CAPSTONE REPORT

Project Title: *Sentiment Analysis for Amazon Reviews Using NLP*

Executive Summary

This project applies Natural Language Processing (NLP) to classify Amazon customer reviews as positive or negative. Because Amazon receives millions of reviews, manual reviews become impossible. The final model—**Logistic Regression combined with SMOTE**—achieved strong balanced performance and demonstrated the ability to automatically detect negative sentiment, enabling early identification of product issues and supporting customer satisfaction strategies.

Business Problem

Amazon and similar ecommerce platforms rely heavily on customer reviews to evaluate product quality and customer satisfaction. However:

- They receive an extremely high volume of reviews.
- Negative reviews often reveal product defects or service issues.
- Without automated tools, insights from reviews may be delayed or missed.

The goal of this project is to provide a reliable sentiment classification model that supports:

- **Automated monitoring of customer satisfaction**
- **Early detection of product or quality problems**
- **Data-driven decision-making for customer support and product teams**

Dataset Description

- **Source:** Kaggle – Amazon Reviews Dataset
- **Format:** CSV
- **Target Variable:** Binary sentiment (positive vs. negative)
- **Challenge:** Severe class imbalance (90% positive, 10% negative)

Each record includes a customer review along with a star rating. Sentiment labels were generated based on the rating.

Data Preprocessing

The following NLP preprocessing steps were used:

- Lowercasing
- Removal of punctuation, URLs, and special characters
- Tokenization
- Stopword removal
- Lemmatization
- TF-IDF vectorization (unigrams & bigrams)
- Additional engineered features:
    - **Word count**
    - **Character count**

These steps standardize and structure the text data for machine learning models.

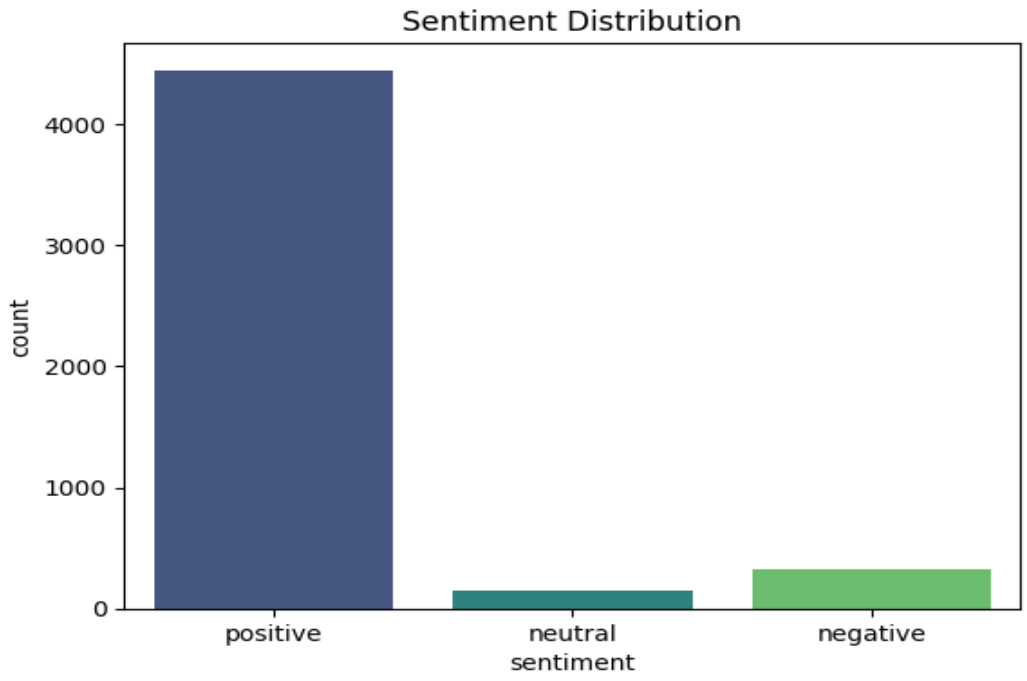
Exploratory Data Analysis (EDA)

Key observations from EDA:

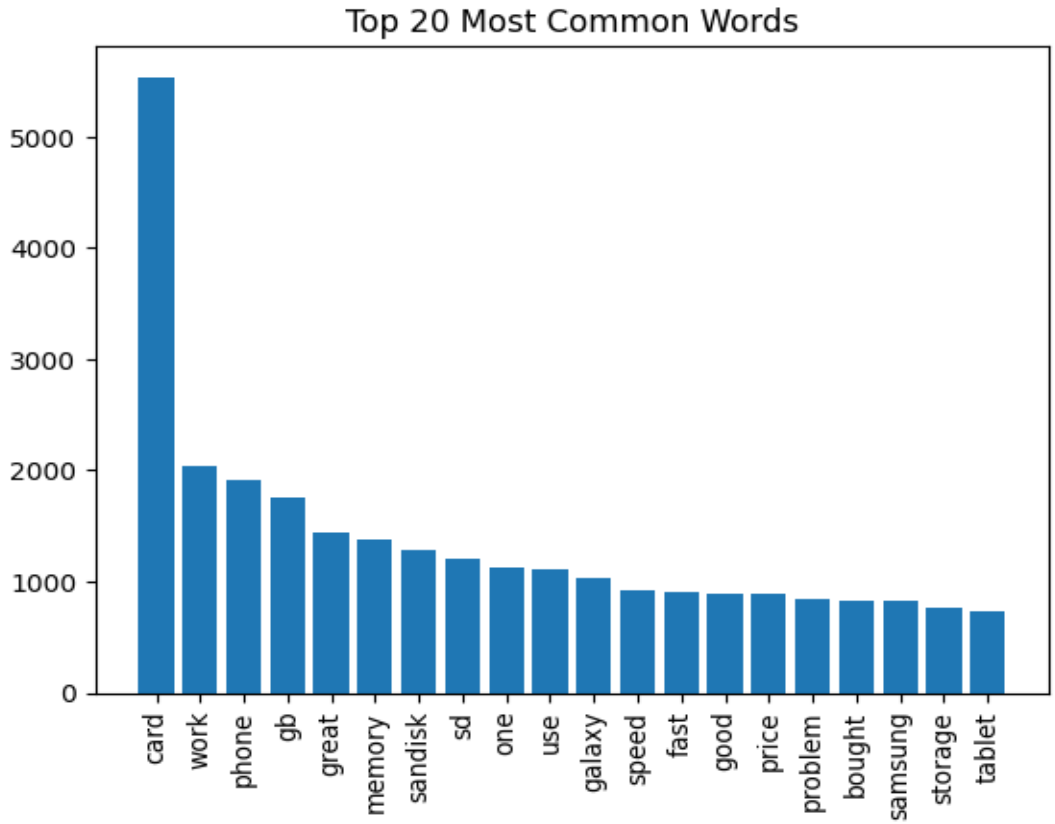
- **Sentiment was highly imbalanced**, with positive reviews dominating.
- Negative reviews frequently included words like: *broken, defective, cheap, doesn't work*.
- Positive reviews commonly mentioned: *excellent, perfect, great, durable*.
- Review lengths differed across sentiments—negative reviews tend to be shorter and more direct.
- Correlation analysis indicated limited relationships among engineered text-length features.

Visuals used in the analysis include:

