2025

Sentiment Analysis of IMDB Reviews: A Text Classification Pipeline for Binary Opinion Detection

PDAN8411 POE

MAX WALSH – ST10203070

27 JUNE 2025

Table of Contents

Introduction

The objective of this analysis is to build a sentiment classification model capable of analysing public sentiment in textual reviews, specifically in the context of customer feedback. The client, being a medical scheme, is seeking to understand how sentiment analysis techniques could help identify common themes and customer satisfaction levels based on large amounts of textual data (OpenAI, 2025). Although the client initially referenced HelloPeter reviews, for methodological robustness and reproducibility, the IMDB Movie Reviews Dataset was selected instead. This dataset contains 50,000 pre-labelled movie reviews, 25,000 positives and 25,000 negatives, sourced from Kaggle and curated by Narasimhan (2020) for sentiment classification research.

This project implemented the full text analytics pipeline, starting with exploratory data analysis (EDA), to preprocessing, to feature engineering via TF-IDF vectorisation, to supervised classification using logistic regression. An emphasis is placed on practical evaluation metrics such as precision, recall, F1 score, accuracy, and AUC to interpret the model's real-world applicability, specifically in classifying subjective sentiment with balanced sensitivity and specificity (Muller & Guido, 2016). After which, regularisation and retraining are considered to assess the model's robustness under changing hyperparameters (OpenAI, 2025).

All in all, this project offers the client a comprehensive demonstration of sentiment modelling using natural language processing (NLP), showing the effective adaptation of techniques for customer-facing feedback systems in the healthcare or insurance sectors (OpenAI, 2025).

Dataset Justification

Source: Kaggle – “IMDB Movie Reviews Dataset” by Lakshmipathi Narasimhan (Narasimhan, 2020).

Size: 50,000 reviews, 2 columns (review, sentiment)

Target Variable: sentiment (binary: positive or negative)

The IMDB Movie Reviews dataset was chosen for this sentiment analysis task for its relevance, balance, quality, and suitability for training a text-based classification model (OpenAI, 2025). Given the client’s goal of identifying customer sentiment trends from review data, the chosen dataset aligns well. Some key reasons for this choice include:

- Relevance to Sentiment Classification Objectives: The dataset contains 50,000 movie reviews, each of which is labelled either positive or negative (Narasimhan, 2020). Although sourced from a film domain, the structure and content directly support the client’s broader goal of building models capable of assessing public sentiment, like customer reviews from HelloPeter (OpenAI, 2025).
- Balanced Target Distribution: The target variable sentiment is evenly split with 25,000 positive and 25,000 negative reviews (Narasimhan, 2020). Such a balance promotes stability during the model training process and avoids bias towards a dominant class, giving fair evaluation across both sentiments (Narasimhan, 2020).
- Clean and Minimalist Structure: The dataset offers two well-defined columns – review (free-text) and sentiment (categorical) (Narasimhan, 2020). No missing values are found, and the data types are appropriate for text-based machine learning pipelines with text strings for reviews and binary labels for classification (Narasimhan, 2020).
- Sufficient Volume and Variability: 50,000 observations ensure the dataset is large enough to support robust training and model generalisation (OpenAI, 2025). Such a scale also offers the opportunity to experiment with different model types.
- Recognition as a Benchmark Dataset: This IMDB dataset is considered a reliable standard for performance benchmarking due to its extensive use in academic literature (OpenAI, 2025). Extensive documentation of the dataset behaviour across a range of algorithms has been conducted, aiding in both validation and reproducibility of results (Narasimhan, 2020).
- Transferability Across Domains: The expressions of sentiment are largely domain agnostic in the dataset, and so models trained on this dataset can be reasonably

adapted to other textual review scenarios like product feedback, complaint analysis, or customer satisfaction tracking (OpenAI, 2025).

The IMDB Movie Reviews Dataset is of high quality, well-balanced, and domain-flexible, offering a strong foundation for developing sentiment classification models (Narasimhan, 2020). Given the methodological credibility and practical relevance, it aligns well with the project goals.

Model Justification

Logistic Regression was selected as the primary algorithm for this sentiment classification task due to its balance between predictive capability, interpretability, and computational efficiency (Muller & Guido, 2016). Logistic Regression, being tailored for binary classification problems, is particularly useful for scenarios where outcomes fall into two distinct categories, which in this case are positive or negative sentiment (Muller & Guido, 2016).

Furthermore, Logistic Regression performs well with sparse, high-dimensional feature spaces, such as those produced by Term Frequency-Inverse Document Frequency (TF-IDF) vectorisation (OpenAI, 2025). During the feature engineering process, each review was transformed into a numerical representation capturing unigram and bigram frequencies, resulting in a 10,000-dimensional feature space (OpenAI, 2025). Logistic Regression can effectively handle such data structures without the requirement of dimensionality reduction, making it well-suited for text classification tasks (Muller & Guido, 2016).

More so, interpretability was a strong consideration in model selection. Logistic Regression offers transparent coefficient estimates for each feature, which is effective for determining which terms correlate to positive or negative sentiment classifications (Muller & Guido, 2016). Such transparency is important in business environments where clients may need justification for classification decisions for audit or policy formulation purposes (OpenAI, 2025).

