THURSDAY 06/02/2025

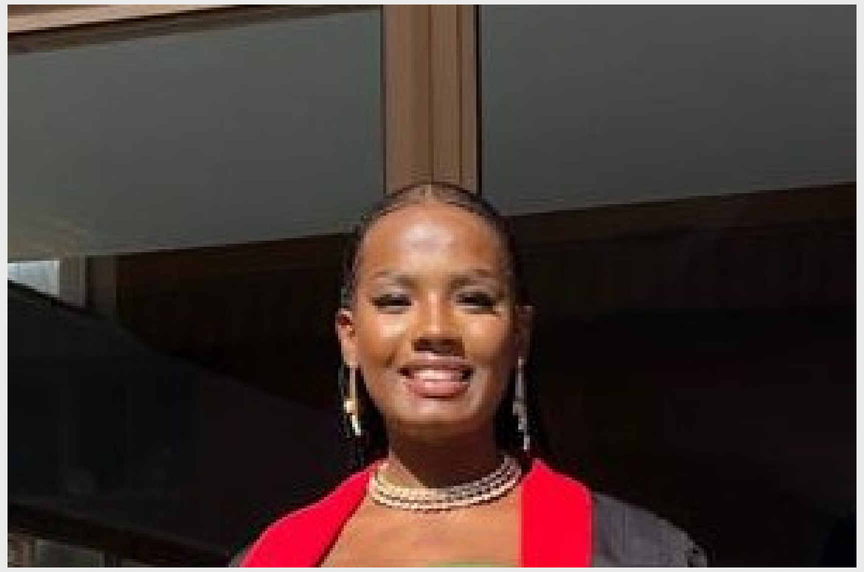
PRESENTED BY TEAM 3

NEWS ARTICLE

CATEGORY PREDICTOR

MEET THE TEAM

Nthabiseng Mokhachane



Lindelwe Mathonsi

Melokuhle Makhwasa

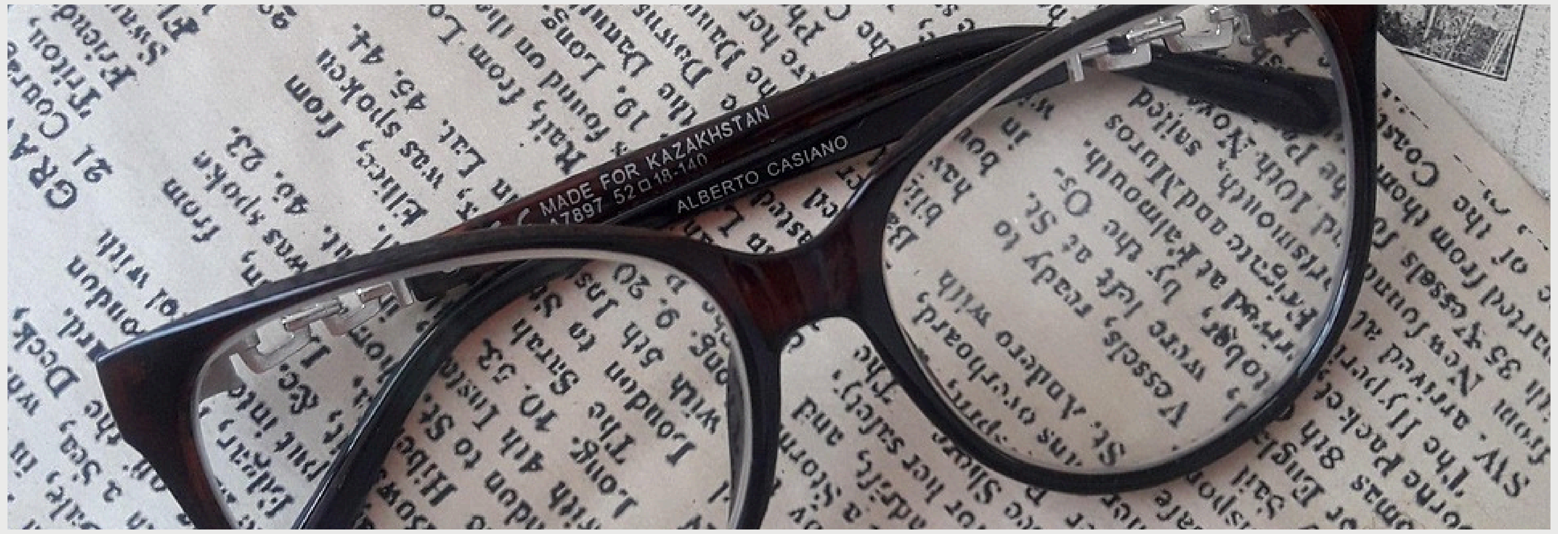
Lamel Kekana

Sanele Bhembe

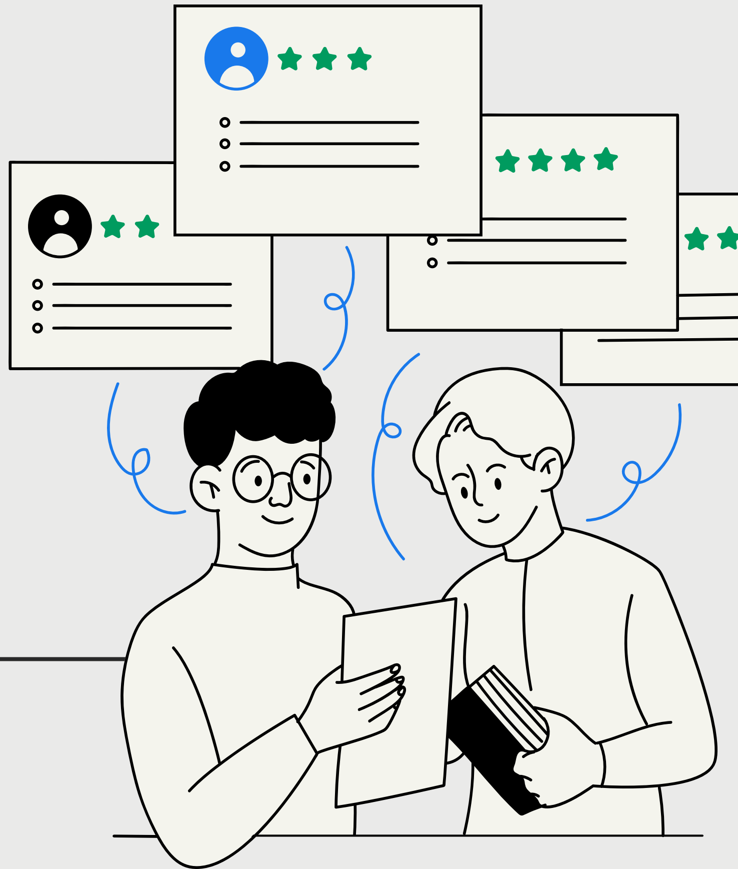
INTRODUCTION

AIM

