


Article

# Sentiment Analysis of Product Reviews Using Machine Learning and Pre-Trained LLM

Pawanjit Singh Ghatora <sup>1</sup>, Seyed Ebrahim Hosseini <sup>1,\*</sup> , Shahbaz Pervez <sup>1</sup> , Muhammad Javed Iqbal <sup>2</sup>  
and Nabil Shaukat <sup>3</sup> 

<sup>1</sup> Whitecliffe School of Information Technology, Auckland 1010, New Zealand; 20231893@mywhitecliffe.com (P.S.G.); shahbazp@whitecliffe.ac.nz (S.P.)

<sup>2</sup> Department of Computer Science, University of Engineering and Technology Taxila, Taxila 47050, Pakistan; javed.iqbal@uettaxila.edu.pk

<sup>3</sup> Institute of Design, Robotics and Manufacturing, School of Mechanical Engineering, University of Leeds, Leeds LS2 9JT, UK; n.shaukat@leeds.ac.uk

\* Correspondence: seyedh@whitecliffe.ac.nz

**Abstract:** Sentiment analysis via artificial intelligence, i.e., machine learning and large language models (LLMs), is a pivotal tool that classifies sentiments within texts as positive, negative, or neutral. It enables computers to automatically detect and interpret emotions from textual data, covering a spectrum of feelings without direct human intervention. Sentiment analysis is integral to marketing research, helping to gauge consumer emotions and opinions across various sectors. Its applications span analyzing movie reviews, monitoring social media, evaluating product feedback, assessing employee sentiments, and identifying hate speech. This study explores the application of both traditional machine learning and pre-trained LLMs for automated sentiment analysis of customer product reviews. The motivation behind this work lies in the demand for more nuanced understanding of consumer sentiments that can drive data-informed business decisions. In this research, we applied machine learning-based classifiers, i.e., Random Forest, Naive Bayes, and Support Vector Machine, alongside the GPT-4 model to benchmark their effectiveness for sentiment analysis. Traditional models show better results and efficiency in processing short, concise text, with SVM in classifying sentiment of short length comments. However, GPT-4 showed better results with more detailed texts, capturing subtle sentiments with higher precision, recall, and F1 scores to uniquely identify mixed sentiments not found in the simpler models. Conclusively, this study shows that LLMs outperform traditional models in context-rich sentiment analysis by not only providing accurate sentiment classification but also insightful explanations. These results enable LLMs to provide a superior tool for customer-centric businesses, which helps actionable insights to be derived from any textual data.

**Keywords:** sentiment analysis; machine learning; large language models (LLMs); OpenAI; GPT-4; text analysis; product reviews



**Citation:** Ghatora, P.S.; Hosseini, S.E.; Pervez, S.; Iqbal, M.J.; Shaukat, N. Sentiment Analysis of Product Reviews Using Machine Learning and Pre-Trained LLM. *Big Data Cogn. Comput.* **2024**, *8*, 199. <https://doi.org/10.3390/bdcc8120199>

Academic Editor: Domenico Ursino

Received: 7 October 2024

Revised: 24 November 2024

Accepted: 10 December 2024

Published: 23 December 2024



**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

With the growing dependence on digital data, developing effective and scalable sentiment analysis systems is vital for any organization, whether business- or product-focused. Analyzing sentiment is also crucial for businesses of all sizes, in knowing and promptly responding to customers' feedback or opinion on several topics. It plays an important role in various areas such as marketing, customer support, and product development by helping businesses accurately understand customer sentiments based on their feedback [1]. Social platforms help professionals from various fields (business, marketing, sports, politics and product evaluation, etc.) rely on available approaches capable of analyzing user sentiments on particular topic of interest [2,3]. This helps in better decision-making and in monitoring and shaping a brand image through online platforms. Effective utilization

of sentiment analysis methods allows companies to preemptively identify and address potential issues, thereby promoting a positive brand image. This research focuses on the utility of sentiment analysis for product-based organizations, highlighting its significance in managing customer relationships based on product reviews and feedback, which can help in changing product strategies and improvements accordingly.

In the digital era, the complexity and volume of online data have grown exponentially, driven by advancements in digital technologies. The types of data gathered online range from various product reviews on e-commerce platforms to comments on social media and discussions on blogs, which pose various types of challenges. These data are often unstructured, sprawling across various formats and languages, including informal slang, which complicates the sentiment analysis process [4]. The volume of these data continues to increase, making manual analysis impractical. These challenges underscore the need for robust, automated sentiment analysis systems that can adapt and scale to keep pace with the evolving digital landscape. These automated systems are essential for businesses to harness insights into consumer behavior, market trends, product reviews, and overall brand perception, which are critical for strategic decision-making.

There are plenty of use-cases across the field, but to narrow down the scope, few use-cases related to product-based organizations are taken into consideration. In the perspective of variable sizes of organizations, sentiment analysis emerges as a crucial tool in this context for several reasons. Sentiment analysis is widely used in various sectors, including marketing, customer support, and product design, as it plays a vital role in understanding customer views and responses [5]. By examining online discussions surrounding brands, companies can preemptively spot and tackle potential issues, thereby developing a favorable brand image [6,7]. This research focuses on the following tasks:

1. Comparing the effectiveness of traditional machine learning models and pre-trained LLMs to analyze sentiment of different product reviews. This involves model performance evaluation based on metrics such as accuracy, precision, recall, and F1 score to identify each model's strengths and limitations.
2. Assessing the ability of LLMs to interpret complex, context-rich text and provide detailed sentiment insights. Specifically, we investigate LLMs' capacity to capture subtle language cues and classify sentiments beyond basic positive, negative, and neutral categories.
3. Evaluating the capabilities of LLMs in providing detailed insights and explainability in sentiment analysis.

To achieve the above, we conducted a series of experiments with both traditional machine learning and LLM-based models on a Flipkart dataset of product reviews. Our results reveal that Support Vector Machines (SVMs) demonstrate high efficiency to analyze shorter context text, which enables models to effectively capture sentiment in short-length comments or reviews. However, we find that LLMs significantly improve classification accuracy for longer, context-rich text. This enables businesses to extract significant insights from customer reviews. Ultimately, this study advances sentiment analysis techniques by allowing us to select the most suitable model based on the complexity and the analytical detail required in specific business contexts.

The remainder of this paper is organized as follows: Section 2 presents a comprehensive literature review on sentiment analysis in consumer reviews, outlines key techniques, models, and research findings. Section 3 provides details about the materials and methods employed in this study. Section 4 discusses the results obtained, followed by an in-depth analysis and discussion of their implications. The novelty and contributions of the proposed approach are highlighted in Section 5. Section 6 includes key findings, limitations, and potential future research directions, and the last section presents the conclusion of this work.

## 2. Literature Review

This section analyses the key concepts and methodologies within Natural Language Processing (NLP) as those applied to sentiment analysis of different product reviews.

NLP is a subset of computer science connected with computational linguistics [8]. It enables seamless communication between humans and machines by teaching computers to understand and interpret human language; hence, through NLP, machines can process and analyze text, offering insights and responses to users [9]. NLP offers the potential to develop models and processes capable of extracting information from both text and audio data [8].

Sentiment analysis represents a branch within the evolution of text mining technology, focusing on extracting opinions from textual content [10]. The main objective of sentiment analysis is to determine the sentiment polarity in natural language texts [11], and that can be performed using machine learning classification techniques.

