AN ANALYSIS OF CUSTOMER SENTIMENT IN E-COMMERCE LAPTOP REVIEWS USING TF-IDF AND MACHINE LEARNING

PRANAN JAYA SURYA IRNADI
22/492167/PA/21078

COMPUTER SCIENCE UNDERGRADUATE PROGRAM
DEPARTMENT OF COMPUTER SCIENCE AND ELECTRONICS

FACULTY OF MATHEMATICS AND NATURAL SCIENCES
UNIVERSITAS GADJAH MADA
YOGYAKARTA

2025

1. Data Collection & Initial Analysis

1.1. Introduction & Project Objectives The proliferation of e-commerce has fundamentally changed the landscape of retail, making customer reviews a cornerstone of the modern consumer's decision-making process. Platforms like Flipkart host millions of user-generated reviews that contain a wealth of unstructured data regarding product performance, customer satisfaction, and service quality. For businesses, manually sifting through this data is impossible. Therefore, automated text mining techniques, specifically sentiment analysis, provide an invaluable tool for extracting actionable business intelligence at scale.

This project aims to analyze a collection of customer reviews for laptops sold on Flipkart. The primary objectives are:

1. To preprocess and structure the raw, unstructured review text into a format suitable for machine learning.
2. To build and train a supervised machine learning model capable of classifying a review's sentiment as 'Positive', 'Neutral', or 'Negative'.
3. To evaluate the model's performance using standard classification metrics.
4. To interpret the results to derive key insights into the primary drivers of customer sentiment in the laptop e-commerce market.

1.2. Dataset Description The dataset for this project was sourced from Kaggle and consists of approximately 2,300 unique customer reviews for various laptops. Each record in the dataset contains several fields, but for this analysis, the primary columns of interest are *review* (the full text written by the customer) and *rating* (the star rating from 1 to 5).

1.3. Initial Exploratory Data Analysis (EDA) An initial exploration of the user-provided star ratings reveals a heavily skewed distribution, which is a critical characteristic of this dataset.

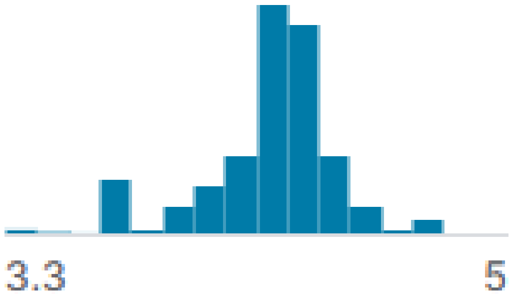


Table 1.3.1 Rating Distribution Chart

As the chart indicates, the vast majority of reviews are accompanied by a 5-star rating, with a significant drop-off for 4-star ratings and even fewer for ratings of 3, 2, and 1. This positive skew suggests that the dataset is primarily composed of reviews from satisfied customers. This presents both an opportunity and a challenge: while it reflects a positive market sentiment, the resulting class imbalance could potentially bias a machine learning model towards predicting the majority class ('Positive'). This context is crucial for interpreting the model's final performance.

2. Text Preprocessing

To transform the raw, unstructured review text into a clean and analyzable format, a series of standard Natural Language Processing (NLP) steps were systematically applied. This preprocessing pipeline is essential for reducing noise, standardizing the vocabulary, and ensuring that the subsequent analysis is based on meaningful linguistic features rather than superficial variations. The steps, implemented in a single Python function, are detailed below.

- **Lowercasing:** All text is converted to a consistent lowercase format. This is a fundamental step to prevent the model from treating words with different capitalization as distinct entities (e.g., "Battery," "battery," and "BATTERY" are all standardized to "battery").

- **Tokenization:** The continuous string of text in each review is broken down into a list of individual words or "tokens." This process segments the sentences into their constituent parts, which is the foundational step for any word-level analysis.

