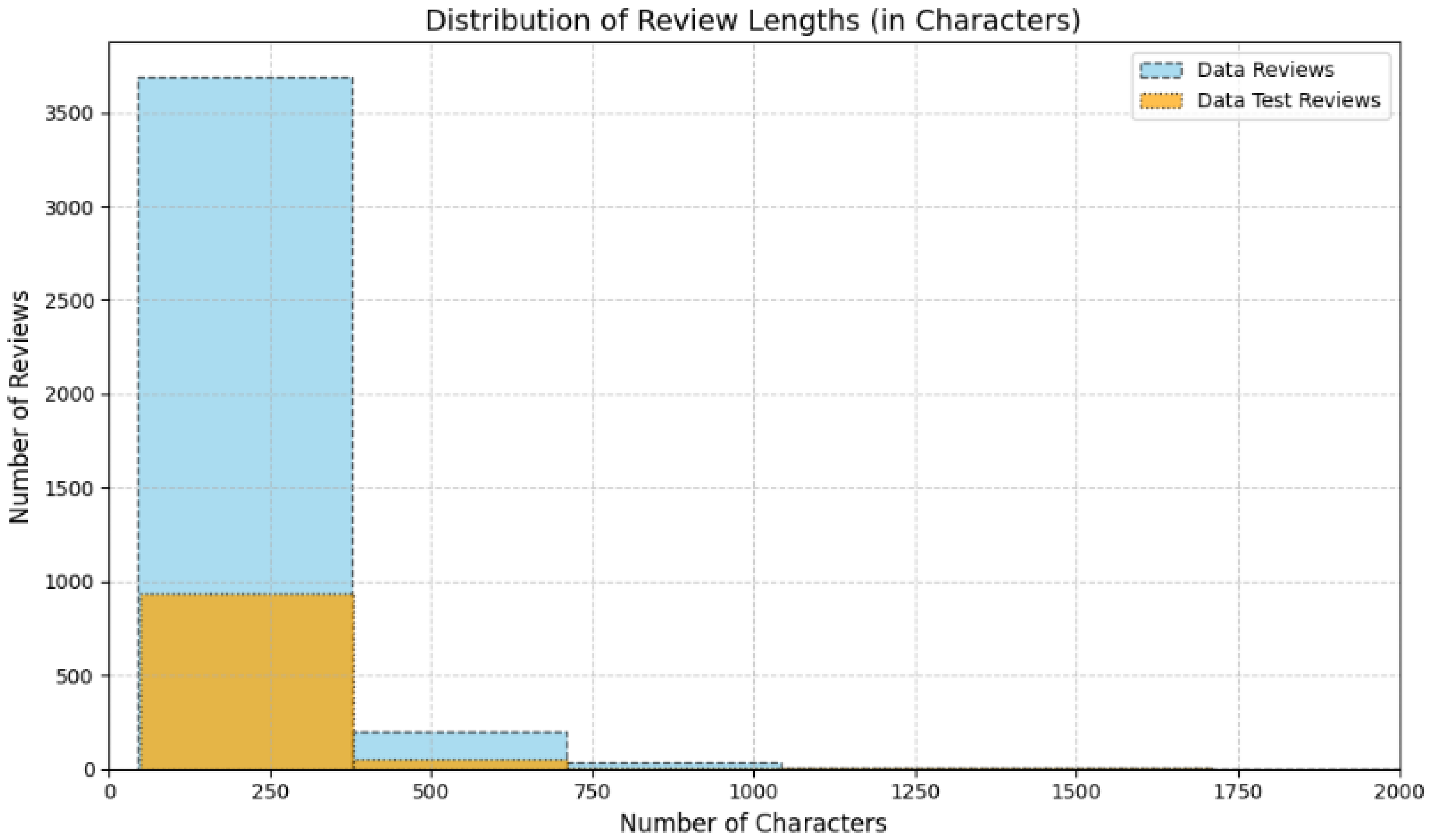
Sentiment Analysis of Product Reviews

- **NAME: MINAL DEVIKAR**
- **INSTITUTION: DIGICHROME ACADEMY**

OVERVIEW

- **Introduction to the Project**:
  - Analyzing customer sentiment from product reviews.
  - Using machine learning algorithms for classification.
  - Objective: Predict the sentiment (Positive, Neutral, Negative) from the reviews.
- **Tools Used**: Python, Scikit-learn, TensorFlow, TextBlob, Keras

