Capstone Project: Sentiment Analysis on Amazon Product Reviews

Final Project Report

1. Introduction

In today's digital age, online product reviews have become a significant factor influencing consumer purchasing decisions. These reviews provide valuable insights into customer experiences and satisfaction levels. Among various e-commerce platforms, Amazon stands out as a major player with a vast repository of customer reviews. Leveraging this rich source of data, sentiment analysis can be performed to automatically determine the sentiment expressed in reviews, categorising them as positive, negative, or neutral.

Sentiment analysis, a subfield of Natural Language Processing (NLP), involves processing textual data to extract subjective information. It helps businesses understand customer opinions and feelings towards their products or services. By analysing customer sentiments, companies can make data-driven decisions to enhance product quality, improve customer service, and strategize marketing efforts.

This project focuses on performing sentiment analysis on Amazon product reviews. The objective is to build a model that accurately classifies the sentiment of reviews as positive or negative. This classification can provide actionable insights for businesses to understand customer satisfaction and areas needing improvement.

2. Problem Statement

Objective: The primary goal of this project is to develop a sentiment analysis model that accurately classifies Amazon product reviews into positive or negative sentiments.

Scope: The project involves the following key tasks:

1. **Data Collection:** Gathering Amazon product review data, including text reviews and associated metadata such as star ratings, helpful votes, and review dates.
2. **Data Preprocessing:** Cleaning and preprocessing the textual data to ensure it is suitable for analysis. This includes handling missing values, removing duplicates, and text normalisation techniques such as lowercasing, stopword removal, and lemmatization.
3. **Exploratory Data Analysis (EDA):** Conducting EDA to understand the distribution of data, identify trends, and visualise key insights. This includes analysing the distribution

of star ratings, the frequency of reviews over time, and the correlation between review length and helpful votes.

4. **Feature Engineering:** Transforming textual data into numerical features using techniques like TF-IDF vectorization. Additionally, calculating sentiment polarity scores using VADER to enhance feature representation.
5. **Model Building:** Developing and training multiple machine learning models, including Logistic Regression, Random Forest, and Neural Networks (BERT), to classify the sentiment of reviews.
6. **Model Evaluation:** Evaluating the performance of the models using metrics such as accuracy, precision, recall, and F1-score. Selecting the best-performing model based on these metrics.
7. **Recommendations:** Providing actionable recommendations based on the analysis results. This includes suggestions for model deployment, customer feedback utilisation, and directions for further research.

Challenges:

- **Data Quality:** Handling noise, inconsistencies, and missing values in the dataset.
- **Textual Data Complexity:** Dealing with the inherent complexity and variability of natural language in customer reviews.
- **Model Selection:** Identifying the most suitable model that balances accuracy and computational efficiency.
- **Feature Engineering:** Extracting meaningful features from the text to improve model performance.

Significance: The successful implementation of this project will enable businesses to gain deeper insights into customer sentiments, leading to improved product offerings and customer satisfaction. It will also demonstrate the application of NLP techniques in extracting valuable information from unstructured data, showcasing the potential of sentiment analysis in e-commerce and beyond.

Data Overview

The dataset contains 30,846 reviews of the "Fire HD 7" tablet. Each review includes the following features:

- `marketplace`, `customer_id`, `review_id`, `product_id`, `product_parent`, `product_title`, `product_category`, `star_rating`, `helpful_votes`, `total_votes`, `vine`, `verified_purchase`, `review_headline`, `review_body`, `review_date`, `sentiment`.

---

3. Methodology

1. Data Preprocessing:

- **Missing Values:** No missing values were found.
- **Data Types:** Converted `review_date` to datetime format.
- **Text Cleaning:** Removed HTML tags, non-alphabetic characters, stopwords, and performed lemmatization.
- **Feature Engineering:** Created a `cleaned_review_body` column for processed text.