- Develop a robust model to correctly categorize a news article based on the analyses of the articles' content.



- A Data loading and inspection
- B Data cleaning
- C Data preprocessing
- D Exploratory data analysis (EDA)
- E Model development
- F Model evaluation
- G Model deployment



OBJECTIVES

DATA CLEANING

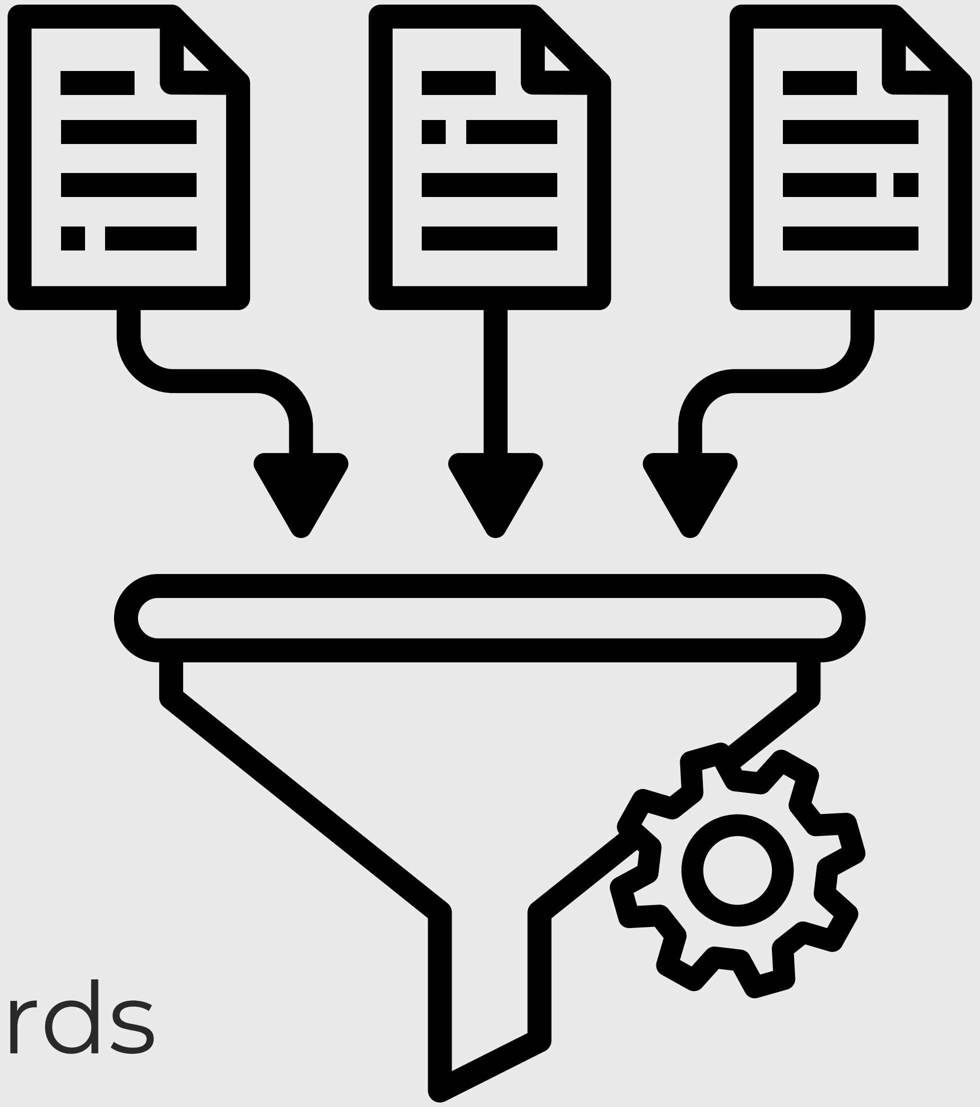
- From a Github source.

