NEWS ARTICLES CLASSIFICATION

Presented by:

Justin Ndivhuwo
Aiden Geluk
Ntokozo Hadebe
Sinenkosi Sikhakhane
Kentse Mphahlele
Obed Segwate Mabowa

TABLE OF CONTENT

- Introduction
- problem statement
- Objectives and Goals
- Data cleaning
- Exploratory Data Analysis
- Model Training
- Model Matrics Evaluation
- Model Performance Analyses
- Conclusion

INTRODUCTION

- Rapid modern changes and developments
- Daily publication of numerous news articles by Maji Ndogo's news24

- Overwhelming quantity of news
- Difficulty for readers in finding relevant information

INTRODUCTION

Challenges in the News Article Industry:

- *Information Overload*

- *Misclassification*

- *Incorrect Trend Analysis*

PROBLEM STATEMENT

- Rapid increase in digital news content.
- Difficulty to efficiently categorize and manage information.

- How accurately can news articles be classified into the different categories?
- Assess the effectiveness of classification algorithms in accurately categorizing news articles

OBJECTIVES AND GOALS

Primary objectives:

- Analyze a dataset of news articles and develop a robust classification model.
- Accurately categorize articles into predefined categories

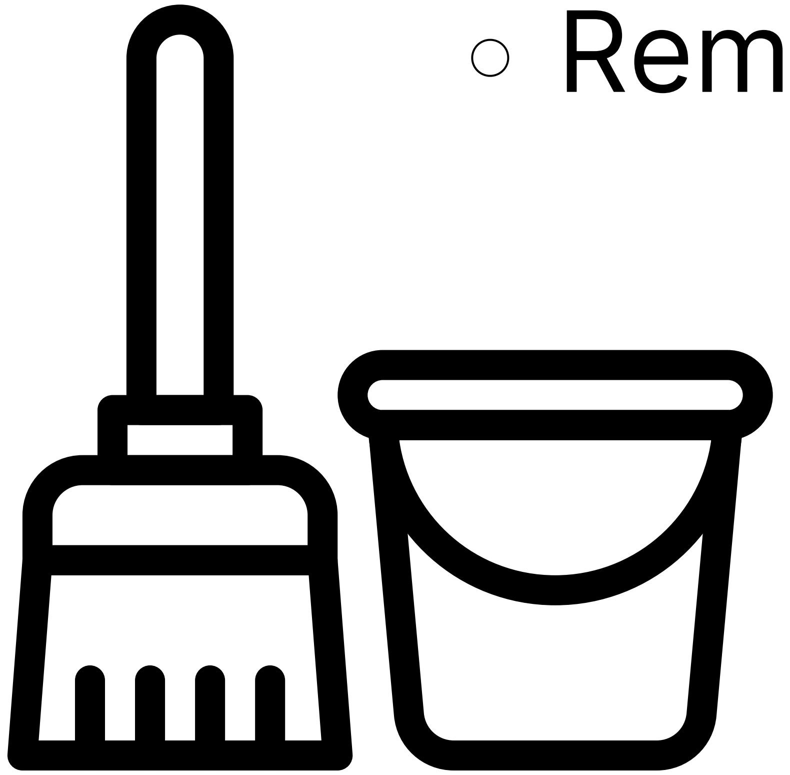
Goal:

- To deliver an effective and efficient news classification application for Maji Ndogo News24.

DATA CLEANING

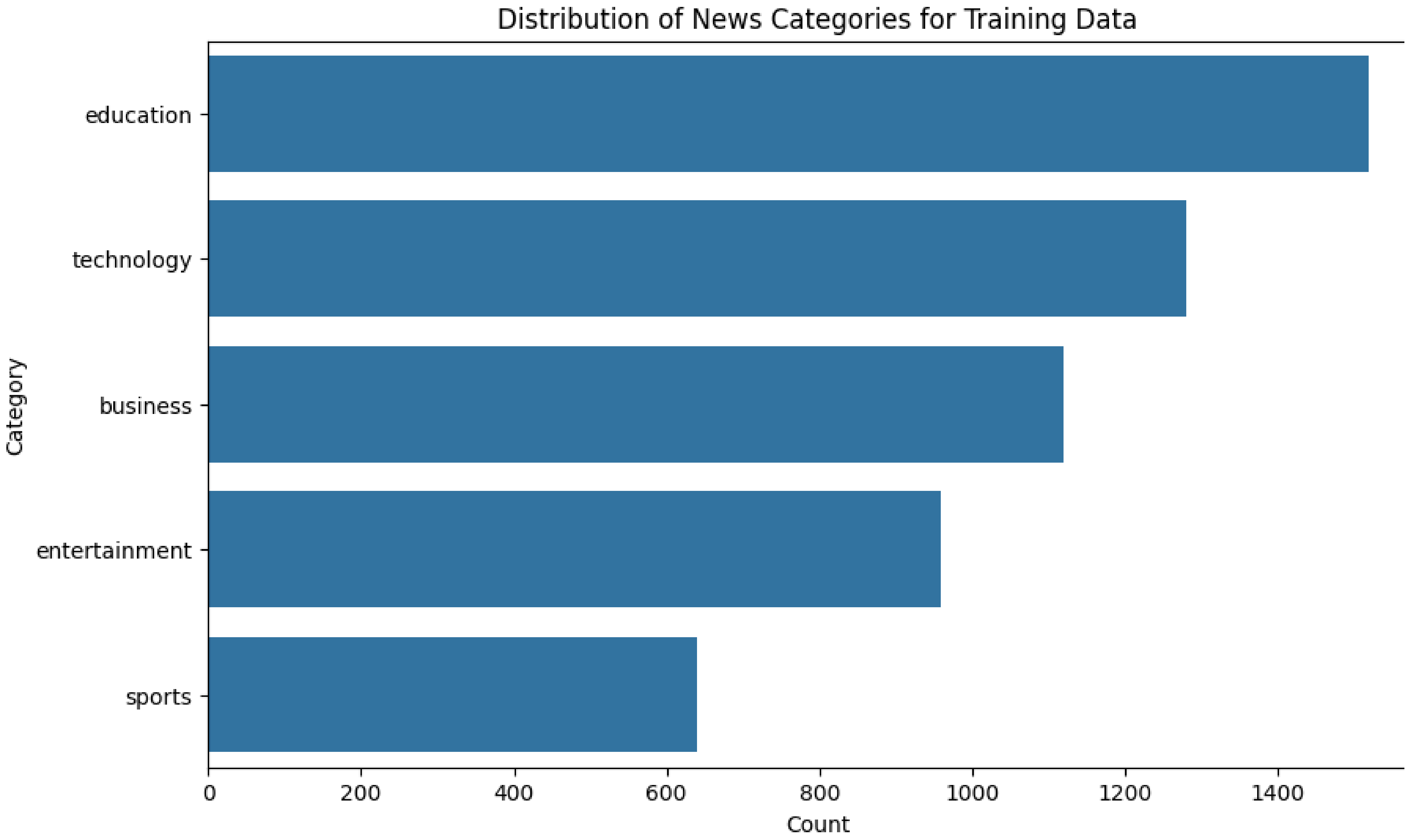
Data cleaning includes:

- Removing Punctuation and Special Characters
- Handling Missing Values
- Removing Non-Textual Elements



EXPLORATORY
DATA ANALYSIS
(EDA)

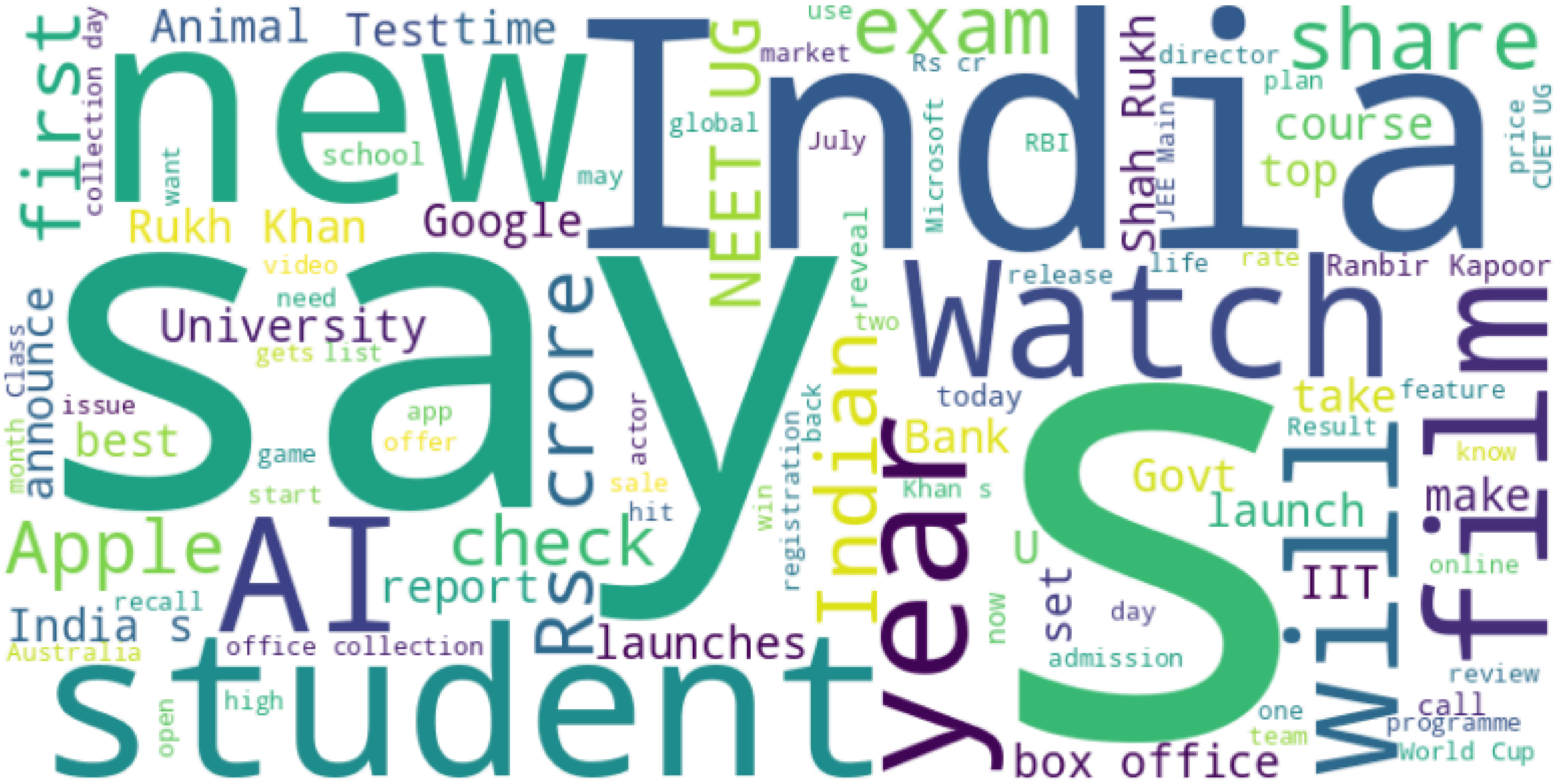
CATEGORY DISTRIBUTION FOR TRAINING DATA



Results: we observed that education has the highest count.

