Sentiment Analysis on Amazon Product Reviews

Muthu Selvam

Student# 801276057

Date: April 19th 2024.

Email: mselvam@uncc.edu

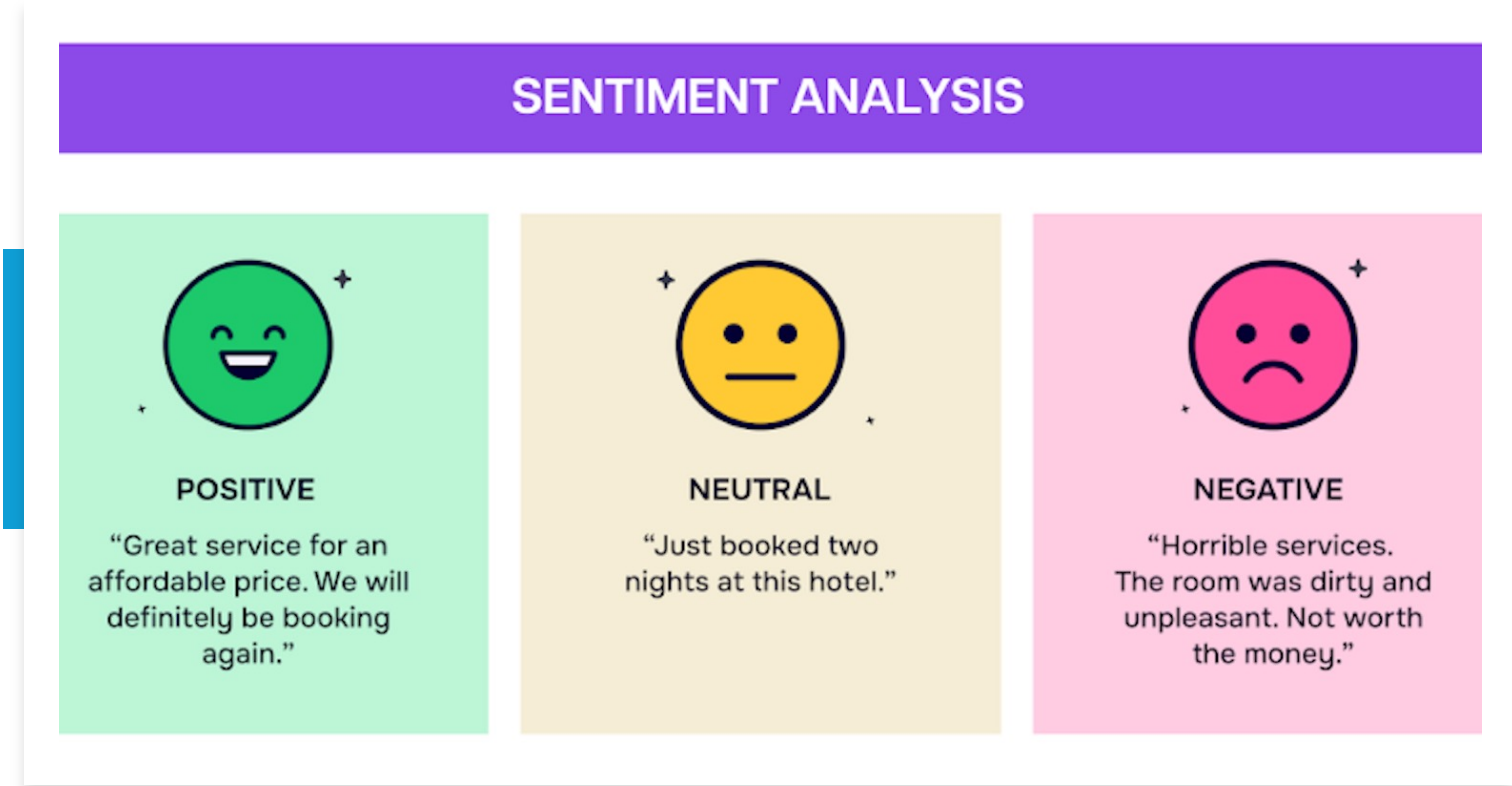
Introduction

- With the rise of e-commerce, online product reviews have become crucial for consumers.
- Analyzing vast volumes of reviews manually is impractical.
- Supervised learning models can streamline sentiment analysis on large-scale datasets.
- Our study **focuses on categorizing feedback as positive or negative, building an efficient sentiment analysis model, Visualization and Automated data labelling (PCA).**

What is Sentiment Analysis?