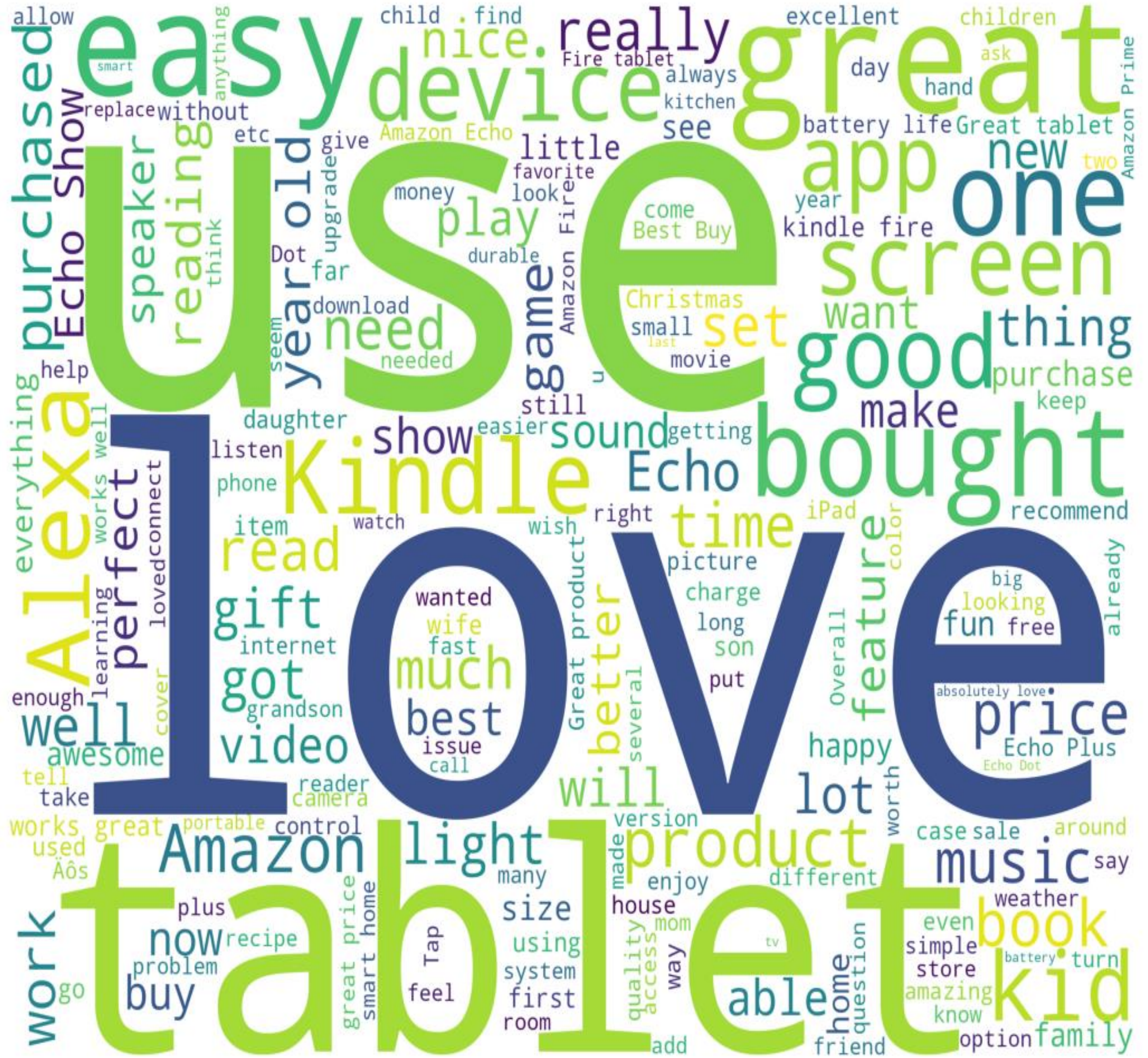
DATA VISUALISATION



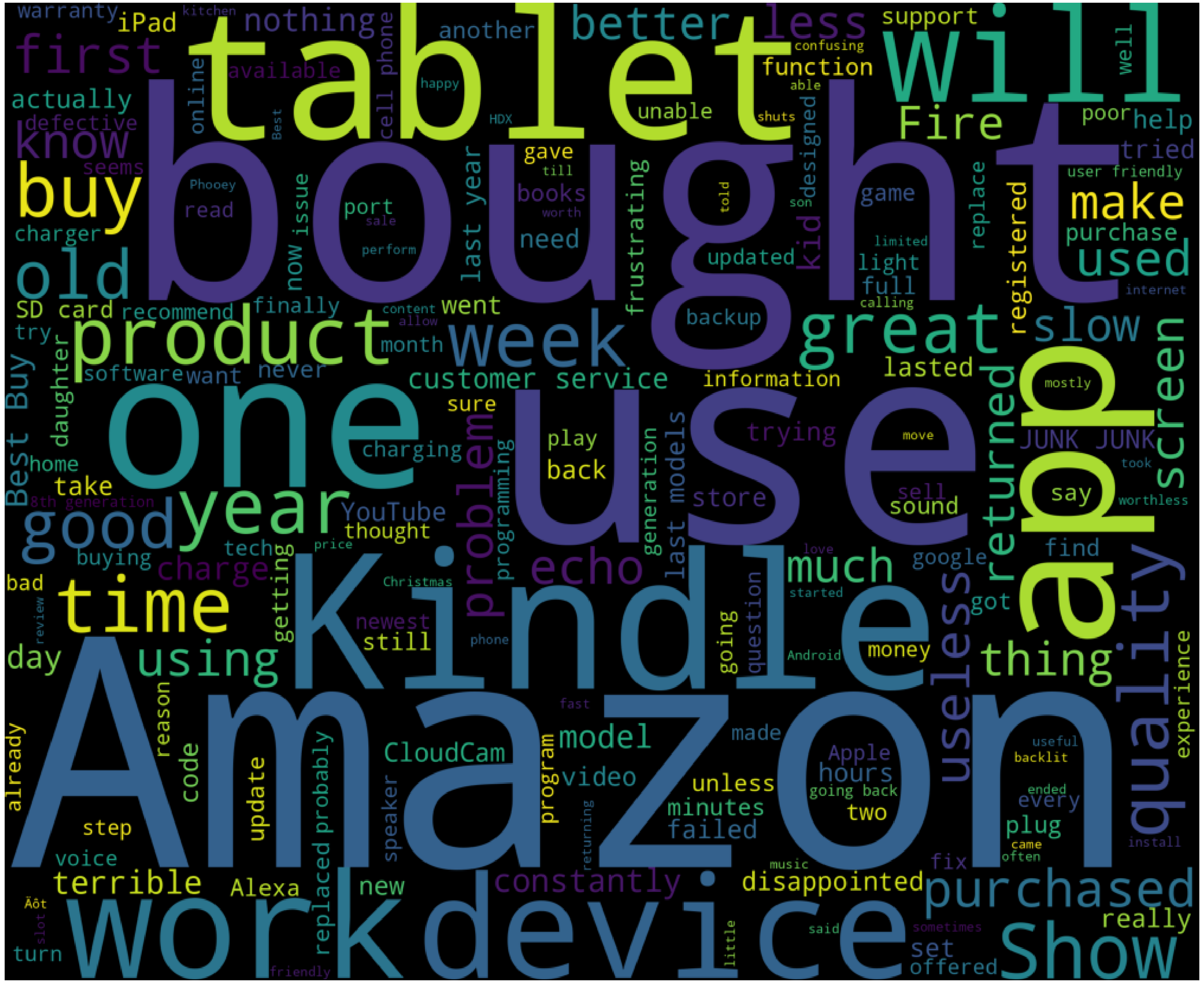
- The distribution of review lengths is heavily skewed towards shorter reviews, with most having fewer than 250 characters.