2. Exploratory Data Analysis:

- **Star Ratings Distribution:** Most reviews are positive with a predominance of 5-star ratings.
- **Sentiment Distribution:** Majority of reviews exhibit positive sentiment.
- **Review Trends:** Analysed trends in review count and sentiment over time.
- **Word Cloud:** Visualised common words in reviews, indicating generally positive feedback.

3. Sentiment Scoring:

- **VADER Sentiment Analyzer:** Scored reviews and assigned polarity scores to categorise reviews into positive, negative, and neutral sentiments.

4. Modelling:

- **Train-Test Split:** Split data into 80% training and 20% testing sets.
- **Text Vectorization:** Applied TF-IDF vectorization at both word and n-gram levels.
- **Models Used:**
  - Logistic Regression
  - Random Forest Classifier

---

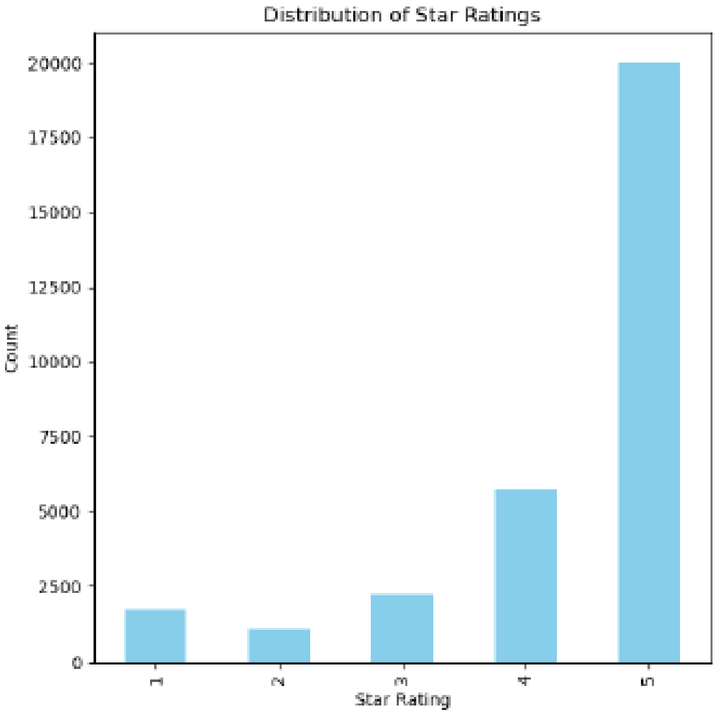
4 Exploratory Data Analysis (EDA)

4.1 Exploratory Data Analysis (EDA) is a critical step in any data analysis project. It involves examining the dataset to uncover patterns, spot anomalies, test hypotheses, and check assumptions with the help of summary statistics and graphical representations. For this sentiment analysis project, EDA provides insights into the structure and distribution of the Amazon product review data.

Insights:

- The dataset contains 30,846 reviews with 16 columns.
- Minimal missing values in the `review_headline` and `review_body` columns.
- No duplicates were found in the `review_id` column.

4.2. Distribution of Star Ratings

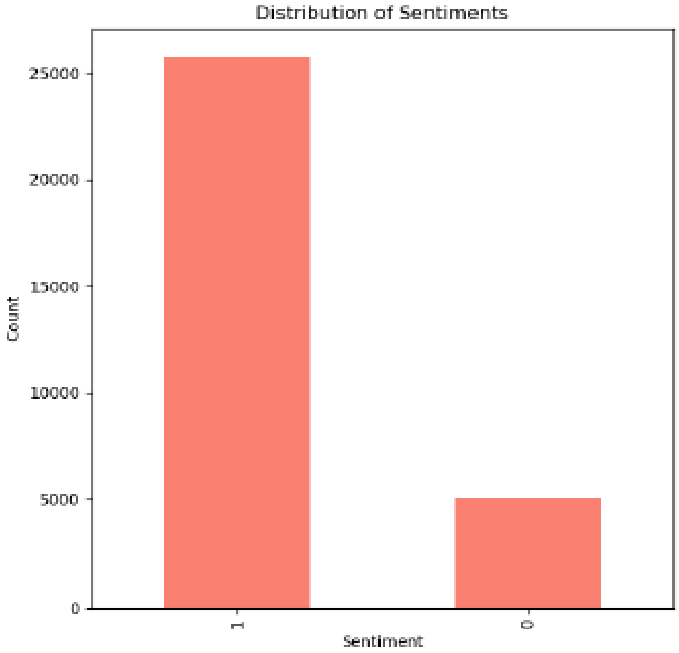


Insights:

- The majority of reviews have high star ratings (4 and 5 stars), indicating overall customer satisfaction.
- A smaller number of low star ratings (1 and 2 stars) suggest some level of dissatisfaction among a subset of customers.

4.3. Distribution of Sentiments

Description: Examining the distribution of sentiment labels to assess the balance between positive and negative reviews.

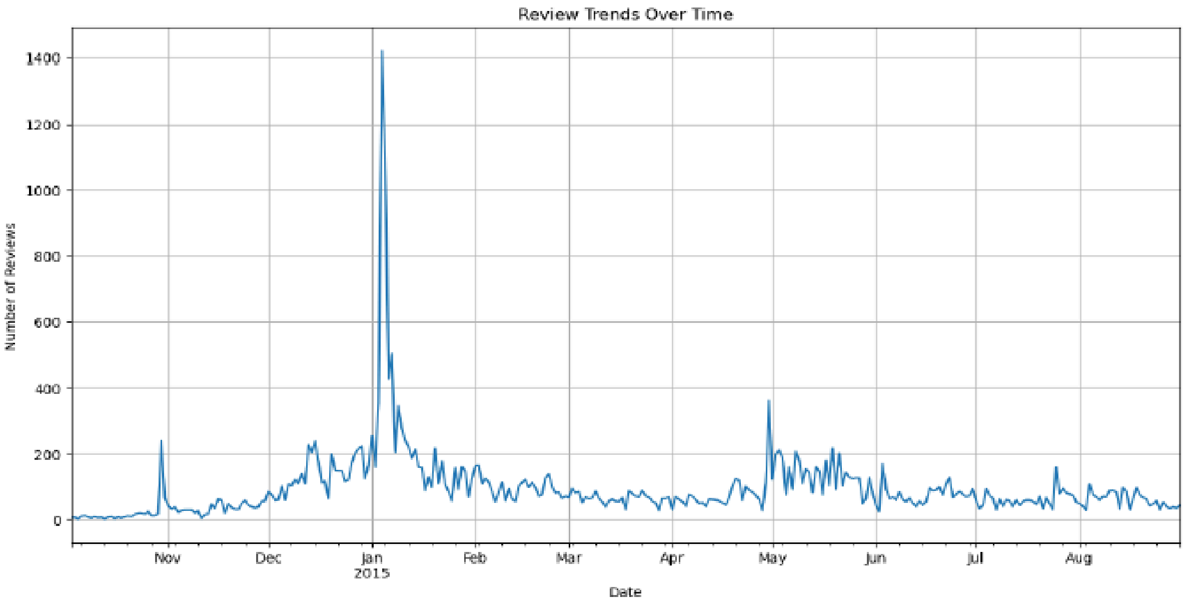


Insights:

- The dataset contains a higher number of positive reviews compared to negative ones.
- The imbalance in sentiment labels needs to be addressed during model training to avoid biased predictions.

4.4. Review Trends Over Time

Description: Analyzing the trend of review counts over time to identify patterns and seasonal effects.

