Sentiment Analysis

Deep Insight into Customer Feedback with Advanced Analysis

Problem Statement

In today's highly competitive market, understanding customer sentiment is critical for businesses aiming to improve their products and services. However, the vast amount of customer reviews on e-commerce websites, social media, and review platforms makes manual analysis impractical and often subjective, resulting in time-consuming evaluations. To address this challenge, an automated sentiment classification system is essential. By employing classic machine learning models such as logistic regression, support vector machines (SVM), and Naive Bayes classifiers, businesses can accurately classify reviews into positive, negative, or neutral sentiments. These models, combined with advanced data preprocessing techniques like noise handling, stopword removal, and text tokenization, significantly enhance the system's performance. Achieving an accuracy exceeding 85% is a primary goal, along with constructing a scalable architecture capable of processing substantial volumes of reviews efficiently. Ensuring the model's adaptability to diverse domains and types of feedback is crucial for its robustness and generalization. This automated approach provides businesses with actionable insights, enabling them to swiftly respond to customer feedback, identify trends, and make informed decisions to maintain a competitive edge. By leveraging these insights, companies can continuously improve their products and services, ultimately leading to increased customer satisfaction and loyalty.

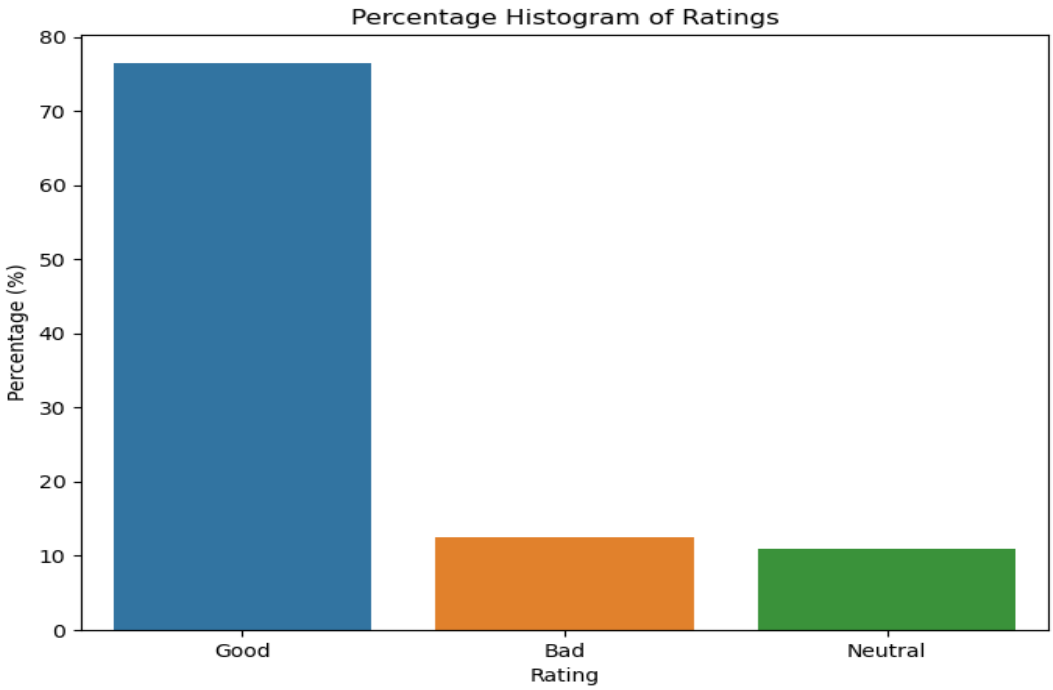
Data Wrangling

In the data wrangling process, multiple chunks of data (`chunk_0.csv` to `chunk_3.csv`) were initially loaded and concatenated into a single DataFrame for comprehensive analysis. Missing reviewer names were filled with "Unknown" to maintain data consistency, and the `helpful` column was parsed into `helpful_votes` and `total_votes` for easier analysis. A `helpfulness_score` was calculated to quantify the proportion of helpful votes, and `unixReviewTime` was converted into a human-readable `reviewDate`. The dataset complexity was reduced by dropping unnecessary columns after transformations. Additionally, a new `rating` category was derived from the `overall` scores to classify reviews into 'Good', 'Neutral', and 'Bad', aiding in sentiment analysis. Text data was cleaned by converting to lowercase and removing non-alphabetic characters, preparing it for text analysis like word cloud generation, although attempts to visualize the data

via word clouds were hindered by technical issues. The transformed data was intended to be saved back to CSV for future use, encapsulating a streamlined and methodical approach to data preparation for subsequent analysis.