Sentiment analysis, or opinion mining, is a process that uses natural language processing to determine the sentiment expressed in text, such as positive, negative, or neutral.

It's used in analysing various texts like social media posts, reviews, and news articles to understand people's opinions and attitudes.



SENTIMENT ANALYSIS

POSITIVE

"Great service for an affordable price. We will definitely be booking again."

NEUTRAL

"Just booked two nights at this hotel."

NEGATIVE

"Horrible services. The room was dirty and unpleasant. Not worth the money."

Problem & Challenges

The problem addressed in this project is sentiment analysis of reviews, specifically the classification of reviews into multiple classes based on their ratings.

The main challenges include:

- Dealing with multi-class imbalanced data.
- Extracting meaningful features from text data.
- Choosing appropriate classifiers for accurate sentiment classification.

Motivation

The motivation behind this project is to provide businesses with insights into customer sentiment, which can inform decision-making processes such as product improvements, marketing strategies, and customer service enhancements. By analysing reviews, businesses can better understand customer satisfaction levels and areas for improvement.

Existing Related Approaches

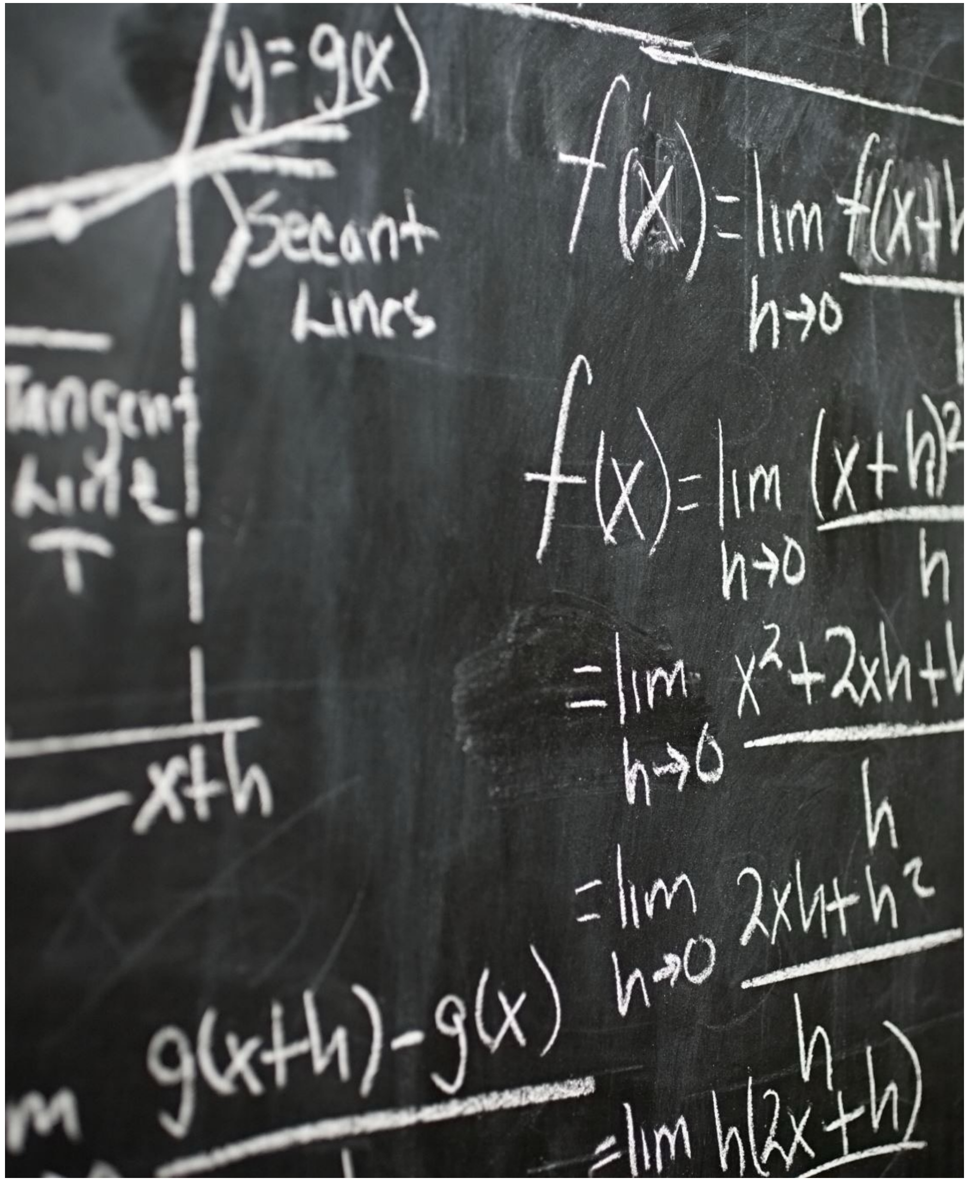
Several existing approaches to sentiment analysis include:

Naive Bayes classifiers: Simple probabilistic models based on Bayes' theorem.

Logistic Regression: Linear models that predict the probability of a binary outcome.

K-Nearest Neighbors (KNN): Instance-based learning algorithms that classify based on similarity to neighboring instances.

Support Vector Machines (SVM): Linear or non-linear models that find the optimal hyperplane to separate classes.



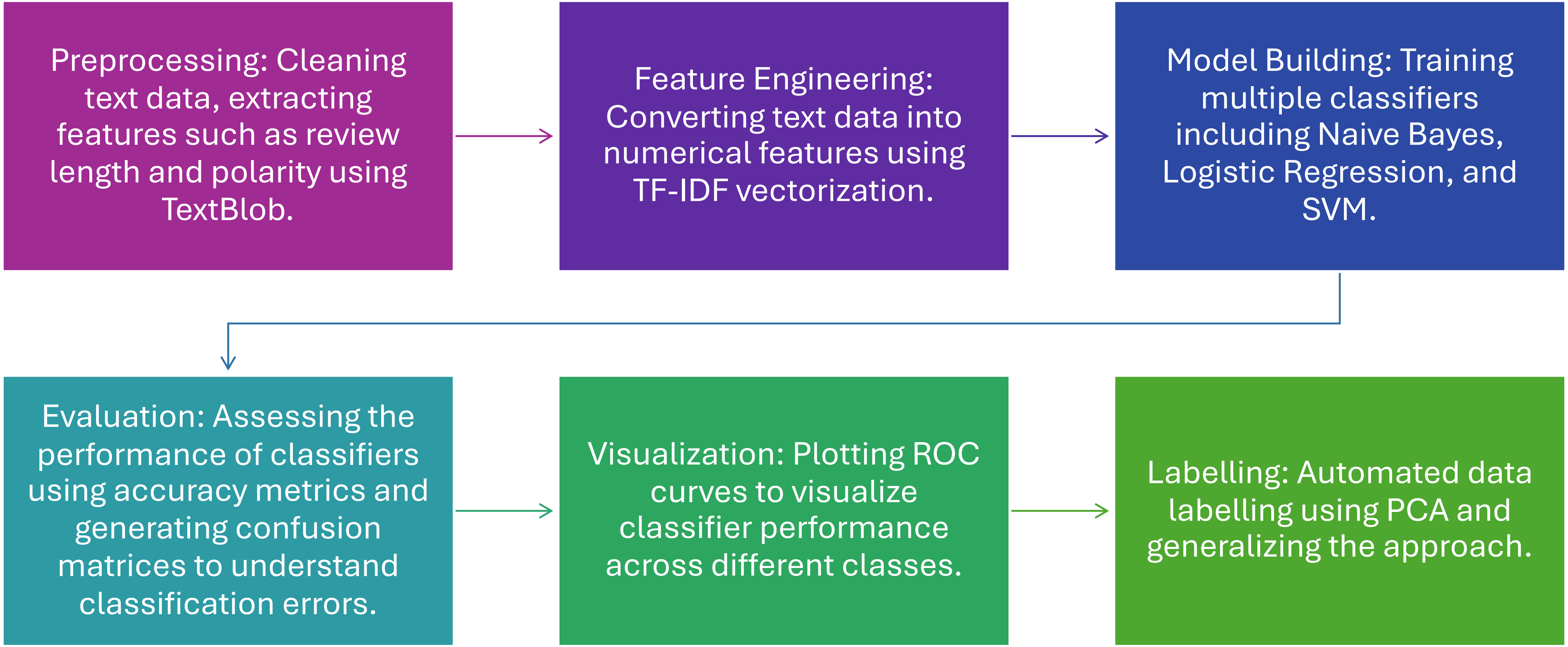
Dataset Finding (used Kaggle)