Additionally, the model is highly efficient and trains quickly on large datasets, offering reliable probability estimates for each prediction (Muller & Guido, 2016). Such probabilistic output can be used to implement thresholds or confidence scoring mechanisms in downstream systems, which is important for practical deployment (OpenAI, 2025).

Whilst some alternative models are more complex, such as Random Forests or Transformer-based language models (BERT, for example), which could offer marginal gains in accuracy, they lack the interpretability required and require significantly more computational resources (Muller & Guido, 2016; Hashemi-Pour, 2024). Given the context of this study, explainability and efficiency are desired over marginal performance gains; therefore, Logistic Regression offers a pragmatic and technically robust solution (OpenAI, 2025).

To conclude, Logistic Regression offers a scalable, transparent, and effective approach for binary text classification whilst aligning with best practices in natural language processing (Muller & Guido, 2016). Serving as a good benchmark for models to be developed in the future as an extension of this work (OpenAI, 2025).

Exploratory Data Analysis (EDA)

Figure 1: Sentiment Distribution Bar Chart

The bar chart illustrated in Figure 1 shows a perfectly balanced class distribution, with 25,000 positive and 25,000 negative reviews. Such seen symmetry is ideal for binary classification tasks, as there is no need for resampling strategies (oversampling or undersampling) and ensures no bias towards a dominant class in accuracy and other metrics (Muller & Guido, 2016). Finally, a more reliable model evaluation and fairer learning across sentiment categories are facilitated (OpenAI, 2025).

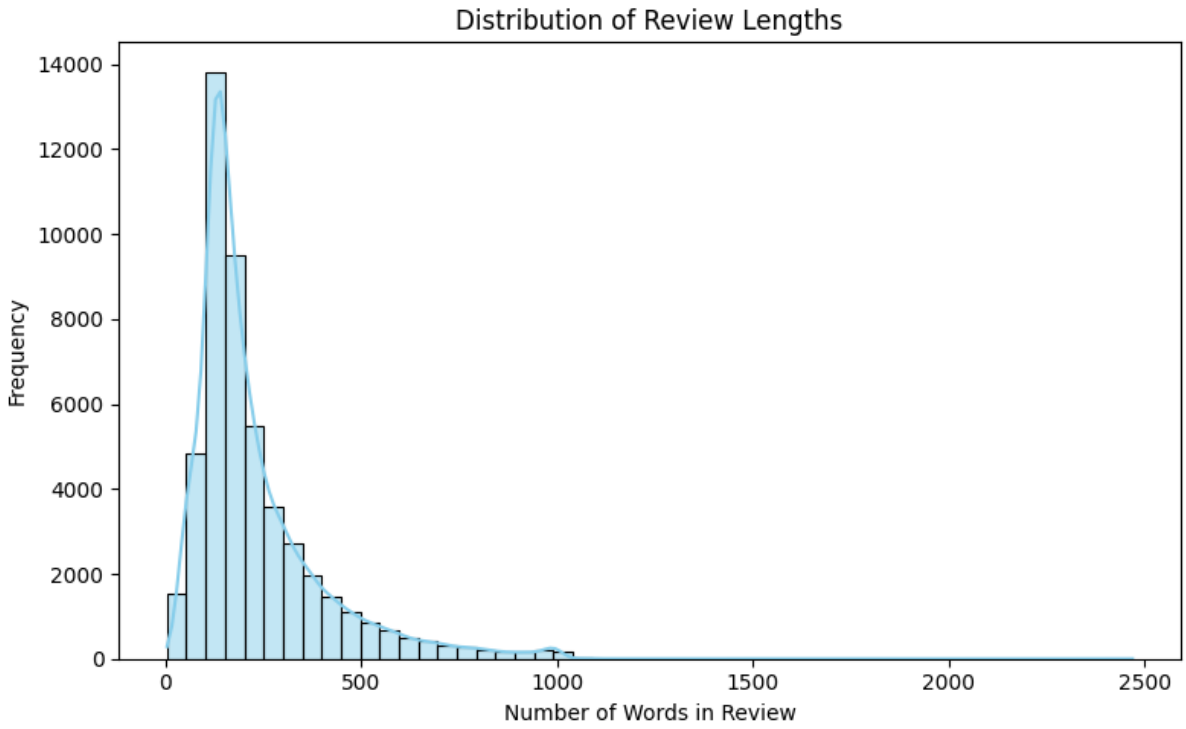

Figure 2: Distribution of Review Lengths Histogram

The histogram seen in Figure 2 reveals a right-skewed distribution of review lengths, with the majority of reviews falling between 100 and 300 words. The average review length is 231 words, with a few being extremely long reviews extending beyond 1,000 words. While most user-generated content is concise, there are significant outliers that may affect model performance, as seen from the long-tail distribution (OpenAI, 2025). The findings in this EDA outline the importance of preprocessing steps like truncation or padding to standardise input lengths, which is necessary when vectorisation techniques like TF-IDF are applied (Narasimhan, 2020).