Headline	Description	Content	Url	Category
text	text	text	text	text
text	text	text	text	text

- Drop duplicate value.
- Standardize column naming convention.

DATA PREPROCESSING

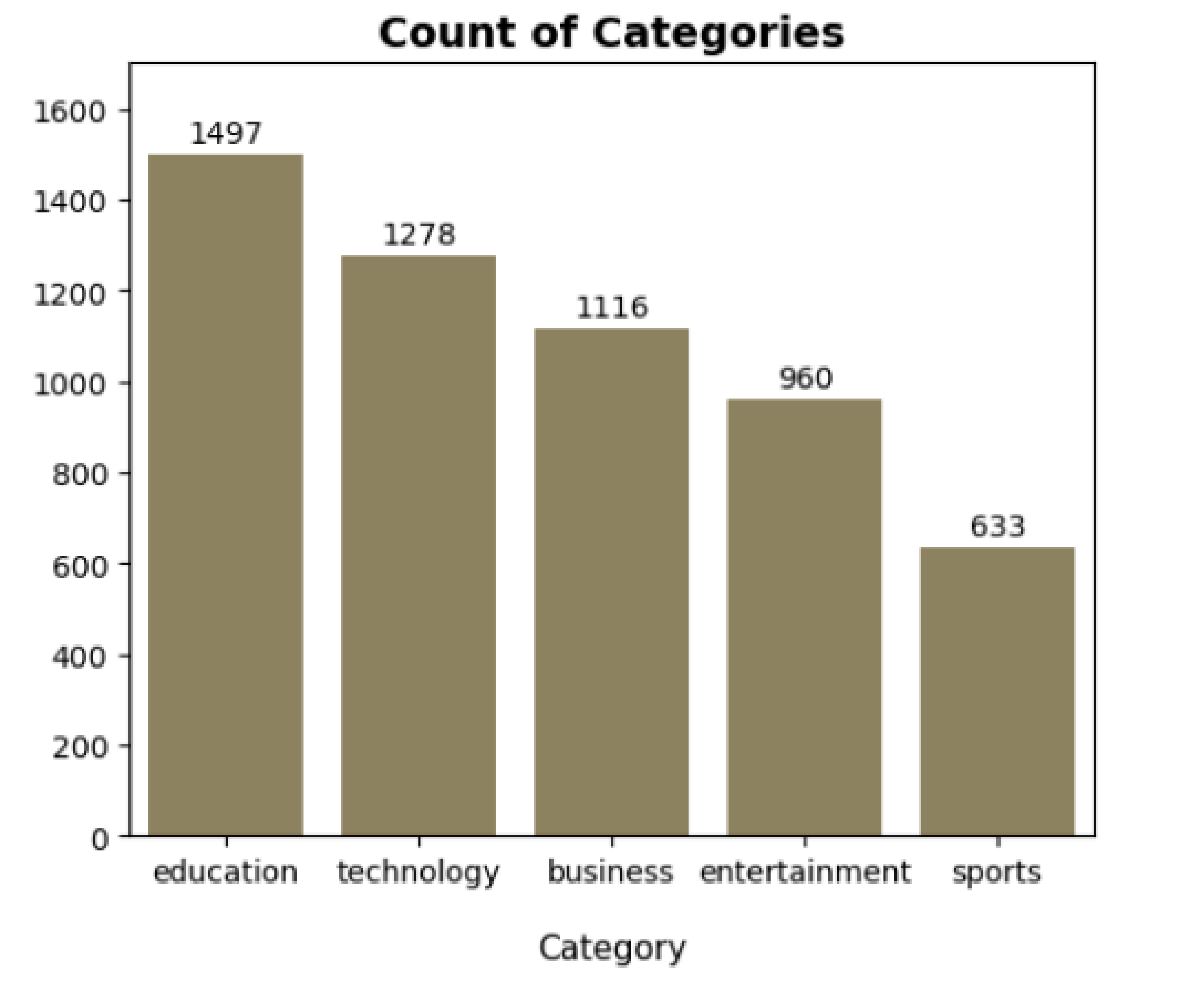
- Text cleaning
- Tokenization
- Remove Stopwords

