NLP Sentiment Analysis

Exploring the power of Natural Language Processing to extract meaningful insights from unstructured text data.

- AMANDEEP
- ARTIKA SINGH
- MR. DEEPESH KUMAR VERMA
- KAVETI NANI KARTIK
- MEDI SANJAY
- MEKALA SHIVAKUMAR
- RAMELLA HIMA BINDU

Business Objective: Sentiment Extraction

Goal

Extracting sentiment from customer product reviews.

Application

Informing product development and marketing strategies.

Value

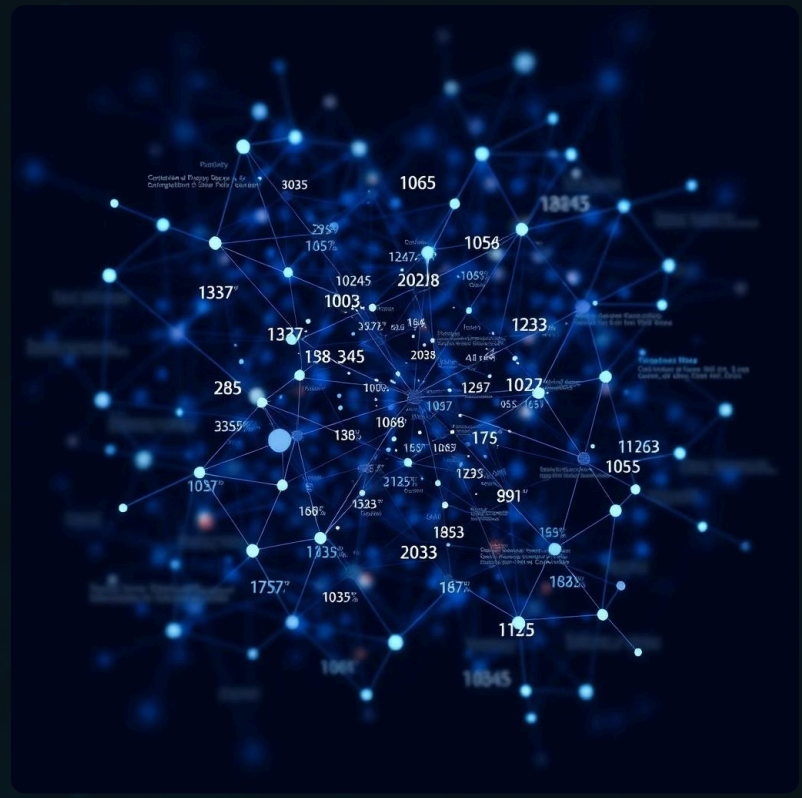
Gaining actionable insights from raw customer feedback at scale.

Dataset Structure: Customer Reviews

The dataset comprises user reviews for a specific product, sourced from a prominent e-commerce platform. Each entry provides a comprehensive view of customer feedback, crucial for granular sentiment analysis.

It contains three key columns:

- **Title:** Concise summary of the review's main point.
- **Rating:** Numerical score reflecting overall satisfaction (e.g., 1-5 stars).
- **Body:** Detailed textual content of the user's review.



Data Pre-processing Pipeline

1

Sentiment Score Calculation and Labeling

Assigning sentiment scores and categorizing reviews.

2

Rating '3' Exclusion

Removing neutral reviews for clearer sentiment distinction.

3

Outlier Detection

Filtering unusual review lengths to enhance model robustness.