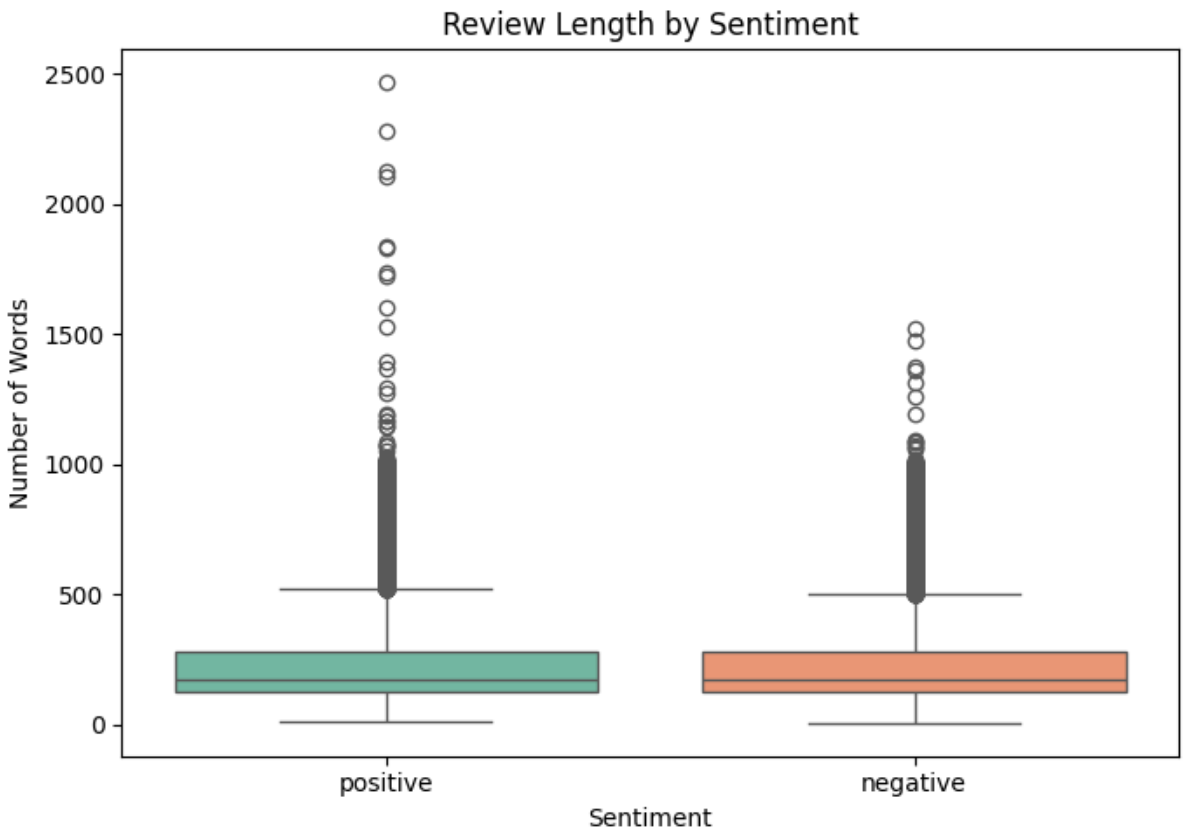

Figure 3: Review Length by Sentiment Box Plot

Figure 3 shows the box plots displaying the distribution of review lengths across the two sentiment classes (positive and negative). Given that both categories show similar median word counts, no major disparity in the typical length of reviews between classes is seen. Although the spread and number of outliers are considerable, specifically for positive reviews, which include several very long entries. This indicates a high degree of variability in how users express sentiment, with some individuals writing extensively regardless of sentiment polarity (OpenAI, 2025). Longer positive reviews might show patterns that the model can learn from during vectorisation, despite the overlap in interquartile ranges (OpenAI, 2025). Overall, the need for careful preprocessing is reinforced, and review length should be included as a potential auxiliary feature (Narasimhan, 2020).

Modelling Process

The modelling process for this sentiment analysis involved a structured pipeline of data preprocessing, feature engineering via text vectorisation, and classification model development using logistic regression. Each step in this process was designed to ensure data quality, enhance model performance, and support the explainability of results.

Data Preprocessing

First, the IMDB dataset was inspected for structural integrity and completeness. The initial confirmation was that there were no missing values across its two columns: review and sentiment, and so no rows were dropped. To enable supervised learning classification, the sentiment column was encoded as binary values (positive = 1, negative = 0) from the originally labelled positive and negative (Suzanne, 2023). To explore the text verbosity as a potential pattern in sentiment, a review_length column was created (OpenAI, 2025).

A custom text preprocessing function was applied to prepare the reviews for machine learning. The function standardised the review text through converting all characters to lowercase, removing HTML tags, URLs, punctuation, and numbers, and eliminating stopwords from the Natural Language Toolkit (NLTK) corpus (Suzanne, 2023).

Feature Engineering and Selection

This project focused on raw text rather than structured numeric variables, and so to convert text into machine-readable input, the TF-IDF (Term Frequency-Inverse Document Frequency) vectorisation technique was applied (Suzanne, 2023). This method allowed the conversion of each review into a sparse matrix of numerical values representing the importance of each word across all reviews (OpenAI, 2025). The following parameters were used to configure the vectorisation: max_features=10000 – to cap dimensionality and focus on the most informative terms, ngram_range=(1, 2) – to include both unigrams and bigrams, enabling recognition of sentiment phrases like ‘not good’, min_df=5, max_df=0.9 – to remove rare and overly common terms that may contribute to noise (OpenAI, 2025). This vectorisation process produced a TF-IDF matrix with 10,000 features and 50,000 observations (OpenAI, 2025). These features are well-suited for logistic regression, which handles sparse matrices efficiently (Muller & Guido, 2016).

