiment analysis of restaurant review to classify the reviews provided by the clients of the restaurant into positive and negative polarities**.** The results of the Bidirectional Long Short-Term Memory (BILSTM) technique produced highest accuracy of 91.35%. This approach acquires satisfactory results compared to other existing technique of Logistic Regression with 80.9%, Decision tree with 81.9%, Random forest with 84.7%, Support vector machine with 88.3% and Naive Bayes with 89.5%.

Ara, et al. [13] analysed restaurant's customer opinion in order to predict restaurant quality and future improvement using Natural Language Processing (NLP) based opinion mining. The experiment result shows that the efficiency of the proposed opinion mining approach is 85.714% for retrieving customer's opinions. The proposed approach and manual observation are both effective for predicting customer opinion.

Abdullah, et al. [14] combined an explicit or direct aspect of food quality, ambience, service and price together with the implicit aspects of delay and delicious in restaurant reviews. The results show a significant improvement in the respective evaluation for precision with 0.87, recall with 0.92, and F1-Score with 0.89.

Eidul, et al. [15], predict ratings of restaurant business based on features to help new entrepreneurs to set up new business using Machine Learning and Neural Network. The result shows that convolutional neural network (CNN) model give an accuracy score of 97.2 and 25 percent which is higher than the following algorithms of Decision Tree, Support Vector Machine and random Forest. The research predict the rating as accurately as possible, after analysing the

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26feb1207

performance and results the model is clearly performing much better than any other model.

Eidul, et al. [16], classified visitor sentiments as Tripadvisor users towards Happy Banana Komodo, MadeInItaly, Mediterraneo, and La Cucina restaurants in Labuan Bajo using Webharvy application, RapidMiner application, Decision Tree (DT), Naïve Bayes (NB), Support Vector Machine (SVM), and K-Nearest Neighbor (K-NN). The result shows that the implementation of the positive and negative sentiment classification method for the comprehensive review data from the Tripadvisor website for the products and services of restaurants Happy Banana Komodo, MadeInITaly, Mediterraneo, and La Cucina Restaurants are relevant with K-Nearest Neighbor (K-NN) with accuracy value of 99.27%, a precision value of 100%, and a recall of 98.53%. The research can be used as a recommendation for restaurant business managers in Labuan Bajo to optimise products and services to improve the positive image of Labuan Bajo restaurants as an essential part that supports the idea of Indonesia's super premium tourist destinations.

Rita, Vong, Pinheiro and Mimoso [1], investigated how online review sentiments towards four key aspects (food, service, ambience and price) change after a restaurant is awarded a Michelin Star using a web crawler and Semantria. The study findings revealed that overall sentiments decreased after restaurants were awarded a Michelin Star, in which service sentiment was the most affected aspect, followed by food and ambience. Yet, price sentiment showed a prominent increase.

Aruna, et al. [17], proposed a sentiment analysis for Zomato restaurant, to analysed user comments and reviews. The study highlighted 10 words that affect the results, after improving the accuracy on precision and recall by 92% such words are: “bad”, “good”, “average”, “best”, “place”, “love”, “order”, “food”, “try”, and “nice”.

Gujrati, et al. [18] analysed the reviews and classify them with respect to category using machine learning. The study investigated the reviews given by the customers of the restaurant with help of Machine Learning model and the model is built using a labeled dataset of restaurant reviews as positive or negative. The model is trained using a Naive Bayes and Support Vector Machine. The research intends to collect more data from various social networking sites and apply deep learning and neural networks for better performance in the future.

Fragko, et al. [19] determined the efficiency of off-the-shelf sentiment analysis APIs in recognising low-resource languages, such as Greek Language using the Meaning Cloud web-based tool. The result found low agreement between the web-based and the actual raters in the food delivery services related data, also the low accuracy of the results highlights the need for specialised sentiment analysis tools capable of recognising only Greek language as a low-resource language.

From the above literature review, studies failed to address some aspect of restaurants such as menu variety and creativity in customer’s review, this has the potential to enhance the quality of restaurant more efficiently. Naive Bayes classifiers have been applied in aspect-based sentiment analysis of restaurant reviews, the integration of SentiWordNet with Naive Bayes in this specific context appears to be less common in recent research. This justify further studies to explore this combination to enhance sentiment analysis accuracy in restaurant reviews. To the best of our knowledge this is the first attempt to explore this approach.

III. MATERIALS AND METHODS

This research employed a hybrid framework methodology that can classify restaurant reviews into positive and negative sentiments based on the required aspect by integrating the strengths of SentiWordNet to potentially improve the accuracy of naïve Bayes classification algorithm in restaurant review using Aspect-based Sentiment Analysis.

This research begins by exploring the restaurant review dataset, then continues with pre-processing - transform case, tokenize, stop word removal, stemming, and punctuation removal, the result obtained from the text preprocessing stage would be split into train and test data by applying dictionary based technique SentiWordNet to automatically classify the data instead of using professional annotators manually based on the required aspect to be used as a train data and passed it to train Naïve Bayes classifier for more accuracy in classifying the sentiment. Then the classification is tested using test data and evaluated using a new dataset to obtain an accuracy value. The research method flowchart can be seen in Figure 1 below.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26feb1207