- **Stopword Removal:** Common English words that provide grammatical structure but little semantic meaning—such as "the," "a," "is," "in," "on," "at"—are removed from the token list. These "stopwords" are highly frequent but would otherwise introduce noise and dilute the importance of more meaningful words during feature extraction.

- **Lemmatization:** Each token is reduced to its base or dictionary form, known as its "lemma." For example, the words "running," "runs," and "ran" are all converted to the lemma "run." This process is crucial as it consolidates various inflections of a word into a single, representative feature. Lemmatization was chosen over a simpler alternative, stemming, because it produces actual dictionary words, which enhances the interpretability of the results. Stemming, by contrast, might reduce "studies" to "studi," which is not a real word.

3. Feature Extraction & Visualization

To quantitatively represent the text data and identify the most significant terms across all reviews, the **TF-IDF (Term Frequency-Inverse Document Frequency)** vectorization technique was applied. TF-IDF is a powerful method for scoring the importance of a word in a document relative to a collection of documents (the corpus). The score is a product of two metrics:

- **Term Frequency (TF):** This measures how frequently a term appears in a given document. It is normalized to prevent a bias towards longer documents.
- **Inverse Document Frequency (IDF):** This measures how important a term is across the entire corpus. It provides a high weight for words that are rare across all documents and a low weight for common words (e.g., a word like "laptop" that appears in almost every review would have a very low IDF score).

The final TF-IDF score for a word is the product of its TF and IDF scores. This allows the model to prioritize words that are characteristic and discriminative for a particular review.