Model Training

The steps taken during the model training process are as follows: The feature matrix X_tfidf and the binary sentiment labels y were split into training and test sets, using an 80/20

ratio, respectively, using stratified sampling. This ensures both sets maintain class balance, which is critical for evaluation integrity (OpenAI, 2025).

Next, a logistic regression model was trained using Scikit-learn's LogisticRegression() with default parameters (C=1.0, penalty='l2', solver='liblinear'). For sparse input and the support of L2 regularisation to avoid overfitting in high-dimensional space, the liblinear solver is well suited (Muller & Guido, 2016). The model was then trained on the training data and then used to predict sentiments on the test set (OpenAI, 2025).

After the training and predictions, the model was evaluated using several metrics, including accuracy, precision, recall, F1 score, AUC (Area Under the ROC Curve), and a confusion matrix. To assess the model's ability to separate the two sentiment classes across various thresholds, the ROC curve and AUC value were visualised (OpenAI, 2025).

Finally, to assess robustness and generalisability, a second logistic regression model was retrained with a strong L2 regularisation parameter (C=0.1) (OpenAI, 2025). This retraining enabled the evaluation of how an increased penalty on model complexity affected performance, specifically in minimising overfitting whilst maintaining accuracy (Muller & Guido, 2016).

Model Evaluation

The trained logistic regression model was evaluated on the 20% holdout test set using a comprehensive set of classification metrics. Metrics used to evaluate include accuracy, precision, recall, F1 score, AUC (Area Under the ROC Curve), and a confusion matrix. To enhance interpretability, the two key visuals produced were a confusion matrix heatmap and the ROC curve.

Initial Model Performance (C=1.0)

- Accuracy: 89.53% - This metric shows the proportion of all correct predictions across the test set (Muller & Guido, 2016). Given that the dataset is balanced, accuracy is a reliable indicator of general model effectiveness (Muller & Guido, 2016).
- Precision 89.00% - Of all reviews predicted as positive, 89% were truly positive. In sentiment analysis, precision helps reduce misclassification of negative reviews as positive, which would potentially bias client reporting (OpenAI, 2025).
- Recall: 90.00% - Indicating that 90% of all actual positive reviews were correctly predicted. A high recall is important in cases where it's critical to identify all positive feedback, for instance, reputation management (OpenAI, 2025).
- F1 Score: 90.00% - The F1 score represents a harmonic mean of precision and recall (Muller & Guido, 2016). Offering a balanced view of false positives and false negatives, both of which affect customer sentiment analysis pipelines (Muller & Guido, 2016).
- AUC Score: 0.96 – This score represents how well the model distinguishes between classes across all thresholds (Muller & Guido, 2016). The score reflects a strong separation between positive and negative sentiment classes across all thresholds (Muller & Guido, 2016).

Threshold Consideration:

These results are calculated using a default decision threshold of 0.5. Meaning that any review with a predicted probability above 50% of being positive is classified as positive (OpenAI, 2025). In deployment, this threshold can be lowered to increase recall for negative reviews.

Confusion Matrix:

- True Positives: 4522
- True Negatives: 4431
- False Positives: 569
- False Negatives: 478