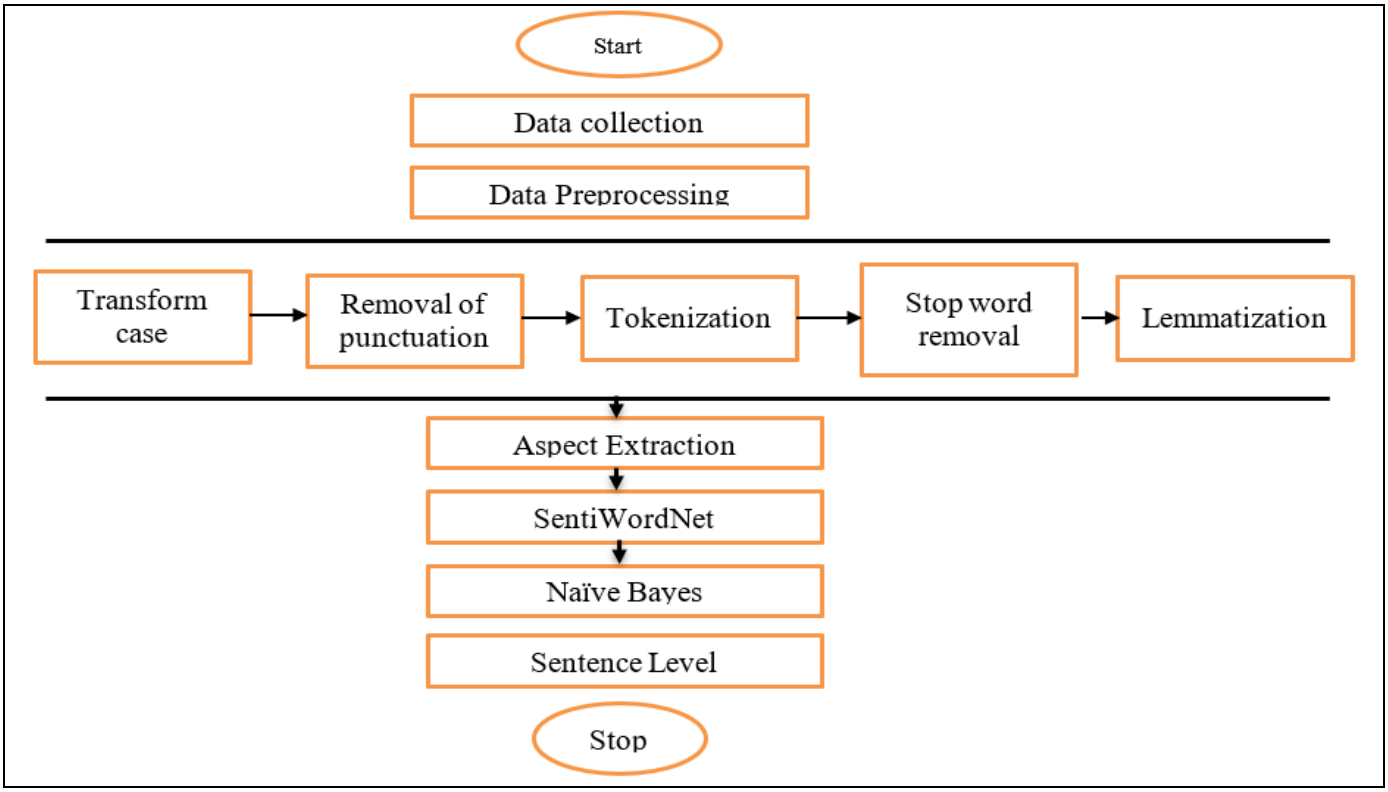

Fig 1 Flowchart of Aspect-Based Sentiment Analysis (ABSA) Process.

A. Environmental Setting

The experiments were conducted in Jupyter Notebook within the Anaconda distribution, chosen for its flexibility in iterative development and strong support for data preprocessing, machine learning, and visualization. Python 3.12.4 was used, along with key libraries including Pandas and NumPy for data manipulation, scikit-learn for model building and evaluation, and Matplotlib for plotting results.

Text processing relied on NLTK, with the SentiWordNet and wordnet corpora, tokenization via word_tokenize, and part-of-speech tagging through pos_tag. Essential NLTK resources, including 'punkt', 'averaged_perceptron_tagger', 'wordnet', and 'SentiWordNet', were downloaded prior to analysis. The experimental pipeline utilized TF-IDF vectorization alongside sentiment features (positive, negative) for GaussianNB.

B. Data Collection

The data used in this research is restaurant reviews dataset comprising of 10,000 reviews [20], the dataset have different aspect, including food quality, service quality, price fairness, ambience, cleanliness, location, menu variety and staff of a restaurant. The selection of aspects is based on the influence of aspect assessment on customer decisions. In terms of price, consumers will know whether the review is cheap or expensive, in terms of location, will know whether there is nice location or not, and in terms of cleanliness, it can be seen from the restaurant whether its clean or not.

Data Pre-Processing is the stage for preparing textual data that will be used at a late stage. Natural Language Toolkit would be used in this research for evaluating the following text pre-processing stage which are transform case, removal of punctuation r, tokenization, stop word removal, and lemmatization.

Here all characters in the data are changed to lowercase, (e.g., "Delicious FOOD" → "delicious food") and this will help standardize words and avoid duplication of tokens due to case differences (e.g., "Food" and "food" are treated the same).

The unnecessary punctuations are then removed in a document (e.g., "The food was great!!!" → "The food was great") as it will help in preventing punctuation marks from being considered as separate tokens, which could affect analysis accuracy.

The text are further split into individual words or tokens. This step is essential before further processing like stemming or lemmatization [21]. E.g., "The food was great" → ["The", "food", "was", "great"]. This will make the review very easier to analyze words separately, particularly when identifying aspects (food, service, ambience).

C. Aspect Extraction

Technique adopted for aspect extraction in this research is the supervised learning. Specifically, the Naïve Bayse classification. Aspect extraction plays a crucial role in enabling us to gain granular insights like ambience, service, and food quality from the restaurant review dataset.

D. SentiWordNet

In SentiWordNet, each synset of WordNet is being assigned the three sentiment numerical scores; positive, negative and objective that are calculated using a set of classifiers, each synset (denoted by sn) of WordNet is associated with three numerical scores, namely Obj(sn),

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26feb1207

Sub(sn), and Neg(sn). In this work SentiWordNet would be used to classify the dataset into positive and negative opinions which, will serve as data train before taking it to Naïve Bayes for general classification.

E. Aspect Extraction Procedure

The implementation process followed these steps:

➢ *Aspect Term Identification*

The first step involved building an aspect dictionary by collecting keywords related to each aspect. For example:

- Food quality: "taste", "fresh", "delicious", "undercooked"
- Service quality: "waiter", "attentive", "rude", "staff"
- Price fairness: "cheap", "expensive", "reasonable", "value"
- Ambience: "quiet", "romantic", "noisy", "decor"
- Menu variety: "options", "selection", "choices", "limited menu"

This dictionary was enhanced by using WordNet synonyms and SentiWordNet polarity scores, ensuring broader coverage of opinion words.

➢ *Part-of-Speech (POS) Tagging*

- Each review sentence was tokenized and POS-tagged using the NLTK library in Python.
- Nouns and noun phrases were treated as candidate aspects (e.g., food, service, menu), while adjectives, verbs, and adverbs around them were treated as opinion words (e.g., delicious, rude, expensive).

➢ *Aspect–Opinion Pairing*

- Dependency parsing was applied to link aspect terms with their closest opinion words.
- Example: "The food was delicious but the service was slow."
- Extracted pairs: (food, delicious), (service, slow).

➢ *Sentiment Scoring using SentiWordNet*

- Opinion words were matched with their polarity scores in SentiWordNet.
- Positive, negative, and objective scores were retrieved.

Example: "delicious" → Positive: 0.875, Negative: 0.0.

Each aspect was assigned the sentiment of its associated opinion word.

➢ *Aspect-Level Classification*

- After sentiment scores were obtained, the Naïve Bayes classifier was trained on these aspect–opinion pairs.
- This allowed classification of each aspect within a review independently (e.g., food = positive, service = negative).

➢ *Storage and Representation*

- Extracted aspects and their sentiments were stored in a structured format for further analysis.
- Example output for one review:

✓ Review: "The food was delicious but the service was slow."
✓ Aspects :(Food: Positive, Service: Negative).

F. TF-IDF Vectorization

TF-IDF evaluates the importance of a term within a document relative to a corpus by combining two measures: Term Frequency (TF) and Inverse Document Frequency (IDF). TF quantifies how frequently a term appears in a document.

G. Naive Bayes

At the sentiment prediction stage, the data will be trained using the Naïve Bayes algorithm, which is a supervised machine learning technique. The result obtained at this stage is the polarity of sentiment from each aspect.