Machine learning is a subset of artificial intelligence (AI), characterized by a machine's capacity to replicate human-like intelligence. AI systems are engineered to solve complex problems by mimicking the strategies humans use to navigate difficulties [12]. There are already incredible applications such as autonomous vehicles, natural language processing, and facial recognition systems utilizing machine learning techniques for its operations [12]. Hence, machine learning provides an effective method of performing sentiment analysis using natural language processing techniques.

Large Language Models (LLMs) like ChatGPT are described as powerful machine learning models capable of generating realistic and meaningful text [13]. These models can perform sentiment analysis as a core task in marketing for understanding consumer emotions and opinions [14].

### 2.1. Analysis of Existing Research

The following subsections outline two different methodologies for sentiment analysis. The first relates to the traditional approach of training and testing machine learning models. The second, which is gaining popularity, involves the use of pre-trained LLMs that can be directly applied to sentiment analysis tasks.

#### 2.1.1. Traditional Approach of Training and Testing Machine Learning Models

Medhat et al. conducted an extensive survey of sentiment analysis methodologies and highlighted the strengths and limitations of each [15]. Their study compared machine learning approaches (e.g., support vector machines, Naive Bayes) with lexicon-based techniques, concluding that hybrid methods often yield superior results. Araque et al. proposed an ensemble approach combining classic machine learning algorithms and deep learning for improved sentiment classification accuracy [16]. This work concluded that hybrid models leveraging word embedding (e.g., Word2Vec, GloVe) offer substantial improvements over standalone methods.

This work is based on the implementation of the Bidirectional Encoder Representations from Transformers (BERT) to enhance natural language understanding. Moreover, this work has shown a strong ability to grasp context and their significance [6]. The study of suspicious online reviewers is critical for e-commerce platforms, employing data analysis to identify fake reviewers [17]. The analysis of hotel reviews employs a range of machine learning models to accurately categorize sentiments, proving invaluable for the hospitality industry [18]. Similarly, product review analysis incorporates diverse analytical techniques, enhancing consumer insight for retailers. The analysis of online movie reviews targets the entertainment sector, offering insights that can significantly impact marketing strategies [19]. Lastly, but not least, Twitter sentiment analysis focuses on brand perception on social media, providing key insights for brand management and marketing [20].

This study provides an overview of NLP techniques for sentiment analysis using pre-trained models like BERT and GPT [21]. The work highlights their effectiveness in

understanding nuanced sentiments compared to traditional ML models. The paper also examines the computational trade-offs involved in using such models.

Existing studies address specific challenges and opportunities for potential future enhancements. The BERT implementation, while powerful, struggles with high computational demands [6]. Improvements include optimizing these models for broader accessibility and suggesting a need for continuously adaptive algorithms [17]. In sentiment analyses of hotel and product reviews, the challenge lies in processing complex and large datasets while accurately interpreting diverse linguistic expressions [18]. Future improvements could integrate more sophisticated linguistic and contextual analyses to enhance accuracy. For movie reviews and Twitter analysis, the rapid evolution of language and mixed sentiments present ongoing challenges, with potential solutions including the incorporation of multi-modal data (such as images, audio and videos) and more dynamic models to better capture the wide range of user sentiments.

### 2.1.2. Pre-Trained LLMs

Bellar et al. discussed the importance of sentiment analysis in e-commerce using deep learning models such as CNN, RNN, and Bi-LSTM, together with advanced embeddings like BERT, FastText, and Word2Vec [22]. It highlighted the usefulness and comparison of model performance on Woman Clothing Reviews datasets through both 5-class and 3-class classification experiments. The mentioned deep learning models show convincing performance in analyzing sentiment of product reviews.

Zhang et al. introduced deep learning techniques, specifically convolutional neural networks (CNNs) and recurrent neural networks (RNNs), for sentiment analysis of product reviews [23]. Their experiments demonstrated that deep learning models significantly outperform traditional machine learning methods in capturing semantic nuances. Devlin et al., proposed BERT which revolutionized sentiment analysis by pre-training on a large corpus and fine-tuning for specific tasks [24]. Their experiments demonstrated BERT's superior performance in understanding contextual sentiment in product reviews, outperforming earlier models on benchmark datasets.

Sun et al. extended BERT's application by exploring domain-specific fine-tuning for sentiment analysis [25]. They showed that fine-tuning BERT on review-specific datasets yielded significant improvements in accuracy and F1 scores, showcasing the model's adaptability to nuanced language in product reviews.

Krugmann and his research team aims to evaluate the proficiency of LLMs, specifically GPT-3.5, GPT-4, and Llama 2, in sentiment analysis within the context of marketing research, comparing their performance against high-performing transfer learning models [14,26]. Krugmann and Hartmann's research paper stands out by exploring the zero-shot capabilities of LLMs like GPT-3.5, GPT-4, and Llama 2 in sentiment analysis, a novel approach compared to traditional models that often require machine learning model training and testing [14]. The strength of the paper lies in its demonstration that LLMs can achieve comparable or even superior accuracy in sentiment classification tasks without the need for extensive pre-training specific to the task. Furthermore, the paper highlights the exceptional performance of LLMs in providing explainable outputs, which is crucial for applications requiring transparency in AI-driven decisions, which is major challenge in traditional AI/ML approaches. This paper highlights challenges with LLMs for sentiment analysis, observing their inconsistent performance across different types of textual data, i.e., structured and unstructured data [14]. LLMs produced better results when performing sentiment analysis of structured data like online reviews. LLMs struggle.

The existing research on sentiment analysis using benchmark data sets highlights several limitations. One major issue is the lack of dataset diversity, with many studies focusing on a few domains, limiting generalizability. On the other hand, handling noisy and imbalanced data, including typos, sarcasm, and informal language, remains a persistent challenge. Advanced deep learning techniques like transformers (e.g., BERT and GPT) often achieve high accuracy but are resource-intensive and lack interpretability. The review

also highlights the evolution of sentiment analysis methodologies from traditional machine learning approaches to the use of pre-trained LLMs. Traditional models are effective when combined with hybrid techniques, face challenges in handling complex and large datasets as well as adapting to the dynamic nature of language. On the other hand, pre-trained LLMs such as BERT and GPT-3.5 represent a major breakthrough, offering superior contextual understanding and adaptability through fine-tuning and zero-shot capabilities. However, these models also encounter issues, including high computational demands and inconsistent performance with unstructured data. This comprehensive review indicates that while LLMs provide significant advantages over traditional models, ongoing research is needed to address their limitations, optimize their computational efficiency and enhance their robustness across diverse types of textual data.

## 2.2. Theoretical Framework

In today's competitive market, organizations heavily rely on order numbers and statistics to estimate their success. However, an equally critical yet often overlooked aspect is understanding customer sentiment related to their products and services. Given the vast scale of information available across various channels, manually reading and interpreting all comments and feedback becomes a challenge. This is where the field of data science, particularly its subset machine learning, becomes invaluable.

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data. It encompasses a wide range of techniques from statistics, data analysis, and computer science to understand and analyze various situations or scenarios with data. Machine learning, a key part of data science, leverages algorithms to use data, learn from it, and then make predictions. By automating the analysis process, machine learning enables organizations to efficiently process and analyze vast amounts of feedback, identifying patterns and sentiments that may not be immediately noticeable.