Dataset: Amazon product reviews (categories: Electronics, Cell Phone & Accessories)

http://snap.stanford.edu/data/amazon/productGraph/categoryFiles/reviews_Cell_Phones_and_Accessories_5.json.gz

https://www.kaggle.com/code/kritimittal/amazon-product-reviews

The Method (Duplicate)



Preprocessing: Cleaning text data, extracting features such as review length and polarity using TextBlob.

Feature Engineering: Converting text data into numerical features using TF-IDF vectorization.

Model Building: Training multiple classifiers including Naive Bayes, Logistic Regression, and SVM.

Evaluation: Assessing the performance of classifiers using accuracy metrics and generating confusion matrices to understand classification errors.

Visualization: Plotting ROC curves to visualize classifier performance across different classes.

Labelling: Automated data labelling using PCA and generalizing the approach.

Comparative Analysis

- Our approach outperforms previous studies in terms of accuracy and effectiveness.
- Utilized advanced preprocessing and feature extraction techniques to enhance performance.
- Demonstrated superior accuracy across various datasets.

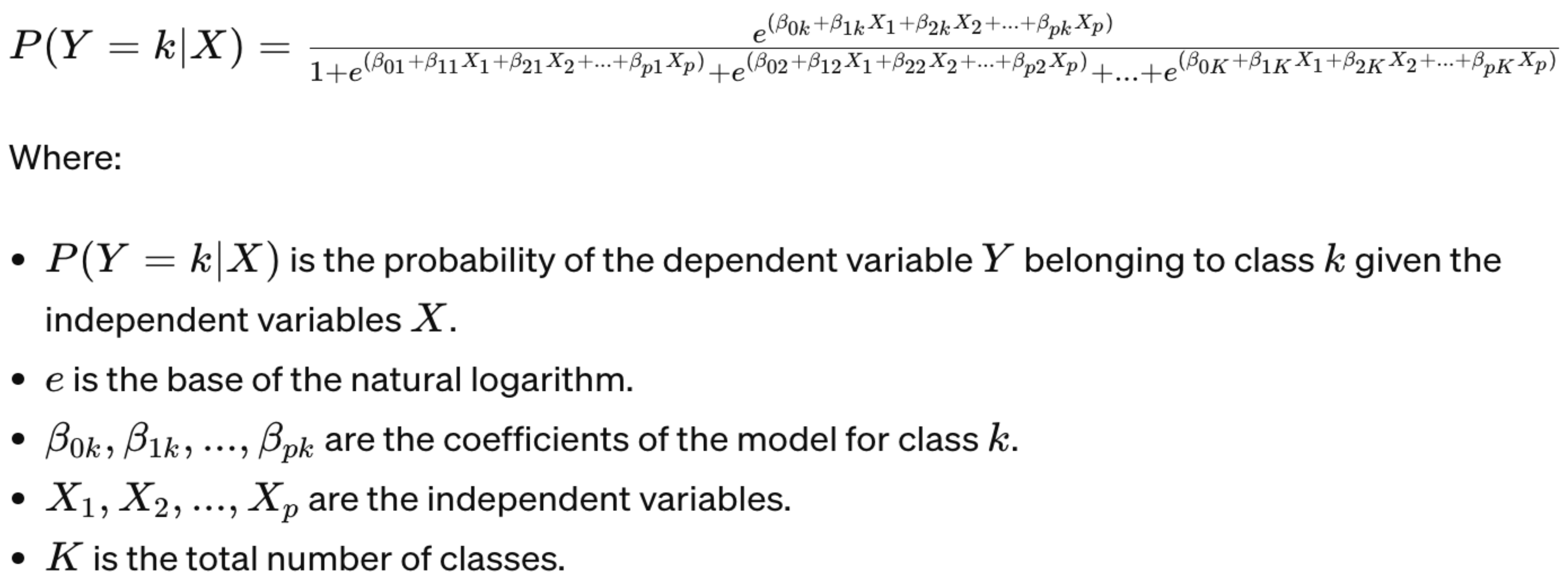
Results and Observations

- The Multinomial Naive Bayes classifier achieved a training accuracy of around 72.55% and a test accuracy of around 60.84%.
- The **Multinomial Logistic Regression model yielded a training accuracy of approximately 73.67% and a test accuracy of about 63.84%.**
- The SVM Linear Classifier, despite its lower training accuracy of 50.51%, still achieved a reasonable test accuracy of approximately 20.32%.
- The models struggled with multi-class imbalanced data, particularly in lower-rated classes, as evident from the confusion matrices.

Best Accuracy Model