H. Mathematical Representation of Naïve Bayes Algorithm

The Naïve Bayes classifier is a probabilistic model based on Bayes' theorem with the assumption of independence among features. Given a dataset with features $X_1, X_2, \ldots X_n$ and a class variable C, the probability of a class given the feature values is:

$$P(C \mid X1, X2, \ldots, Xn) = \frac{P(C)P(X1,X2,\ldots,Xn|C)}{P(X1,X2,\ldots,Xn)} \quad \ldots\ldots\ldots\ldots\ldots(1)$$

Since the denominator $P(X_1, X_2, \ldots X_n)$ is constant for all classes, we can express:

$$P(C/X_1, X_2, \ldots X_n) \propto P(C)P(X_1, X_2, \ldots X_n /C \ldots\ldots\ldots\ldots(2)$$

I. The Naïve Bayes assumption simplifies the joint probability:

$$P(X_1, X_2, \ldots X_n /C) = P(X_1/C)\, P(X_2/C), P(X_n/C) \ldots(3)$$

Thus, the final classification rule becomes:

$$P(X_1, X_2, \ldots X_n /C) \propto P(C) \prod P(X_i/C) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots(4)$$

For classification, we choose the class C^* that maximizes this ability:

$$C^* = arg\ max\ P(C) \prod P\left(\frac{X_i}{C}\right). \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots(5)$$

J. Variants and Handling Negative Values

An ensemble of two variants used in the proposed approach are:

➢ Multinomial Naïve Bayes: Ideal for text features such as TF-IDF, which are non-negative.
➢ Gaussian Naïve Bayes: Handles continuous features, including negative values.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26feb1207

K. Multinominal Naïve Bayse

Multinominal Naïve Bayse uses a discrete count features such as word count and TF-IDF without negatives. Mathematically it is represented as follows:

For class C_k and x_i which denotes term i, the likelihood is:

$$P(x_i|C_k) = \frac{N_{ik}+\propto}{N_k+\propto m}, \quad \text{.....(6)}$$

Where:

- $N_{ik} = number\ of\ times\ feature\ i\ appears\ in\ documents\ of\ class\ C_k$
- $N_k = \sum_i N_{ik} = total\ count\ of\ all\ features\ in\ class\ C_k$
- $m = number\ of\ features$
- $\propto = Laplace\ smoothing\ parameter$

For a document with counts $x_1, x_2, \dots, x_n$:

$$P(X|C_k) \propto \prod_{i=1}^{n} P(x_i|C_k)^{x_i}, \quad \text{.....(7)}$$

To avoid underflow, the Log form is used and expressed as follows:

$$\log P(C_k|X) = \log P(C_k) + \sum_{i=1}^{n} x_i \log P(x_i|C_k) \quad \text{.....(8)}$$

L. Gaussian Naïve Bayse

The Gaussian Naïve Bayse is used for continuous features such as real values and suitable to handle negative values.

For each feature x_i, is assumed it follows a normal distribution in each class:

$$P(x_i|C_k) = \frac{1}{\sqrt{2\pi\sigma_{ik}^2}} \exp\left(-\frac{(x_i-\mu_{ik})^2}{2\sigma_{ik}^2}\right), \quad \text{.....(9)}$$

Where:

- $\mu_{ik} = mean\ of\ feature\ i\ in\ class\ C_k.$
- $\sigma_{ik}^2 = variance\ of\ feature\ i\ in\ class\ C_k.$

Then:

$$P(X|C_k) \propto \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_{ik}^2}} \exp\left(-\frac{(x_i-\mu_{ik})^2}{2\sigma_{ik}^2}\right) \quad \text{.....(10)}$$

The log form is:

$$\log P(C_k|X) = \log P(C_k) - \frac{1}{2}\sum_{i=1}^{n} \log(2\pi\sigma_{ik}^2) - \sum_{i=1}^{n} \frac{(x_i-\mu_{ik})^2}{2\sigma_{ik}^2}. \quad \text{.....(11)}$$

In this study, the feature set includes:

- SentiWordNet polarity scores (may contain negative values).
- TF-IDF features (normalized but non-negative).

Since Multinomial NB cannot process negative values, a hybrid approach is proposed:

- Apply MultinomialNB to text features (TF-IDF).
- Apply GaussianNB to sentiment features (SentiWordNet).
- Combine outputs at the probability level, allowing each model to operate on its optimal feature space.

Additionally, the proposed pipeline will automatically:

- Detects negative values in feature sets.
- Routes them to the appropriate model (Gaussian for negatives, Multinomial for non-negative).
- Blends the results to avoid manual intervention.

IV. RESULTS AND DISCUSSION

This section presents results of the proposed Aspect Based Sentiment of restaurant reviews. These include data implementation, followed by aspect extraction, also next subsection discussed sentiment classification, evaluation metrics and result were also discussed including comparative analysis, complexity analysis, and the test for statistical significant, discussion and conclusion.

- *Data Implementation*

The experiment was conducted on a restaurant reviews dataset comprising of ten thousand reviews, the dataset has different aspects in a review, including food quality, service quality, price fairness, ambience, cleanliness, location, and menu variety. Figure 2 below, describes a sample of reviews from the dataset.

	Restaurant	Reviewer	Review	Rating	Metadata	Time	Pictures
0	Beyond Flavours	Rusha Chakraborty	The ambience was good, food was quite good . h...	5	1 Review , 2 Followers	5/25/2019 15:54	0
1	Beyond Flavours	Anusha Tirumalaneedi	Ambience is too good for a pleasant evening. S...	5	3 Reviews , 2 Followers	5/25/2019 14:20	0
2	Beyond Flavours	Ashok Shekhawat	A must try.. great food great ambience. Thnx f...	5	2 Reviews , 3 Followers	5/24/2019 22:54	0
3	Beyond Flavours	Swapnil Sarkar	Soumen das and Arun was a great guy. Only beca...	5	1 Review , 1 Follower	5/24/2019 22:11	0
4	Beyond Flavours	Dileep	Food is good.we ordered Kodi drumsticks and ba...	5	3 Reviews , 2 Followers	5/24/2019 21:37	0
5	Beyond Flavours	Nagabhavani K	Ambiance is good, service is good, food is aPr...	5	1 Review	5/24/2019 15:22	0
6	Beyond Flavours	Jamuna Bhuwalka	Its a very nice place, ambience is different, ...	5	1 Review	5/24/2019 1:02	0
7	Beyond Flavours	Sandhya S	Well after reading so many reviews finally vis...	4	1 Review	5/23/2019 15:01	0
8	Beyond Flavours	Akash Thorat	Excellent food , specially if you like spicy f...	5	1 Review , 1 Follower	5/22/2019 23:12	0
9	Beyond Flavours	Smarak Patnaik	Came for the birthday treat of a close friend....	5	1 Review , 1 Follower	5/22/2019 22:37	0

Fig 2 A Sample of Reviews from the Dataset.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26feb1207

The dataset was downloaded and put into a data frame using Pandas. Pandas is a powerful open-source Python library used for data manipulation and analysis. Pandas makes handling large datasets efficient and intuitive, and used to import, clean, transform, filter, group, merge, and visualize data with just a few lines of code.