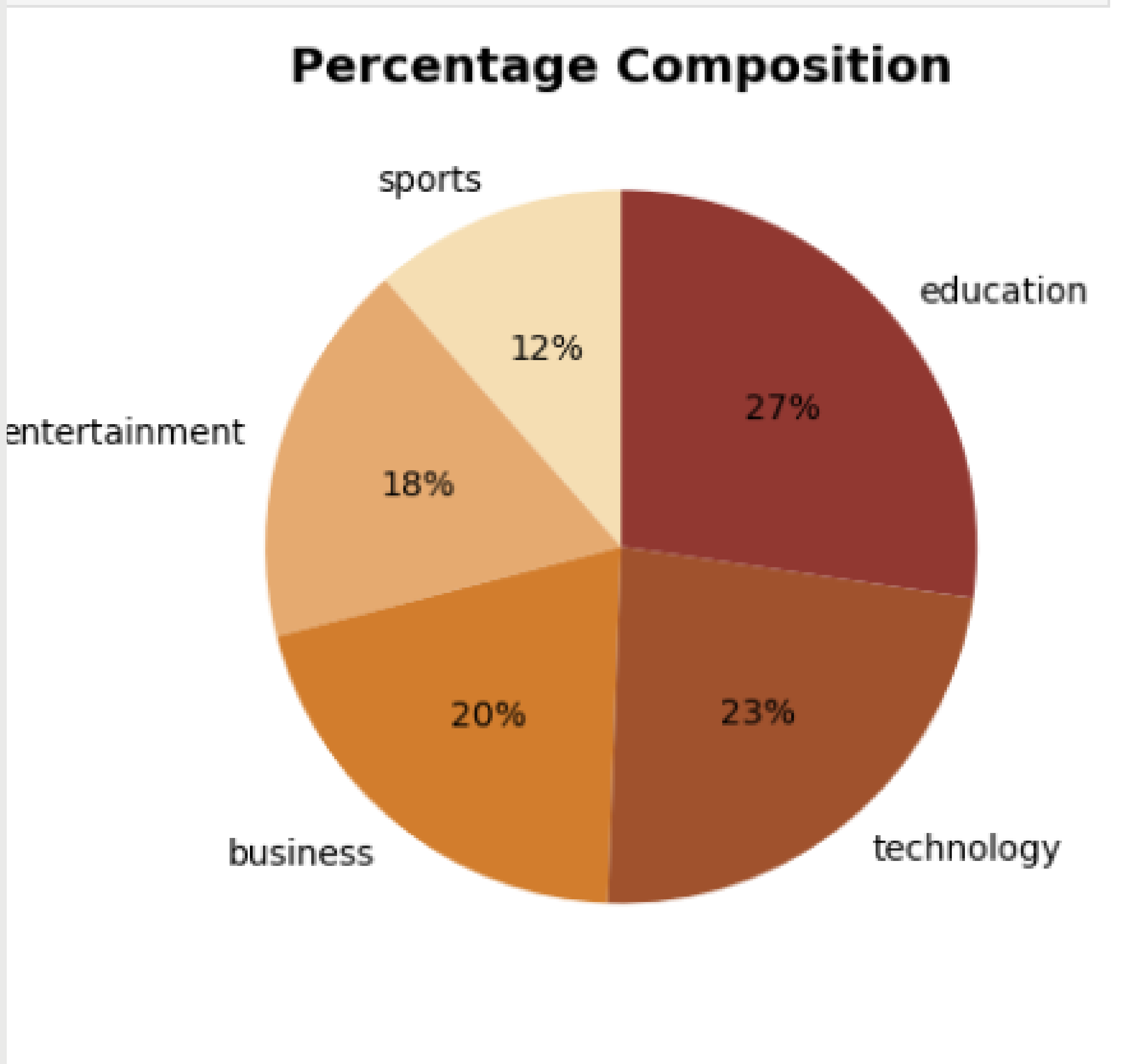


- Lemmatization
- Feature Extraction
- Vectorization

EXPLORATORY DATA ANALYSIS

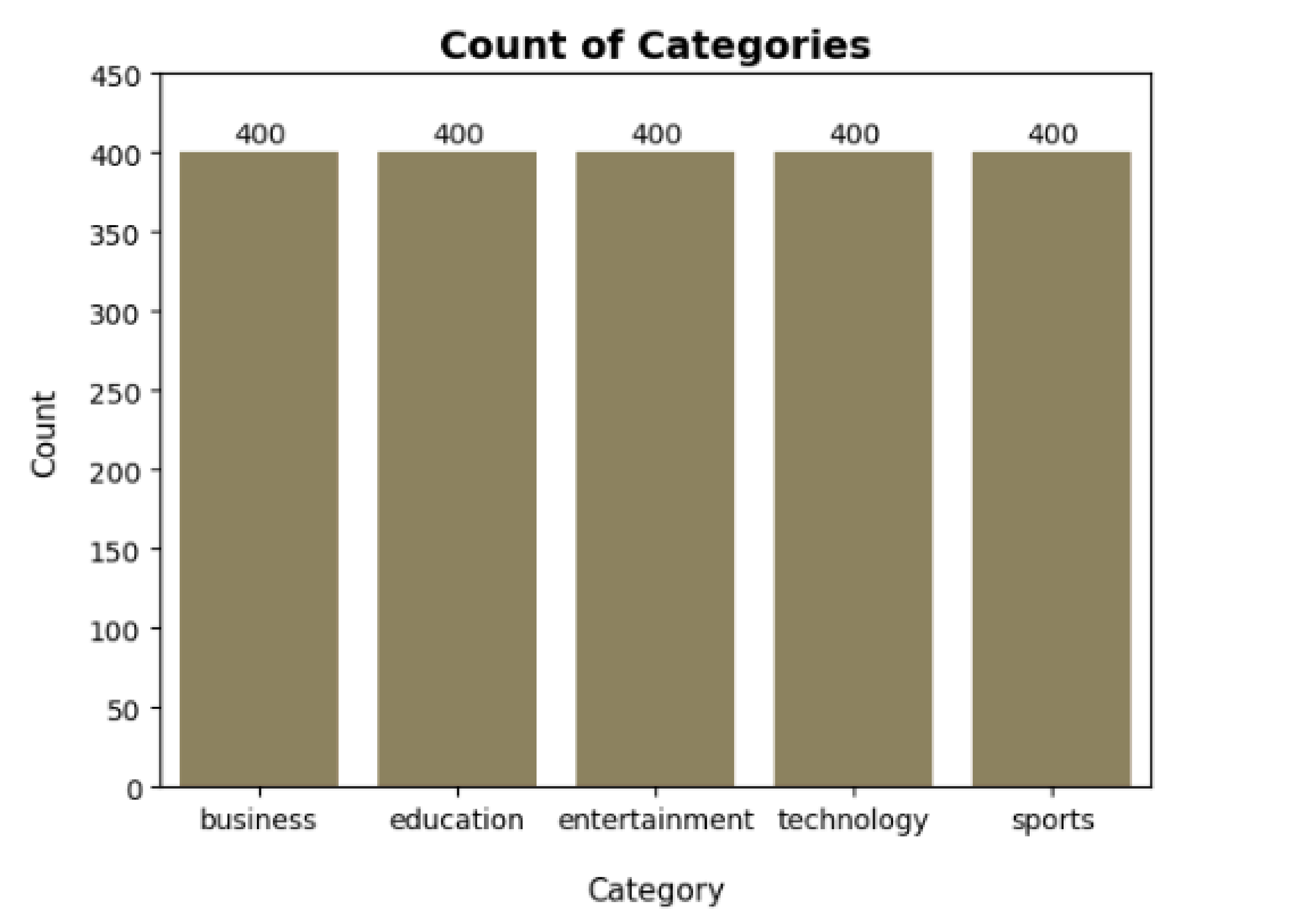