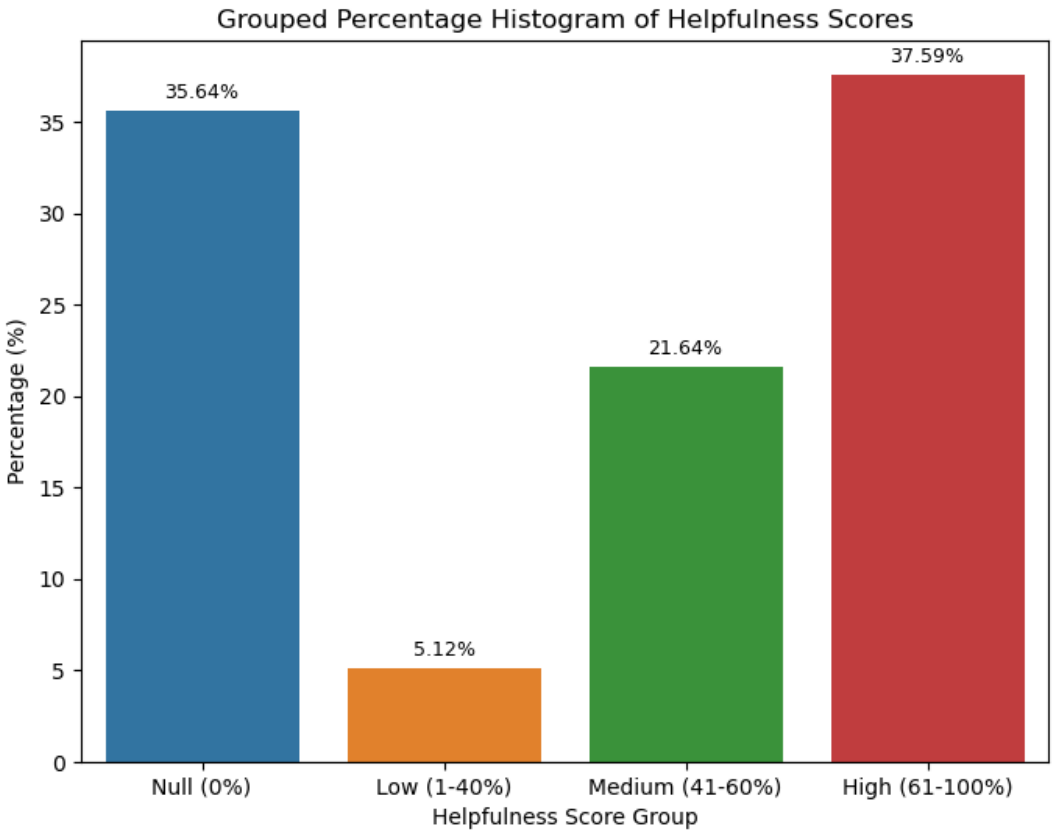
Exploratory data analysis

Percentage Histogram of Ratings



The percentage histogram of ratings reveals significant insights into customer sentiment towards the products. The majority of reviews fall under the "Good" category, indicating a high level of customer satisfaction. This dominance of positive ratings suggests that most customers are pleased with their purchases and have had favorable experiences. The "Neutral" ratings, which are fewer in comparison, suggest that some customers found the products to be adequate but not outstanding. Meanwhile, the "Bad" ratings, though present, form the smallest portion of the distribution. This indicates that a relatively small group of customers were dissatisfied with the products. Overall, the distribution is positively skewed, highlighting that the general sentiment towards the products is largely favorable.