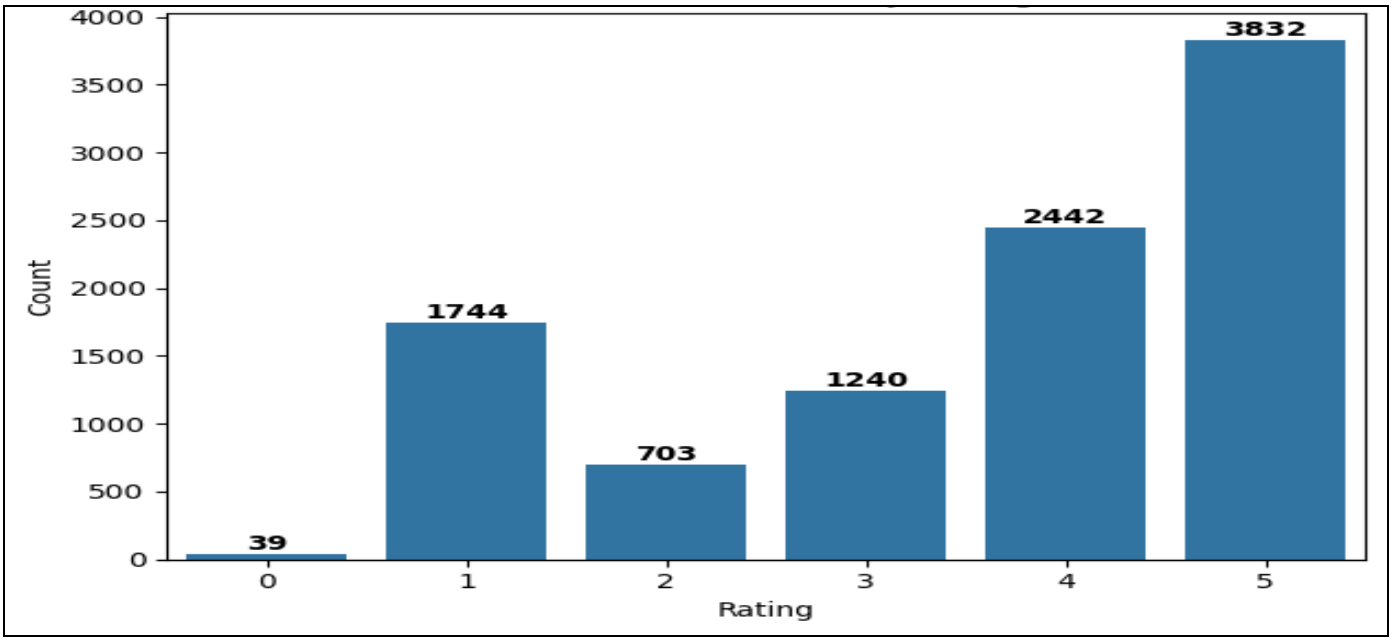

Fig 3 The Number of Reviews by Rating.

In order to explore the dataset, a bar chart was plotted with ratings as independent variable and counts as the dependent variable as shown in Figure 3. From the ratings of 0 to 5, 0 being the lowest rating and 5 being the highest rating. Out of the 10,000 reviewers, 39 rated their experience 0, 1744 rated their experience as 1703, 1240, 2442, 3832 rated as 2, 3, 4, and 5, respectively.

➢ *Data Preprocessing*

To pre-process the data, spaCy for text preprocessing in Natural Language Processing (NLP) was used. The medium English language model (en_core_web_md) was first loaded and a pre-process function defined. Inside the function, the input text was processed into a Doc object using the spaCy pipeline. Each token was then iterated through, performing lemmatization - converting words to their base form, e.g., "running" → "run", converting to lowercase, and stripping whitespace from tokens. Stop words (common words like "the," "is") and punctuation were filtered, leaving only meaningful words, as shown in Figure 4. below, "processed_reviews". This produced a normalized text for abstract based sentiment analysis and classification.

	Restaurant	Reviewer	Review	Rating	processed_reviews
0	Beyond Flavours	Rusha Chakraborty	The ambience was good, food was quite good . h...	5	ambience good food good saturday lunch cost ef...
1	Beyond Flavours	Anusha Tirumalaneedi	Ambience is too good for a pleasant evening. S...	5	ambience good pleasant evening service prompt ...
2	Beyond Flavours	Ashok Shekhawat	A must try.. great food great ambience. Thnx f...	5	try great food great ambience thnx service pra...
3	Beyond Flavours	Swapnil Sarkar	Soumen das and Arun was a great guy. Only beca...	5	soumen das arun great guy behavior sincerety g...
4	Beyond Flavours	Dileep	Food is good.we ordered Kodi drumsticks and ba...	5	food good.we order kodi drumstick basket mutto...
5	Beyond Flavours	Nagabhavani K	Ambiance is good, service is good, food is aPr...	5	ambiance good service good food apradeecp subr...
6	Beyond Flavours	Jamuna Bhuwalka	Its a very nice place, ambience is different, ...	5	nice place ambience different food order tasty...
7	Beyond Flavours	Sandhya S	Well after reading so many reviews finally vis...	4	read review finally visit place ambience good ...
8	Beyond Flavours	Akash Thorat	Excellent food , specially if you like spicy f...	5	excellent food specially like spicy food court...
9	Beyond Flavours	Smarak Patnaik	Came for the birthday treat of a close friend....	5	come birthday treat close friend perfect place...
10	Beyond Flavours	Saubhagya Bhuyan	The service was great and the food was awesome...	5	service great food awesome service staff manab...
11	Beyond Flavours	Srivaths07	Very good ambience, amazing food ,good service...	5	good ambience amazing food good service friend...
12	Beyond Flavours	Kunj Mishra	Food was very good. Soup was as expected. In s...	5	food good soup expect starter order honey chil...
13	Beyond Flavours	Pradeep Vetapalem	Food is too good. Telangana kodiak fry is must...	5	food good telangana kodiak fry try mutton biri...
14	Beyond Flavours	Kankaria.ritu	We ordered corn cheese balls, manchow soup and...	1	order corn cheese ball manchow soup paneer sha...
15	Beyond Flavours	Abhay Sharma	Food and ambience is fantastic.. Waiter Manav ...	5	food ambience fantastic waiter manav maji quic...
16	Beyond Flavours	Shubham Jaiswal	Came here for lunch and the food was good and ...	4	come lunch food good tasty try buffet item veg...
17	Beyond Flavours	Srijani Mukherjee	The best thing about this place is the food. M...	5	good thing place food favorite dish definitely...
18	Beyond Flavours	Hari Jangam	Polite and friendly staff. Nice ambience and g...	5	polite friendly staff nice ambience good sprea...
19	Beyond Flavours	Suneet Soni	Food is really good. We had vegetarian items i...	5	food good vegetarian item include paneer mushr...

Fig 4 Preprocessed Text Review in Natural Language Processing Using spaCy

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26feb1207

To further explore the dataset, spaCy and TextBlob were integrated to compute sentiment analysis scores (polarity and subjectivity) and adds them as a custom extension attribute to spaCy's Doc objects, extracting polarity (how positive/negative the text is) and subjectivity (how subjective/objective the text is). The average polarity and subjectivity of the reviews was computed as shown in Figure 5. The 10,000 reviews show an overall slightly positive sentiment (average polarity 0.26) and are moderately subjective (subjectivity 0.60), indicating that feedback is largely opinion-based rather than factual. This suggests customers are generally satisfied, though not highly enthusiastic, and a deeper look into the distribution could uncover specific improvement areas.