Sentiment Score Calculation & Labeling

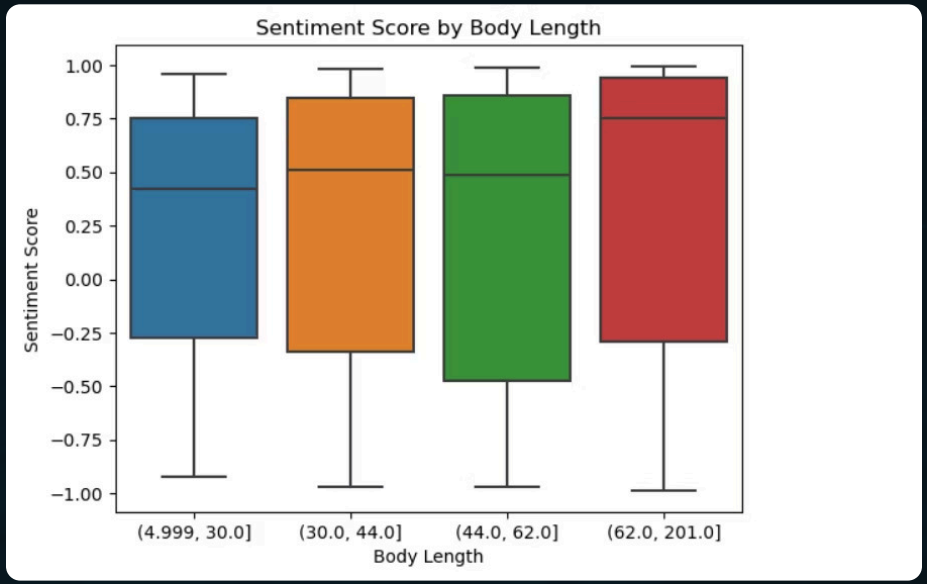
Reviews were processed using the VADER sentiment analyzer, generating a score from -1 (negative) to +1 (positive). A custom rule-based function then categorized each review as either 'positive' or 'negative', transforming raw text into actionable sentiment categories for subsequent analysis. This binary classification streamlines downstream modeling efforts.

Excluding Neutral Ratings (Rating '3')

Reviews with a rating of '3' were strategically excluded to eliminate ambiguity. These neutral ratings often dilute the distinction between truly positive and negative sentiments. This exclusion ensures a cleaner binary classification, significantly improving the precision and recall of the sentiment model by focusing on clear emotional polarity.

Outlier Detection: Text Length Analysis

Outlier detection focused on review text length. Exceptionally short or long reviews, often indicative of irrelevant content or data entry errors, were identified and removed. This crucial step prevents skewed model training, ensuring that the dataset contains only high-quality, representative data. The interquartile range (IQR) method was employed for robust outlier identification.

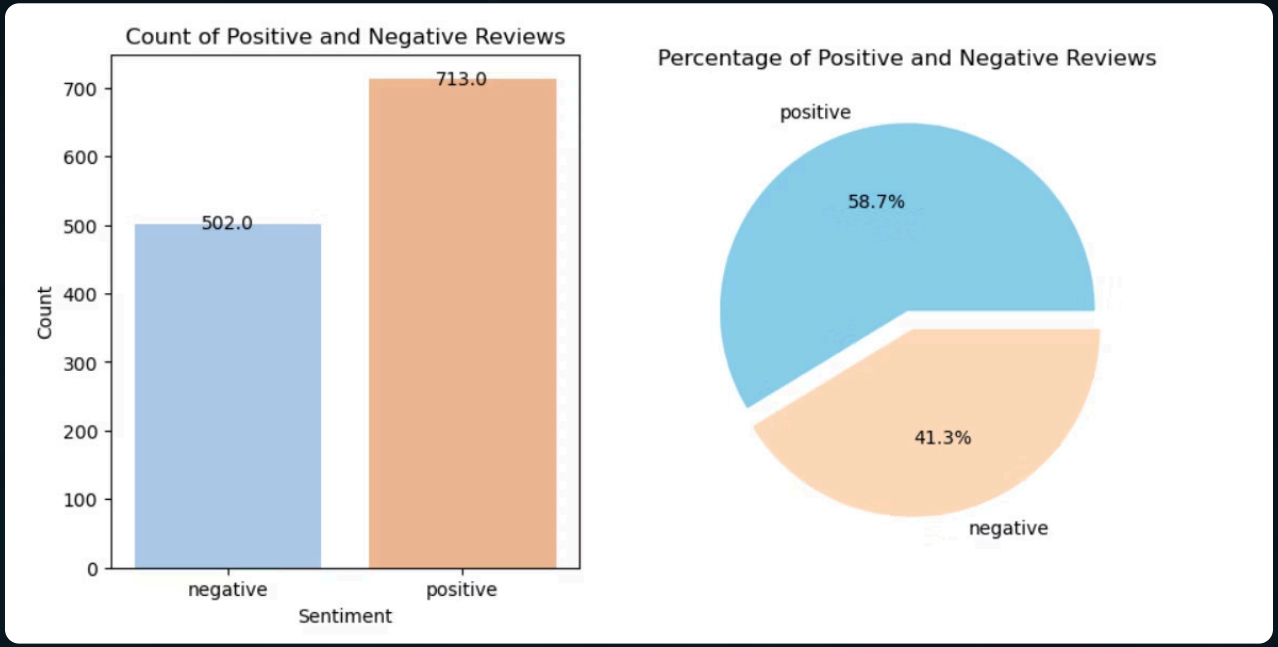


Exploratory Data Analysis (EDA)