This research focuses on to investigate the effectiveness of machine learning approach and LLMs for sentiment analysis of product reviews [27]. The approach and conclusion derived from this research can be adopted and applied in any product-based organization. Implementing machine learning techniques and LLMs for sentiment analysis allows organizations to gain deep insights into how customers perceive their products. This can reveal areas that need improvement, highlight what is working well, and provide a better understanding of customer needs and preferences. Ultimately, leveraging machine learning and LLM for sentiment analysis not only enhances customer satisfaction but also drives strategic business decisions, developing a more customer-centric approach in today's data-driven world. Hence machine learning and LLM are leveraged for experimentation in this research. The following high-level strategy is followed for this research, as shown in Figure 1.

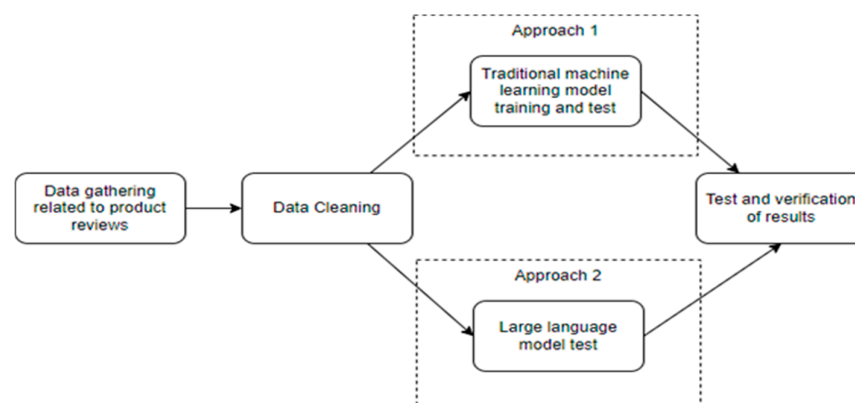


Figure 1. High level strategy.



Data were gathered related to product reviews. The objective was to prepare a dataset that was diverse and representative of the different sentiments (e.g., positive, negative, and neutral) customers express about products. This diversity helps in training a robust machine learning model.

Once the data were collected, the next crucial step was to clean it. Data cleaning is performed to remove redundant and not required data (such as HTML tags, irrelevant symbols, or irrelevant text segments) and standardize the format to ensure that the data used for training and testing machine learning models are of high quality to achieve the desired outcome.

The following two approaches were followed for sentiment analysis on product reviews:

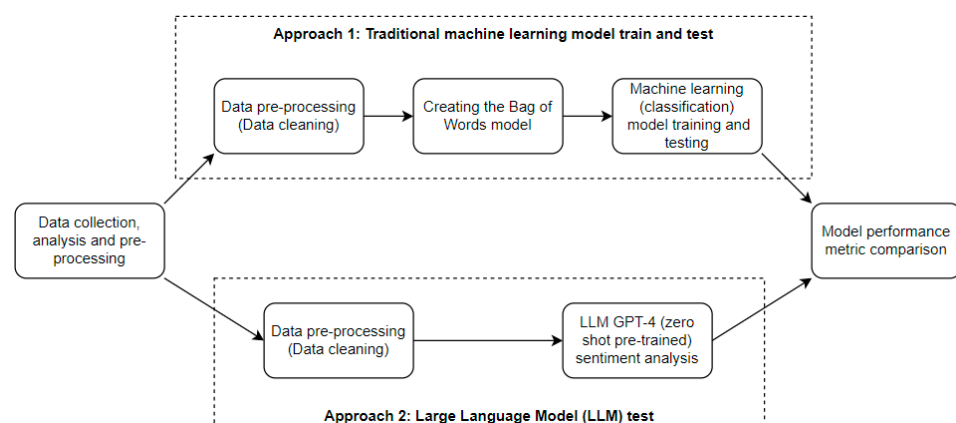
- Machine learning models used for classification were leveraged in this research and were trained to classify the sentiment of product reviews. Multiple models were used for this purpose, i.e., Random Forest, Naive Bayes and Support Vector Machines (SVM). These models were trained on a subset of the dataset (training set) and validated on another subset (i.e., test set) to verify the model performance.
- LLM for sentiment analysis: An LLM, specifically OpenAI GPT-4, was utilized for the sentiment analysis classification. This powerful model is pre-trained and leverages advanced machine learning techniques to accurately interpret and classify a wide range of emotions and sentiments expressed in textual data. By deploying OpenAI GPT-4, we aimed to achieve a deeper and more nuanced understanding of customer sentiments, enhancing our ability to derive meaningful insights from product review dataset.

Test and verification of the machine learning models and LLM were performed, and the model performance was compared on the basis performance metrics, i.e., Accuracy, Precision, Recall and F1 Score, to identify which machine learning model performs better in predicting sentiments.

This approach integrates data science and machine learning techniques to effectively analyze sentiments expressed in product reviews, providing valuable insights that can inform business strategies, product improvements, and customer service practices.

### 3. Materials and Methods

This section presents the strategy utilized for conducting sentiment analysis on product reviews. Additionally, the methodology is shown graphically in Figure 2, and each step is detailed further to enhance understanding.



**Figure 2.** Proposed methodology.

As illustrated above, two distinct experimental approaches were adopted. The first approach involves the traditional machine learning methodology, where models are trained and tested to evaluate their performance accurately. The second approach leverages pre-trained LLMs, which offers the advantage of utilizing sophisticated, already-trained models to bypass the lengthy and resource-intensive training phase. This approach can

significantly accelerate the analysis process and improve scalability. Despite the differences in these methodologies, both approaches share a common initial step, i.e., the collection of data. This critical phase ensures that both the traditional models and the pre-trained LLMs operate on the same robust and comprehensive dataset, facilitating a comparison of their respective capabilities in sentiment analysis.

### 3.1. Data Acquisition, Analysis and Pre-Processing

Data acquisition is the key step because the quality of the data will influence the results derived from it. The intent was to experiment and see the effectiveness of machine learning techniques in sentiment analysis and that can also be achieved using a public dataset. Hence, a public dataset from Kaggle was leveraged, which was taken from Kaggle, which was created by Vaghani and Thummar [28]. The machine learning approach was generalized, and this can be leveraged on private datasets in the future. Data extraction was performed on Kaggle, and it was in comma-separated values (CSV) format. The public dataset prepared by Vaghani and Thummar is referred to as Kaggle in this research analysis [28].

#### 3.1.1. Dataset Features

This dataset includes the reviews related to different type of products such as such as electronic items, clothing of men, women and kids, home decor items, automated systems and others. The following are the six columns or features in the dataset:

- i. Product name: Name of the product.
- ii. Product price: Price of the product.
- iii. Rate: Customer's rating on product (between 1 to 5).
- iv. Review: Customer's review of each product.
- v. Summary: This column includes descriptive information of customers' thoughts on each product.
- vi. Sentiment: This column contains 3 labels, Positive, Negative and Neutral (which was given based on summary).

The data type of various features is stated as follows:

- i. Product name: String.
- ii. Product price: Integer.
- iii. Rate: Integer.
- iv. Review: String.
- v. Summary: String.
- vi. Sentiment: String.