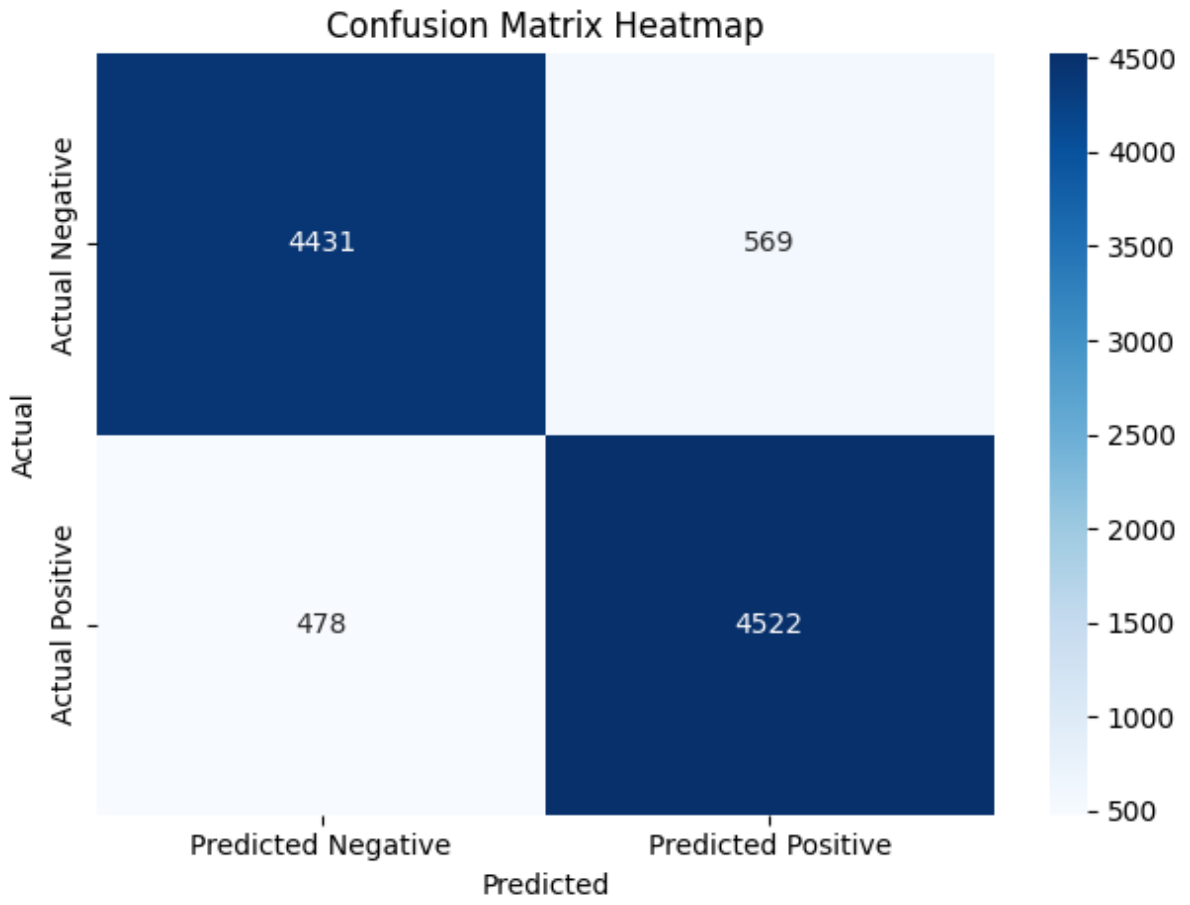

Figure 4: Confusion Matrix Heatmap – Initial Model

The confusion matrix seen in Figure 4 demonstrates strong balance in predictive accuracy across both classes, with near-symmetrical performance (Wagavkar, 2024). The relatively lower number of false predictions suggests the model can effectively distinguish sentiment polarity (OpenAI, 2025).

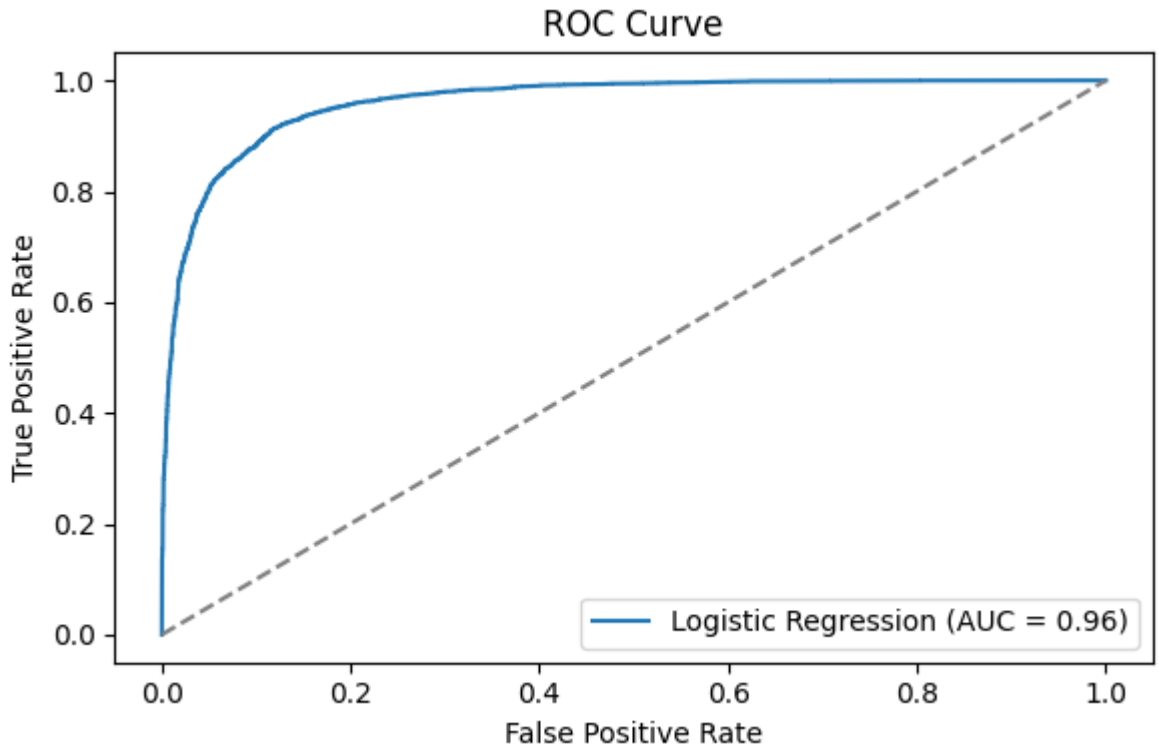

Figure 5: ROC Curve – Initial Model

The ROC curve seen in Figure 5 shows a steep rise and an AUC of 0.96, meaning a high true positive rate across most thresholds (Dash, 2022). The strong area under the curve shows excellent sensitivity and specificity (Dash, 2022). It can also be assumed from the strong AUC score that the model is flexible in adjusting thresholds to meet various business goals (OpenAI, 2025).