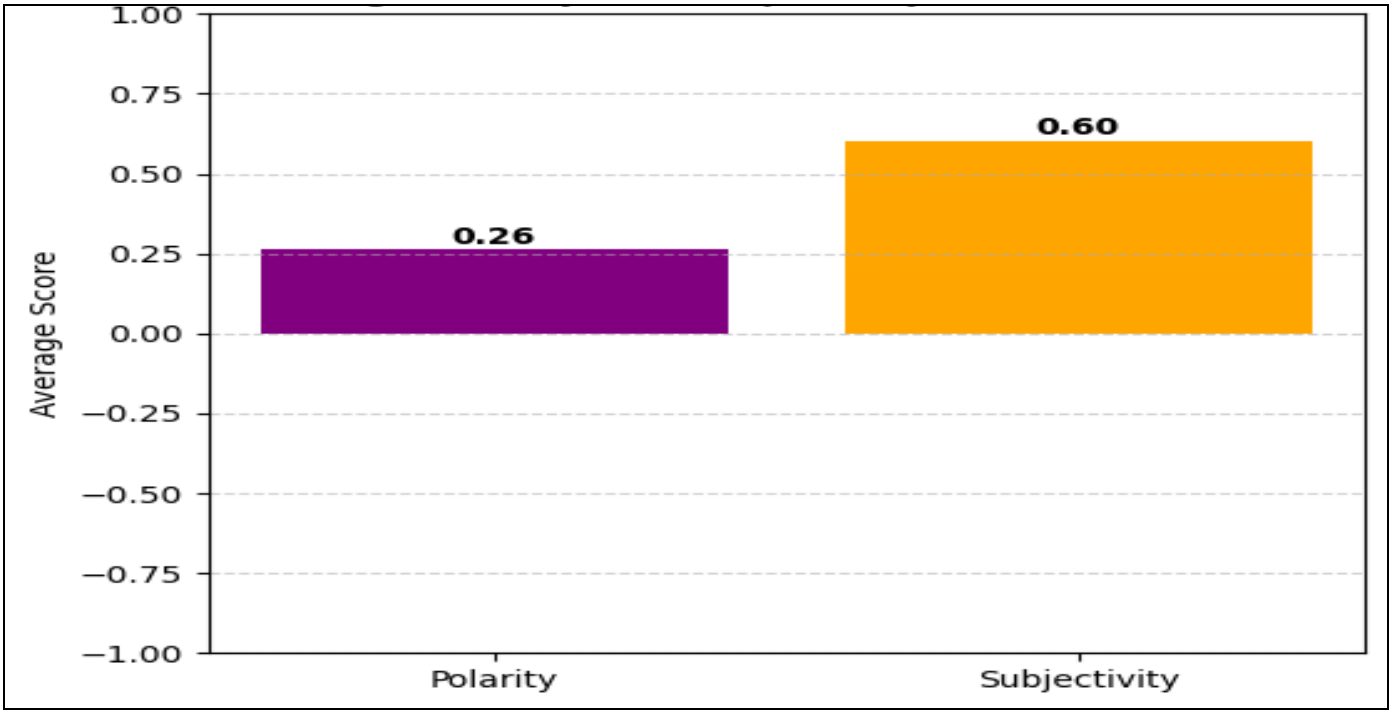

Fig 5 Average Polarity and Subjectivity Across All Reviews.

A further analysis of the sentiment components, shown in Figure 6 and 7 below, breaks down the reviews into positive, negative, and polarity scores. The results indicate an average positive sentiment of 0.12, a negative sentiment of 0.04, and a polarity score of 0.63, suggesting that the majority of reviews are largely objective with only mild positive expressions and minimal negativity. This reinforces the earlier observation that customer feedback is generally factual with slight approval, highlighting the need for a closer examination of individual reviews to identify specific opportunities for improving customer satisfaction.

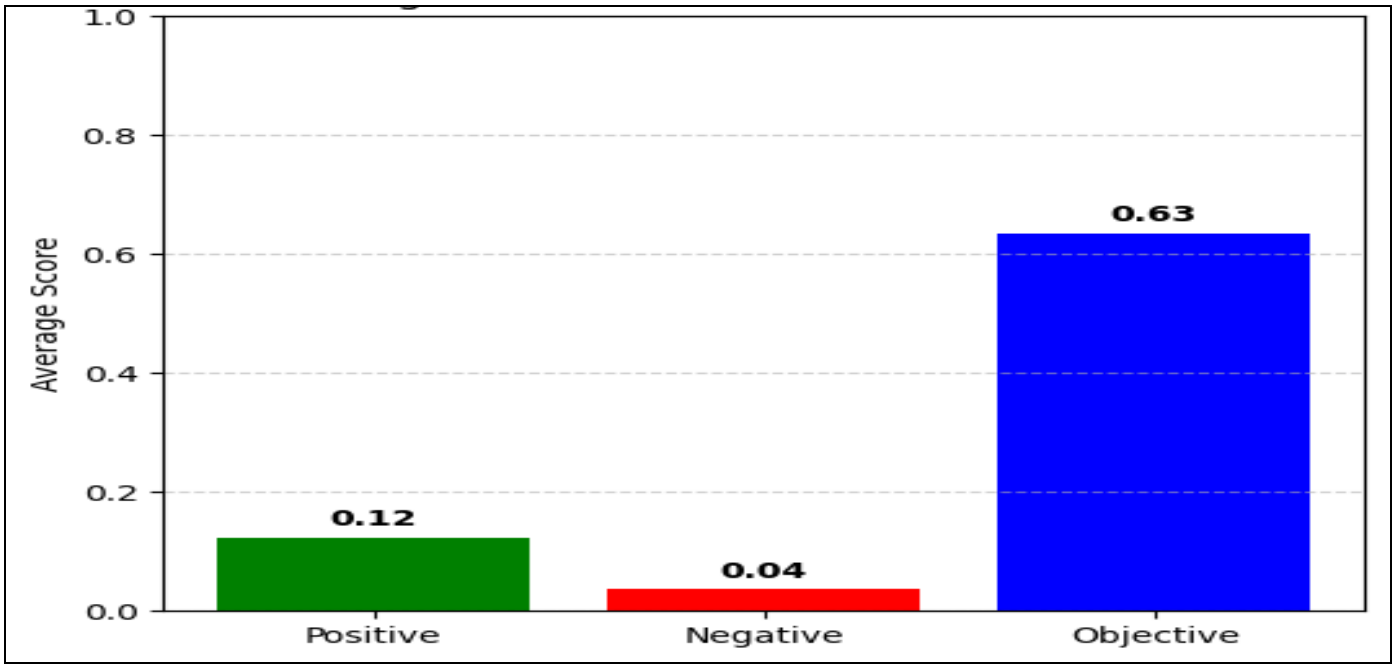

Fig 6 Average Sentiment Scores Across All Reviews.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26feb1207

➢ *Sentiment Classification*

Fig 7 Sentiment Classification

In the sentiment classification process, SentiWordNet was employed to obtain sentiment scores for words linked to extracted aspects in a review. For instance, in the sentence "the food was good but the service mix up orders", each opinion word was assigned a sentiment score, with "good" carrying a strong positive score and "mix up order" indicating a negative sentiment. These scores were then transformed into features, such as the average positive or negative score associated with a specific aspect like "food" or "service.", as shown in Table 1 below.

Table 1 Table Showing the Combine Bag-of-Words with SentiWordNet Score.

Restaurant	Reviewer	Review	Rating	processed_reviews	Sentiment	swn_score	sentiword_score	sentiment_label
Beyond Flavours	Rusha Chakraborty	The ambience was good, food was quite good . had Saturday lunch , which was cost effective . Good place for a sate brunch. One can also chill with friends and or parents. Waiter Soumen Das was really courteous and helpful.	5	ambience good food good saturday lunch cost effective good place sate brunch chill friend parent waiter soumen das courteous helpful	1	0.21	0.15	1
Beyond Flavours	Anusha Tirumalaneedi	Ambience is too good for a pleasant evening. Service is very prompt. Food is good. Over all a good experience. Soumen Das - kudos to the service	5	ambience good pleasant evening service prompt food good good experience soumen das kudos service	1	0.25	0.16	1
Shah Ghouse Hotel & Restaurant	Tyson Sairaj	The contact number in the true caller shows best cheaters in the world... Good I had ordered for mutton biryani nd I got noodles which are ugly... Wow I am so happy for that Good keep it up(bullshit) How careless they are for delivering this type of service to customer who orders for food This is really disgusting and bullshit I suggest everyone not to order any food from this restaurant(Shah ghouse gachibowli) Better remove service	1	contact number true caller show good cheater world good order mutton biryani nd get noodle ugly wow happy good up(bullshit careless deliver type service customer order food disgusting bullshit suggest order food restaurant(shah ghouse gachibowli well remove service zomato friend issue restaurant	0	0.06	0.05	1

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26feb1207

		from zomato Many of our friends had many issues from this restaurant, like delivering different food from ordered food, and delivering one shawarma instead of two and others Zomato authority should take of all these problems which are spoiling the name and fame		like deliver different food order food deliver shawarma instead zomato authority problem spoil fame				

These features (i.e., sentiment, swn_score, sentiword_score, and sentiment_label) served as input to a Naive Bayes classifier, which predicted the overall sentiment for each aspect, classifying "food" as positive and "service" as negative in this example.