Our EDA encompassed a deep dive into review lengths, sentiment labels, and ratings. Visualizations like histograms, box plots, count plots, and pie charts revealed critical patterns: how review length correlates with sentiment, the distribution of positive vs. negative reviews, and rating spread across sentiment categories. This holistic view provided a solid foundation for effective model training and informed feature engineering.

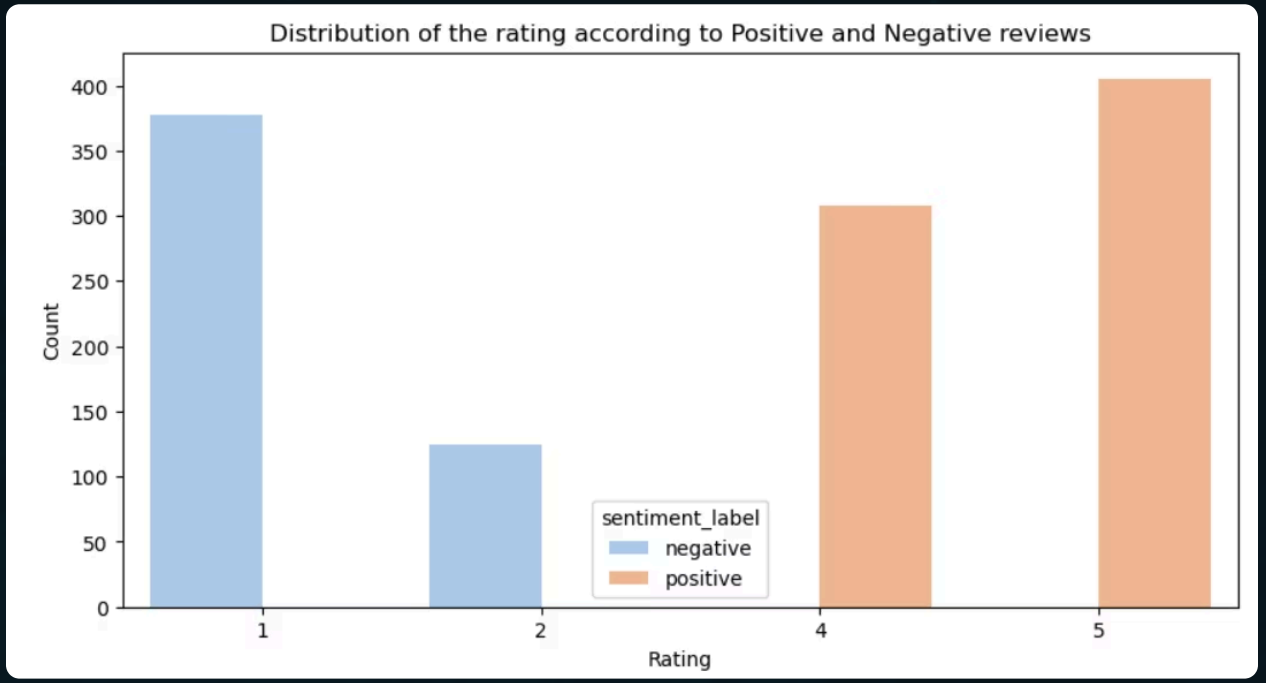
Sentiment Distribution

The bar chart clearly illustrates the dominant presence of positive reviews, outnumbering negative ones significantly. This imbalance is common in product review datasets and highlights the need for stratified sampling or oversampling techniques during model training to prevent bias towards the majority class.



Rating Distribution by Sentiment

This visualization shows that higher ratings (4 and 5 stars) are almost exclusively associated with positive sentiment, while lower ratings (1 and 2 stars) correspond to negative sentiment. This confirms the effectiveness of our sentiment labeling process post-VADER analysis and reinforces the data's suitability for binary classification.