Model Robustness and Limitations

Although the performed well on the test data, it should be noted that the dataset is pre-cleaned and evenly balanced, which are ideal conditions for machine learning (Narasimhan, 2020). In real-world applications like HelloPeter, sentiment mining may find misspellings, sarcasm, mixed languages, or imbalanced classes, which can all affect the model's performance (OpenAI, 2025). To further enhance real-world generalisation, it would be wise to apply data augmentation and advanced text embeddings like BERT (Hashemi-Pour, 2024).

Conclusion

To conclude, the logistic regression model trained on TF-IDF features demonstrated solid performance in classifying review sentiment. The high accuracy and AUC, along with the balanced precision and recall, underscore the model’s reliability as a foundational model for sentiment classification (Muller & Guido, 2016). So this model is considered a viable candidate

for deployment or even a benchmark for comparison with future iterations of more complex models (OpenAI, 2025).

Model Retraining

Despite the strong and balanced performance from the initial model, model retraining was conducted with an adjusted regularisation parameter, C=0.1, down from the default C=1.0. The retraining was down for two key reasons: To test the model's robustness and generalisation under increased regularisation, and to satisfy the evaluation protocol of the project.

Regularisation can be described as a technique used in machine learning to reduce model complexity and prevent overfitting by penalising large coefficients (Muller & Guido, 2016). Lowering the regularisation by reducing the regularisation parameter to 0.1 in the second model can promote simpler, more generalisable decision boundaries (Muller & Guido, 2016). To ensure an isolated effect of the regularisation parameter adjustment, no changes were made to the feature set, preprocessing pipeline, or data splits. Additionally, the threshold for classification was maintained at the default value of 0.5. The same metrics and visuals were used to interpret the retrained model, including the Receiver Operating Characteristic (ROC) curve and the confusion matrix heatmap.

Retrained Model Performance (C=0.1)

- Accuracy: 87.56% - This accuracy value shows a slight drop in predictive performance as compared to the initial model's accuracy of 89.53%. The reduced accuracy may be an indication of underfitting, a common result from stronger regularisation (Muller & Guido, 2016).
- Precision 86.32% - Based on this precision score, the retrained model was more conservative, producing fewer false positives. This might be considered desirable in situations where incorrectly labelling negative reviews as positive could mislead stakeholders (OpenAI, 2025).
- Recall: 89.46% - The retrained model detected true positive reviews at a slightly reduced rate compared to the initial model, indicating a slight sensitivity trade-off (Muller & Guido, 2016).
- F1 Score: 87.86% - A strong balance between precision and recall was maintained with the retrained model, even though the overall value reduced slightly from 90%, which can be expected with the stronger regularisation (OpenAI, 2025).
- AUC Score: 0.95 – While the retrained model's AUC score was marginally lower than the original model's AUC score of 0.96, the retrained model's score indicates excellent class separability and strong generalisation across thresholds (OpenAI, 2025).

Confusion Matrix:

- True Positives: 4473
- True Negatives: 4283
- False Positives: 717
- False Negatives: 527