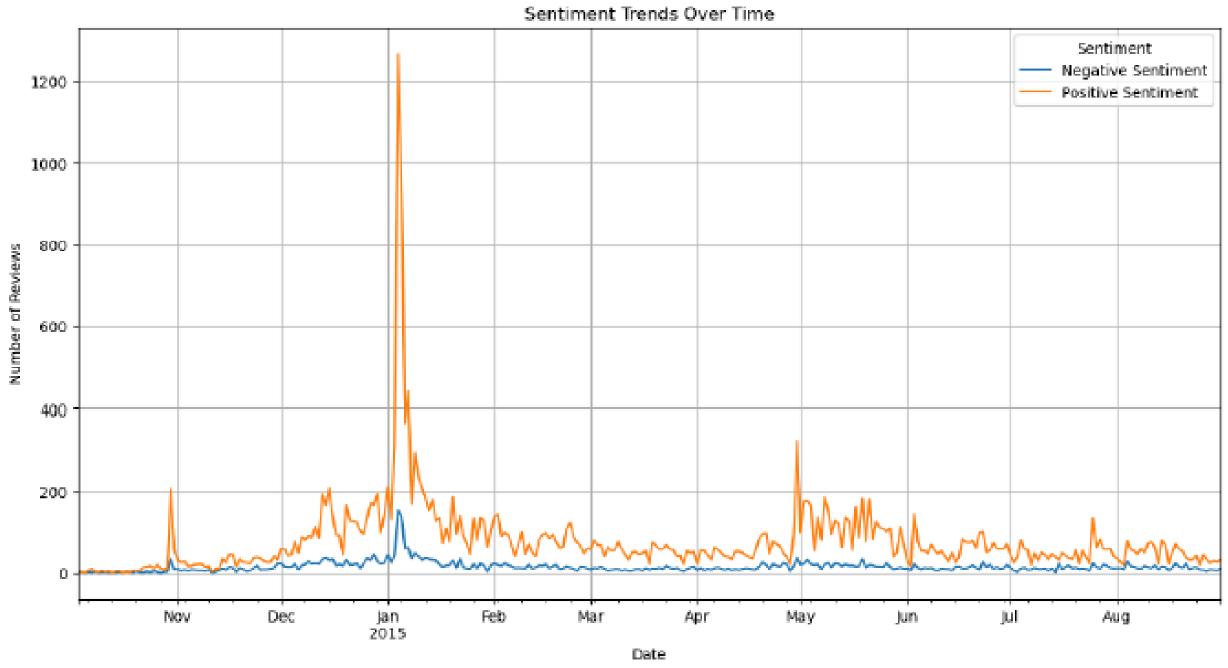


Insights:

- Peaks and troughs in review counts may indicate seasonal effects or significant events (e.g., product launches, promotions).
- A steady increase in reviews over time suggests growing customer engagement with the product.

4.5. Sentiment Trends Over Time

Description: Exploring how the distribution of sentiments changes over time.