Data Cleaning Pipeline

1 Merge Title and Body

Combined review titles and bodies into a single text field to ensure comprehensive sentiment capture.

2 Remove HTML Tags

Stripped all HTML elements (e.g.,

,

) from the text to eliminate irrelevant markup.

3 Clean Punctuation & Special Characters

Eliminated all punctuation marks and special characters using regex for simplified text processing.

4 Remove Stopwords

Filtered out common, low-information words (e.g., "the", "is") to focus on salient content.

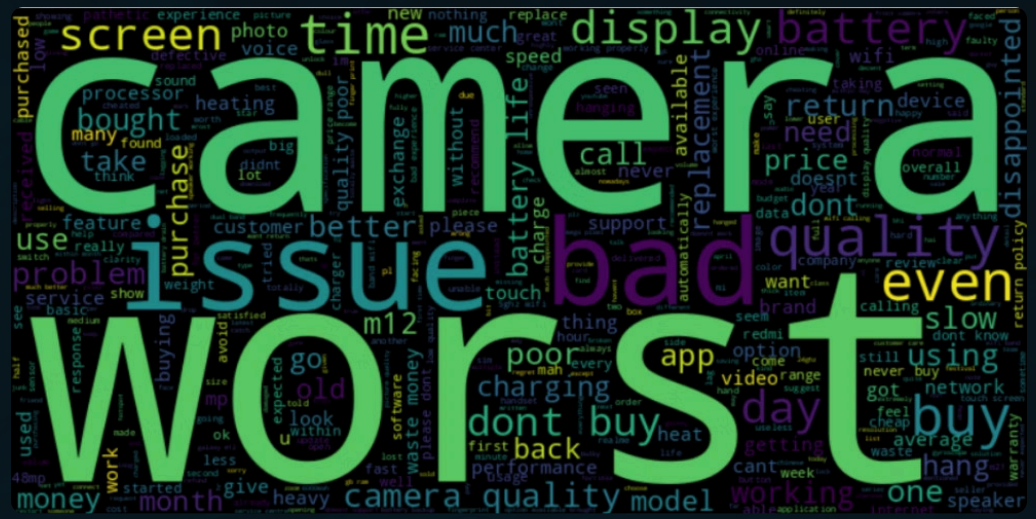
5 Apply Lemmatization

Reduced words to their base forms (e.g., "running" to "run") to normalize vocabulary and improve analysis.

Sentiment Visualization: Word Clouds

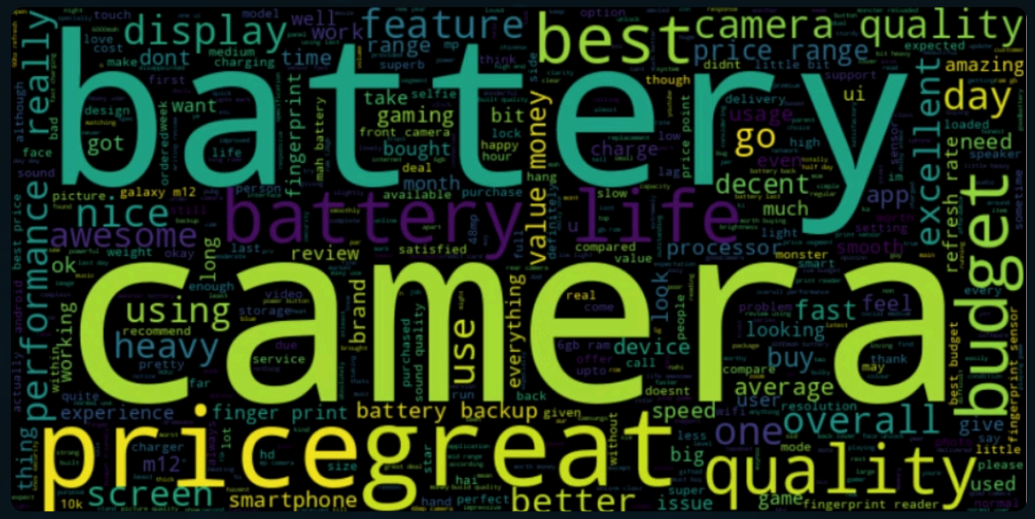
Word clouds were generated for both positive and negative sentiment categories to highlight frequently occurring terms, offering a quick visual summary of key themes.