To enhance classification accuracy, two variants of Naïve Bayes were integrated in the proposed pipeline. Multinomial Naïve Bayes was applied to text-based features such as TF-IDF, as it is designed for discrete, non-negative values. In contrast, Gaussian Naïve Bayes was employed to handle continuous features, including those with negative values such as sentiment polarity scores derived from SentiWordNet. In this way, TF-IDF features were routed to the Multinomial model, while sentiment features were processed by the Gaussian model.

➢ *Evaluation metrics*

To assess the performance of the proposed sentiment classification model, standard evaluation metrics in Natural Language Processing (NLP) and machine learning were first employed. These include Precision, Recall, F1-score, and Accuracy:

- Precision measures the proportion of correctly predicted positive samples among all predicted positives. It quantifies the classifier's ability to avoid false positives.
- Recall (or Sensitivity) is the proportion of correctly predicted positive samples among all actual positives. It reflects the classifier's ability to capture relevant results.
- F1-Score is the harmonic mean of precision and recall, providing a balanced measure when both are important, especially in imbalanced datasets.
- Accuracy measures the overall proportion of correctly classified samples.

In addition, macro average and weighted average scores were reported. The macro average computes the mean metric score across all classes, treating each class equally, while the weighted average accounts for class imbalance by weighting each class by its support.

➢ *Classification Results*

The classification results of the baseline and proposed models are presented in Figure 8. The total number of test samples of data used for testing is twenty percent of the ten thousand samples (i.e., 10,000 × 0.2 = 2,000). The test dataset has a total of 497 samples class 0 or negative reviews and 1,503 the number of class 1 or positive reviews in the test set. The class distribution in the test set is imbalanced (roughly 1:3), which matches the overall dataset distribution.

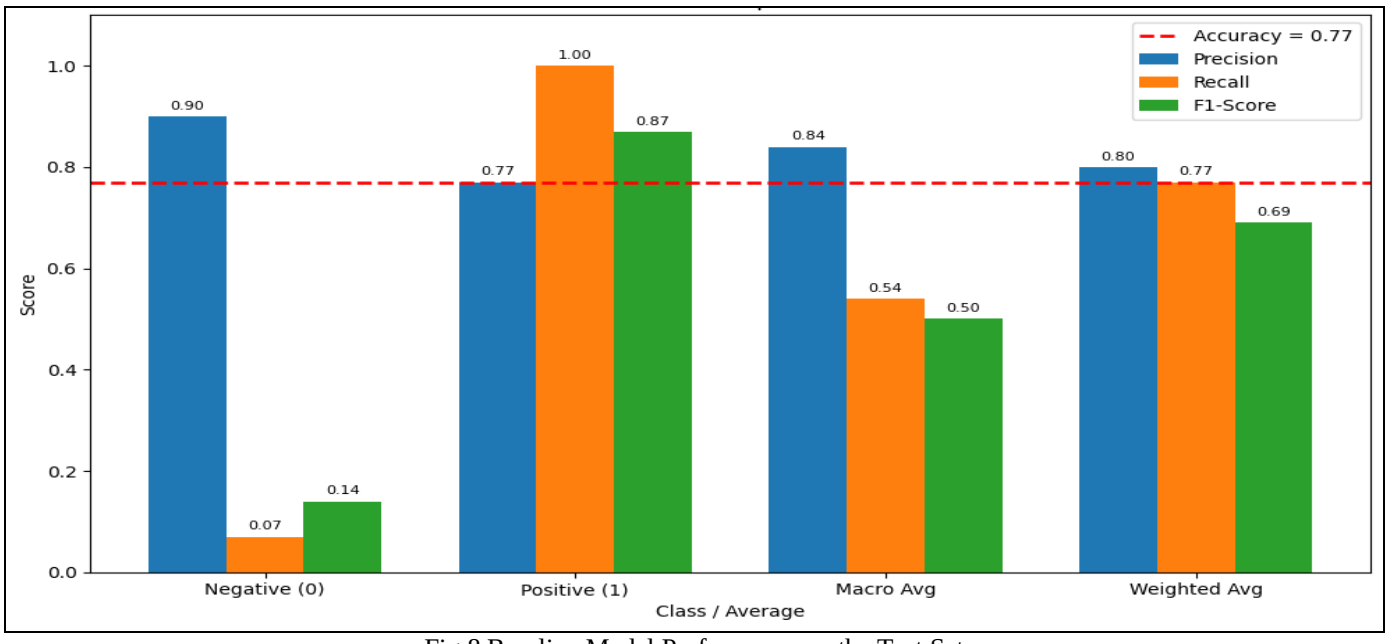

Fig 8 Baseline Model Performance on the Test Set

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26feb1207

The result in Figure 8 print a grim picture of the performance of the SentiWordNet and Naïve Bayse approach. This is particularly true because, for the negative sentiment classification with precision 90%, recall 7%, and F1-Score 14%, suggests the model is predicting very few negatives correctly (low recall). Most negatives are misclassified as positive. Precision is high because when it does predict negative, it's usually correct. The positive sentiment on the one hand has precision 77%, recall 100%, and F1-Score 87%, implying the model catches almost all positives (recall 1.0), but has some false positives (precision 0.77). However, the F1-score is high. The average of precision, recall, F1 across classes without considering class imbalance**,** shows the model struggles with minority class – negative sentiments. The average of precision, recall, F1 weighted by class support, reflects overall accuracy better considering the dataset imbalance with the rates 80, 77, and 69 for precision, recall, and F1-score, respectively.

Although overall accuracy appears acceptable at 77%, it is misleading since it largely reflects the dominance of the positive class, masking the model's inability to handle the minority class effectively. These shortcomings stem from the dataset imbalance, the limited discriminative power of GaussianNB on sentiment scores, and the TF-IDF with weighted voting scheme favouring the majority class. One potential improvement is threshold adjustment, where probability thresholds for the negative class are fine-tuned rather than relying solely on np.argmax, which could help recover more true negatives and improve recall.

➢ *Model Performance and Limitations*

These weaknesses in Figure 8 above stem from both the skewed dataset and modeling choices, such as GaussianNB's limited ability to capture signals for negative reviews and the TF-IDF with weighted voting scheme, which tends to reinforce the majority class. To Address these issues, threshold adjustment strategy is employed, where probability cutoff is shifted to improve recall for the minority class, with the aim of creating a more balanced performance.