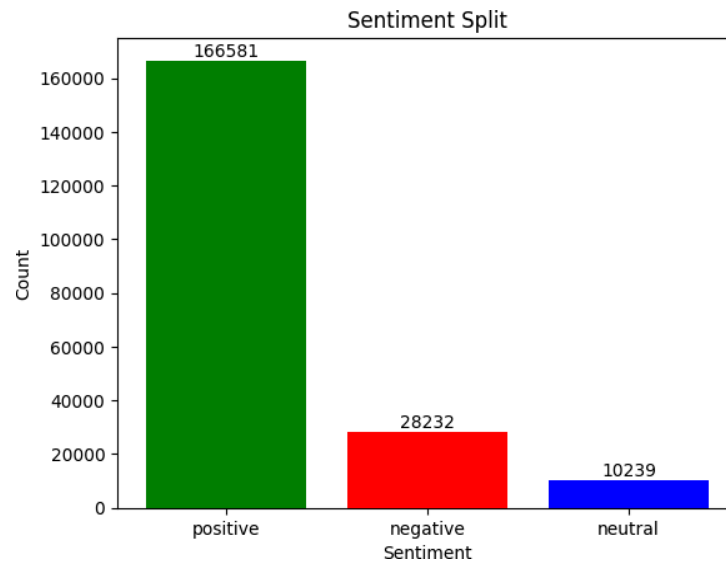
#### 3.1.2. Dataset Description

This dataset has 205,052 records, and it is a diverse dataset covering negative, neutral and positive sentiments, and its record count is shown in Table 1.

**Table 1.** Dataset record counts split by sentiment.

Sentiment	Count
Negative	28,322
Neutral	10,239
Positive	166,581
Grand Total	205,052

Graphical presentation of the dataset counts is shown in Figure 3.



**Figure 3.** Dataset record counts split by sentiment.

### 3.1.3. Dataset Analysis

The following are the favorable observations related to the dataset:

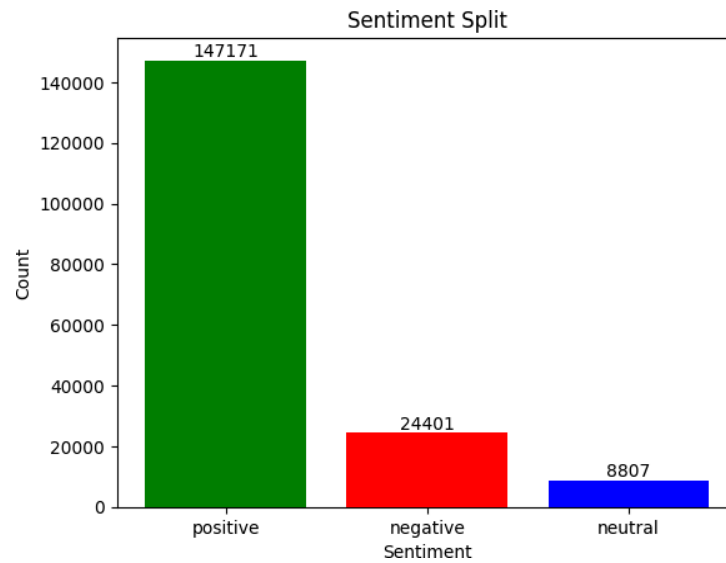
- The dataset has a high record count, i.e., 205,052, which will be good for training
- “Review” and “Summary” are the main features because this feature has customer comments that can be easily used to deduce the customer sentiments about the product. This text-based feature holds the qualitative data expressed by customers in their own words, providing exact and clear understanding of their sentiments. These features are selected based on the importance of these in the consumer review dataset to know about customer feedback.
- This dataset does not have any personal identifiable information; hence, it does not pose any challenges related to privacy violation.
- Drawbacks related to the dataset and how those are addressed are as follows:
- The dataset has a high record count of positive sentiments as compared to other sentiments (i.e., negative and neutral); hence, this can lead to biased or inaccurate predictions, hence impacting the model performance. This was addressed by taking an equal record count for each sentiment.
- The datasets have product names but do not have categories; hence, it was hard to map the sentiment directly to the product category to deduce the areas of improvements related to each product, but that is not a major limitation to this study because this research is about proving the efficacy and usability of machine learning and LLM techniques for sentiment analysis.

### 3.1.4. Data Pre-Processing

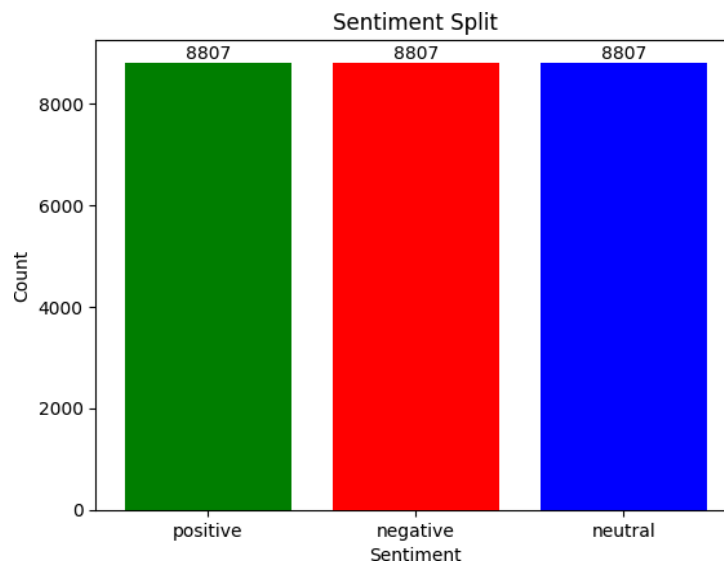
Reviews were the primary data used for sentiment analysis. However, the dataset contained 24,664 records with null entries in the review field. These null records were removed to ensure the integrity and effectiveness of the sentiment analysis process. Retaining such records would pose challenges in both machine learning and LLM execution, potentially leading to inaccuracies or errors during model training and analysis. By eliminating these null entries, the dataset was optimized for more accurate and reliable sentiment analysis. Record counts after removing null records are shown in Figure 4.

As is evident in Figure 5, neutral sentiments had the lowest record count, i.e., 8807; hence, the same number of records was taken for other sentiments, i.e., positive and negative, so that there was no data bias. The record counts are shown in Figure 5 after removing data bias.





**Figure 4.** Dataset record counts split by sentiment without null records.



**Figure 5.** Dataset record counts after addressing bias.

### 3.2. Approach 1: Traditional Machine Learning Model Training and Test for Sentiment Analysis on Product Reviews

This section explains the experimentation steps and outcome of the traditional machine learning model training and test approach followed for sentiment analysis on product reviews.

#### a) Data cleaning

This step involves data pre-processing, i.e.:

- Tokenizing involves splitting a sequence of characters or sentences into individual words. By transforming sentences into tokens, where each token is assigned a unique identifier, the complexity of the text data is reduced to more manageable units.
- Normalization focuses on refining a sentence by removing extraneous elements such as special characters, hashtags, and URLs. This process is critical in preparing text data by eliminating or modifying parts that are incorrect, incomplete, or irrelevant.
- Stemming reduces words to their root form, addressing different morphological variants of a word. This process simplifies complex words to their basic forms, thereby reducing the variety of word forms present in the text.