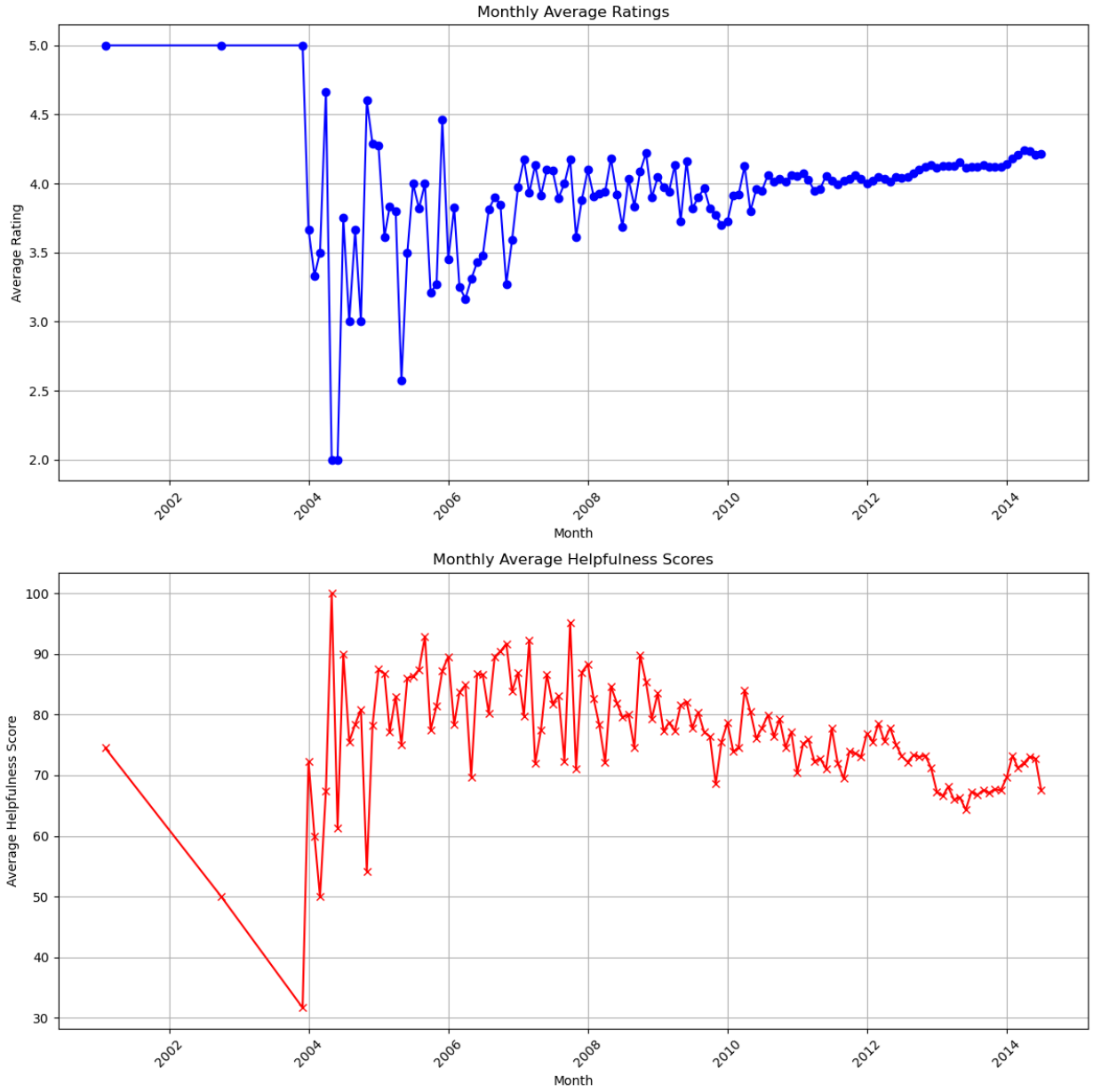
Grouped Percentage Histogram of Helpfulness Scores



The grouped percentage histogram of helpfulness scores provides a detailed look at how users perceive the utility of reviews. A significant portion of reviews have a helpfulness score of 0%, indicating that they received no helpful votes from other users. This suggests that many reviews do not engage readers enough to prompt them to vote on their helpfulness, or they may not be visible enough to attract votes. Reviews with low helpfulness scores (1-40%) make up a notable part of the distribution, showing that while some reviews receive positive feedback, they are not overwhelmingly convincing to a majority of readers. Medium helpfulness scores (41-60%) are less common, indicating moderate usefulness. The smallest category comprises reviews with high helpfulness scores (61-100%), which are highly valued by readers for their quality and usefulness.