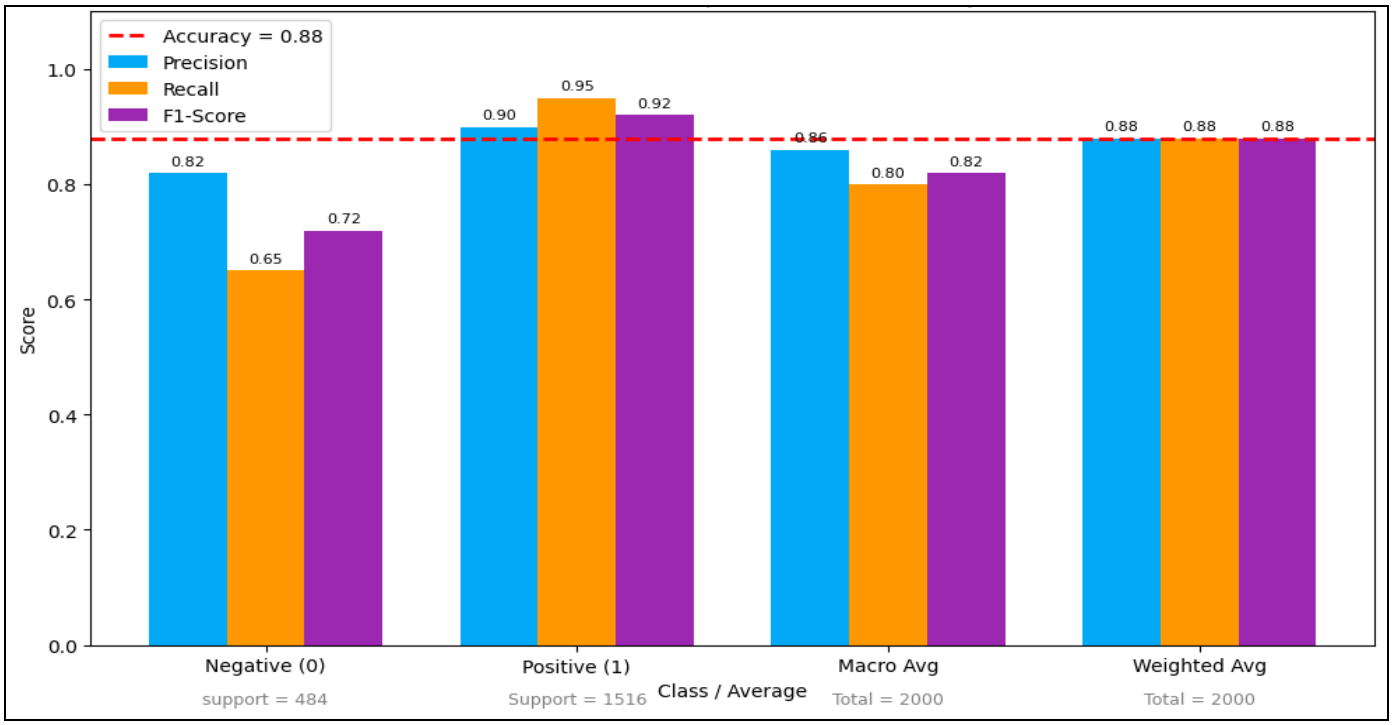

Fig 9 Ensemble Classification Report with Threshold Adjustment.

The ensemble model with threshold adjustment in Figure 9, demonstrates a marked improvement in handling class imbalance compared to the earlier results. For the negative class, precision (0.82) and recall (0.65) are substantially higher than before, leading to a stronger F1-score of 0.72, indicating that the model now recovers a meaningful portion of true negatives without excessively misclassifying positives. The positive class continues to perform strongly, with precision at 0.90, recall at 0.95, and an excellent F1-score of 0.92, showing that the adjustment preserved its predictive strength. Overall accuracy rises to 88%, but unlike the earlier misleading accuracy, this value better reflects balanced performance across both classes. The macro average scores (precision = 0.86, recall = 0.80, F1 = 0.82) suggest improved fairness between classes, while the weighted averages (0.88 across all metrics) confirm consistency given the larger support of positive reviews. These results highlight how threshold adjustment has mitigated the bias toward the majority class, yielding a more robust and reliable model for the sentiment classification.

➢ *Comparative Analysis*

This section compares the performance of the proposed approach with other methods using precision, recall, and the F1-Score as evaluation metrics. These metrics shows the performance of the metrics in classifying both positive and negative review. The precision, recall, and F1-score results for methods presented by [7,22,23], and contrasted below.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26feb1207

Table 2 Comparing the Performance of the Proposed Method with Other Approaches

Metrics	Naïve Bayes	ELMo Wikipedia & SentiCircle	Naïve Bayes & SentiWordNet	Ensemble Proposed Method
Precision	60.00	79.00	71.00	86.00
Recall	29.03	78.00	67.00	80.00
F1-score	20.20	80.00	66.00	82.00

The experimental results shows that the proposed method attained Precision 86%, Recall 80%, and F1 score 82%, outperforming several previously reported baselines. For example, compared with [23] Naïve Bayes & SentiWordNet, the proposed method improves precision by 15 points (71 → 86), recall by 13 points (67→80), and F1 by 16 points (66→82), Compared with [7] ElmoWikepedia & senticircle, observed improvements of 7, 2 and 2 points in precision, recall and F1, respectively. The gains relative to [22], Naïve Bayes classifier are larger; however, Kan et al. report an unusually low F1 (20.20) despite precision of 60, indicating either a different evaluation protocol (for example macro vs micro F1-Score) or substantial class imbalance in their evaluation dataset, therefore direct comparison is interpreted with caution. The balanced improvements in both precision and recall indicate that the proposed approach reduces both false positives and false negatives, yielding a substantially higher F1-score.

➢ *Test for Statistical Significance*

The Wilcoxon statistic is a non-parametric statistical test used to compare paired or independent samples when the data does not necessarily follow a normal distribution [24]. It is commonly used in machine learning and statistical analysis to compare the performance of two algorithms. The Wilcoxon Signed-Rank Test was used in this study to determine whether the performance of the proposed method and the previous method were statistically significant. This non parametric test is appropriate because it does not assume normality and is well suited for comparing paired data such as classification metrics across the same test folds. The Wilcoxon test was conducted between the proposed ensemble approach and Naïve Bayes methods using precision, recall, and F1-Score evaluation metrics. The Wilcoxon test results indicate a statistically significant difference between the proposed ensemble method and Naïve Bayes. Specifically, the results shows that the proposed ensemble outperform Naïve bayes statistically with a p-value of 1.91e-06 in terms of precision, recall, and F1-Score.

Furthermore, the proposed ensemble method outperformed both Naïve Bayes and the SentiWordNet method in terms of precision (p = 5.7220e-06). In terms of recall, the ensemble method outperformed Naïve Bayes and SentiWordNet with a p value= 0.32998. For the F1-score, the ensemble method significantly outperformed Naïve Bayes with a p value = 6.2943e-05. Finally, the Wilcoxon test results revealed that the proposed ensemble method significantly outperformed the hybrid Elmo-Wikipedia method in terms of precision with a p value = 1.91e-05.

➢ *Discussion of Comparative Analysis*

The comparative evaluation shows that the proposed ensemble method outperforms previously reported approaches in terms of precision, recall, and F1-score. The method achieved 86% precision, 80% recall, and an 82% F1-score, demonstrating a balanced performance that effectively reduces both false positives and false negatives.

When compared with Naïve Bayes & SentiWordNet[23], the ensemble achieved notable gains of +15 precision**,** +13 recall, and +16 F1-score. Improvements over ELMo Wikipedia & SentiCircle [7] were smaller but consistent, with +7 precision, +2 recall, and +2 F1-score**.** The largest differences were observed relative to the Naïve Bayes model [22], which reported a particularly low F1-score.