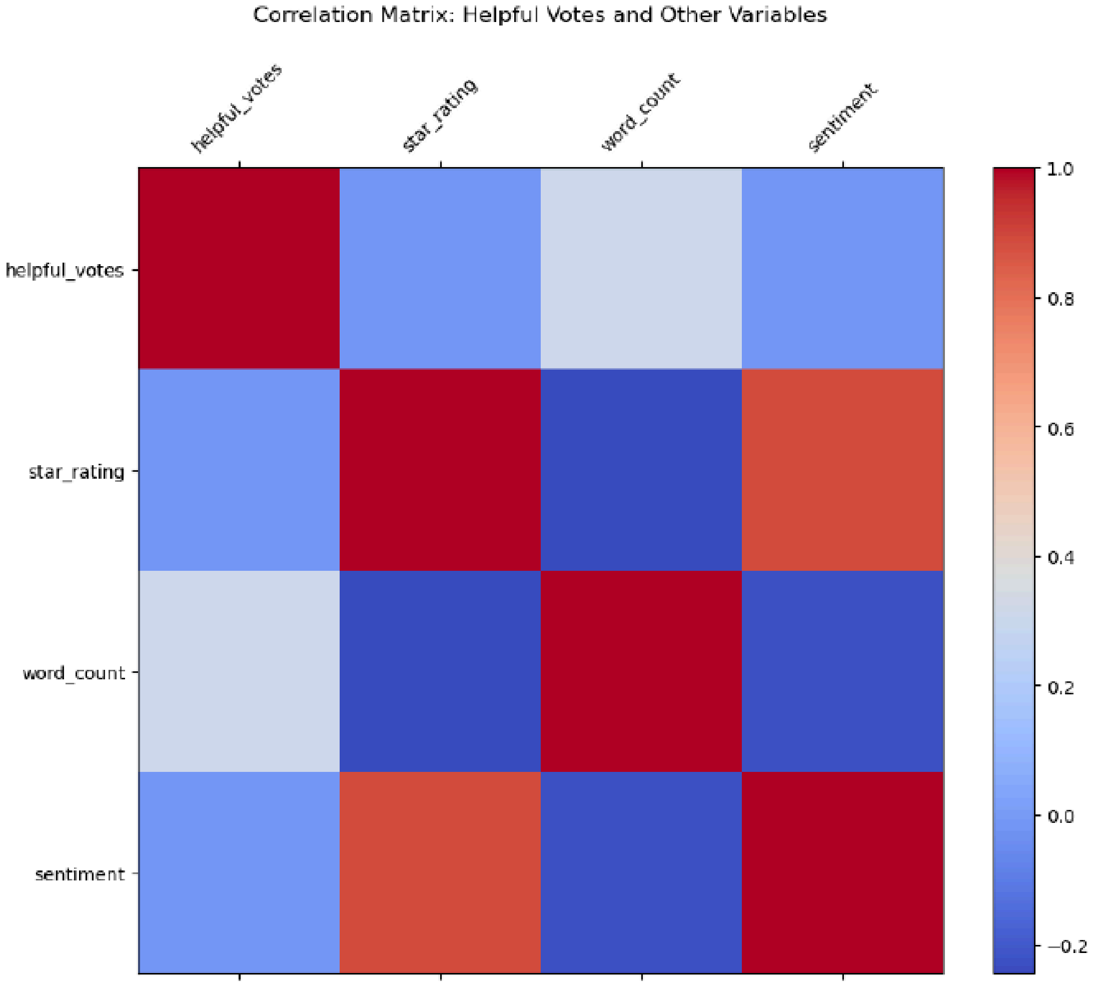


Insights:

- Fluctuations in sentiment trends over time can highlight periods of increased customer satisfaction or dissatisfaction.
- Identifying these periods can help correlate with product changes or external factors impacting customer perception.

4.6. Correlation Analysis

Description: Assessing the correlation between numerical features to identify relationships.