- The majority of the dataset consists of "Data Reviews," while "Data Test Reviews" form a smaller subset.
- Few reviews exceed 500 characters, and almost none are longer than 1000 characters.
- The plot indicates a need for handling class imbalance in review length during text preprocessing.

POSITIVE WORDS WORDCLOUD



NEGATIVE WORDS WORDCLOUD



Data Collection & Preprocessing

- **Data Collection**:
- Source: Product reviews from an online marketplace.
- Columns: Review Text, Review Title, Sentiment Labels (Positive, Neutral, Negative)
- **Preprocessing Steps**:
- Text Cleaning (removal of stop words, punctuation, and unnecessary characters).
- Tokenization and Lemmatization.
- Sentiment label encoding and vectorization.

CLEANED DATA

	reviews.text	reviews.text
0	purchased black fridaypros great price even sa...	purchased black fridaypros great price even sa...
1	purchased two amazon echo plus two dots plus f...	purchased two amazon echo plus two dots plus f...
2	average alexa option show things scren stil li...	average alexa option show things scren stil li...
3	god product exactly wanted god price	god product exactly wanted god price
4	3rd one ive purchased ive bought one al nieces...	3rd one ive purchased ive bought one al nieces...
...	...	...
3937	fun family play may get boring newnes wears we...	fun family play may get boring newnes wears we...
3938	love kindle great product reduces eye strain e...	love kindle great product reduces eye strain e...
3939	loking blutoth speaker use phone didnt want wo...	loking blutoth speaker use phone didnt want wo...
3940	second amazon fire 7 tablet purchased time col...	second amazon fire 7 tablet purchased time col...
3941	satisfied tablet fast eficient	satisfied tablet fast eficient

DATE TIME EXTRACTION

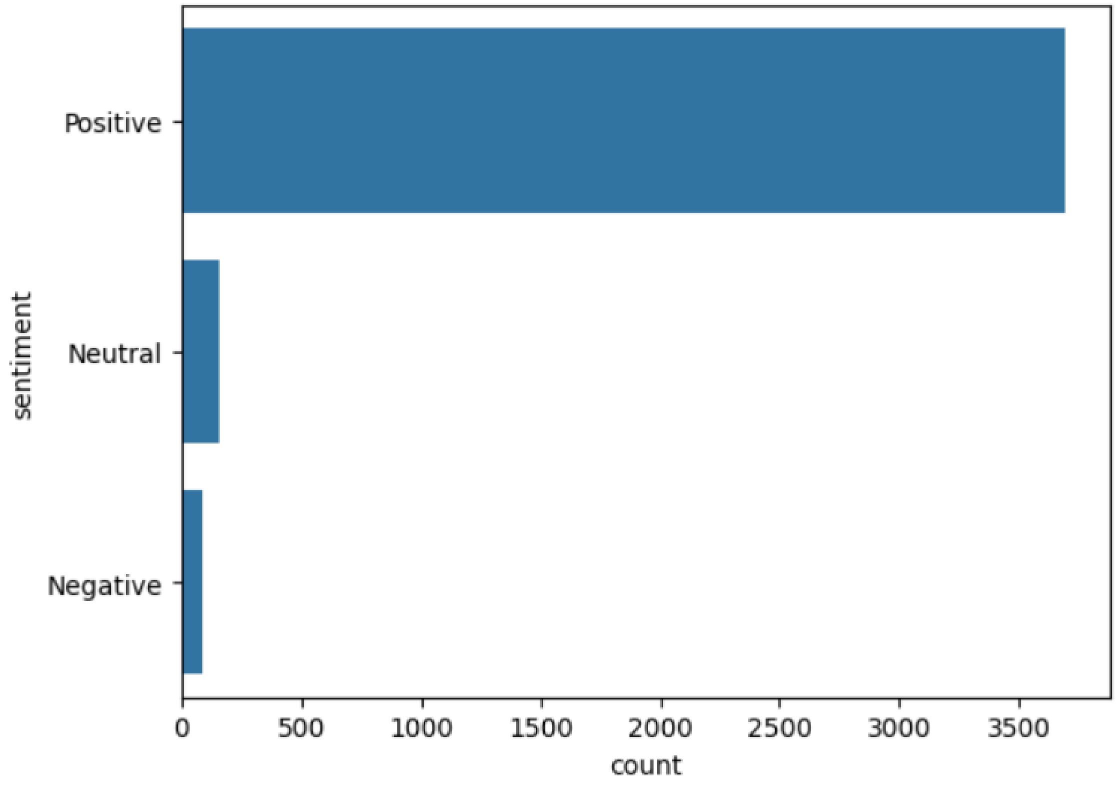
reviews.text	reviews.title	sentiment	reviews_day	reviews_month	reviews_year
purchased black fridaypros great price even sa...	powerful tablet	2	26	12	2016
purchased two amazon echo plus two dots plus f...	amazon echo plus awesome	2	17	1	2018
average alexa option show things scren stil li...	average	1	20	12	2017
god product exactly wanted god price	great	2	4	8	2017