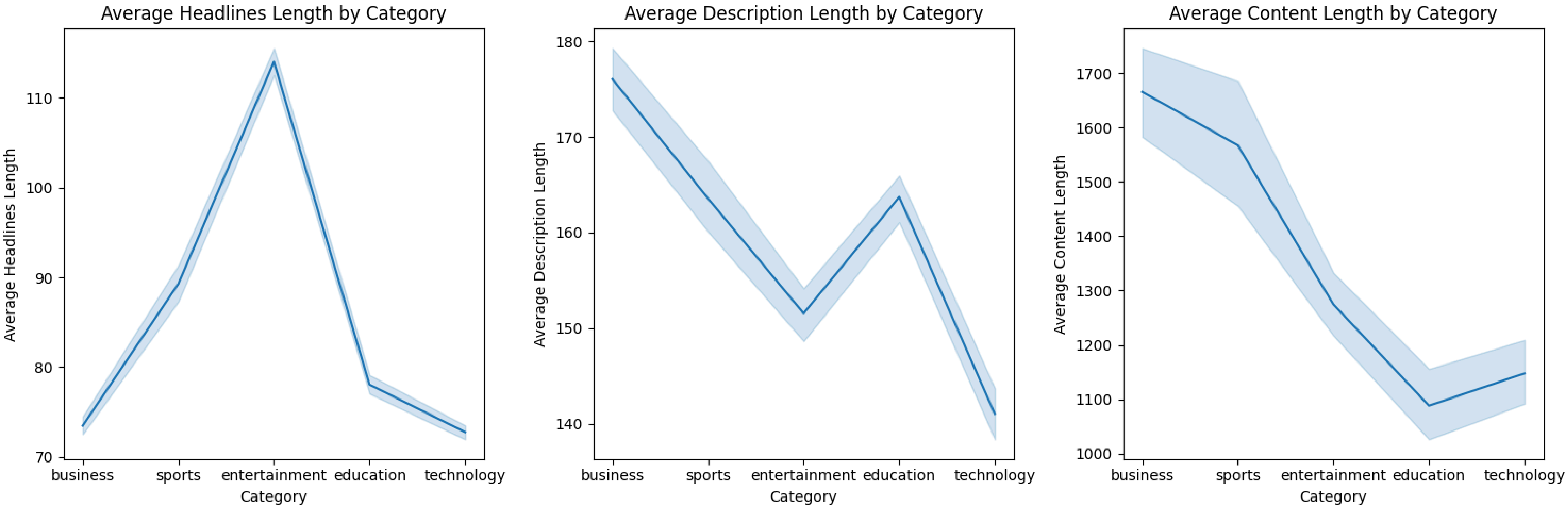
CATEGORY DISTRIBUTION FOR TEST DATA

Word Cloud for Headlines



Headline Word Cloud

TEXT LENGTH DISTRIBUTION BY CATEGORY



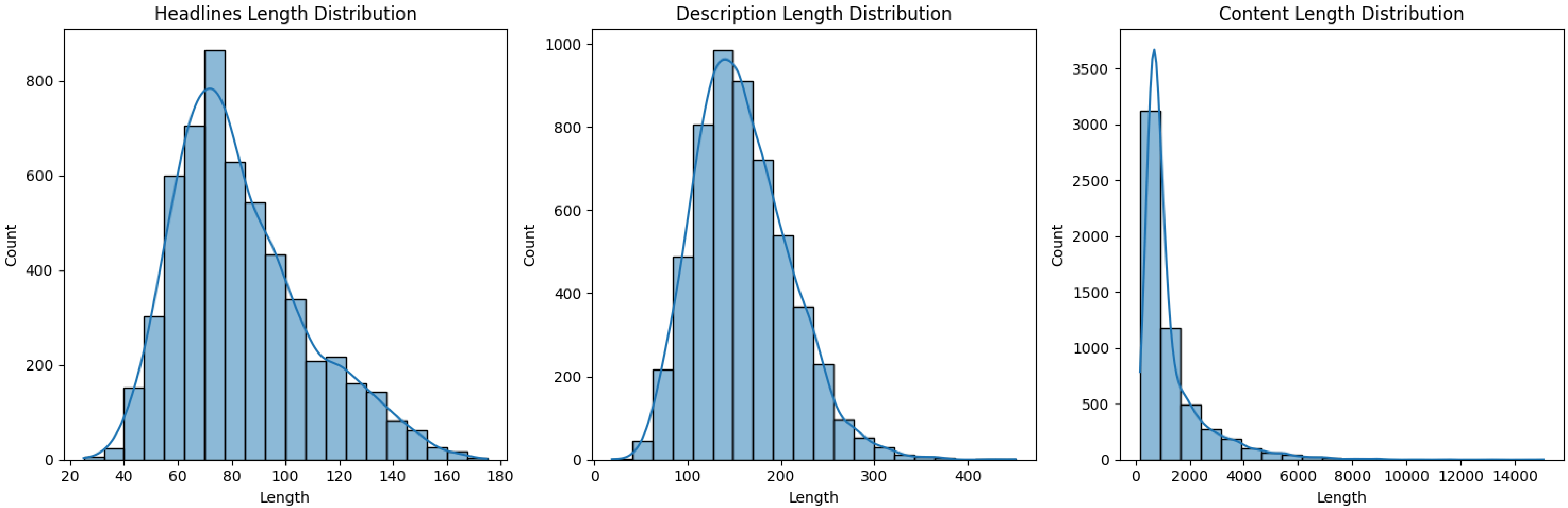
Average Headlines Length by Category

Average Description Length by Category

Average Content Length by Category

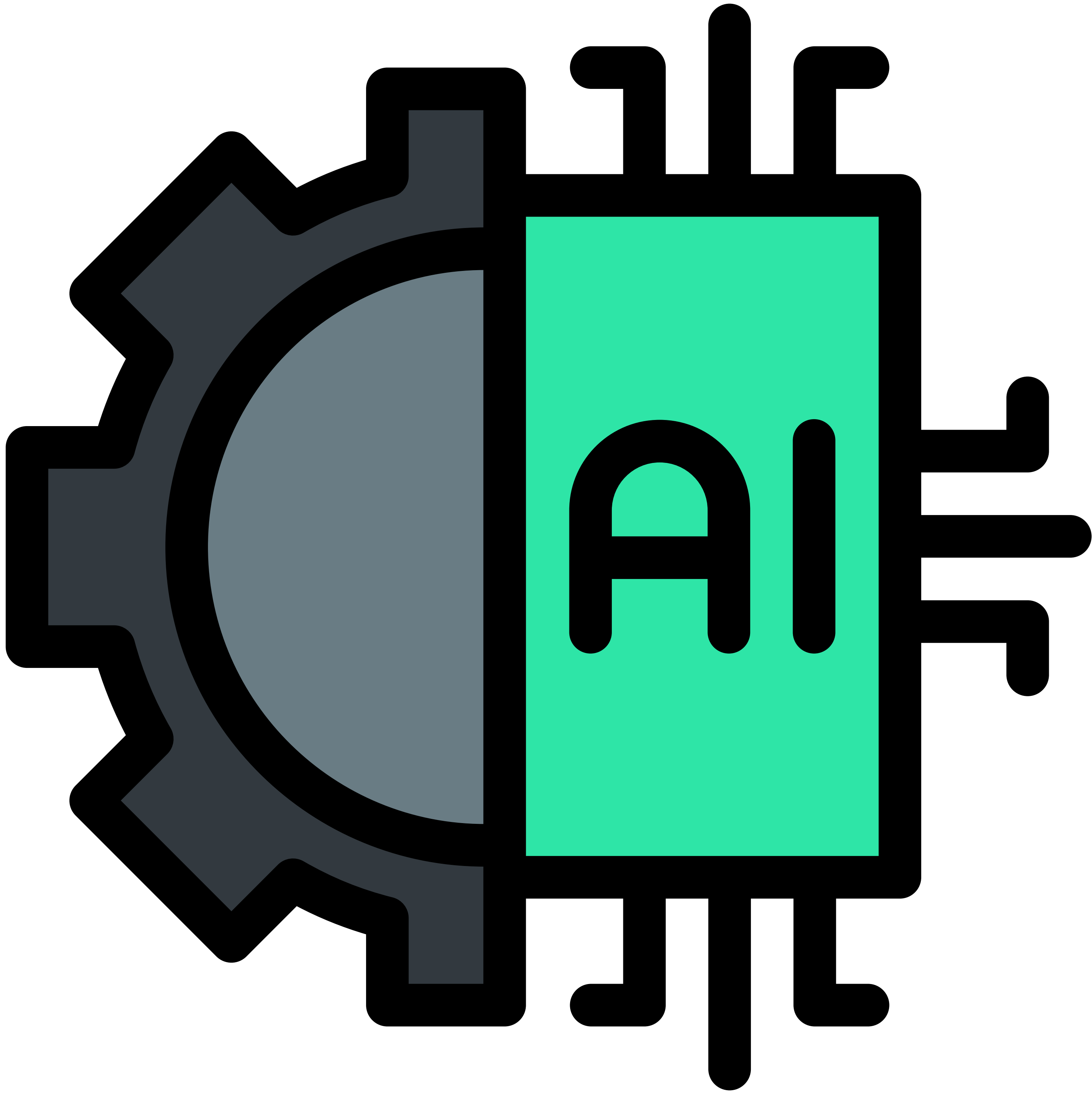
The line plots visualize the average length of headlines, descriptions, and content across different categories of news articles.

TEXT LENGTH ANALYSIS



The histogram plots visualize the total length of headlines, descriptions, and content across different categories of news articles.