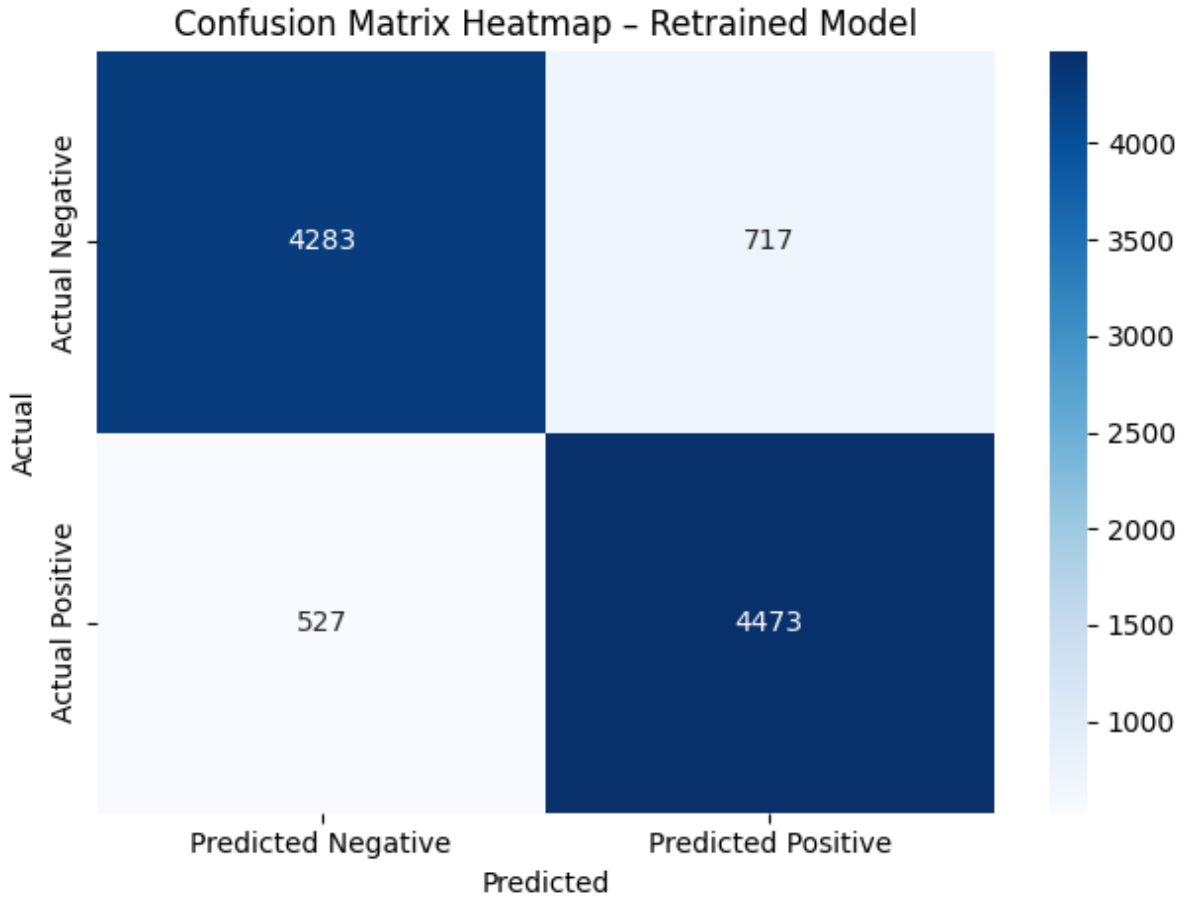

Figure 6: Confusion Matrix Heatmap – Retrained Model

From the heatmap seen in Figure 6, we can conclude that the retrained model predicted a slightly increased number of false positives and false negatives compared to the original model. Although the balance is still deemed as good, this confirms that the stricter regularisation has come at a small cost in classification accuracy and sensitivity (OpenAI, 2025).

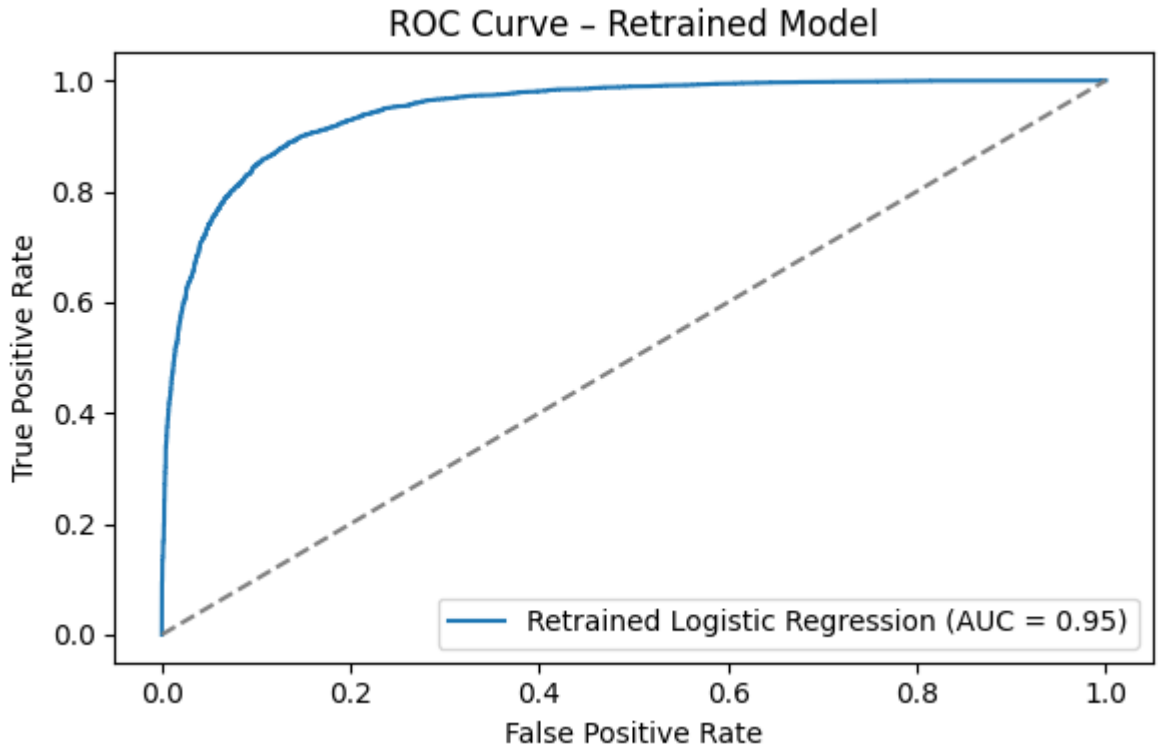

Figure 7: ROC Curve – Retrained Model

The ROC curve seen above in Figure 7 shows that the retrained model displays a strong diagonal rise with an AUC of 0.95. The AUC is slightly lower than the original model's AUC of 0.96, but it still shows high performance in distinguishing between positive and negative sentiment, underscoring the robustness of logistic regression even with stronger penalisation (Dash, 2022).