Negative Sentiment



Common words associated with negative reviews, indicating areas for improvement.

Positive Sentiment



Frequently used terms in positive feedback, highlighting strengths and popular features.

Feature Engineering & Imbalance Handling

During the feature engineering phase, the dataset was split into training and testing sets. We then applied the Synthetic Minority Over-sampling Technique (SMOTE) to address the class imbalance, specifically due to the higher volume of positive reviews. This ensures that the model is not biased towards the majority class and can accurately predict both positive and negative sentiments.

SMOTE generates synthetic samples for the minority class, preventing overfitting and improving classification accuracy on imbalanced datasets.

Model Selection & Training

A diverse set of machine learning and deep learning models were rigorously tested to determine the most effective approach for sentiment classification.

1

Traditional ML

- Logistic Regression
- Multinomial Naive Bayes
- Support Vector Machine (SVM)
- Random Forest
- Ridge Classifier

2

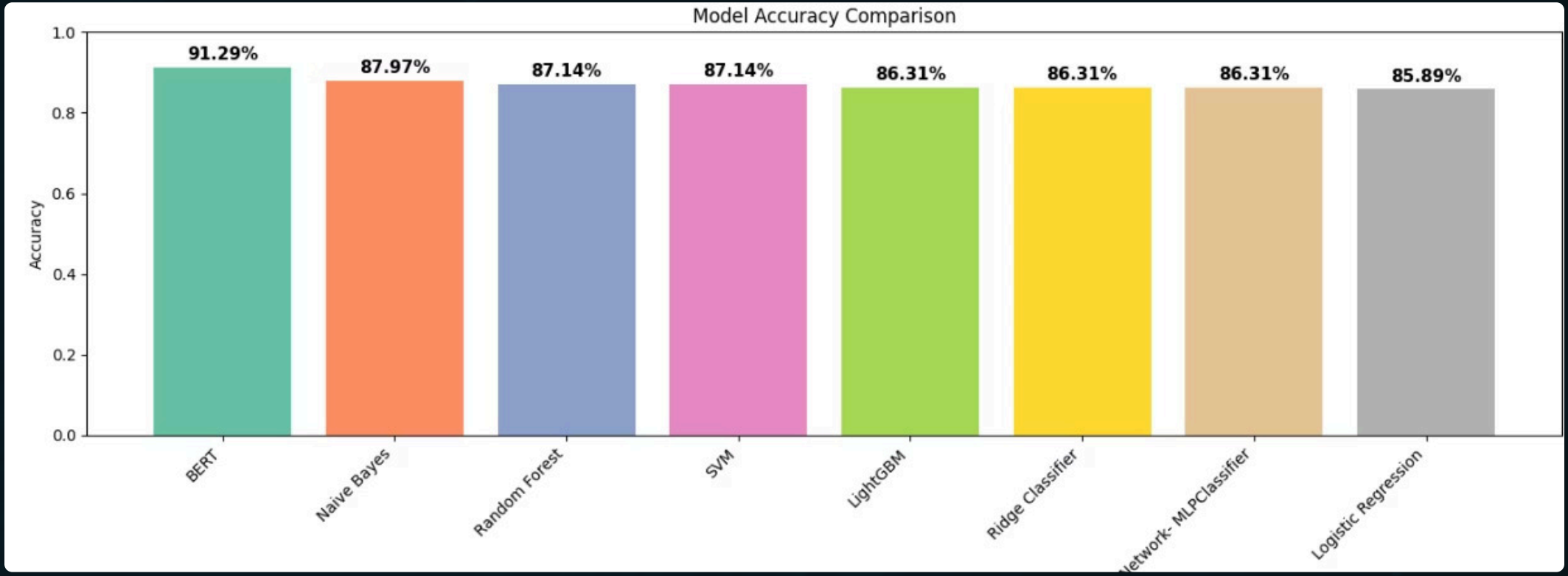
Advanced ML & DL

- Neural Network (MLPClassifier)
- LightGBM
- BERT (Bidirectional Encoder Representations from Transformers)