- Visualising the training data

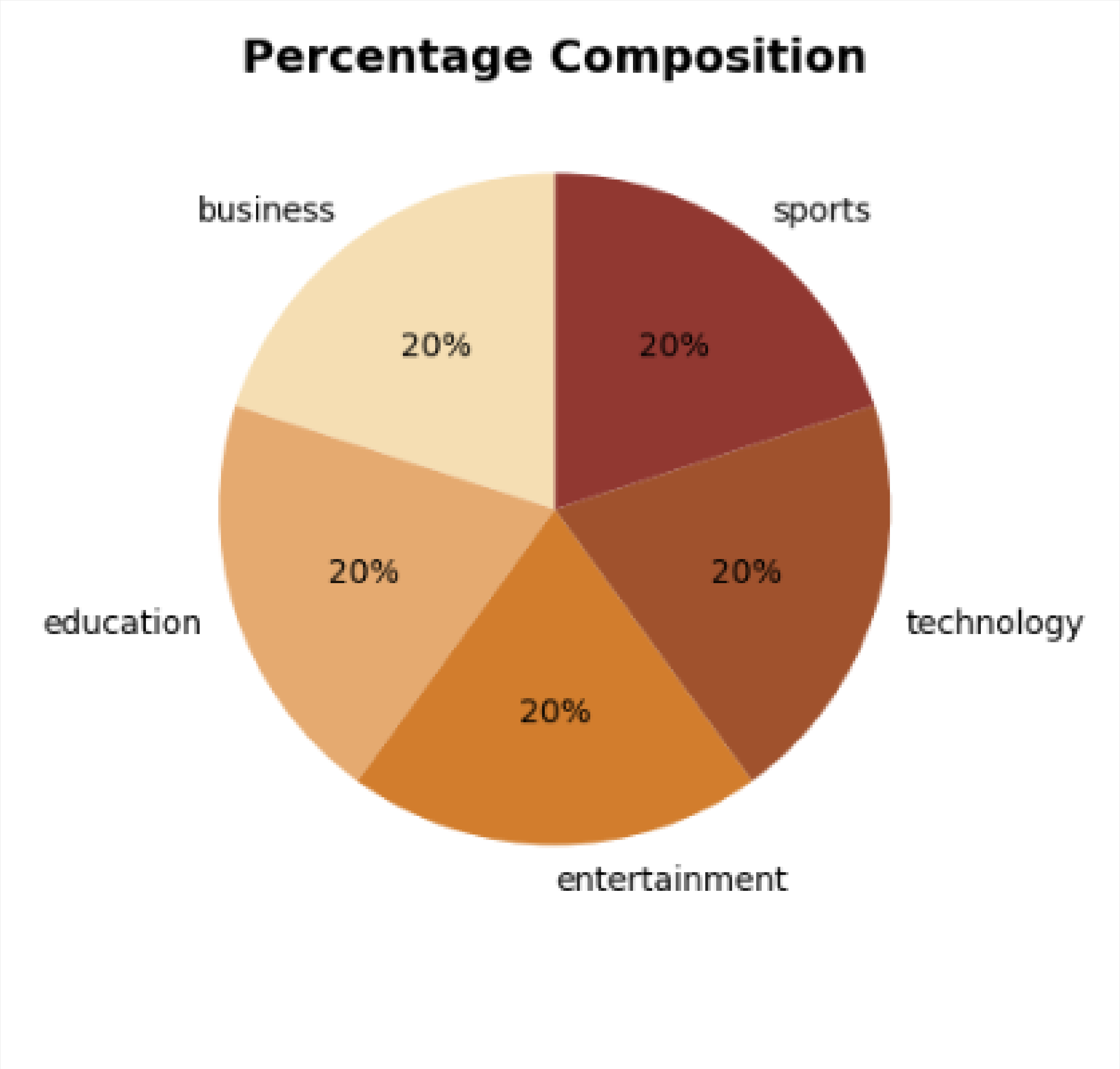




EXPLORATORY DATA ANALYSIS

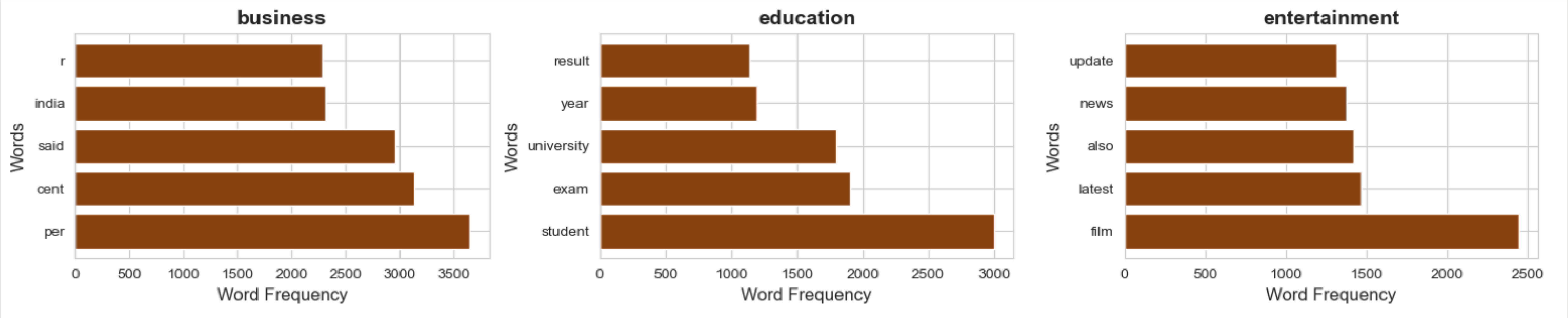