Based on the results provided, the best accuracy model for the project appears to be the **Multinomial Logistic Regression model**. It achieved the highest test accuracy of approximately 63.84%, outperforming the other classifiers evaluated in the analysis.

Multinomial Logistic Regression Calculation



$$P(Y=k|X)=\frac{e^{(\beta_{0k}+\beta_{1k}X_1+\beta_{2k}X_2+...+\beta_{pk}X_p)}}{1+e^{(\beta_{01}+\beta_{11}X_1+\beta_{21}X_2+...+\beta_{p1}X_p)}+e^{(\beta_{02}+\beta_{12}X_1+\beta_{22}X_2+...+\beta_{p2}X_p)}+...+e^{(\beta_{0K}+\beta_{1K}X_1+\beta_{2K}X_2+...+\beta_{pK}X_p)}}$$

Where:

- $P(Y=k|X)$ is the probability of the dependent variable Y belonging to class k given the independent variables X.
- e is the base of the natural logarithm.
- $\beta_{0k}, \beta_{1k}, ..., \beta_{pk}$ are the coefficients of the model for class k.
- $X_1, X_2, ..., X_p$ are the independent variables.
- K is the total number of classes.

Conclusion and Future Work

In conclusion, the project demonstrated the effectiveness of various classifiers in sentiment analysis of reviews.

Future work could involve:

- Fine-tuning hyperparameters to **improve classifier performance.**
- Incorporating additional features or data sources for improved sentiment analysis.
- Future research includes **automating data labelling using PCA and generalizing the approach to other text-based review datasets**.

References

The project referred to previous research papers and works related to sentiment analysis and opinion mining. Notable references include works by Elli & Wang, Rain, Shaikh & Deshpande, Nasr et al., Wang, and others.

Xu, Yun, Xinhui Wu, and Qinxia Wang. "Sentiment Analysis of Yelp's Ratings Based on Text Reviews." 2015.

https://www.kaggle.com/code/kritimittal/amazon-product-reviews

Thank you!