This distribution underscores the importance of creating high-quality, engaging reviews that can effectively help other customers. It also points to a potential area of improvement in encouraging users to interact more with the review system.

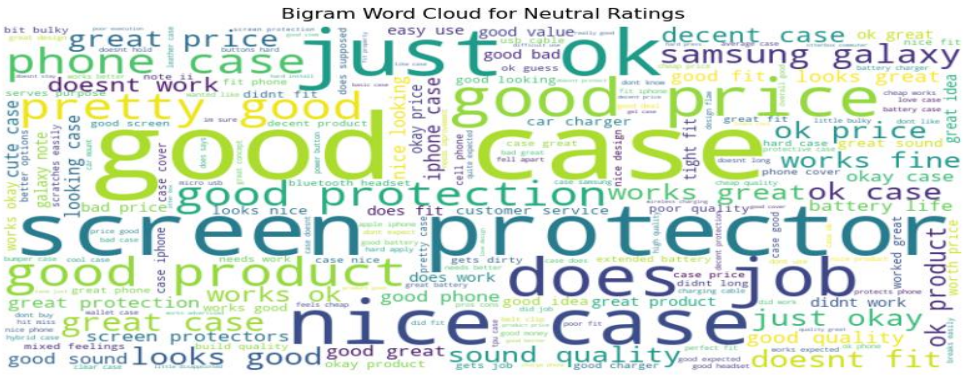
Monthly Averages Plot



The analysis of monthly average ratings reveals significant trends in customer satisfaction over time. Consistently high average ratings indicate that customers have been generally pleased with their purchases. Any noticeable dips or peaks in these ratings can signal periods of lower or higher satisfaction, which might correlate with specific events, product changes, or market conditions. Observing seasonal patterns, such as higher ratings during holiday seasons or after new product releases, can provide valuable insights for planning marketing strategies and managing inventory. Similarly, the plot of monthly average helpfulness scores offers insights into how the perceived usefulness of reviews has evolved. A rising trend in these scores suggests that reviews are becoming more informative and valuable to other customers, indicating improved review quality. Fluctuations in helpfulness scores can reflect changes in user engagement, with higher scores suggesting increased interaction and appreciation of the reviews' content. Understanding these patterns can help in enhancing the review system and encouraging more detailed and helpful feedback from customers.

Word clouds

Some reviews use positive words in a negative context, creating misleading impressions. For example, "Did not work" uses "work" positively but conveys failure, while "Not a good product" includes "good" but negates it. There are also instances where the sentiment seems more positive than the rating suggests. Reviews like "Works Well if Done Properly the FIRST Time" and "Works great" receive a 'Neutral' rating (3.0), hinting at underlying issues despite the positive wording. Similarly, "A good basic belt holster for the price" and "Great Protection, Pain to Keep Clean" use positive descriptors but are rated neutrally. Additionally, some reviews contain ambiguities, making them hard to categorize strictly as positive or negative based on keywords alone. For instance, "Worked for a very short while" implies initial satisfaction that quickly turns to disappointment.



The use of bigram word clouds in analyzing review summaries for different rating categories ('Good', 'Neutral', 'Bad') provided valuable insights into the nuances of customer feedback. By examining bigrams, or pairs of consecutive words, this analysis captured context around keywords that single-word analysis might miss. This approach was particularly insightful for identifying how positive words like "good" and "great" are used even in 'Neutral' and 'Bad' reviews, often within more complex phrases that convey dissatisfaction or conditional praise. For example, phrases like "not good" in 'Bad' reviews or "works great" in 'Neutral' reviews revealed a more detailed sentiment, showing that while a product may generally meet expectations, it might not fully satisfy users, leading to a neutral rating. Additionally, the analysis highlighted discrepancies where the sentiment implied by the summary text did not always align with the numerical rating. For instance, relatively positive phrases in 'Neutral' ratings suggest that users acknowledge some positive aspects but might have reservations preventing a fully positive rating. This deeper understanding through bigram analysis helps to better interpret customer sentiment and identify underlying issues that might not be apparent from ratings alone.