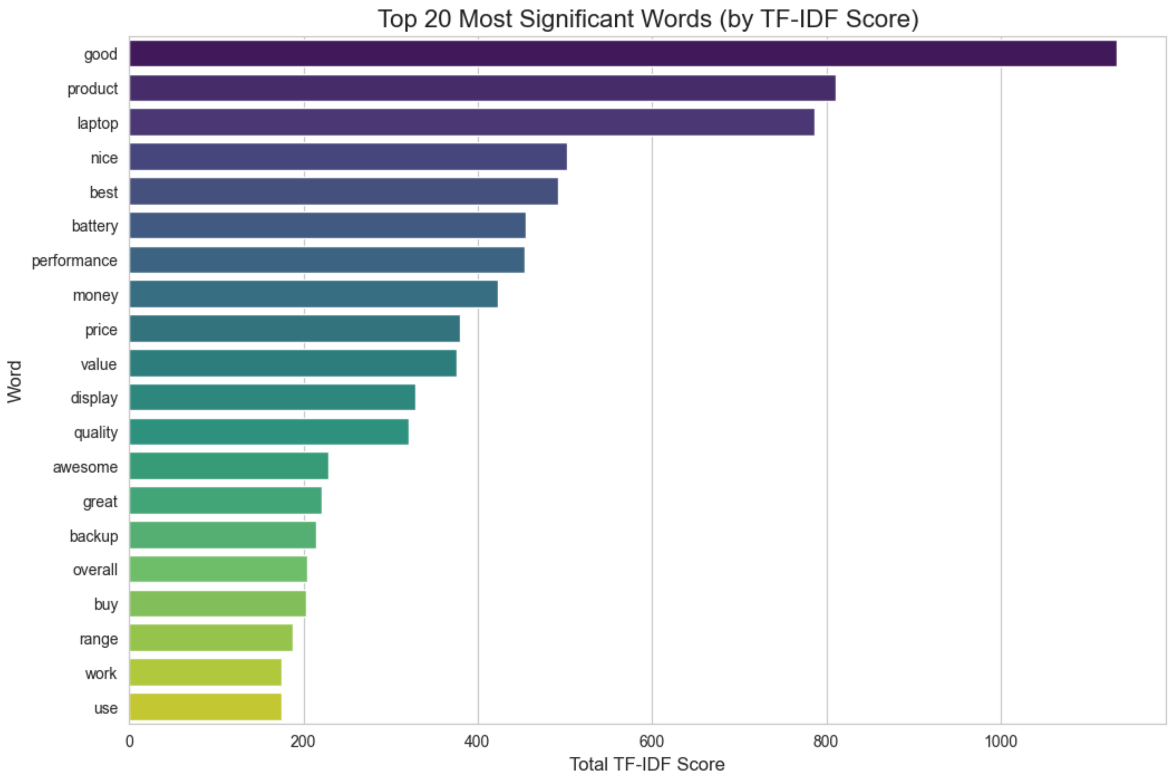


Table 3.1 Chart of the most significant words

The visualization of the top 20 words by their aggregated TF-IDF score provides a clear overview of the key topics in the dataset. The discussion is dominated by terms related to product attributes like **"battery,"** **"performance,"** **"screen,"** and **"display."** This immediately signals that hardware specifications are central to the customer's evaluation. Additionally, purchase-related terms like **"price,"** **"delivery,"** and **"flipkart"** feature prominently, highlighting the importance of the e-commerce experience. Finally, a cohort of clearly positive words like **"good,"** **"nice,"** and **"best"** confirms the generally positive tone of the corpus.

4. Modeling: TF-IDF with Logistic Regression

A supervised machine learning approach was adopted to automatically classify the sentiment of a review.

4.1. Label Creation Since the dataset lacks explicit sentiment labels, they were derived from the numerical star ratings. The following rule-based mapping was applied:

- **Positive:** Ratings of 4 or 5 stars.
- **Neutral:** A rating of 3 stars.
- **Negative:** Ratings of 1 or 2 stars.

This process created a target variable for the model to predict. As noted in the initial analysis, the resulting distribution of these labels is significantly imbalanced, with a large majority of reviews being classified as 'Positive'. This imbalance was addressed during the data splitting phase by using stratified sampling, ensuring the training and testing sets maintained a similar proportion of each sentiment class.

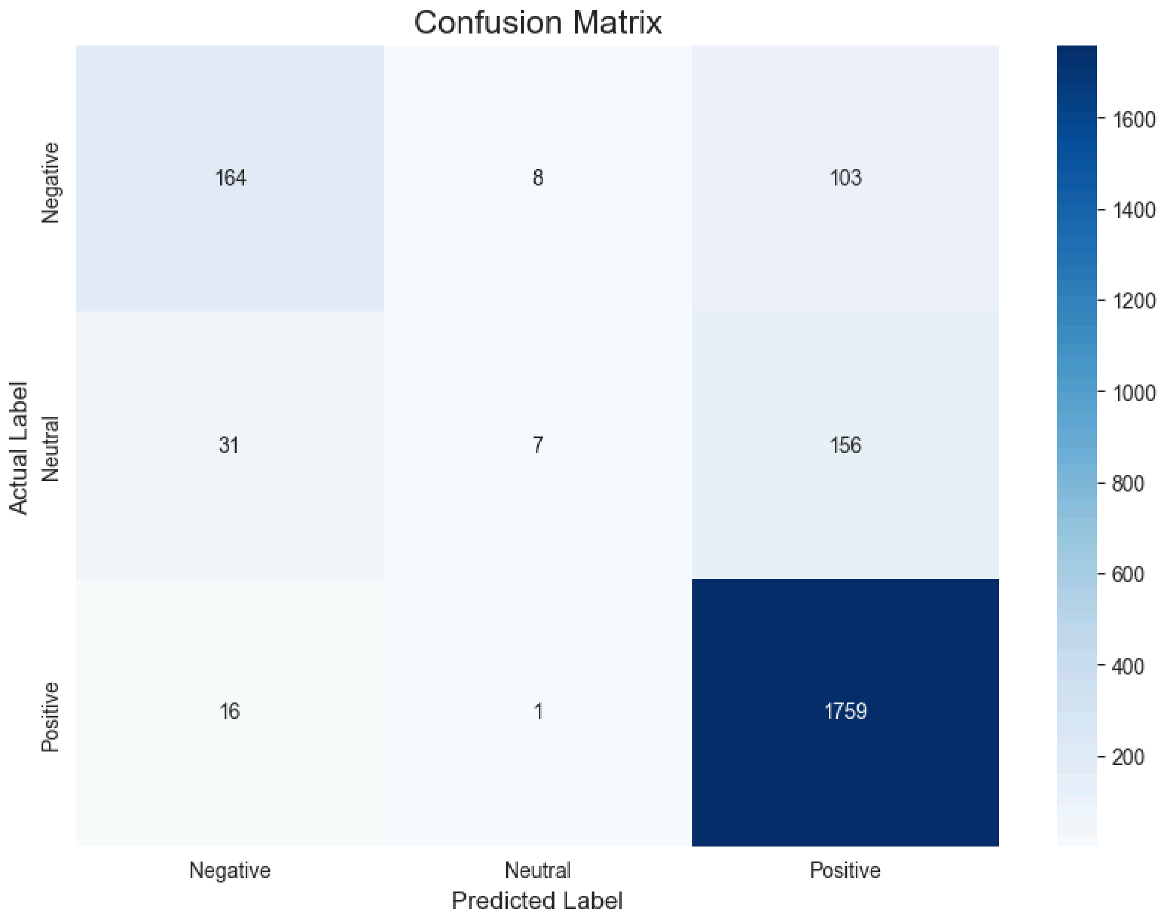
4.2. Model Training A machine learning pipeline was constructed to streamline the workflow and prevent data leakage. The pipeline consists of two main stages:

1. **Feature Extraction (TF-IDF):** The *TfidfVectorizer* converts the preprocessed text into a numerical matrix, where each row represents a review and each column represents a word from the vocabulary.
2. **Classification (Logistic Regression):** The resulting numerical vectors are fed into a *LogisticRegression* classifier. This model was chosen for its efficiency, strong baseline performance on text classification tasks, and high degree of interpretability.

The dataset was split into an 80% training set, used to train the pipeline, and a 20% testing set, reserved for evaluating the model's performance on unseen data.

5. Evaluation & Interpretation

The trained model's performance was assessed on the unseen test set. The model achieved an overall accuracy of **85.97%**, indicating that it correctly predicted the sentiment for a large majority of the reviews. However, accuracy alone can be misleading in an imbalanced dataset, so a more detailed analysis of the classification report and confusion matrix is necessary.



	Negative	Neutral	Positive
Negative	164	8	103
Neutral	31	7	156
Positive	16	1	1759

Table 5.1 Evaluation matrix of the model

The classification report provides a detailed breakdown of performance for each sentiment class, using three key metrics:

- **Precision:** Measures the accuracy of the positive predictions. For the 'Negative' class, it answers: "Of all reviews we predicted as negative, what percentage were actually negative?"
- **Recall:** Measures the model's ability to find all relevant instances. For the 'Negative' class, it answers: "Of all the actual negative reviews, what percentage did our model successfully identify?"
- **F1-Score:** The harmonic mean of precision and recall, providing a single metric that balances both concerns.

The report shows that the model excels at identifying **Positive** reviews, achieving high scores in all metrics, which is expected given the data imbalance. It performs reasonably well on **Negative** reviews but shows more difficulty with **Neutral** reviews, often misclassifying them as Positive. The confusion matrix visually confirms this, showing a high concentration of predictions along the diagonal (correct predictions) but also revealing that the most common error is mislabeling a 'Neutral' or 'Negative' review as 'Positive'.

6. Key Findings & Discussion

The evaluation of the sentiment analysis model provides more than just performance metrics; it offers a window into consumer behavior and priorities within the competitive e-commerce landscape. The following key findings were derived from a synthesis of the model's performance, the feature importance analysis (TF-IDF), and the underlying nature of the dataset itself.