- Visualising the testing data





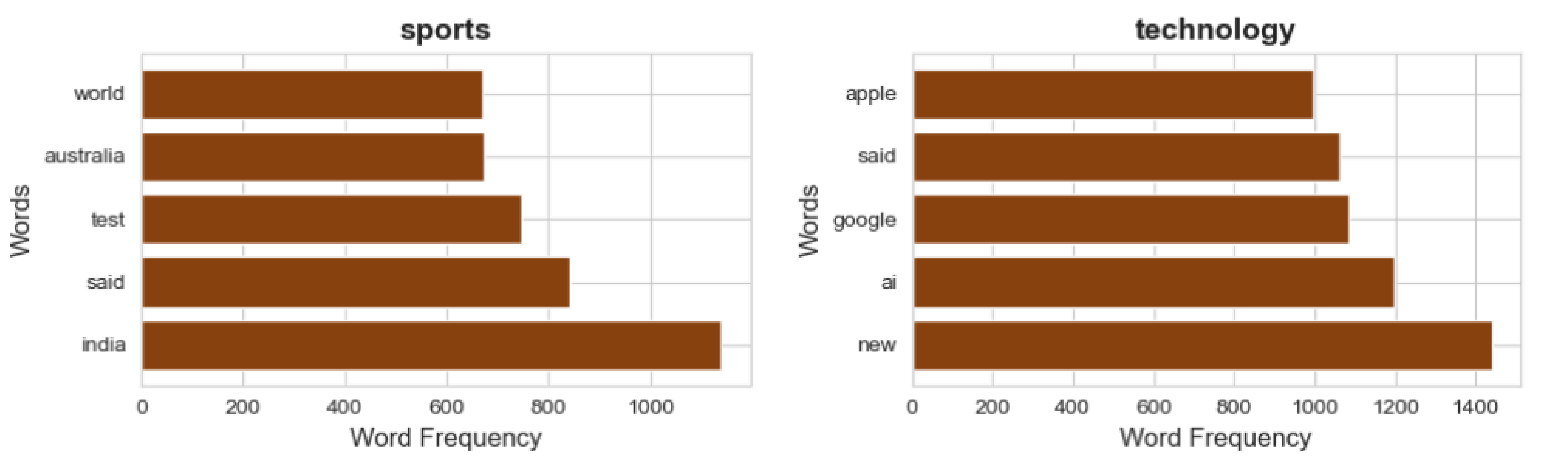
EXPLORATORY DATA ANALYSIS

- Top 5 frequent words per category

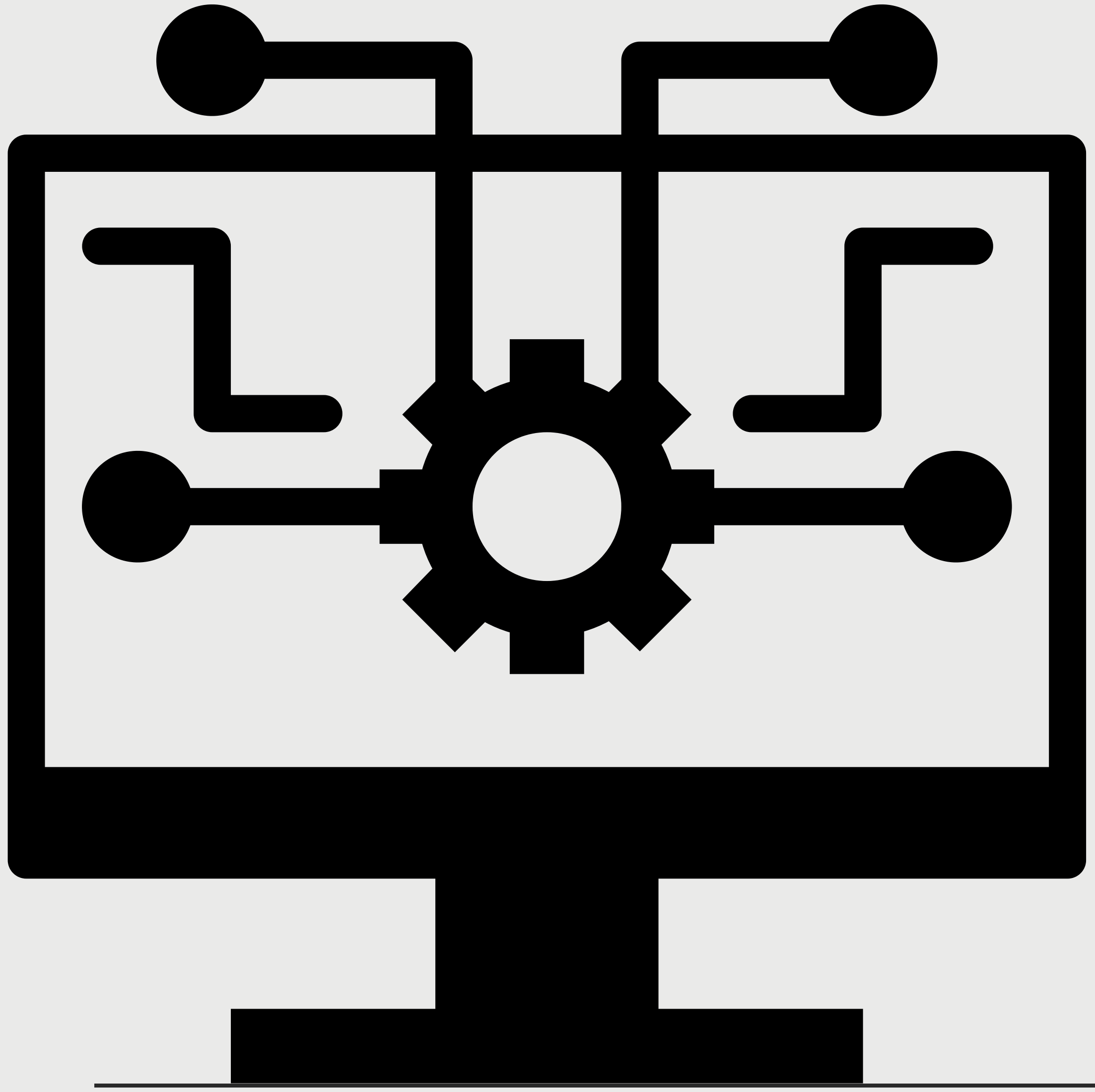