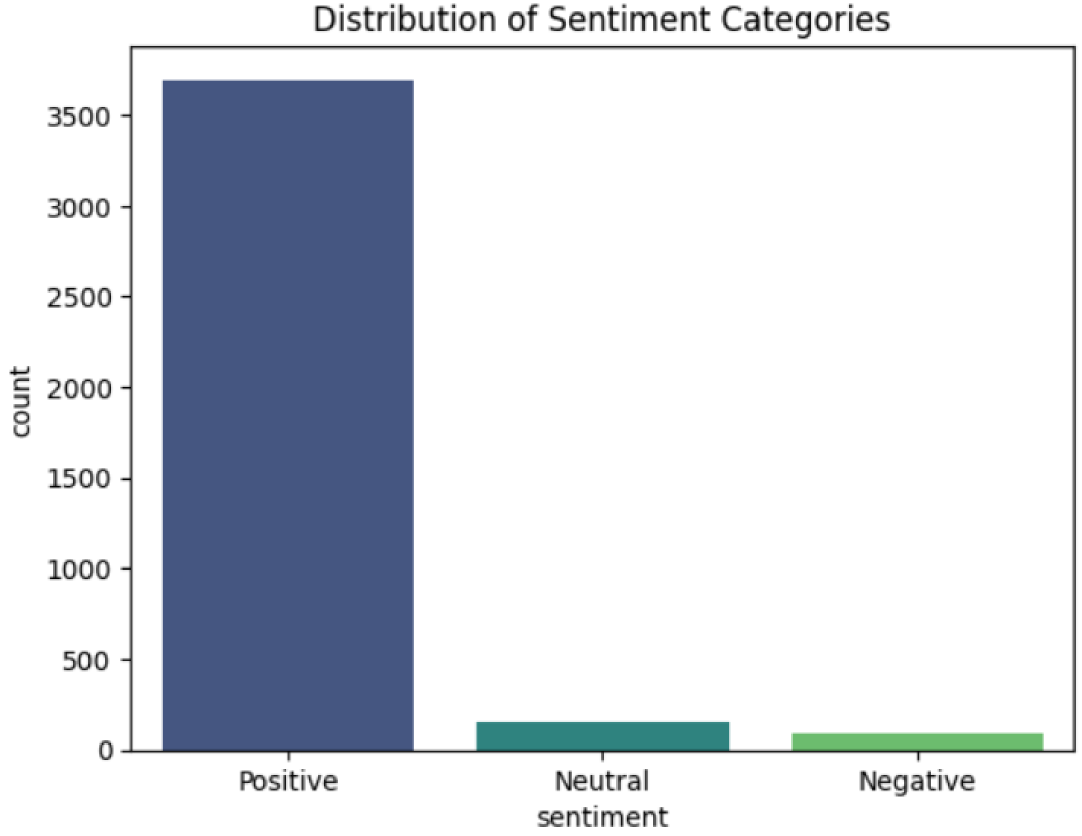
Text Vectorization Techniques

- **TF-IDF VECTORIZER**:
- Converts text data into a matrix of TF-IDF features.
- Weights words according to their frequency and relevance.
- **COUNT VECTORIZER**:
- Converts text data into a matrix of token counts.

HANDLING CLASS IMBALANCE

<Axes: xlabel='count', ylabel='sentiment'>

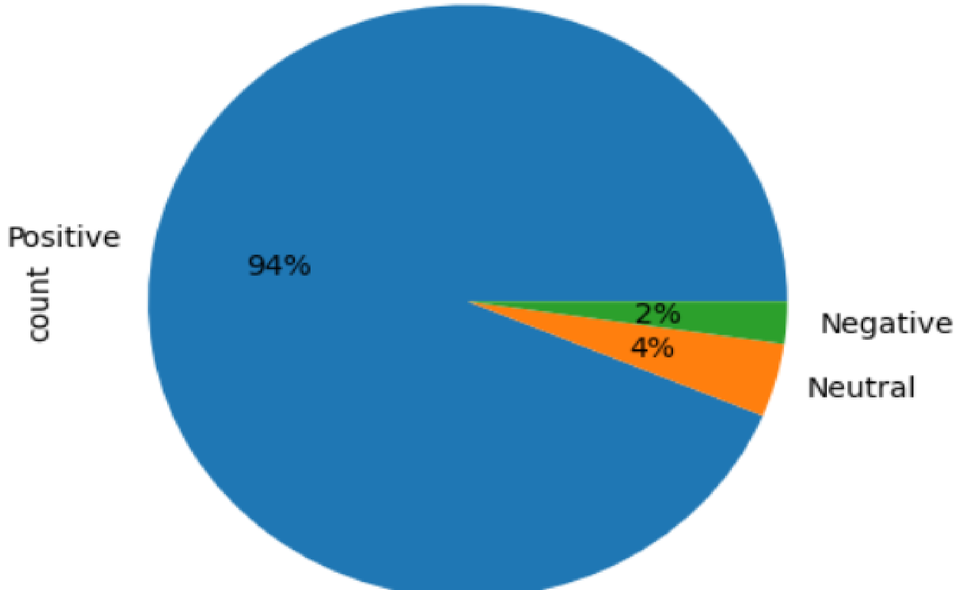




The class distribution of the sentiment dataset is as follows:

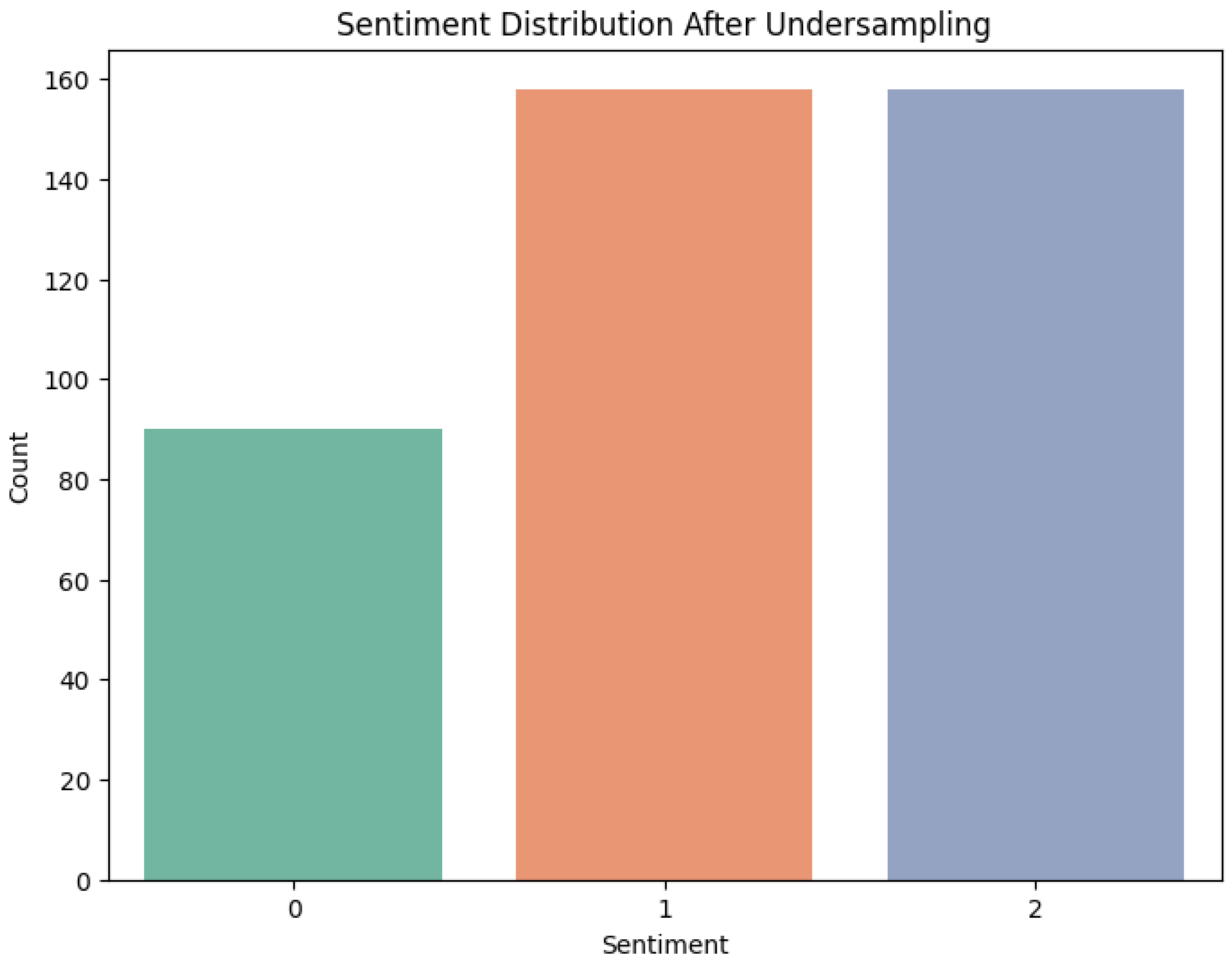
- **Positive** - 3749 instances
- **Neutral -** 158 instances
- **Negative -** 93 instances