Model Robustness and Limitations

Although the retrained model achieved strong performance overall, the increased regularisation introduced a slight drop in recall and AUC, showing the conservative nature of the decision boundary (Muller & Guido, 2016). Such behaviour helps confirm that the model is more cautious in predicting positive reviews, which could be potentially useful in scenarios where false praise must be avoided (OpenAI, 2025). Although genuine positive customer feedback may be overlooked, this may limit customer experience initiatives (OpenAI, 2025). Although the retained model is mildly less prone to overfitting and more generalisable, the reduced sensitivity may not justify the performance trade-off, considering the importance of capturing sentiment nuances (OpenAI, 2025). For future iterations, threshold tuning or contextual embeddings like BERT should be considered (Hashemi-Pour, 2024).

Conclusion

In conclusion, a delicate balance between regularisation and performance in sentiment classification has been highlighted in this retraining exercise. Strong generalisation and slightly reduced performance metrics were seen in the stricter model, as expected with increased regularisation (Murel, 2023). Under domain-specific requirements, application of this model should be thought out, specifically if missing positive or negative sentiment has real-world reputational effects (OpenAI, 2025).

Recommendations & Conclusion

Recommendations based on findings are as follows:

1. Deploy Initial Model (C=1.0) for Production Use: The initial model achieved strong and balanced results, confirming its suitability as a baseline model for text-based sentiment classification.
2. Avoid Over-regularisation Without Compensating Adjustments: Given the slightly reduced performance results of the retrained model with its stronger penalty, particularly in recall, it can be assumed that the increased regularisation may have led to underfitting, causing the model to overlook subtle sentiment signals (Murel, 2023). So, in future deployments, to compensate for sensitivity loss, mechanisms like threshold tuning or ensemble smoothing should be incorporated when reducing complexity (Murel, 2023).
3. Implement Threshold Tuning Based on Business Needs: The default classification threshold may not be aligned with the operational priorities. In the instance of managing reputation, a higher recall might be desirable to ensure all negative reviews are caught (OpenAI, 2025). It is recommended to evaluate the precision-recall trade-off across all various thresholds and adjust accordingly (Muller & Guido, 2016).
4. Monitor Input Length and Consider Pre-Truncation: Given the wide variety of review lengths seen in the EDA results, a maximum token limit or review summarisation pipeline should be considered, as extreme lengths may dilute important features (OpenAI, 2025).
5. Advanced Models for Future Iterations: For future projects, this logistic regression model should be considered as a benchmark, and more advanced architectures like Random Forests or Transformer-based models (BERT, for example) should be developed (Hashemi-Pour, 2024). These more advanced models may improve performance on nuanced sentiment tasks, although at the cost of interpretability and computational complexity (Muller & Guido, 2016).
6. Continuous Model Evaluation and Feedback Loop: Given the nature, sentiment language may evolve, so a feedback mechanism should be incorporated which will support periodic retraining with updated datasets, giving insights into new performance metrics (Muller & Guido, 2016).

Conclusion

To summarise, this project successfully demonstrated the effectiveness of logistic regression for binary sentiment classification on a benchmark Natural Language Processing dataset (Narasimhan, 2020). The approach resulted in an interpretable and performant model that offers a practical solution for real-world applications involving customer feedback analysis. The importance of model tuning and domain alignment has been highlighted through regularisation and retraining (Murel, 2023). A solid foundation has been provided by these insights, enabling expansion into more complex or domain-specific sentiment analysis tasks (OpenAI, 2025).

Disclosure of AI Use

Sections: POE.

Name of the tool used: ChatGPT4.

Purpose behind use: Outlines, summaries, Python analysis code, queries, evaluations, and suggestions.

Date used: 07/05/2025.

Link to chat: https://chatgpt.com/share/6800b44d-98b4-8004-a81b-7855cb0b5000

References

Dash, S., 2022. *Understanding the ROC and AUC Intuitively.* [Online]
Available at: https://medium.com/@shaileydash/understanding-the-roc-and-auc-intuitively-31ca96445c02
[Accessed 13 June 2025].

Hashemi-Pour, C., 2024. *BERT language model.* [Online]
Available at: https://www.techtarget.com/searchenterpriseai/definition/BERT-language-model
[Accessed 13 June 2025].

Muller, A. C. & Guido, S., 2016. *Introduction to Machine Learning with Python.* 1st ed. Sebastopol: O'Reilly Media.

Murel, J., 2023. *What is regularization?.* [Online]
Available at: https://www.ibm.com/think/topics/regularization
[Accessed 13 June 2025].

Narasimhan, L., 2020. *Kaggle - Sentiment Analysis of IMDB Movie Reviews.* [Online]
Available at: https://www.kaggle.com/code/lakshmi25npathi/sentiment-analysis-of-imdb-movie-reviews
[Accessed 06 June 2025].

OpenAI, 2025. *Open AI ChatGPT4.* [Online]
Available at: https://chatgpt.com/share/6800b44d-98b4-8004-a81b-7855cb0b5000
[Accessed 10 April 2025].

Suzanne, 2023. *Data Pre-Processing for Linear Regression in Machine Learning.* [Online]
Available at: https://medium.com/@sds152/data-pre-processing-for-linear-regression-in-machine-learning-4b73ec48392a
[Accessed 17 April 2025].

Wagavkar, S., 2024. *Introduction to the Correlation Matrix.* [Online]
Available at: https://builtin.com/data-science/correlation-matrix#:~:text=For%20example%2C%20let's%20say%20you,the%20relationship%20is%20positively%20strong.
[Accessed 20 April 2025].