In summary, the proposed ensemble method consistently improves classification performance over existing baselines and demonstrates statistically validated advantages, particularly in precision and overall classification balance**.** These results highlight its effectiveness and reliability for sentiment analysis applications, where reducing both false positives and false negatives is essential.

➢ *Complexity Analysis*

The overall complexity of the sentiment analysis pipeline is dominated by the number of reviews (n), the average tokens per review (m), and the vocabulary size (V) from TF-IDF vectorization. Preprocessing with spaCy and SentiWordNet scoring both scale linearly with the total number of tokens, giving a time complexity of O(n·m). TF-IDF vectorization and dense feature combination contribute O(n·V) time and space complexity, which often dominates memory usage. Model training and prediction are relatively efficient, with MultinomialNB scaling as O(n_train·V) and GaussianNB as O(n_train·f), where f is the number of sentiment features. Overall, the pipeline's computational cost is primarily driven by text preprocessing and high-dimensional feature representation, while ensemble evaluation adds only minor overhead.

V. CONCLUSIONS

The comparative analysis demonstrates that the proposed ensemble method delivers clear improvements over existing sentiment classification approaches. With 86% precision, 80% recall, and an 82% F1-score, the ensemble achieved balanced performance, outperforming Naïve Bayes, Naïve Bayes & SentiWordNet, and ELMo Wikipedia & SentiCircle baselines. The gains were most pronounced against lexicon-based and traditional classifiers, though measurable improvements were also observed over previous methods.

The statistical evaluation using the Wilcoxon Signed-Rank Test confirmed that these improvements are significant, particularly for precision and F1-score, reinforcing the reliability of the ensemble's performance. While not all recall improvements were statistical significance, the method

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26feb1207

consistently showed stable classification across positive and negative reviews.

The ensemble approach demonstrates both effectiveness and robustness, offering a more balanced and statistically validated improvement over prior methods. These results establish the proposed method as a competitive alternative for sentiment analysis tasks.

REFERENCES

[1]. Rita, P.; Vong, C.; Pinheiro, F.; Mimoso, J. A sentiment analysis of Michelin-starred restaurants. EJMBE 2022, 32, 276-295.

[2]. Pan, M.; Li, N.; Huang, X. Asymmetrical impact of service attribute performance on consumer satisfaction: an asymmetric impact-attention-performance analysis. . Inf Technol Tourism 2022, 4, 221-243.

[3]. Hossain, N.; Bhuiyan, M.R.; Tumpa, Z.N. Sentiment Analysis of Restaurant Reviews using Combined CNN-LSTM. 2020, 1-6.

[4]. Luo, M.; Mu, X. A New Ontology for restaurant Review Sentiment Analysis. Proceedinds of the Association for information Science and Technology 2023, 60, 1068-1070.

[5]. Amalia, P.R.; Winarko, E. Aspect-based sentiment analysis on indonesian restaurant review using a combination of convolutional neural network and contextualized word embedding. . IJCCS (Indonesian Journal of Computing and Cybernetics Systems) 2021, 15, 285-294.

[6]. Abdullah, R.; Suhariyanto; Sarno, R. Aspect Based Sentiment Analysis for Explicit and Implicit Aspects in Restaurant Review using Grammatical Rules, Hybrid Approach, and SentiCircle. International Journal of Intelligent Engineering & System 2021, 295-305. .

[7]. Nurifan, F.; Sarno, R.; Sungkono, K.R. Aspect based sentiment analysis for restaurant reviews using hybrid elmo-wikipedia and hybrid expanded opinion lexicon-senticircle. . International Journal of Intelligent Engineering and Systems 2019, 12, 47-58. .

[8]. Govindarajan, M. Sentiment Analysis of Restaurant review using Hybrid classification method. . In Proceedings of the Proceedings of 2nd IRF International Conference, 2014; pp. 127-133.

[9]. Aye, Y.M.; Aung, S.S. Sentiment Analysis for Reviews of Restaurant in Myanmar Text. IEE COMPUTER SOCIETY 2017, 321-326.

[10]. Adnan, M.; Sarno, R.; Sungkono, M. Sentiment Analysis of Restaurant Review with Classification Approach in the Decision Tree-J48 Algorithm. 2019, 121-126.

[11]. Sharif, O.; Hoque, M.M.; Hossain, E. Sentiment Analysis of Bengali Texts on Online Restaurant Reviews Using Multinomial Na¨ıve Bayes. 1st International Conference on Advances in Science, Engineering and Robotics Technology 2019, 1-6.

[12]. Hossain, N.; Bhuiyan, M.R.; Tumpa, Z.N. Sentiment Analysis of Restaurant Reviews using Combined CNN-LSTM. . 2020, 1-6.

[13]. Ara, J.; Hasan, M.T.; Omar, A.A.; Bhuiyan, H. Understanding Customer Sentiment: Lexical Analysis of Restaurant Reviews. In Proceedings of the 2020 IEEE Region 10 Symposium (TENSYMP), 2020; pp. 295-299.

[14]. Abdullah, R.; Suhariyanto; Sarno, R. Aspect Based Sentiment Analysis for Explicit and Implicit Aspects in Restaurant Review using Grammatical Rules, Hybrid Approach, and SentiCircle. International Journal of Intelligent Engineering & System 2021, 295-305.

[15]. Eidul, T.S.; Imran, M.A.; Das, A.K. Restaurant Review Predictionusing Machine Learning and Neural Network. International Journal of Innovative Science and Research Technology 2022, 7, 1388-1392.

[16]. Eidul, T.S.; Imran, M.A.; Das, A.K. Restaurant Review Predictionusing Machine Learning and Neural Network. . International Journal of Innovative Science and Research Technology 2022, 7, 1388-1392.

[17]. Aruna, M.T.; Devi, P.M.; RenukaSai, M.; Raju, T.G.; Patchimala, M.C.; Kumar, K.S. Sentiment analysis for zomata using user comments review. UGC Care Group I Journal 2023, 13, 172-176.

[18]. Gujrati, M.S.; Tahakik, S.S.; Nahar, S.S.; Gujar, D.S.N. Sentiment analysis of restaurant reviews using machine learning. . International Research Journal of Modernization in Engineering Technology and Science 2023, 5, 8369-8372. .

[19]. Fragko, N.; Liapakis, A.; Ntaliani, M.; Ntalianis, F.; Costopoulou, C. A Sentiment Analysis Approach for Exploring CustomerReviews of Online Food Delivery Services. A Greek Case. PREPRINT 2024, 1-14.

[20]. Kaggle. restaurant-reviews. Available online: https://www.kaggle.com/datasets/joebeachcapital/restaurant-reviews (accessed on March 15).

[21]. Pant, V.K.; Sharma, R.; Kundu, S. An overview of stemming and lemmatization techniques. Advances in Networks, Intelligence and Computing 2024, 308-321.

[22]. Kan, Z.; Luliang, T.; Jie, G.; Chang, R.; Zhang, X.; Xue, Y. Detecting and evaluating urban clusters with spatiotemporal big data. Sensors 2019, 19, 461.

[23]. Rajeswari, A.; Mahalakshmi, M.; Nithyashree, R.; Nalini, G. Sentiment analysis for predicting customer reviews using a hybrid approach. In Proceedings of the 2020 Advanced Computing and Communication Technologies for High Performance Applications (ACCTHPA), 2020; pp. 200-205.

[24]. Sijtsma, K.; Emons, W. Nonparametric statistical methods. International encyclopedia of education 2010, 7, 347-353.