This shows a significant **class imbalance**, where the "Positive" sentiment dominates the dataset

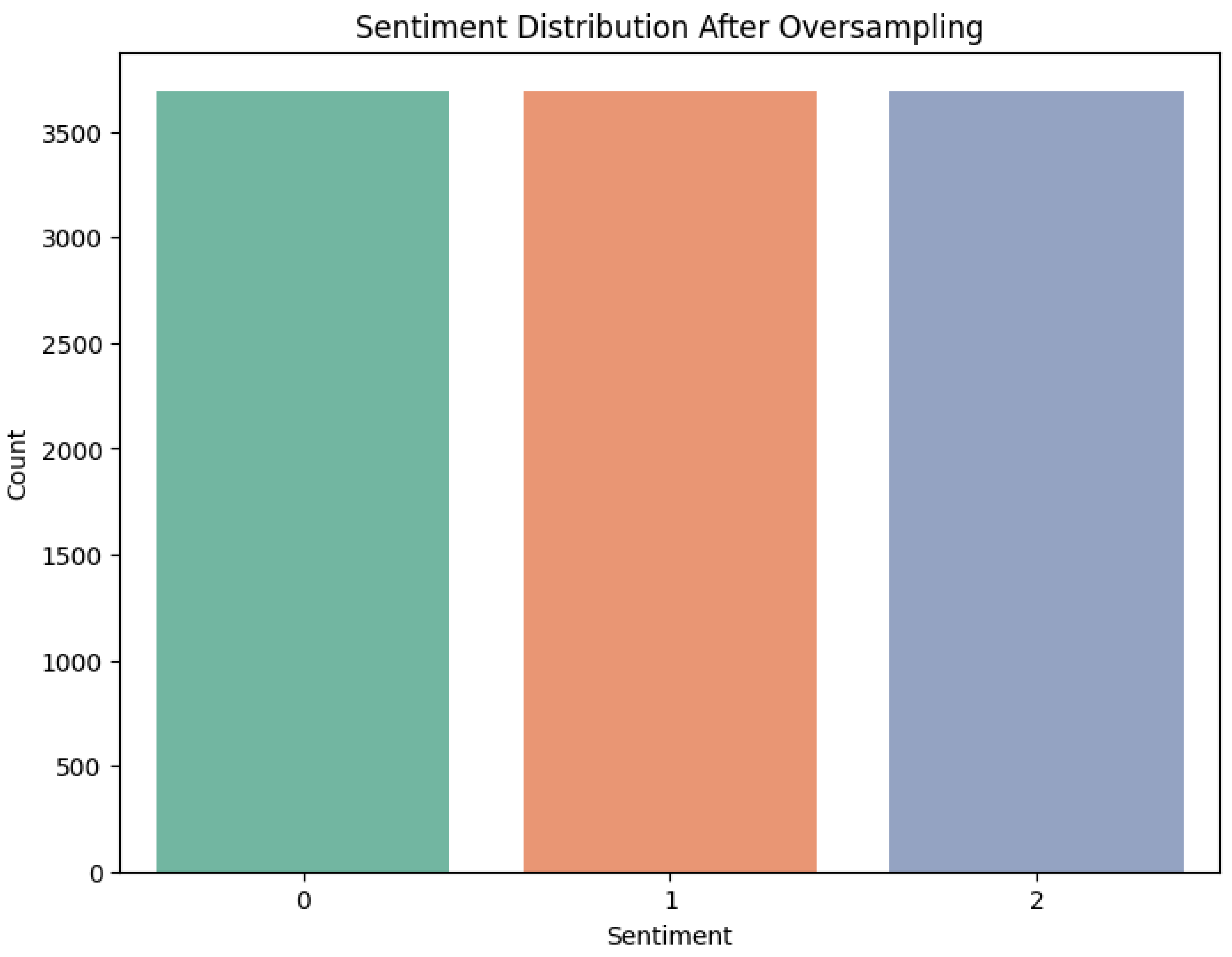


TECHNIQUES TO BALANCE DATA

1. UnderSampling