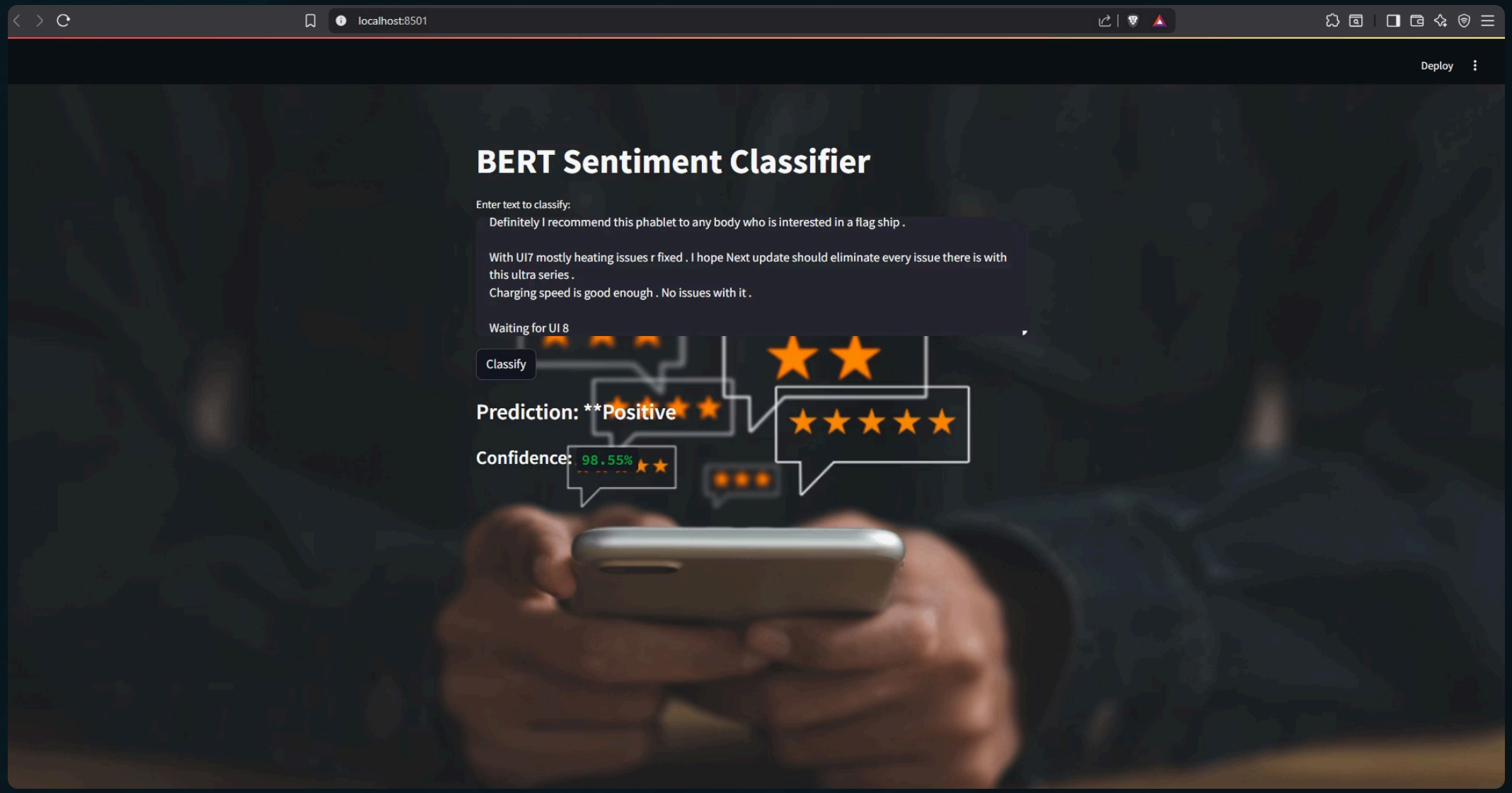
Model Performance: Accuracy Metrics

The performance of each trained model was evaluated based on accuracy scores. This chart illustrates the comparative effectiveness of each algorithm in correctly classifying sentiment.

BERT, a transformer-based model, generally outperforms traditional ML models due to its superior contextual understanding of text.



Sentiment Analysis Application



Detailed Output Insights

The system provides a comprehensive breakdown of each review, including the original text, predicted sentiment label (Positive/Negative), and the confidence score for the prediction. This granular data allows for deeper analysis and targeted actions.

This example output demonstrates the detailed information provided for each customer review, which is critical for operational feedback loops.