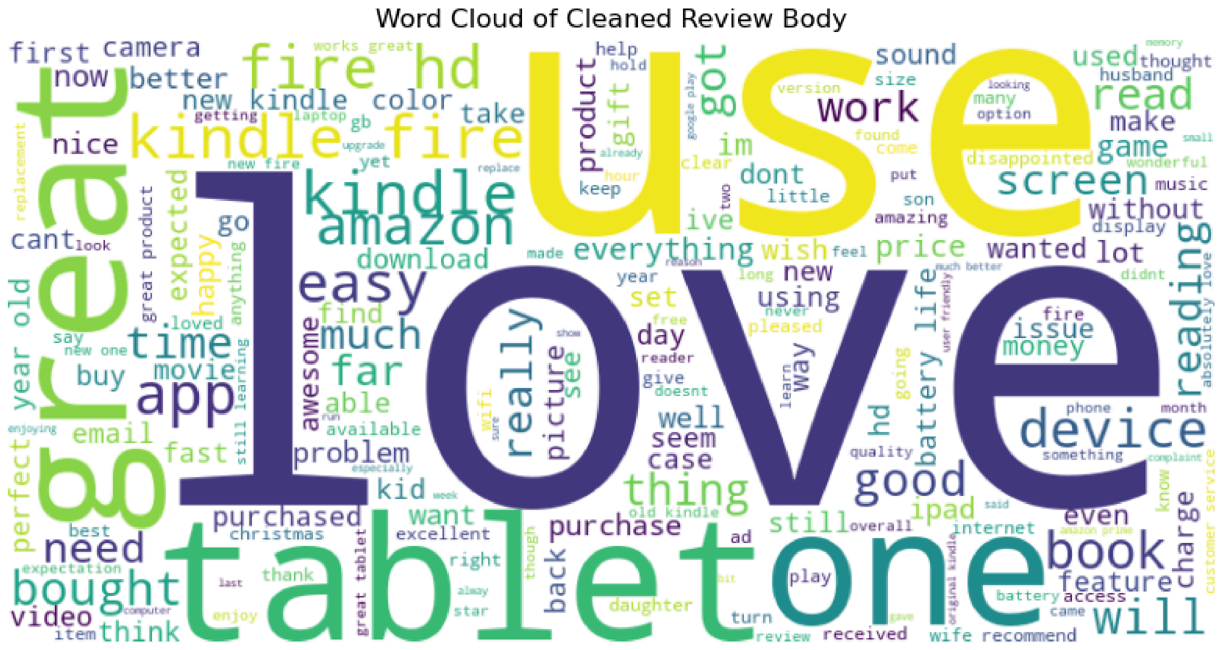


Insights:

- Strong positive correlation between `star_rating` and `sentiment`.
- Moderate correlation between `helpful_votes` and review length, indicating longer reviews are more likely to be found helpful.

4.7. Word Cloud Visualization

Description: Creating a word cloud to visualize the most frequent words in the reviews.