- Stop words removal leads to removing stop words that appear too frequently in text is essential because these words often carry minimal meaningful content and can skew the analysis of text data.

b) Creating the Bag of Words model

The bag of words approach helps with sentiment analysis by converting text into a numerical structure that machine learning systems can understand. This is done through vectorization, where text is changed into numerical vectors, with each different term being a separate attribute and the related number showing the term's frequency in the text. These numerical representations then become the input data for machine learning model training and testing [29,30]. This vectorization translates text into a numerical structure where each dimension of the vector represents a specific word from the dictionary, and the value in each dimension represents the frequency of that word in the document.

c) Machine Learning Model Training

After data are pre-processed, data are used for training machine learning models such as random forest, naïve Bayes and support vector machine for sentiment analysis [31]. Multiple machine learning techniques were leveraged for sentiment analysis because there are unique strengths and approaches of each algorithm to ensure comprehensive and effective analysis. The following is the rationale why the following machine learning based methods were chosen:

- Random Forest

- (a) Strengths: This method utilizes an ensemble of decision trees to ensure a robust and stable prediction, mitigating the risk of overfitting that is typical with single decision trees. Its ability to handle large data sets and high dimensionality makes it suitable for the diverse and complex data structures encountered in NLP.
- (b) Application in Sentiment Analysis: Random Forest can efficiently process textual data to determine sentiments by capturing intricate patterns in the data, which may not be linearly separable. The ensemble approach allows for a majority voting system, enhancing the accuracy of sentiment classification.

- Naive Bayes

- (a) Strengths: This probabilistic classifier is based on Bayes Theorem and is highly efficient with large volumes of data. Naive Bayes excels in calculating the likelihood of outcomes based on the presence of features, making it particularly suited for text classification.
- (b) Application in Sentiment Analysis: It assesses the likelihood of sentiments based on the frequency of words, making it highly effective for analyzing texts where the presence of certain words strongly indicates a particular sentiment.

- Support Vector Machine (SVM)

- (a) Strengths: SVM can create a clear boundary between data classes, even in high-dimensional spaces, which is common in text data. It's known for its effectiveness in handling complex classification problems with clear margin separation.
- (b) Application in Sentiment Analysis: SVM's ability to find the optimal hyperplane allows it to classify intricate and subtle variations in text data, making it highly effective for distinguishing between positive, negative and neutral sentiments, even when the differences are not immediately obvious.

### 3.3. Approach 2: LLM Test for Sentiment Analysis on Product Reviews

The primary purpose of employing an LLM like GPT-4 for sentiment analysis on product reviews is to leverage its advanced natural language understanding capabilities. GPT-4 can interpret and analyze the nuances of human language found in product reviews, allowing it to accurately identify and categorize sentiments expressed by customers, ranging from positive, negative, to neutral. Another significant benefit of using GPT-4, one

of the most advanced LLMs currently in the market, is its accessibility and ease of use. GPT-4 can be easily integrated through APIs that allow developers to access its capabilities without needing to manage the complexities of model training or infrastructure. This ease of integration accelerates the deployment of advanced NLP features. Being at the forefront of AI research, GPT-4 benefits from continuous updates and improvements from OpenAI, ensuring that users always have access to the most advanced tools for their NLP needs.

### 3.3.1. Data Cleaning

- Remove brackets
  - Cleaning out this non-essential text helps focus the analysis on the content of interest and prevents irrelevant data from skewing the results.
- Remove URLs
  - Eliminating URLs simplifies the text and reduces noise, making the dataset cleaner for text processing tasks like sentiment analysis or topic modeling.
- Remove HTML tags
  - This step prevents any markup language from being interpreted as content, thereby maintaining the integrity of the text data.
- Remove new lines
  - This ensures that the text is in a single continuous block, which is easier to process and analyze for patterns or sentiments without unnecessary segmentation. The following subsection explains the experimentation steps and outcome of the LLM test approach followed for sentiment analysis on product reviews.

### 3.3.2. LLM (OpenAI GPT-4): Sentiment Analysis

The OpenAI GPT-4 model was utilized to perform sentiment analysis on product reviews. This process involves submitting the text to be analyzed directly to the model, accompanied by a specific prompt: “What is the sentiment? (Positive, Negative, or Neutral)”.

Based on the input text, GPT-4 evaluates the content and context, subsequently providing a classification that identifies the sentiment of the review. This sophisticated analysis allows for an understanding of the emotional tone conveyed in the text, whether it is positive, negative, or neutral.

### 3.4. Model Performance Metric Comparison

The performance of various machine learning models will be compared using the performance metrics, i.e., Accuracy, Precision, Recall and F1 score.

Accuracy of a model is the total number of correct predictions divided by the total number of predictions [32,33].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

where, TP = True Positives, TN = True Negatives, FP = False Positives, FN = False Negatives.

Precision is the number of true positive results divided by the number of all positive results [34].

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

Recall (Sensitivity) is the number of true positive results divided by the number of all samples that should have been identified as positive [32,34].

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

F1 Score is the harmonic mean of precision and recall [32,34]

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

#### 4. Results and Discussion

Sentiment analysis was conducted through two separate experiments, each focusing on different textual features, i.e., the "Review" field and the "Summary" field. The "Review" field typically contains brief, high-level comments, while the "Summary" field provides more detailed descriptions of these comments. These features are chosen based on their significance and nature of the problem. These experiments were designed to test and evaluate the effectiveness of two distinct approaches, i.e., traditional machine learning techniques and LLMs. The objective was to determine which approach is better suited for analyzing sentiment in each type of text. The performance of each method is evaluated based on key metrics: Accuracy, Precision, Recall, and F1 Score.

##### 4.1. Sentiment Analysis on "Review" Feature

The comparative analysis of these approaches is presented in Table 2. The results clearly demonstrate that when the text size is extremely brief, consisting only of a few words rather than complete sentences, the SVM shows superior efficacy compared to other machine learning and LLM approaches. Although the LLM was employed with zero-shot techniques, it still delivered promising results, closely trailing the performance metrics of the SVM. The macro average of performance metrics was calculated and compared because it helps to reveal whether the model performs uniformly across all classes. It is calculated by first determining the performance metric (like precision, recall, or F1 score) for each class independently and then taking the average of these values.

**Table 2.** Comparison of different machine learning techniques for sentiment analysis on "Review" feature.

	Random Forest	Naive Bayes	SVM	LLM (GPT-4)
Accuracy	0.680	0.449	0.682	0.635
Macro-average Precision	0.677	0.471	0.678	0.630
Macro-average Recall	0.680	0.446	0.682	0.635
Macro-average F1 Score	0.672	0.347	0.673	0.612

To eliminate data bias, an equal number of records for each sentiment category was selected. Since the neutral sentiment category had the lowest record count at 8807, this number was used as the standard for sampling records in the positive and negative categories as well. This approach ensures that the analysis is balanced and free from skew due to uneven data distribution.