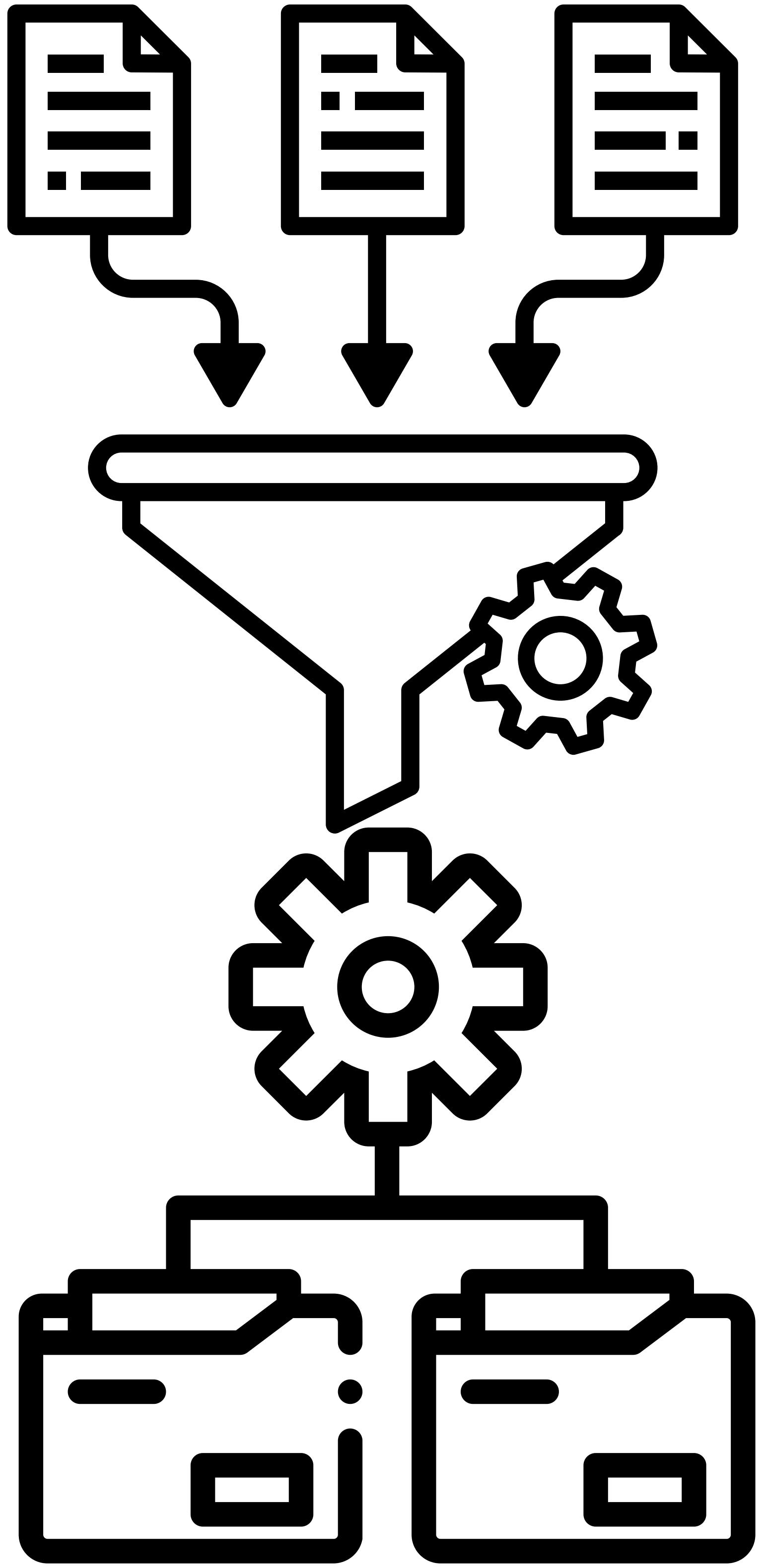
MODEL TRAINING



DATA PREPROCESSING

Data preprocessing includes:

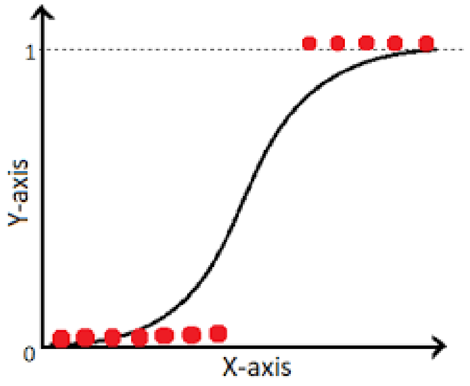
- Lowercasing
- Tokenization
- Removing Stop Words
- Lemmatization
- Handling Imbalanced Data
- Vectorization (TDIF)



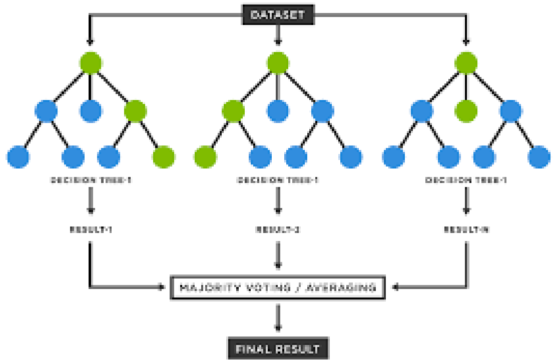
MODEL TRAINING

Models Trained:

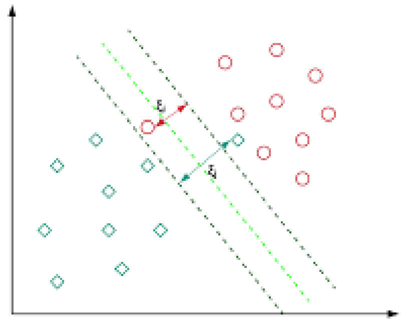
- Logistic Regression



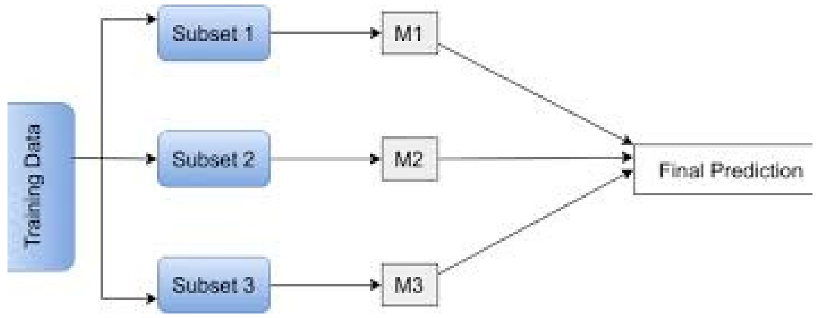
- Random Forest



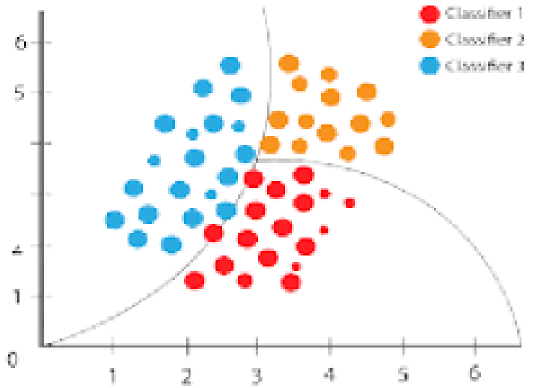
- Support Vector Machine (SVM)