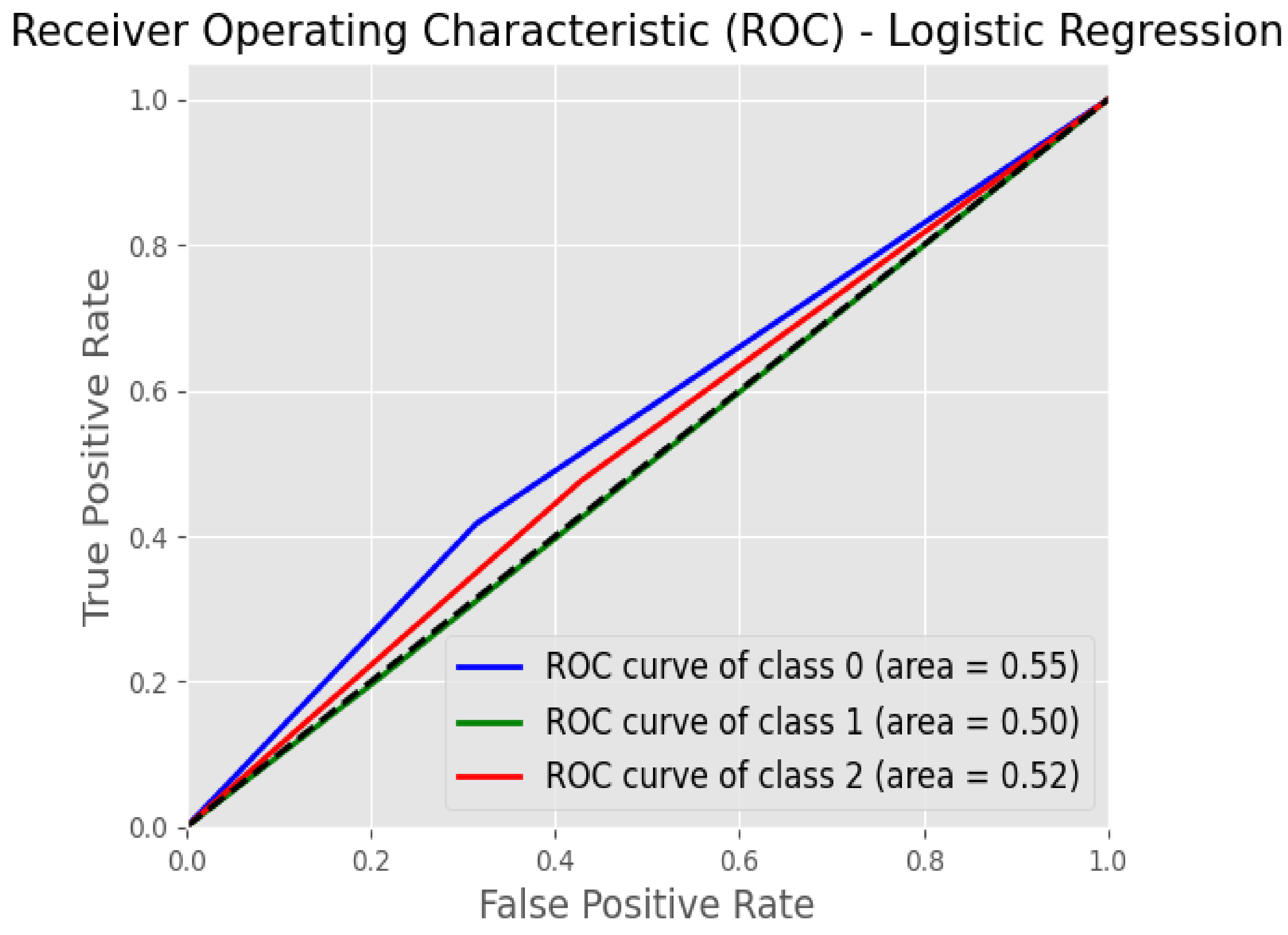


2. OverSampling



Logistic Regression for under-sampled data

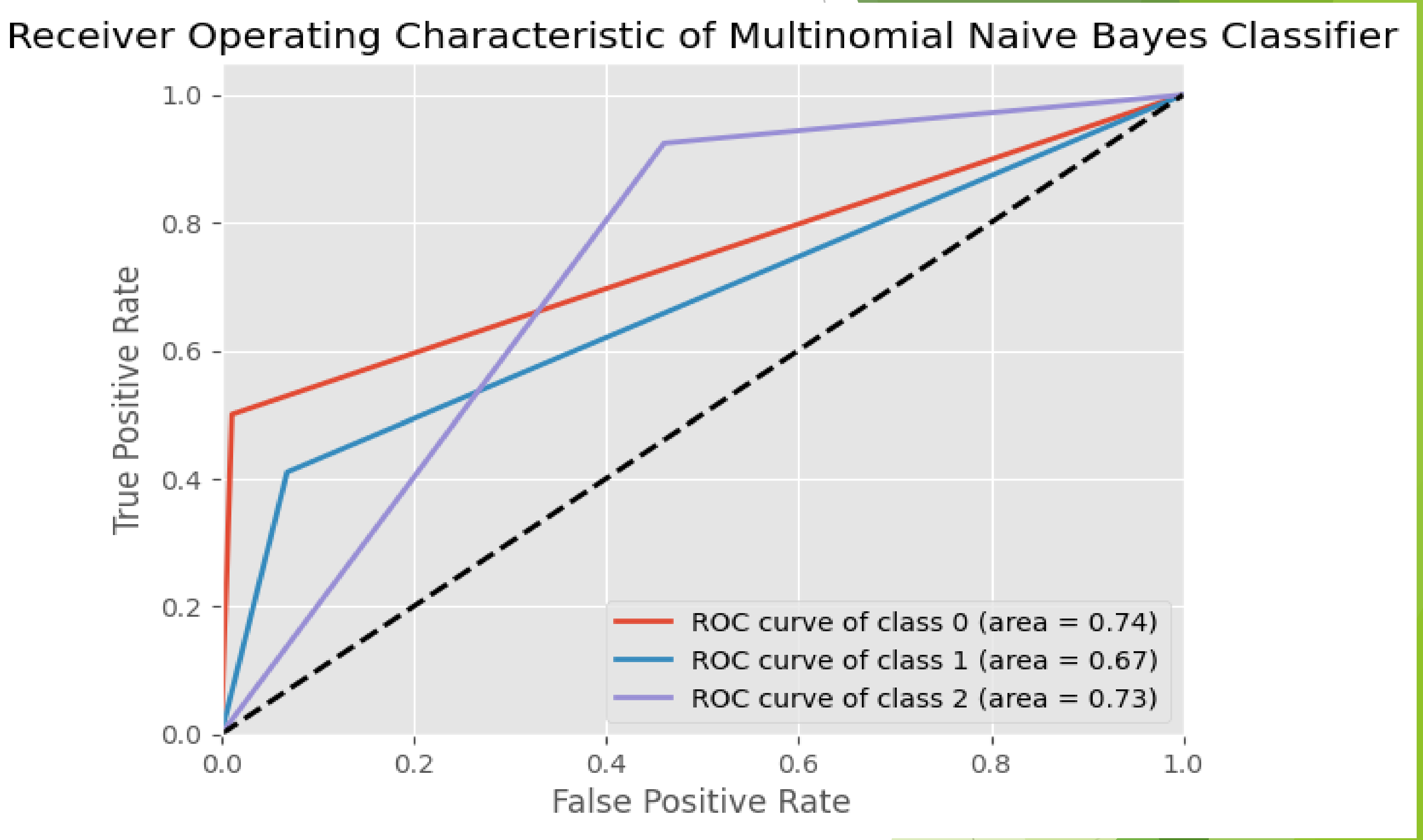
Logistic Regression for over-sampled data