##### 4.2. Sentiment Analysis on "Summary" Feature

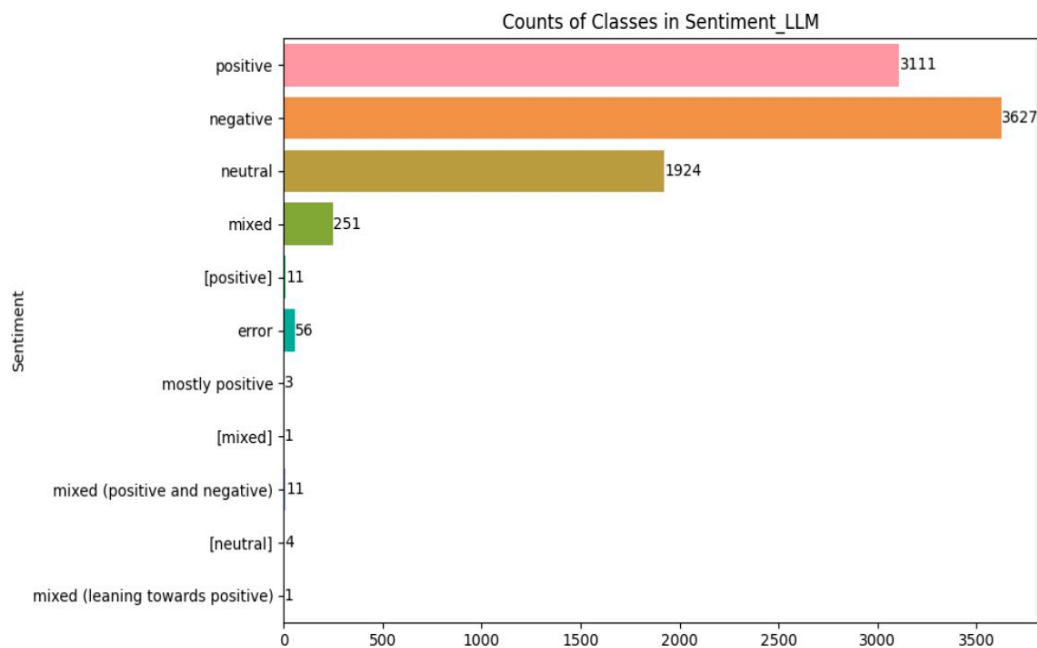
The comparative analysis of this approach is detailed in Table 3. The results clearly indicate that when dealing with longer texts, which provide more context and consist of full statements rather than brief snippets, LLMs demonstrate superior efficacy compared to other machine learning approaches. This is particularly notable even when zero-shot LLM is employed, highlighting its robustness in handling more complex textual data.

To eliminate data bias, an equal number of records 3000 each for the negative, neutral, and positive sentiment categories was selected. This strategy ensures that the analysis remains balanced and free from skewness caused by uneven data distribution. The decision to standardize the record count was also influenced by cost considerations. Analyzing longer texts with the OpenAI GPT-4 model incurs higher costs, so limiting the number of records analyzed was a practical approach to managing expenses while maintaining data integrity.

**Table 3.** Comparison of different machine learning techniques for sentiment analysis on “Summary” feature.

	Random Forest	Naive Bayes	SVM	LLM (GPT-4)
Accuracy	0.767	0.426	0.768	0.816
Macro-average Precision	0.766	0.508	0.768	0.823
Macro-average Recall	0.768	0.421	0.769	0.814
Macro-average F1 Score	0.766	0.380	0.768	0.806

During this experiment, several additional insights related to the execution of text analysis with the LLM were observed which is also shown graphically in Figure 6:



**Figure 6.** LLM classification on Summary feature.

1. The “Summary” feature, which contains longer statements providing more context and insights was analyzed by the LLM. The model did more than just classify the text into three sentiments, i.e., negative, neutral, and positive. Instead, it also identified a “mixed” category. This category captures instances where customers mentioned both positive and negative aspects, which does not necessarily equate to a neutral sentiment. The model further determined the predominant sentiment in mixed cases, whether it leaned more towards positive or negative.
2. While using the LLM for sentiment analysis via API calls executed through Python scripts, some records resulted in errors and were subsequently excluded from the performance metric calculations. This step ensured that only successful analyses contributed to the final evaluation.
3. Some responses from the LLM contained unwanted characters, such as square brackets. These characters were removed before calculating the performance metrics to maintain the integrity and accuracy of the data analysis.

#### 4.2.1. LLM Explainability

LLMs also excel beyond traditional machine learning approaches in terms of explainability. When prompted, LLMs can articulate the reasoning behind their classifications, offering clear and detailed explanations for their decisions and allowing users to understand the logic applied in the sentiment analysis process. Examples of such explanations are shown in Table 4, and demonstrate how LLMs provide insights that go beyond do-

ing classification, thereby adding value through transparency and trust in automated decision-making.

**Table 4.** Examples of LLM explanation on “Summary” feature classification.

Summary	LLM Classification	LLM Explanation
Not saying too bad but after using 1 week the product is not working then after someday it will on, I buy this for hostel use but anyway not recommending	negative	The reviewer expresses dissatisfaction with the product, mentioning that it stopped working after only one week of use. They also mentioned that the product occasionally turns on. The product was specifically purchased for use in a hostel, but it did not meet their expectations. The reviewer ends the review by suggesting others not buy the product, which confirms the negative sentiment.
I use this camera for more than 2 months now it's simple and very comfortable to use for beginners	positive	The review is positive as the consumer has been using the camera for more than two months, suggesting satisfaction and consistent usability. They mention that the camera is simple and very comfortable to use, specifically praising its user-friendliness for beginners. This implies that they found the product easy to use and accommodating for individuals who may not be familiar with camera gear, hinting at its simplicity of design and function. Hence, the user's overall experience with the camera has been positive.
Doing the job faster than other wet grinders slightly getting overheated	mixed	The text presents a mixed sentiment, the user appreciates the product's ability to work more quickly than other wet grinders, which is a positive sentiment reflecting satisfaction with the product's efficiency. However, there is also dissatisfaction, a negative sentiment expressed, as the user mentions that the product gets slightly overheated. This could imply potential disappointment if the overheating issue is serious enough to affect performance or longevity.

LLMs demonstrated superior capabilities in explainability, a critical aspect for adopting AI technologies in business settings. They provided clear, articulate explanations for their sentiment classifications, enhancing the transparency and trustworthiness of automated sentiment analysis.

The result in the above experiments demonstrates that when LLMs are employed with a zero-shot learning approach, they hold significant potential for sentiment analysis. Particularly when analyzing texts that are longer and provide more context, LLMs can perform classifications beyond the standard categories. They can accurately identify nuanced sentiment categories such as 'mixed' sentiments, which involve both positive and negative elements. In scenarios involving complex and detailed text, LLMs not only exceed the capabilities of traditional machine learning models but also enhance the depth and granularity of sentiment analysis.

#### 4.2.2. Performance Comparison Between Traditional ML Models and Pre-Trained LLMs

Based on the above experimentation and analysis, the summary of performance comparison between traditional ML models and pre-trained models is the following.

The preprocessing steps for traditional models and LLMs for sentiment analysis of consumer reviews differ significantly. Traditional ML models mainly focus on lightweight steps, such as: text tokenization, stop-word removal, stemming or lemmatization, feature extraction. These methods are computationally efficient but often fail to capture semantic nuances and context which limit their ability to handle complex sentences. The LLM models rely on minimal preprocessing, such as: data cleaning, tokenization, special tokens, lowercase/punctuation handling, standardized input. The LLMs prefer to use pre-trained embedding techniques such as: word2vec, fasttext, GloVe and ELMo. These embedding techniques are computationally intensive due to their transformer-based architecture.



Key findings in terms of experimental results indicate that traditional machine learning techniques like SVMs perform well on small or concise texts datasets with simple patterns but lack the nuanced understanding that LLMs bring to longer, context-rich text. These models perform faster training but have limited accuracy on complex datasets. However, these often struggle with context and word disambiguation. Pre-trained LLMs, particularly when using zero-shot learning, not only matched but occasionally surpassed traditional methods in accuracy and detail, proving especially effective in interpreting complex sentiments and providing detailed classifications beyond simple positive, negative, or neutral categories. These models capture contextual information, handle long and nuanced sentences effectively. Moreover, these show better generalization to unseen data due to their pre-training on massive data and computationally expensive but provide state-of-the-art results.