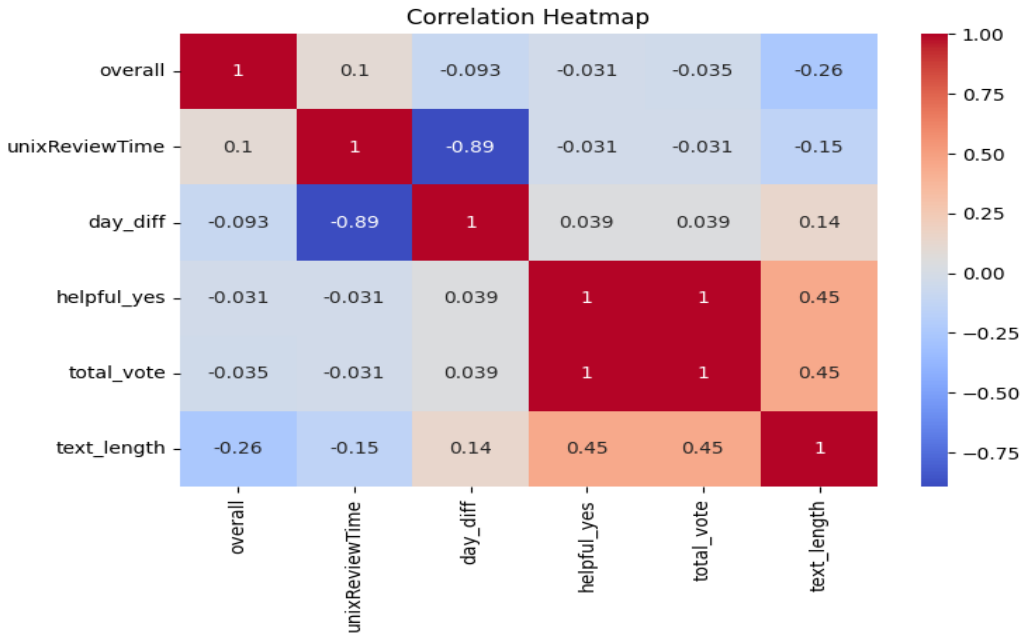
- Sentiment distribution plot



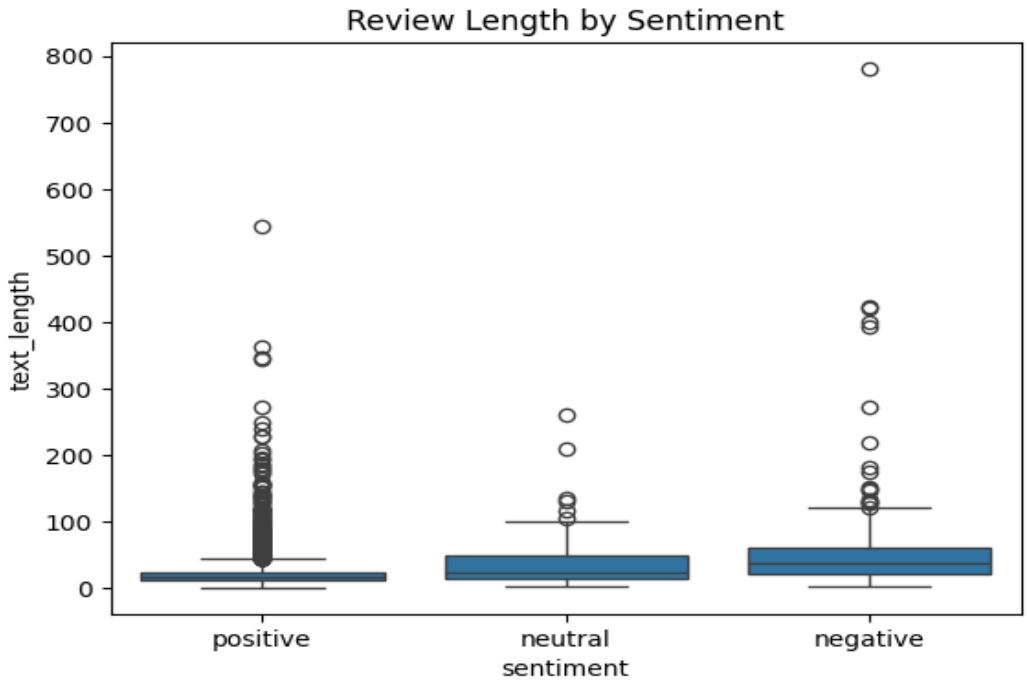
- Word frequency distributions



- Correlation heatmap



- Review length boxplots



(*You will insert visuals manually into the slide deck and/or report as needed.*)

Model Development

Multiple machine learning models were evaluated:

1. **Logistic Regression (class-weight balanced)**
2. **Logistic Regression + SMOTE**
3. Linear Support Vector Machine (SVM)
4. Random Forest (with hyperparameter tuning)
5. Multinomial Naive Bayes

Models were compared using:

- Accuracy
- Precision (macro)
- Recall (macro)
- F1-score (macro)
- Confusion matrix

Macro-averaged metrics were emphasized due to the class imbalance.

Model Performance Comparison

Summary of best-performing models:

Model	Accuracy	Macro Precision	Macro Recall	Macro F1
LogReg + SMOTE	0.9257	0.5815	0.5873	**0.5827**
SVM (Balanced)	0.9359	0.5757	0.5299	0.5488
LogReg (Balanced)	0.9186	0.5713	0.5962	0.5820
Random Forest (Tuned)	0.9247	0.5416	0.4878	0.5081
Naive Bayes	0.9054	0.3018	0.3333	0.3167

Final Model Selection

Final Model Chosen: *Logistic Regression + SMOTE*

Reasons for Selection

- **Best macro F1 performance**, indicating superior balance across classes.
- **Strongest detection of negative sentiment**, the project's priority.

- **Interpretable coefficients**, making insights explainable to stakeholders.
- **Efficient and scalable**, suitable for production systems.
- SMOTE effectively corrected the dataset imbalance.

Key Findings

- Negative reviews most often contained product defect language.
- Positive reviews emphasized quality, satisfaction, and reliability.
- The dataset is significantly skewed toward positive experiences.
- Textual structure varies by sentiment—useful for future modeling.

\Business Recommendations

1. Implement Automated Review Monitoring
Use the model to track customer sentiment in real-time and detect emerging product problems earlier.

2. Prioritize Product Improvements Based on Negative Keywords
Words like *broken*, *cheap*, *defective* signal design or quality issues requiring attention.

3. Integrate with Customer Support Workflows
Flag and route negative reviews to support agents for early intervention and improved customer experience.

Future Work

- Expand sentiment classification to include a **neutral** class.
- Adopt **transformer models** (BERT, RoBERTa) for deeper contextual understanding.
- Use **topic modeling** to identify defect-related patterns.
- Deploy a **real-time inference API** for high-volume review streams.

Conclusion

This capstone demonstrates how NLP techniques can transform large volumes of customer reviews into actionable insights. **Logistic Regression + SMOTE** provides a powerful combination of accuracy, balance, explainability, and business relevance.

The model supports improved product quality, customer satisfaction, and strategic decision-making across the business.