Logistic Regression on over-sampled data is perfroming better than under-sampled data

Model Training & Evaluation

Multinomial Naive Bayes model

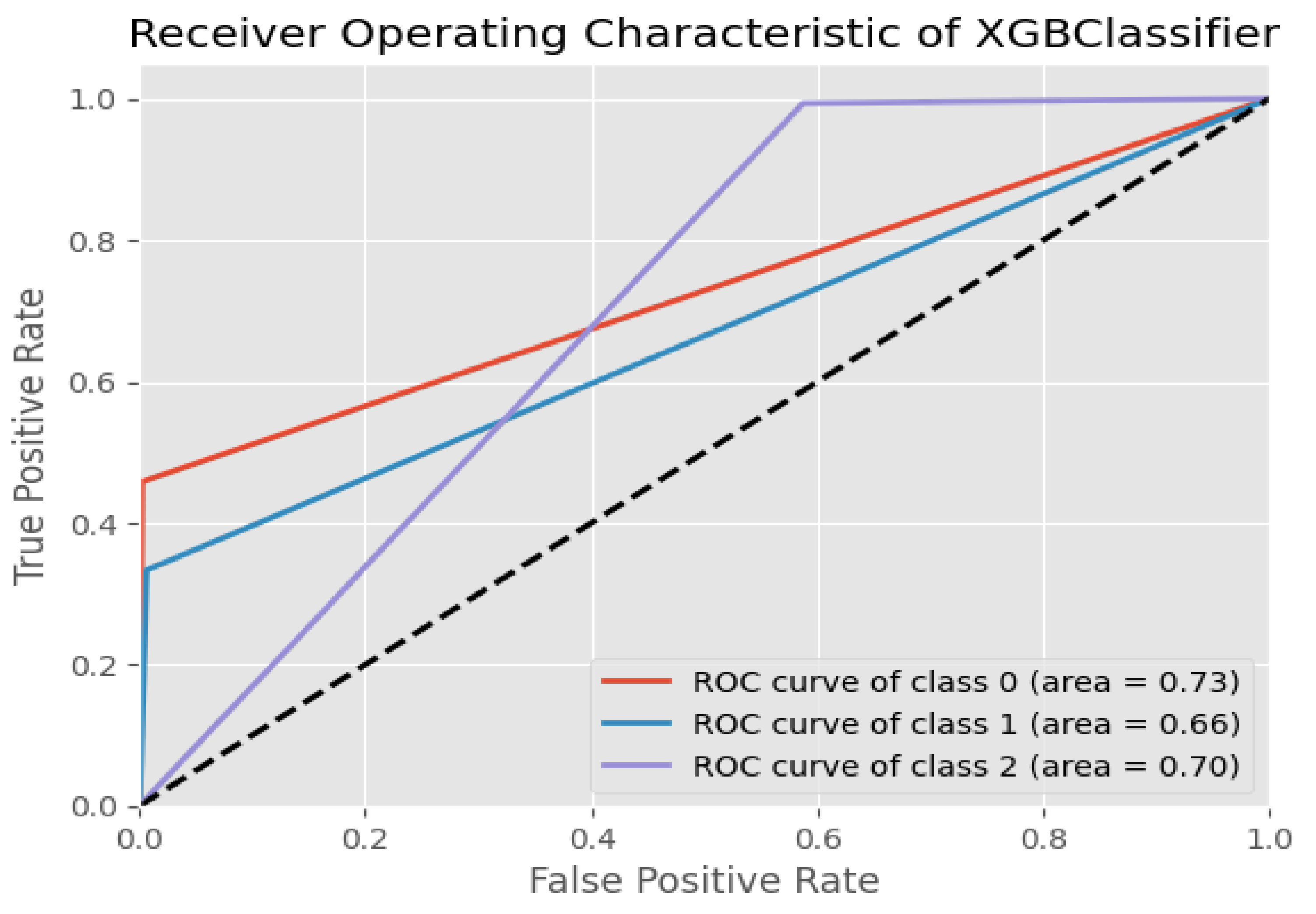


•The ROC curve illustrates the performance of the Multinomial Naive Bayes classifier across three classes.

•Area Under the Curve (AUC) shows moderate performance, with Class 0 (AUC = 0.74) outperforming the others.