The computational requirements of traditional machine learning models for consumer sentiment analysis, such as SVMs, Naive Bayes, and random forest require moderate computational resources. These models rely on less computationally expensive feature extraction techniques like TF-IDF or Bag-of-Words but may struggle with contextual understanding. This makes these more suitable for small-scale applications or resources constrained environment. On the contrary, LLMs like GPT demand significant computational resources due to their dependence on large-scale pre-training, fine-tuning, and the use of transformer architectures with billions of parameters. During training, these models require high-performance GPUs and large memory capacities which make them resource intensive. However, fine-tuned pre-trained LLMs require considerably fewer computational resources than training models from scratch. Computational efficiency can be further enhanced through model pruning, knowledge distillation, and inference optimization.

The bias in these models is also an important concern. Like machine learning models, LLMs can inherit biases from the training data. Since these models are often trained on large corpora of text, which can include biased or skewed content. As a result, LLMs might produce biased outcomes when applied to the task of sentiment analysis for consumer reviews, which can influence sentiment classification results.

This work has focused particularly on a consumer review dataset to evaluate the performance analysis of traditional machine learning and pre-trained LLM models. In future, the author has plans to work on some diverse datasets and case studies from multiple industries, such as healthcare, retail, travel, education, and hospitality, etc.

## 5. Novelty of Approach

This study introduces a unique comparative analysis between pre-trained LLMs, and traditional machine learning techniques specifically tailored for sentiment analysis on product reviews. Unlike conventional NLP pipelines, which often focus exclusively on one type of model or employ extensive custom training, our approach leverages a state-of-the-art pre-trained LLM, GPT-4, without additional fine-tuning, thereby assessing its zero-shot learning capabilities for sentiment analysis. This novel use of a pre-trained LLM provides significant advantages in understanding nuanced sentiment without requiring large amounts of task-specific data or training resources.

Our approach is distinct in its systematic comparison of sentiment analysis efficiency across different text lengths. By examining both short, concise reviews and longer, context-rich text samples, we aim to establish the optimal model type for each scenario. This study demonstrates that SVM are effective for shorter text, while LLMs, particularly GPT-4, outperform traditional models in analysing more complex, lengthier texts.

Furthermore, this study introduces an analysis of LLM explainability within the sentiment analysis domain. Unlike traditional models, which function as “black boxes” and provide limited insight into their classifications, GPT-4 offers clear explanations of its sentiment predictions. This explainability feature is crucial for business applications, as it enhances transparency, builds trust in AI-driven decision-making, and allows organizations

to understand the rationale behind sentiment classifications, particularly in nuanced cases such as mixed sentiments.

These novel elements contribute to a deeper understanding of how pre-trained LLMs can be effectively integrated into business strategies for sentiment analysis, highlighting the balance between computational efficiency, interpretability, and accuracy in model selection.

## 6. Limitations and Future Improvements

This section covers the limitations of the existing work which is performed in this research and suggests improvements for future work.

### 6.1. Limitations

The research primarily utilized a public dataset, which may not fully capture the intricacies and specific characteristics of private datasets that organizations might use when applying these methods in-house. Using public datasets limits the ability to test the model under varied, real-world conditions that are more tailored to specific business needs, potentially affecting the generalization of the approach.

The implementation of sentiment analysis using OpenAI's GPT-4 involves significant costs, which increase as the record count increases. This cost factor can become a major barrier, especially for extensive data analysis in larger organizations or for applications requiring frequent updates and processing of large volumes of data. The choice between the two approaches depends on the dataset size, computational resources, and the complexity of the sentiment analysis task.

### 6.2. Future Improvements

To manage costs more effectively, the research could explore the use of open-source LLMs like Llama2. While hosting these models may incur some costs, overall, the usage expenses would be significantly reduced compared to proprietary models like GPT-4. This approach would also allow for greater customization and flexibility in model tuning and application.

The analysis time with LLMs could be improved by adopting a multi-threading approach and employing more powerful computing resources. This would not only speed up the processing time but also enhance the efficiency of the analysis, allowing for real-time or near-real-time sentiment analysis applications.

To further refine the performance and adaptability of the LLMs, few-shot training techniques could be implemented. This method would enable the models to better adapt to specific linguistic nuances or industry-specific jargon with minimal training examples, improving accuracy and reducing the need for extensive datasets.

These future directions aim to address the current limitations by reducing costs, enhancing processing capabilities, and increasing the adaptability of sentiment analysis models to better meet the diverse needs of various organizations and applications.

## 7. Conclusions

This research has successfully demonstrated the application of traditional machine learning and LLM techniques in sentiment analysis. It provides valuable insights for companies to understand customer perceptions of their products. The study explored two different data features the "Review" and "Summary" and utilized both traditional machine learning methods and advanced LLMs to analyze sentiments expressed in product reviews.

These advancements present a compelling case for integrating LLMs into business strategies for sentiment analysis. By adopting these models, companies can achieve a deeper and more accurate understanding of customer feedback, which can lead to better informed business decisions and improved product offerings.

This research underscores the potential of machine learning in transforming data into actionable business insights, paving the way for more sophisticated analysis techniques that can dynamically adapt to the complexities of human language and sentiment. As

technology evolves, the integration of such AI tools will undoubtedly become a cornerstone in the strategic development of customer-centric business models.

Future research could focus on cost-effective sentiment analysis by utilizing open-source LLMs like LLaMA 2, which, despite some hosting costs, would reduce overall expenses compared to proprietary models like GPT-4 while allowing for customization and flexibility. Enhancing computational efficiency through multi-threading and more robust computing resources could also improve analysis speed, enabling near-real-time applications. Additionally, incorporating few-shot training techniques would refine model adaptability, helping LLMs better understand specific linguistic nuances or industry jargon with minimal examples, thus boosting accuracy and reducing the reliance on extensive datasets. Together, these improvements aim to reduce costs, enhance efficiency, and increase adaptability to meet diverse organizational needs.

**Author Contributions:** Conceptualization, P.S.G., S.E.H., S.P., N.S. and M.J.I.; methodology, P.S.G., S.P. and M.J.I.; software, P.S.G., S.E.H., S.P. and M.J.I.; validation, N.S., S.P. and M.J.I.; formal analysis, P.S.G., S.E.H., N.S., S.P. and M.J.I.; investigation, P.S.G., S.E.H., S.P. and M.J.I.; resources, S.E.H., N.S., S.P. and M.J.I.; data curation, P.S.G. and M.J.I.; writing—original draft preparation, P.S.G., S.E.H., S.P. and M.J.I.; writing—review and editing, P.S.G., S.E.H., N.S., S.P. and M.J.I.; visualization, P.S.G., S.E.H., N.S., S.P. and M.J.I.; supervision, N.S., S.P. and M.J.I.; project administration S.E.H., N.S., S.P. All authors contributed equally to this work. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Data Availability Statement:** The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author(s).