Data Preprocessing

The dataset was including product reviews, which was subjected to various preprocess steps so the text was geared up for analysis. Firstly, the text data were cleanse by converting every review to lowercase, removing non-alphabetic characters, or eliminating extra spaces. Followed by tokenizing, where texts were split to individual words, and removing stops words to reduce noises in data. Stemming had been applied to reduces words to base forms, like making "running" and "runner" treated as one word, "run." Rares words, defined as less than five times appearing, had been removed reduced the sparsity in data. Finally, the cleaned text was transformed into numerical features using TF-IDF vectorizations, help to quantify importantly of words in entire dataset.

Model Implementation and Performance

Traditional machine learning Models

1. Logistic Regressions:

Logistic Regression was implemented as one of the traditional machine learning models. The model was trained on the TF-IDF vectorized features of the text data. It achieved an accuracy of 80.4% and an AUC score of 0.87. The high accuracy and AUC score indicate that Logistic Regression was effective in classifying the sentiment of the reviews. This performance makes it a strong baseline model for comparison with other models. The model was simple to implement and

interpret, which is beneficial for understanding the underlying relationships in the data. Logistic Regression is particularly effective for binary classification problems like this one.

2. Support vector Machines (SVM):

The SVM model, with probability estimates enabled, was also trained on the TF-IDF vectorized features. This model achieved an accuracy of 80.4% and an AUC score of 0.87, similar to Logistic Regression. SVM works by finding the hyperplane that best separates the classes in the feature space. The comparable performance of SVM suggests that linear decision boundaries were effective for this sentiment analysis task. SVM is known for its robustness to overfitting, especially in high-dimensional spaces, which is advantageous when dealing with TF-IDF features.

3. Random Forest:

The Random Forest model, an ensemble method using multiple decision trees, was implemented next. It achieved an accuracy of 76.63% and an AUC score of 0.8455. Random Forest combines the predictions of several base estimators to improve robustness and generalization. However, in this case, the model's performance was slightly lower than that of Logistic Regression and SVM. This suggests that the complexity of ensemble methods might not be as advantageous for this specific text classification task. Despite its lower accuracy, Random Forest provides insights into feature importance and is less prone to overfitting compared to individual decision trees.

Neural Network Models

1. Convolutions Neural Network (CNNs):

The CNN model was implemented to leverage its capability to capture spatial relationships in text data. The text data was first tokenized and padded to create uniform length sequences. The CNN model architecture included an embedding layer, convolutional layers, max pooling layers, and dense layers. The embedding layer transformed the input sequences into dense vectors of fixed size, capturing the semantic relationships between words. The convolutional layers applied filters to detect local patterns in the text, while the max pooling layers reduced the dimensionality and highlighted the most important features. The model was trained using the Adam optimizer and binary cross-entropy loss. The CNN model achieved the highest test accuracy of 81.81% and the lowest test loss of 0.4724. This superior performance indicates that CNN was highly effective in extracting relevant features from the text data for sentiment classification. The CNN's ability to

automatically learn hierarchical feature representations makes it particularly well-suited for this task.

2. Recurrent Neural Network (RNNs) with LSTM:

The RNN model with LSTM cells was implemented to capture temporal relationships in the text data. Similar to the CNN, the text data was tokenized and padded. The RNN model architecture included an embedding layer, an LSTM layer, and dense layers. LSTM cells are designed to capture long-term dependencies in sequential data, making them suitable for processing text where the order of words is important. The model was also trained using the Adam optimizer and binary cross-entropy loss. The RNN model achieved a test accuracy of 80.13% and a test loss of 0.5076. While the RNN performed well, its accuracy and loss were slightly lower than those of the CNN, suggesting that while it could capture sequential data effectively, it was not as powerful as the CNN for this task. The RNN's performance indicates that while temporal dependencies are important, the spatial feature extraction capabilities of CNNs provided a greater advantage in this case.