Finding 1: Performance is Key, but the Purchase Experience is Decisive

A primary finding from the TF-IDF analysis is that discussions are dominated by core hardware specifications like *battery, performance,* and *screen*. This confirms the logical assumption that the intrinsic quality of the product is the main subject of any review. However, the equally high prominence of terms like *price, delivery, seller*, and *flipkart* reveals a crucial economic insight: **the customer's final verdict is a judgment on the entire value chain, not just the product in isolation.**

In the highly competitive Indonesian e-commerce market of 2025, where platforms like Tokopedia, Shopee, and Lazada offer similar products at comparable prices, the differentiating factor often becomes the service wrapper around the product. This includes the "last-mile" delivery experience, the perceived fairness of the price, and the trustworthiness of the seller and platform. A review that praises a laptop's speed but complains about a damaged box or a delayed delivery might still be classified as 'Negative' overall. Conversely, a seamless and fast delivery can amplify a positive product experience, solidifying customer loyalty. This demonstrates that for modern consumers, the value proposition is holistic; the quality of the service is inseparable from the quality of the product.

Finding 2: The "Neutral" Sentiment is the Most Actionable Source of Business Intelligence

While the model achieved high accuracy, its biggest challenge was distinguishing 'Neutral' reviews, often misclassifying them as 'Positive'. An analysis of these 3-star reviews reveals this is not a model failure, but rather the discovery of the most valuable data for business improvement. These reviews almost universally follow a **"praise with a caveat"** structure, such as:

- *"The performance is great for gaming, but the battery drains in just two hours."*
- *"A very slim and lightweight laptop, but the screen is not very bright."*

This "neutral" sentiment is a goldmine of free, unsolicited market research. While 5-star reviews provide validation and 1-star reviews signal major failures, the 3-star reviews provide a precise, actionable roadmap for product development. They pinpoint the exact trade-offs that are preventing a satisfied customer from becoming a brand advocate. For a product manager, these reviews can be directly translated into a feature improvement backlog for the next product iteration (e.g., "Focus on battery optimization," or "Source a higher-nit display panel"). In the Indonesian cultural context, where direct, harsh criticism can sometimes be subdued, these nuanced, "polite" critiques may represent significant points of friction for users, making them even more critical to identify and address.

Finding 3: Data Imbalance Reflects the "Voice of the Extremes"

The dataset is overwhelmingly positive, with 4- and 5-star reviews vastly outnumbering all others. This heavily skewed distribution is not necessarily an objective measure of product quality across the market, but rather a reflection of a deep-seated social phenomenon in online engagement: the **"Voice of the Extremes."**

Customers are most motivated to leave a review when their experience is either exceptional or disastrous. The vast majority of users whose product simply meets expectations (the "silent majority") are less likely to invest the time to provide feedback. This creates a reporting bias where the most visible data comes from the emotional poles of the user base. For a business or a potential buyer, this can be misleading. An artificially high average star rating can mask underlying issues that affect a large number of moderately satisfied users.

This creates a form of information asymmetry, where the aggregated public data does not fully represent the typical user experience. The implication for our model is that its high accuracy is partly a function of being very good at predicting the most common outcome ('Positive'). The true challenge, and a direction for future work, lies in developing models that can better understand and predict the sentiments of the underrepresented 'Neutral' and 'Negative' voices, as they often carry the most weight for strategic decision-making.

7. Conclusion & Future Work

This project successfully developed and evaluated a machine learning model for sentiment analysis of e-commerce laptop reviews. By transforming unstructured text into meaningful features using TF-IDF and training a Logistic Regression classifier, the model was able to predict customer sentiment with high accuracy. The key findings revealed that while core product features like performance and battery life are paramount, the overall e-commerce experience significantly influences customer satisfaction. The analysis also highlighted the unique role of neutral, 3-star reviews as a source of rich, actionable feedback for product improvement.