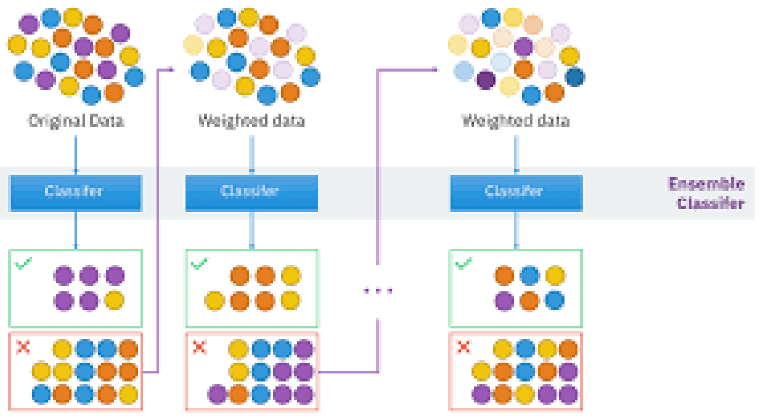
- Bagging



- Naive Bayes



- AdaBoost



MODEL TUNING

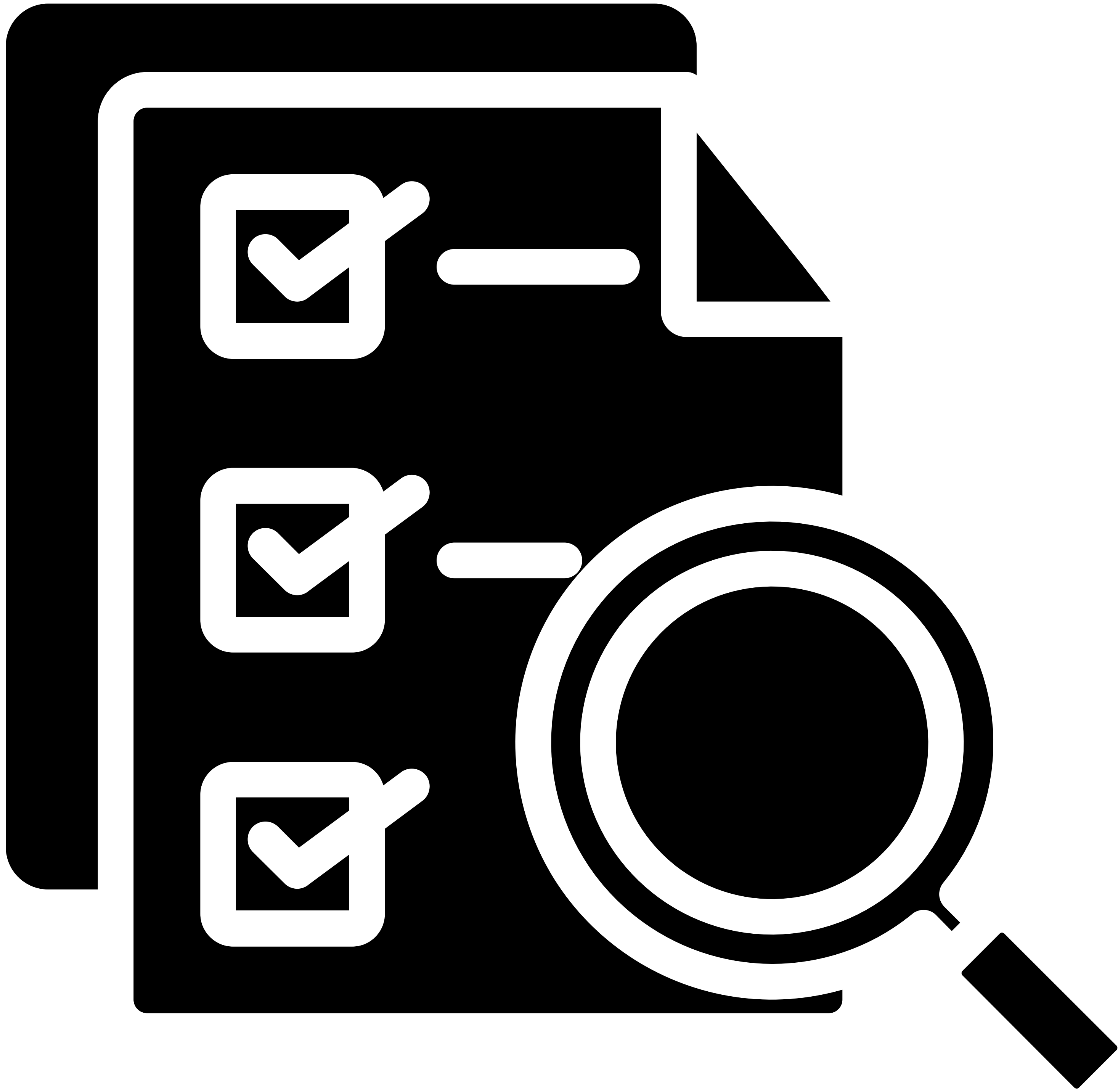
We Tune Models to:

- Optimize Performance, prevent Overfitting/Underfitting, improve Efficiency, and adapt to Data.

Tuning Method Used:

- We tuned our model using RandomSearchCV

MODEL METRICS EVALUATION

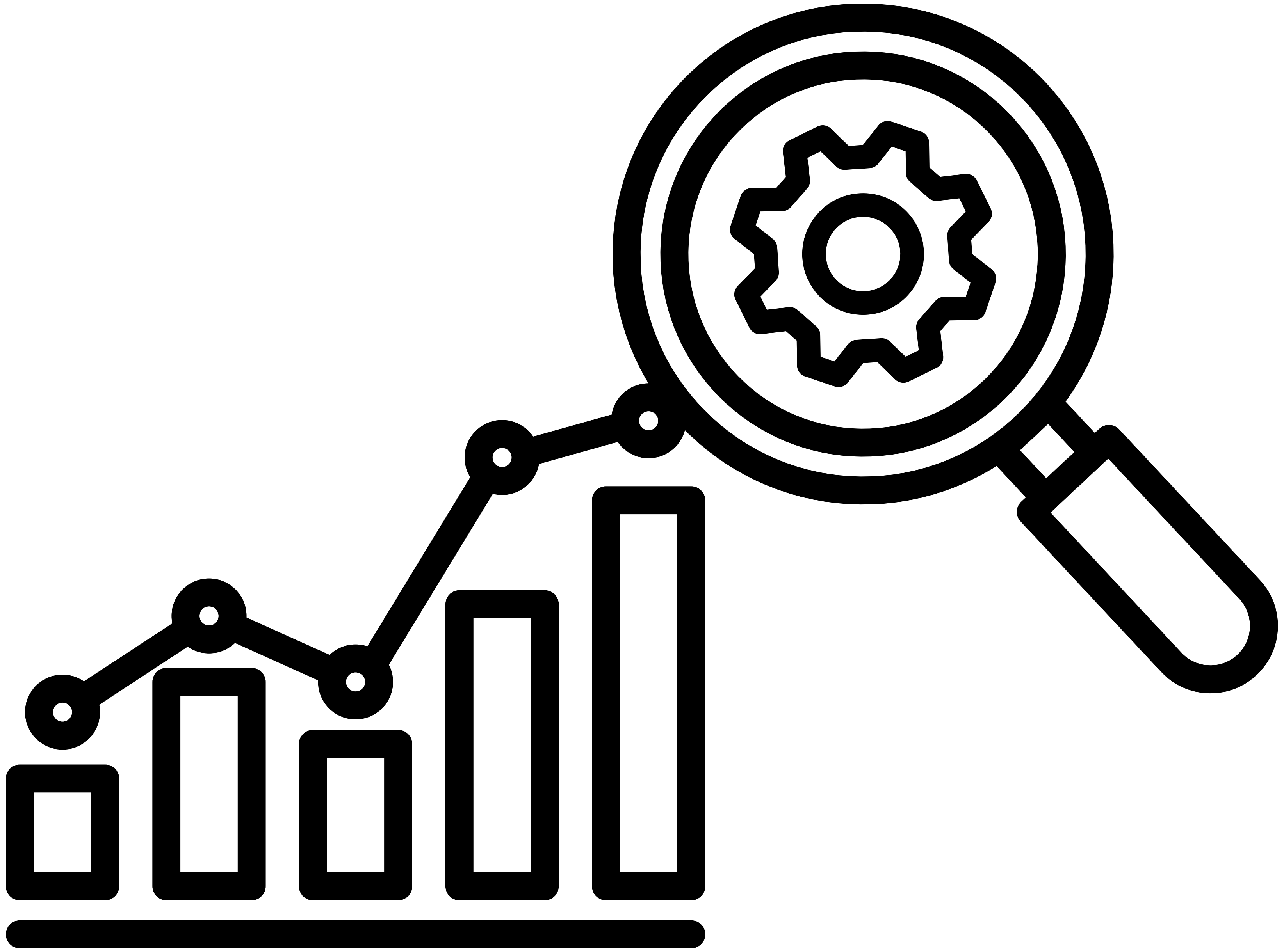


MODEL METRICS EVALUATION

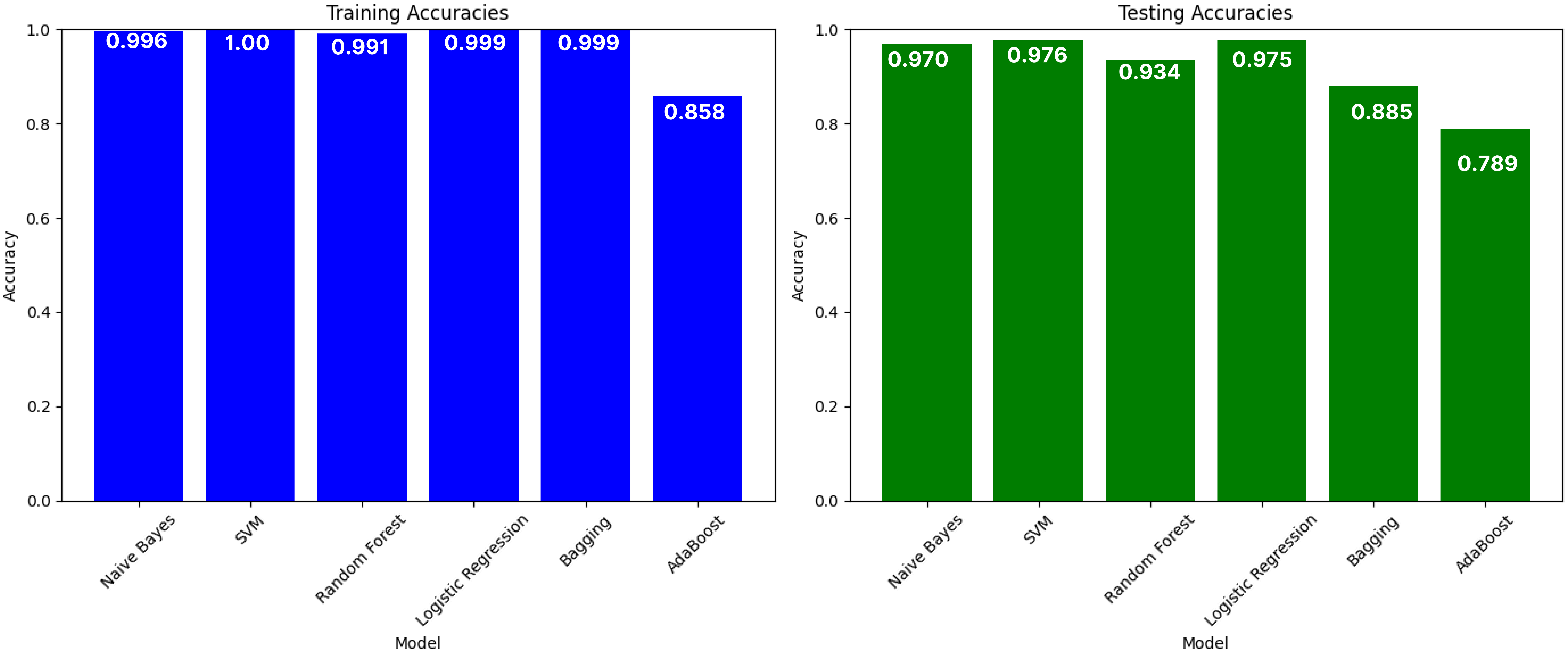
Evaluation Metrics used:

- Accuracy
- F1 Score
- Confusion Matrix
- Classification Report

MODEL PERFORMANCE ANALYSIS

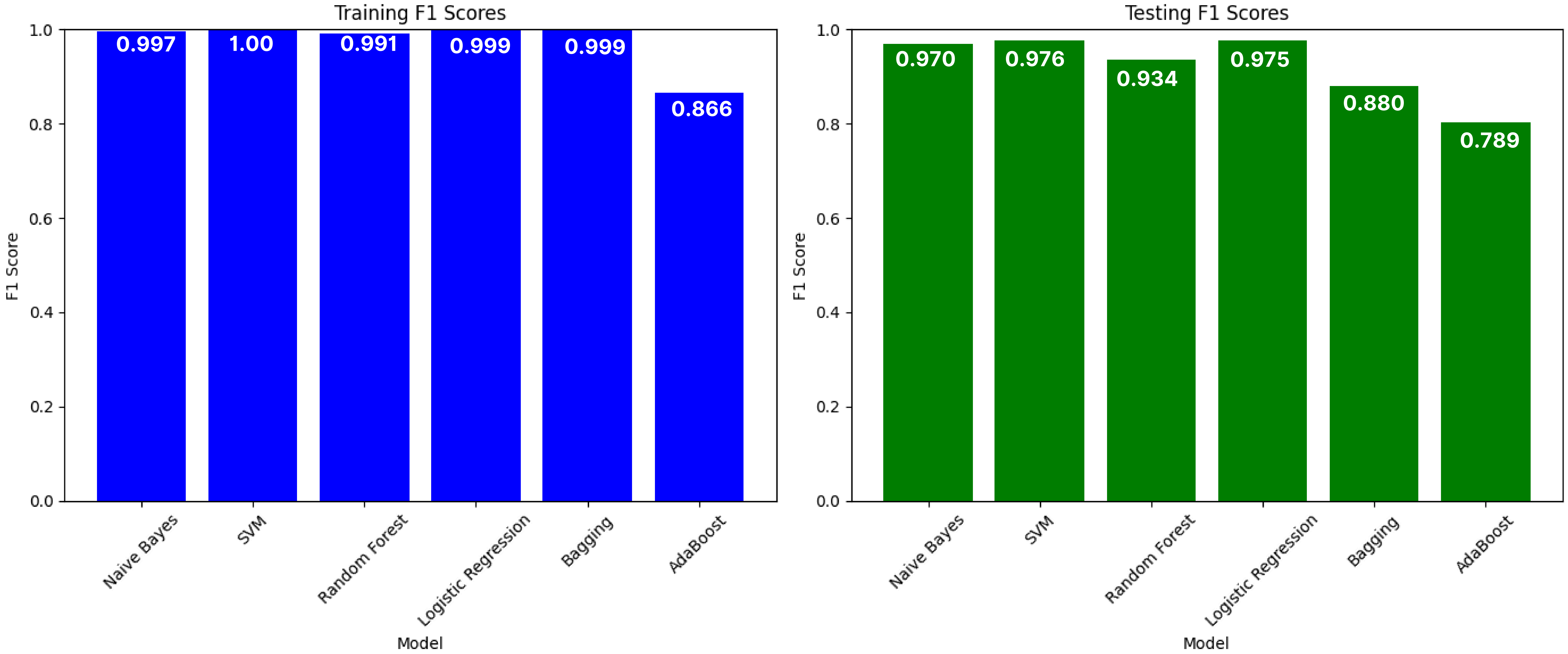


MODEL COMPARISON



The bar plot visually compares the performance of various models on the test data based on their accuracy scores for the training and test data.

MODEL COMPARISON

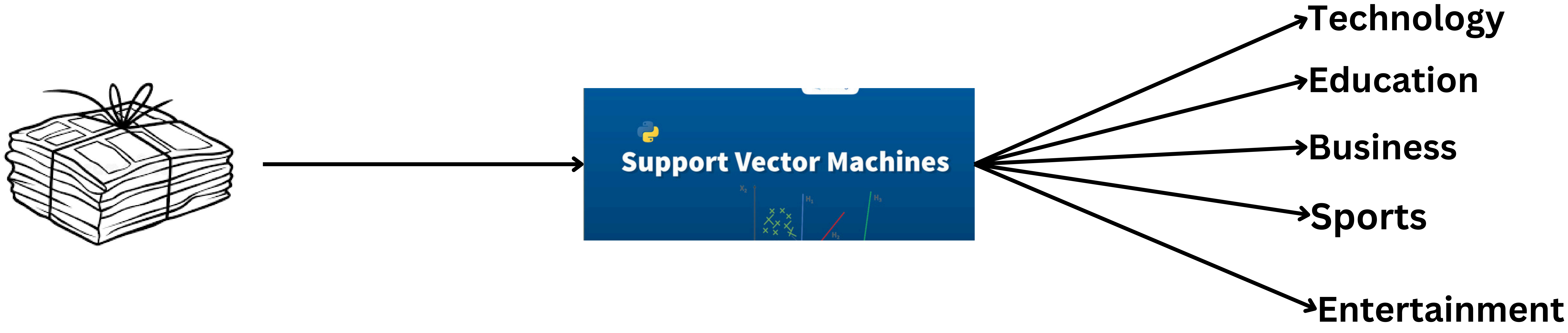


The bar plot visually compares the performance of various models based on their F1 scores on the training and test data.

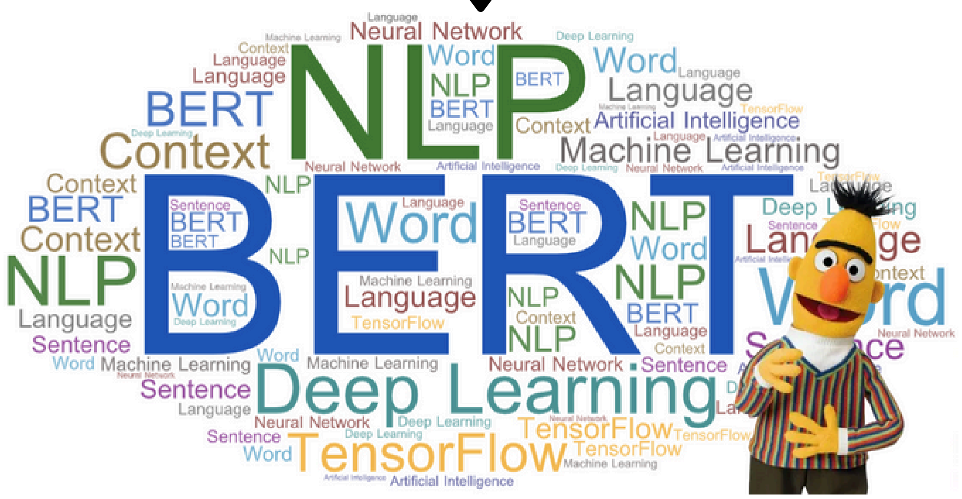
STREAMLIT APPLICATION

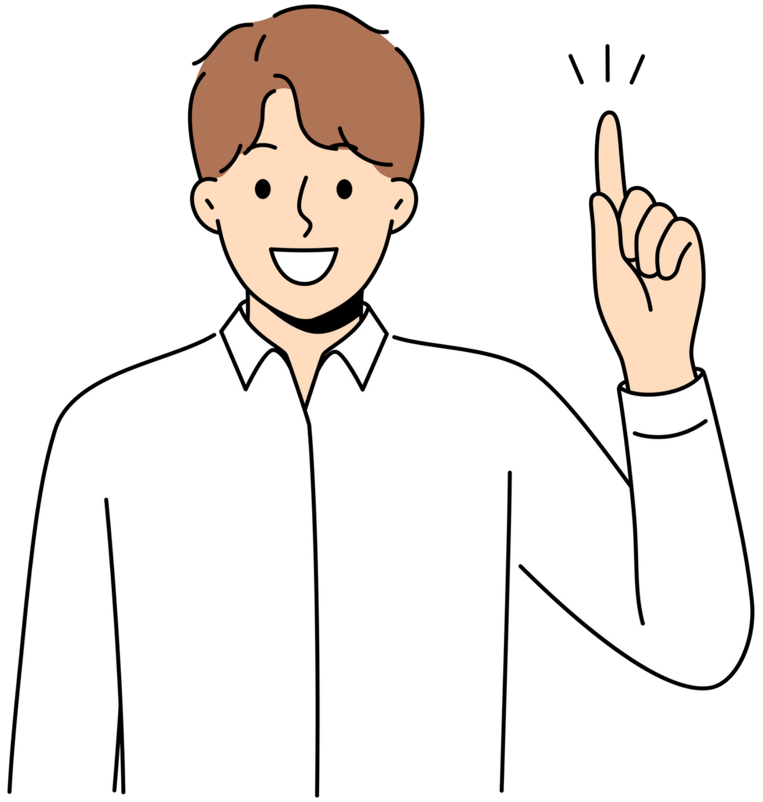
https://justin9503-news-classifier-base-app-h5anzs.streamlit.app/