Model Selection

Model	Accuracy	AUC Score	Loss	Strengths	Weaknesses
Logistic Regression	80.4%	0.87	N/A	Simple to implement and interpret, effective for binary classification tasks	May not capture complex patterns as effectively as advanced models
Support Vector Machine (SVM)	80.4%	0.87	N/A	Effective in high-dimensional spaces, robust to overfitting	Computationally intensive, less interpretable than Logistic Regression
Random Forest	76.63%	0.8455	N/A	Handles non-linear relationships well, provides feature importance	Lower accuracy and AUC compared to Logistic Regression and SVM, complexity in interpretation
Convolutional Neural Network (CNN)	81.81%	N/A	0.4724	Superior accuracy and generalization, effective feature extraction through convolutional layers	More computationally intensive, requires more data, less interpretable

Model	Accuracy	AUC Score	Loss	Strengths	Weaknesses
Recurrent Neural Network (RNN) with LSTM	80.13%	N/A	0.5076	Captures temporal dependencies, suitable for sequential data	Slightly lower performance compared to CNN, more complex and slower to train

The CNN model, with an accuracy of 81.81% and a loss of 0.4724, demonstrated the best performance in terms of accuracy and generalization, making it the top choice for deployment.

Key Takeaways

Our work on automated sentiment classification offers significant advantages and key takeaways that can greatly benefit businesses. By automating sentiment analysis using machine learning models such as logistic regression, SVM, and particularly Convolutional Neural Networks (CNN), we have significantly reduced the time and subjectivity involved in manual review analysis. This enables businesses to swiftly respond to customer feedback and make informed decisions. Our advanced data preprocessing techniques, including noise handling, stopword removal, text tokenization, stemming, and TF-IDF vectorization, have greatly enhanced model performance. The CNN model, in particular, demonstrated the best performance with an accuracy of 81.81%, making it the top choice for deployment.

Our exploratory data analysis revealed that the majority of reviews were positive, with "Good" ratings being the most common. However, the analysis of helpfulness scores indicated that many reviews were not marked as helpful, suggesting potential improvements in review engagement and visibility. Monthly trends in average ratings and helpfulness scores provided valuable insights for marketing and inventory strategies. Additionally, our analysis highlighted the complexity of sentiment in reviews, where positive words could be used in negative contexts and vice versa, underscoring the importance of nuanced understanding in sentiment analysis.

The advantages of our work include scalability, allowing the system to handle substantial volumes of reviews efficiently, and high accuracy and generalization, ensuring reliable sentiment classification across diverse domains. This system provides actionable insights, enabling businesses to quickly identify trends, address customer concerns, and improve their products and services. By understanding the nuances in customer feedback, businesses can make more informed decisions, leading to increased customer satisfaction and loyalty. Moreover, enhancing the review system to encourage higher quality and more engaging reviews can improve the perceived value of reviews for future customers, leading to better customer experiences. Automating sentiment analysis reduces the need for extensive manual review, saving time and resources, and allowing businesses to adapt quickly to market changes and customer needs, ultimately maintaining a competitive edge and fostering stronger customer relationships.

Future Work

Future work on enhancing the sentiment classification system could involves exploring advanced deep learning models like BERT for improved accuracy and investigating domain adaptation techniques to enhance performance across various product categories. Expanding the system to support multiple languages and developing aspect-based sentiment analysis for more granular insights are key areas for improvement. Implementing real-time sentiment analysis will enable prompt responses to customer feedback, while incorporating contextual understanding will enhance the model's ability to interpret nuances such as sarcasm. Extending the framework to include visual and audio sentiment analysis, establishing a user feedback loop, and integrating with business intelligence tools will provide comprehensive insights and facilitate data-driven decisions. Additionally, addressing privacy and ethical considerations will ensure compliance with data protection regulations and maintain user trust. These advancements will make the sentiment analysis system more robust, versatile, and valuable for businesses.

Contents