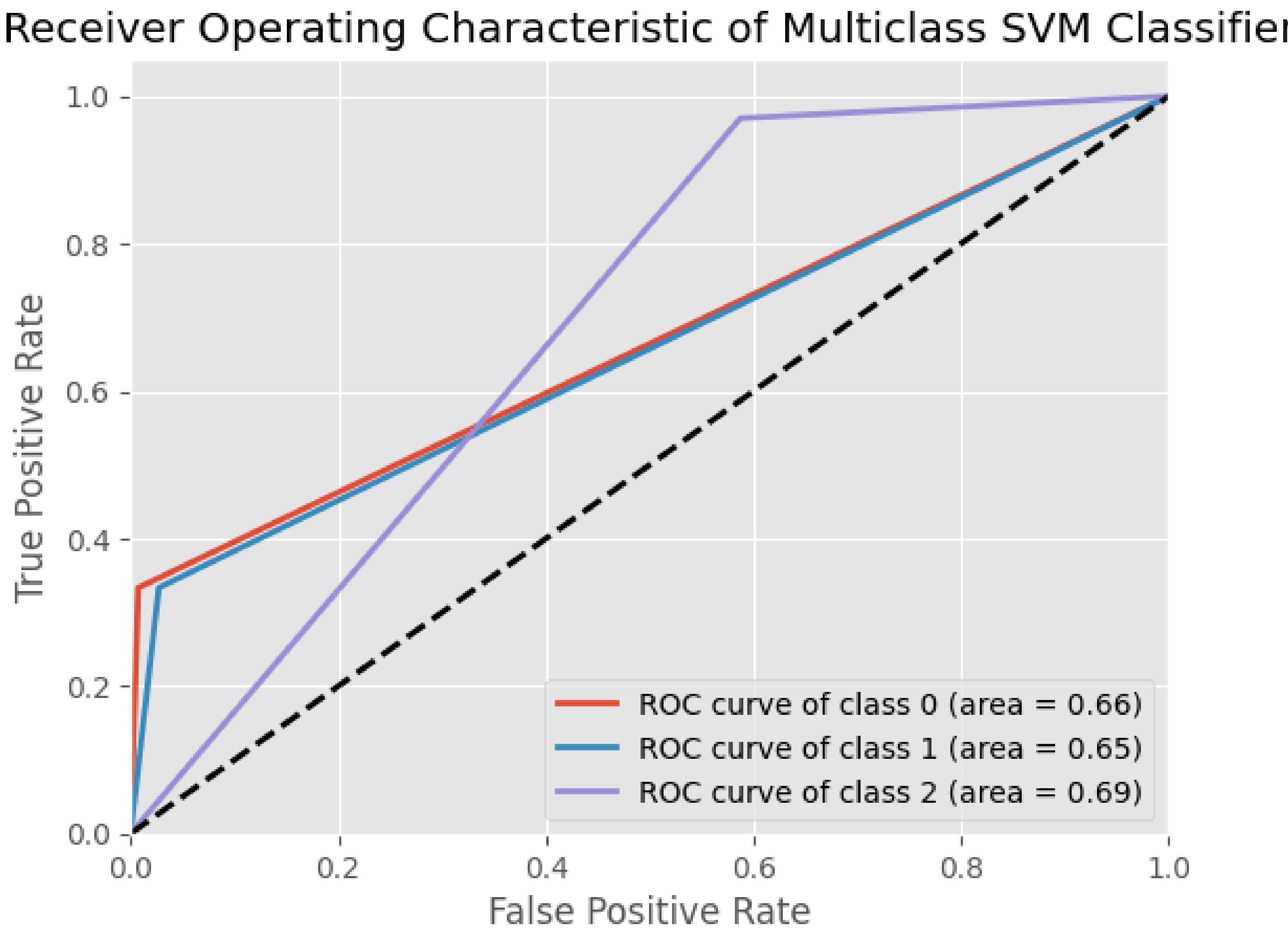
•The classifier demonstrates limited ability to distinguish between certain classes, as seen in the lower AUC for Class 1 (AUC = 0.67).

XGBClassifier



XGBoost performs better in predicting all the classes, with a more balanced performance across each class. The model demonstrates strong class separation and higher AUC, especially for underrepresented classes.

Multiclass SVM classifier



•The ROC curve for the Multiclass SVM classifier indicates moderate performance, with AUC values ranging from 0.65 to 0.69 across classes.

•Class 2 shows the best separability (AUC = 0.69), while Class 1 has the lowest (AUC = 0.65).

•The model struggles with distinguishing certain classes, reflecting a need for better optimization or data balance.

Machine Leaning Model

Model	Matrix	Accuracy	Accuracy	Avg)	Avg)	Avg)	Remarks
Multinomial Naive Bayes	[12 4 8] [2 16 21] [9 62 866]	N/A	89%	0.56	0.61	0.57	Best for quick baseline modeling.
Random Forest Classifier	N/A	100%	95.5%	0.93	0.89	0.91	Handles data complexities effectively.
XGBoost Classifier	[11 1 12] [1 13 25] [1 5 931]	N/A	95.5%	0.83	0.60	0.67	High accuracy due to boosting techniques.
Multi-class SVM	[8 3 13] [2 13 24] [5 23 909]	93%	93%	0.61	0.55	0.57	Performed well on imbalanced datasets.

•**Confusion Matrices**: SVM and XGBoost confusion matrices highlight differences in minority class performance (e.g., Class 0 and Class 1).

•**Accuracy Comparison**: XGBoost achieved the highest validation accuracy (95.5%), while SVM also performed well at 93%.

•**Remarks on SVM**: Effective for imbalanced datasets but lower precision for minority classes.

•**XGBoost**: Excels in both overall accuracy and precise minority class predictions.

Deep Learning Models

- **Artificial Neural Networks (ANN):**
  - Multi-layer perceptron for non-linear relationships.
  - Key Layers: Dense, Dropout.
- **LSTM (Long Short-Term Memory):**
  - Sequential model for text classification.
  - Key Layers: LSTM, Embedding, Spatial Dropout.

	senti_score
0	(0.4333333333333333, 0.7226190476190476)
1	(0.38125000000000003, 0.4145833333333333)
2	(0.25, 0.25)
3	(0.25, 0.25)
4	(0.6000000000000001, 0.725)

dtype: object

Key Findings

- XGBoost and Random Forest showed superior performance (95.5%).
- SVM is a strong choice but slightly behind in accuracy (93%).
- MNB struggled with imbalanced data but is computationally efficient

Conclusion & Future Work

- **Conclusion**:
- XGBoost excels in both accuracy and class balance.
- Deep learning models can be further fine-tuned for complex patterns.
- **Future Work**:
- Explore Transformer models like BERT for improved text understanding.
- Deploy models to production for real-world applications.