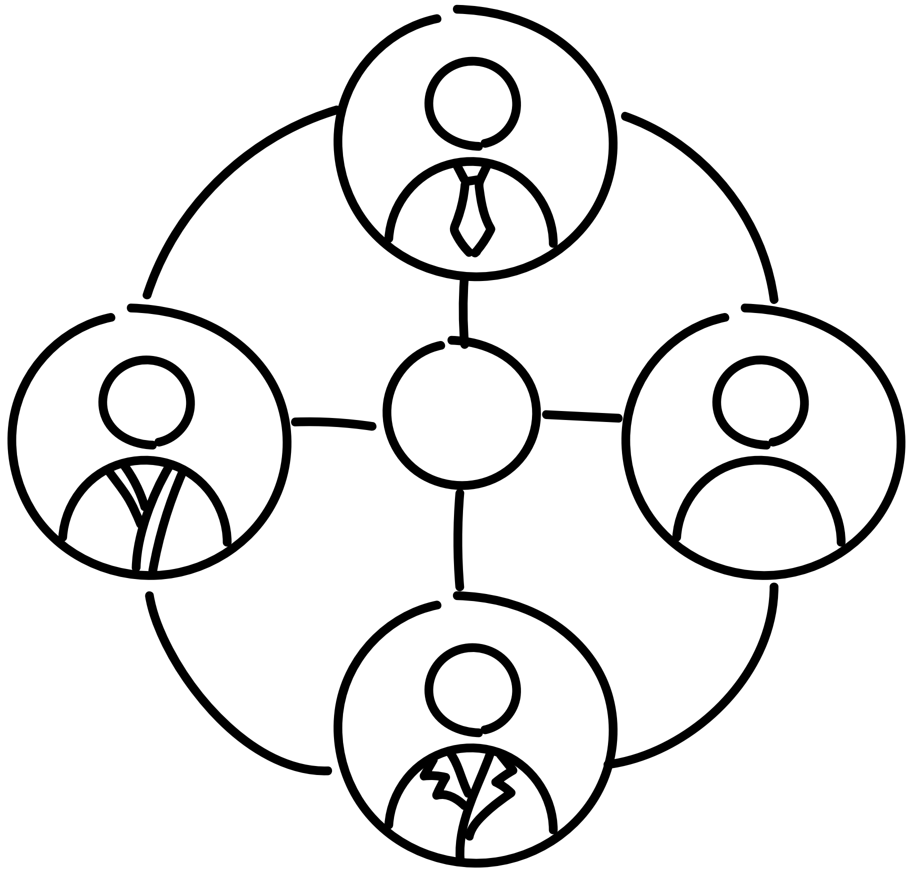
IMPLICATIONS OF THE FINDINGS AND SUGGESTIONS FOR FUTURE WORK



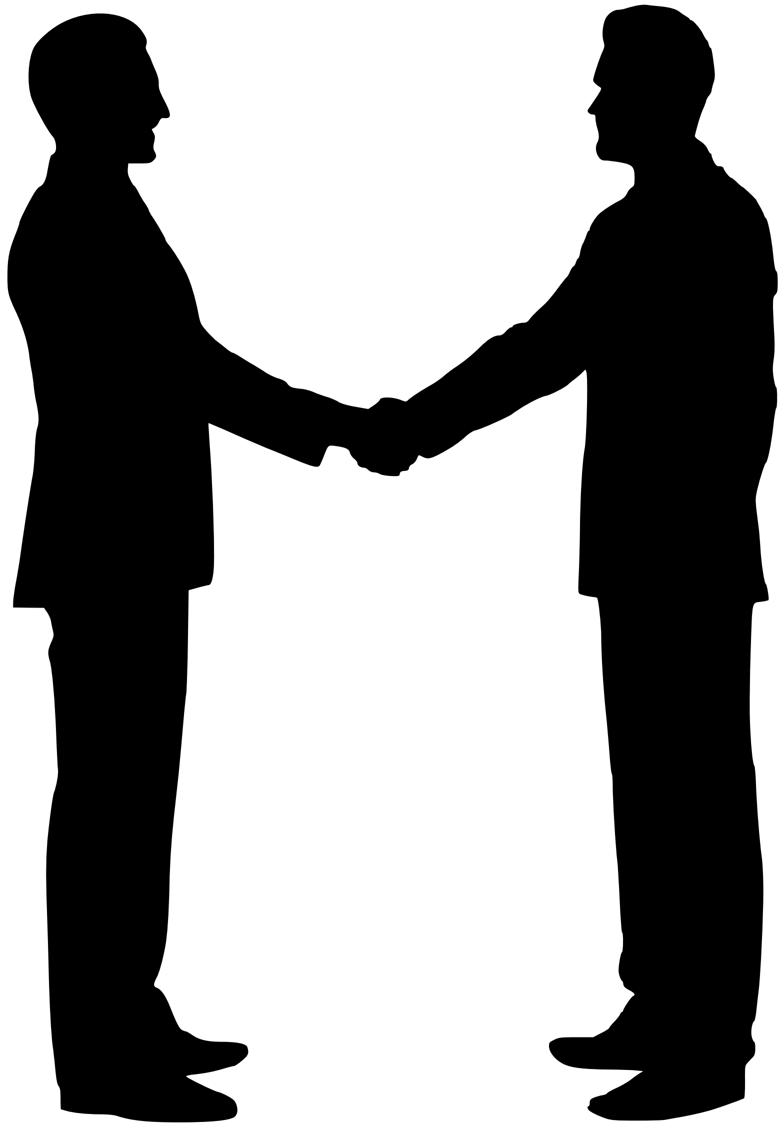
- Using deep learning techniques like BERT



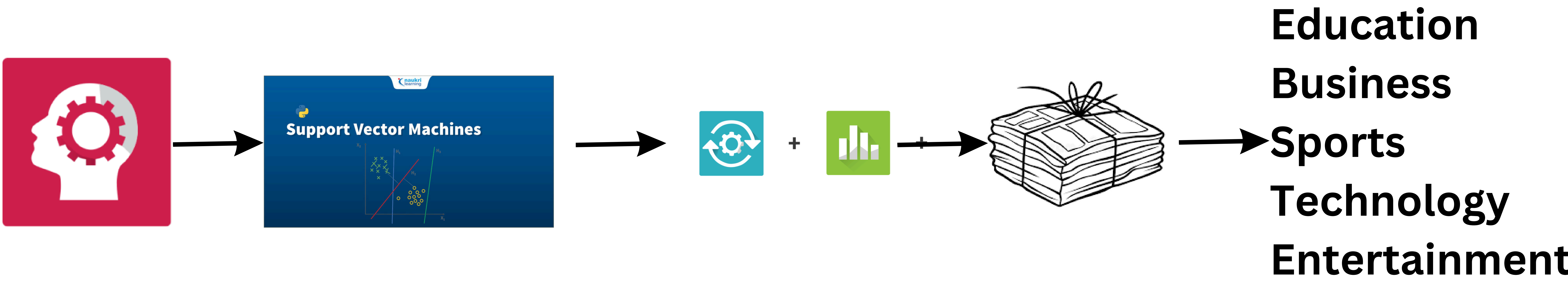




CONCLUSION



CONCLUDING THOUGHTS



Education
Business
Sports
Technology
Entertainment

- The excellent performance of SVM, Naive Bayes, and Logistic regression, are highly effective in classifying news articles.

- These models can be utilized by platforms to tag and categorize articles automatically.

THANK YOU!

QUESTIONS ARE WELCOME