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Chaturvedi, S.; Mishra, V.; Mishra, N. Sentiment analysis using machine learning for business intelligence. In Proceedings of the 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI), Chennai, India, 21–22 September 2017; pp. 2162–2166.
2. Bonifazi, G.; Cauteruccio, F.; Corradini, E.; Marchetti, M.; Terracina, G.; Ursino, D.; Virgili, L. A framework for investigating the dynamics of user and community sentiments in a social platform. *Data Knowl. Eng.* **2023**, *146*, 102183. [[CrossRef](#)]
3. Cauteruccio, F.; Kou, Y. Investigating the emotional experiences in eSports spectatorship: The case of League of Legends. *Inf. Process. Manag.* **2023**, *60*, 103516. [[CrossRef](#)]
4. Wankhade, M.; Rao, A.C.S.; Kulkarni, C. A survey on sentiment analysis methods, applications, and challenges. *Artif. Intell. Rev.* **2022**, *55*, 5731–5780. [[CrossRef](#)]
5. Schmidt, R.K. Automatic Document Classification in Technical Logbooks: A Comparative Study of Supervised, Weakly Supervised and Unsupervised Machine Learning Approaches. Master's Thesis, Universidade NOVA de Lisboa, Lisboa, Portugal, 2024.
6. Abdussalam, M.F.; Richasdy, D.; Bijaksana, M.A. BERT implementation on news sentiment analysis and analysis benefits on branding. *J. Media Inform. Budidarma* **2022**, *6*, 2064–2073. [[CrossRef](#)]
7. Cambria, E.; Schuller, B.; Xia, Y.; Havasi, C. New avenues in opinion mining and sentiment analysis. *IEEE Intell. Syst.* **2013**, *28*, 15–21. [[CrossRef](#)]
8. Fanni, S.C.; Febi, M.; Aghakhanyan, G.; Neri, E. Natural language processing. In *Introduction to Artificial Intelligence*; Springer: Berlin/Heidelberg, Germany, 2023; pp. 87–99.
9. Khurana, D.; Koli, A.; Khatter, K.; Singh, S. Natural language processing: State of the art, current trends and challenges. *Multimed. Tools Appl.* **2023**, *82*, 3713–3744. [[CrossRef](#)] [[PubMed](#)]
10. Loukili, M.; Messaoudi, F.; El Ghazi, M. Sentiment analysis of product reviews for e-commerce recommendation based on machine learning. *Int. J. Adv. Soft Comput. Appl.* **2023**, *15*, 1–13.
11. Mercha, E.M.; Benbrahim, H. Machine learning and deep learning for sentiment analysis across languages: A survey. *Neurocomputing* **2023**, *531*, 195–216. [[CrossRef](#)]
12. Luthra, R.; Bisht, G.S.; Sharma, V.K. *Cardiovascular Diseases (CVD) Prediction Models: A Systematic Review*; Jaypee University of Information Technology: Himachal Pradesh, India, 2023.
13. Julianto, I.T.; Kurniadi, D.; Septiana, Y.; Sutedi, A. Alternative text pre-processing using chat GPT open AI. *J. Nas. Pendidik. Tek. Inform. JANAPATI* **2023**, *12*, 67–77. [[CrossRef](#)]
14. Krugmann, J.O.; Hartmann, J. Sentiment Analysis in the Age of Generative AI. *Cust. Needs Solut.* **2024**, *11*, 3. [[CrossRef](#)]

15. Medhat, W.; Hassan, A.; Korashy, H. Sentiment analysis algorithms and applications: A survey. *Ain Shams Eng. J.* **2014**, *5*, 1093–1113. [CrossRef]
16. Araque, O.; Corcuera-Platas, I.; Sánchez-Rada, J.F.; Iglesias, C. Enhancing Deep Learning Sentiment Analysis with Ensemble Techniques in Social Applications. *Expert Syst. Appl.* **2017**, *77*, 236–246. [CrossRef]
17. Machova, K.; Mach, M.; Vasilko, M. Comparison of machine learning and sentiment analysis in detection of suspicious online reviewers on different type of data. *Sensors* **2021**, *22*, 155. [CrossRef]
18. Priya, C.S.R.; Deepalakshmi, P. Sentiment analysis from unstructured hotel reviews data in social network using deep learning techniques. *Int. J. Inf. Technol.* **2023**, *15*, 3563–3574. [CrossRef]
19. Steinke, I.; Wier, J.; Simon, L.; Seetan, R. Sentiment Analysis of Online Movie Reviews using Machine Learning. *Int. J. Adv. Comput. Sci. Appl.* **2022**, *13*, 618–624. [CrossRef]
20. Rasool, A.; Tao, R.; Marjan, K.; Naveed, T. Twitter sentiment analysis: A case study for apparel brands. *J. Phys. Conf. Ser.* **2019**, *1176*, 022015. [CrossRef]
21. Mathew, L.; Bindu, V.R. A Review of Natural Language Processing Techniques for Sentiment Analysis using Pre-trained Models. In Proceedings of the 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 11–13 March 2020; pp. 340–345.
22. Bellar, O.; Baina, A.; Ballafkih, M. Sentiment Analysis: Predicting Product Reviews for E-Commerce Recommendations Using Deep Learning and Transformers. *Mathematics* **2024**, *12*, 2403. [CrossRef]
23. Zhang, L.; Wang, S.; Liu, B. Deep learning for sentiment analysis: A survey. *WIREs Data Min. Knowl. Discov.* **2018**, *8*, e1253. [CrossRef]
24. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**. [CrossRef]
25. Chi, S.; Qiu, X.; Xu, Y.; Huang, X. How to Fine-Tune BERT for Text Classification? *arXiv* **2019**. [CrossRef]
26. Upadhye, A. Sentiment Analysis using Large Language Models: Methodologies, Applications, and Challenges. *Int. J. Comput. Appl.* **2024**, *186*, 30–34. [CrossRef]
27. Maceda, L.L.; Llovido, J.L.; Artiaga, M.B.; Abisado, M.B. Classifying Sentiments on Social Media Texts: A GPT-4 Preliminary Study. In Proceedings of the 2023 7th International Conference on Natural Language Processing and Information Retrieval, Seoul, Republic of Korea, 15–17 December 2023; pp. 19–24.
28. Vaghani, N.; Thummar, M. Flipkart Product Reviews with Sentiment Dataset. 2023. Available online: <https://www.kaggle.com/dsv/4940809> (accessed on 16 May 2024).
29. Luo, M.; Greenberg, C. Comparing Bag-of-Words, SBERT, and GPT-3 for Bias Detection. *J. Stud. Res.* **2024**, *13*, 20–25. [CrossRef]
30. Jin, P.; Zhang, Y.; Chen, X.; Xia, Y. Bag-of-embeddings for text classification. In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, New York, NY, USA, 9 July 2016; pp. 2824–2830.
31. Kadhim, A.I. Survey on supervised machine learning techniques for automatic text classification. *Artif. Intell. Rev.* **2019**, *52*, 273–292. [CrossRef]
32. Iram, M.; Rehman, S.U.; Shahid, S.; Mehmood, S.A. Anatomy of Sentiment Analysis of Tweets Using Machine Learning Approach. *Int. J. Comput. Sci. Netw. Secur.* **2023**, *23*, 97–106. [CrossRef]
33. Japkowicz, N.; Shah, M. Performance Evaluation in Machine Learning. In *Machine Learning in Radiation Oncology: Theory and Applications*; El Naqa, I., Li, R., Murphy, M.J., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 41–56. [CrossRef]
34. Shah, D. Top Performance Metrics in Machine Learning: A Comprehensive Guide. May 2023. Available online: <https://www.v7labs.com/blog/performance-metrics-in-machine-learning> (accessed on 16 May 2024).

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.