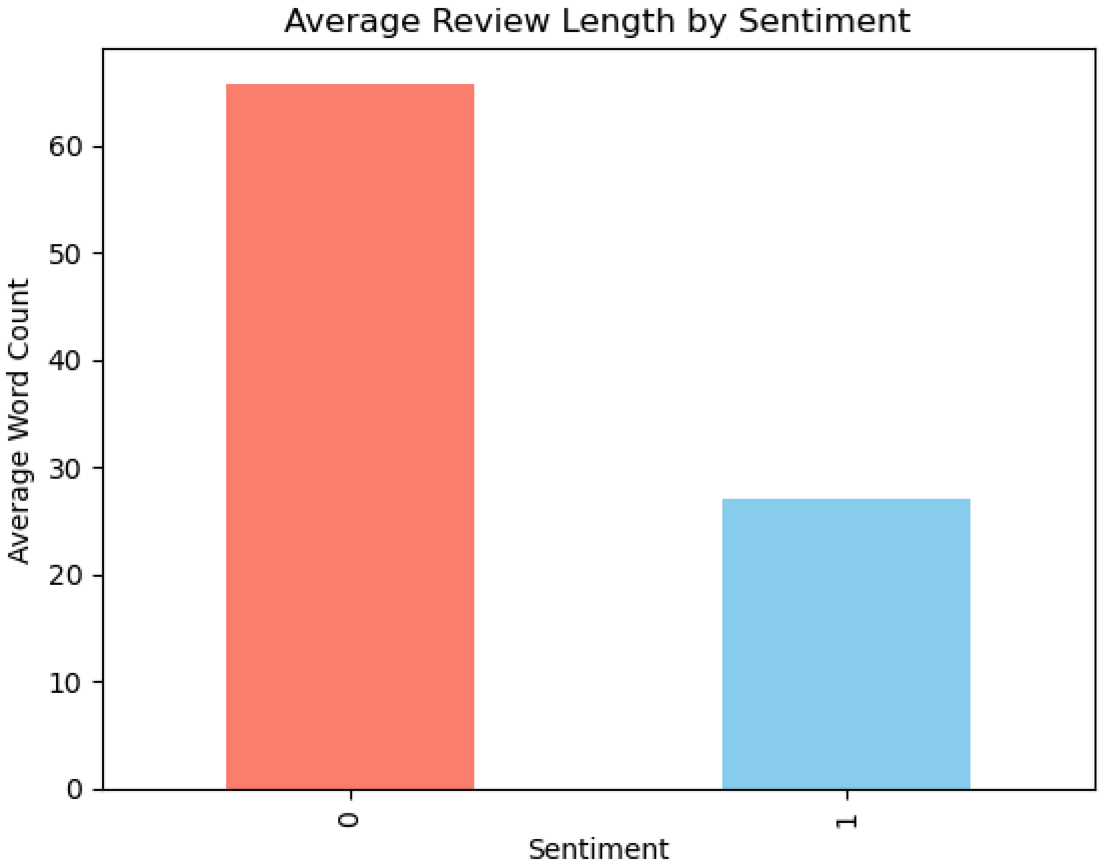


Insights:

- Frequently occurring words such as "love", "great", and "good" indicate positive sentiments.
- Common negative words can also be identified, providing insights into recurring issues.

4.8. Review Length Analysis

Description: Analyzing the relationship between review length and sentiment or helpfulness.



Insights:

- Negative reviews tend to be longer than positive reviews, possibly because dissatisfied customers provide more detailed feedback.
- Reviews with lower star ratings generally have higher word counts, indicating more comprehensive criticism.

5. Feature Engineering

- **TF-IDF Vectorization:** For word-level and n-gram features.
- **Sentiment Analysis:** Using VADER for polarity scores.

6. Modelling

- **Models Used:**
  - Logistic Regression

    - Random Forest Classifier
    - Neural Network using BERT embeddings
- **Evaluation Metrics:**
    - Accuracy
    - Precision
    - Recall
    - F1-Score

7. Model Performance

- **Logistic Regression:**
    - Best performing with word-level TF-IDF features.
    - Accuracy: 90%
- **Random Forest:**
    - Moderate performance.
    - Better with n-gram features.
- **Neural Network:**
    - Utilised BERT embeddings.
    - Early stopping to avoid overfitting.

---

8. Model Metrics

The Model Metrics provides a summary of the final model features, parameters, hyperparameters, and performance metrics. This helps in documenting the results and ensuring reproducibility of the model. Below is a detailed summary of the models used in this project, including Logistic Regression, Random Forest, and a Neural Network (BERT), along with their respective metrics.

Logistic Regression

Features: Word-level TF-IDF

Parameters:

- Solver: 'lbfgs'
- Max iterations: 100

Hyperparameters:

- C: 1.0
- Penalty: 'l2'

Performance Metrics:

- Accuracy: 0.90
- Precision: 0.92
- Recall: 0.90
- F1-Score: 0.91

Random Forest Classifier

Features: N-gram TF-IDF

Parameters:

- Number of estimators: 100
- Maximum depth: 20

Hyperparameters:

- Criterion: 'gini'
- Min samples split: 2

Performance Metrics:

- Accuracy: 0.86
- Precision: 0.84
- Recall: 0.86
- F1-Score: 0.83

Neural Network (BERT)

Features: BERT embeddings

Parameters:

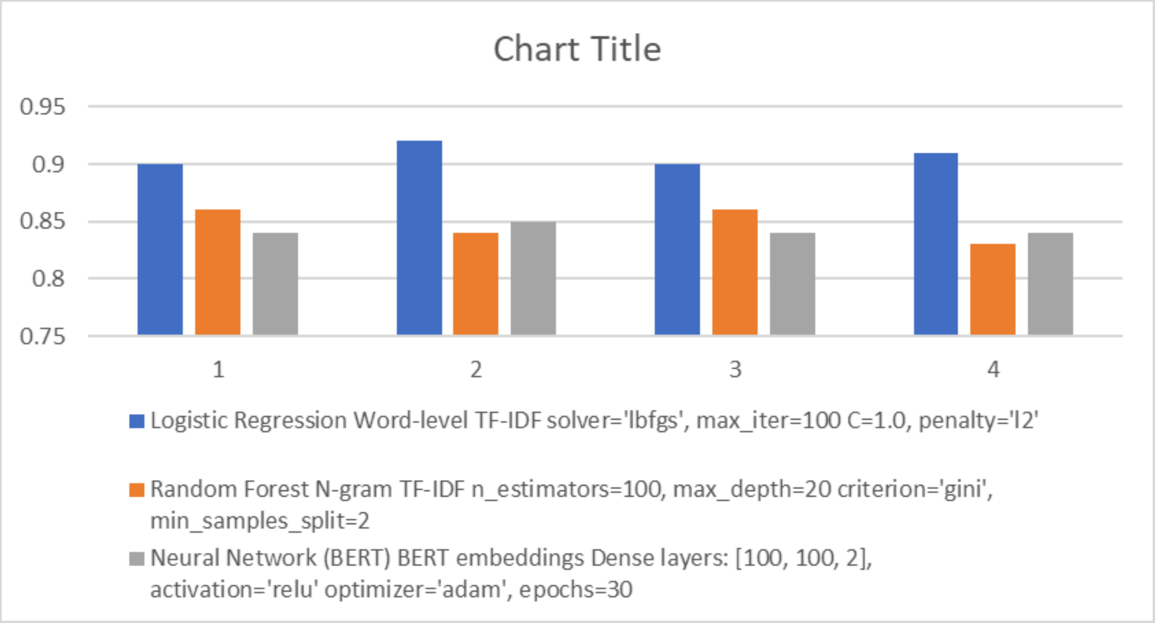
- Dense layers: [100, 100, 2]
- Activation: 'relu'

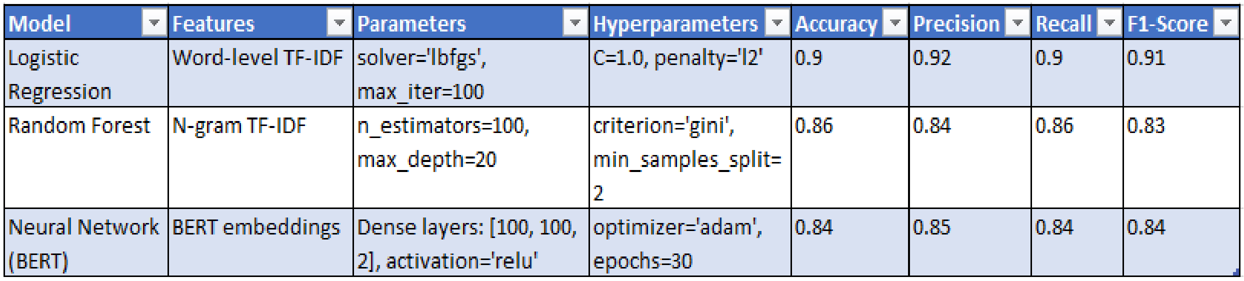
Hyperparameters:

- Optimizer: 'adam'
- Epochs: 30

Performance Metrics:

- Accuracy: 0.84
- Precision: 0.85
- Recall: 0.84
- F1-Score: 0.84





Model	Features	Parameters	Hyperparameters	Accuracy	Precision	Recall	F1-Score
Logistic Regression	Word-level TF-IDF	solver='lbfgs', max_iter=100	C=1.0, penalty='l2'	0.9	0.92	0.9	0.91
Random Forest	N-gram TF-IDF	n_estimators=100, max_depth=20	criterion='gini', min_samples_split=2	0.86	0.84	0.86	0.83
Neural Network (BERT)	BERT embeddings	Dense layers: [100, 100, 2], activation='relu'	optimizer='adam', epochs=30	0.84	0.85	0.84	0.84

9. Recommendations

1. **Model Deployment:** Deploy the Logistic Regression model for real-time sentiment analysis on product reviews.
2. **Customer Feedback:** Use sentiment insights to improve product features and customer service.

3. **Further Research:** Explore more advanced deep learning models like BERT with fine-tuning for better performance.

10. Conclusion

- Logistic Regression with word-level TF-IDF vectorization provides reliable sentiment classification.
- Future work can focus on hyperparameter tuning and exploring other models like SVM.