EXPLORATORY DATA ANALYSIS

- Top 5 frequent words per category



MODELS



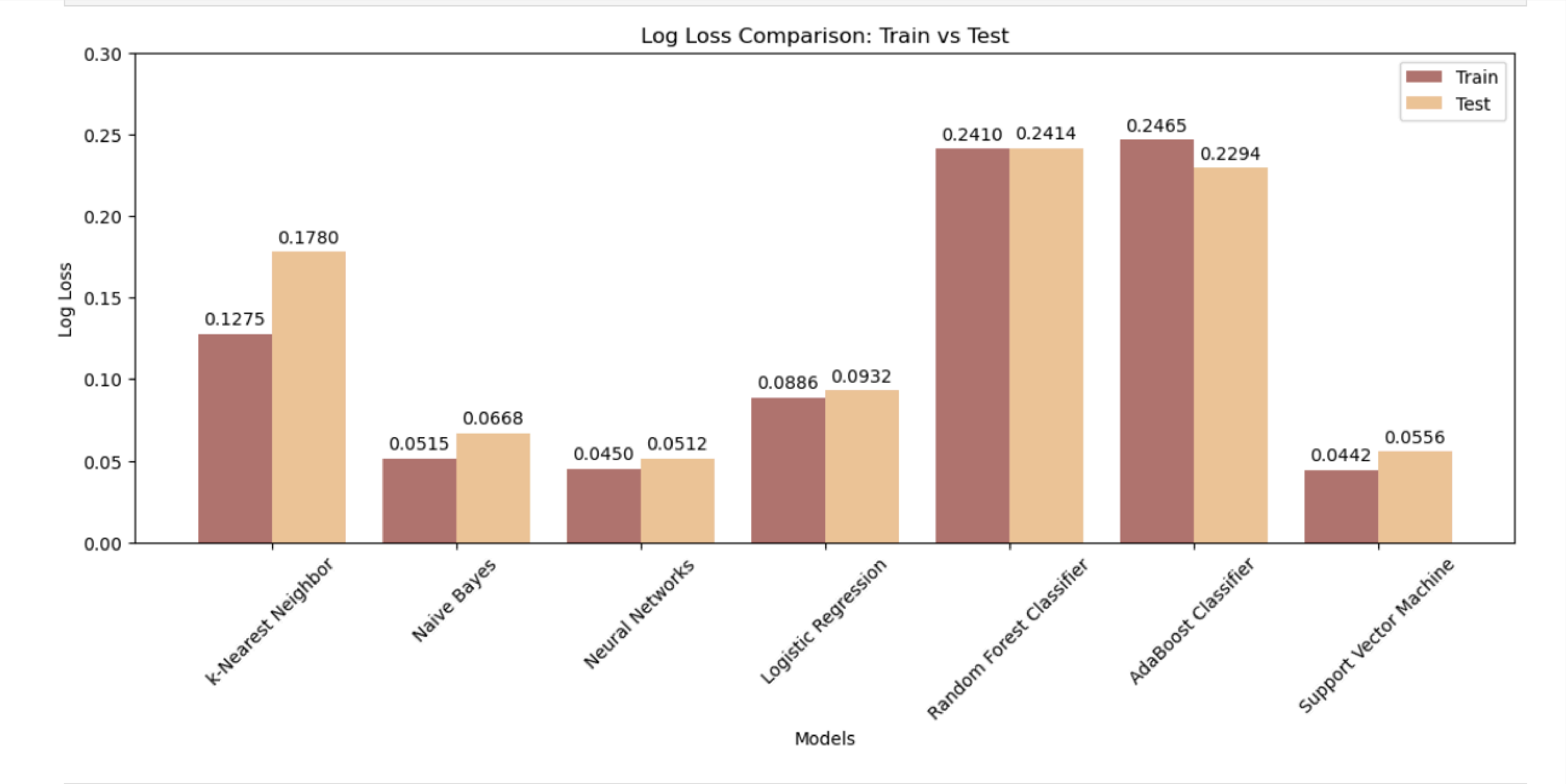
- K-Nearest Neighbor
- Naive Bayes
- Neural Networks
- Logistic Regression
- Random Forest Classifier
- AdaBoost Classifier
- Support Vector Machine (SVM)

VALIDATION

Model	Hyperparameters	Log loss
K-Nearest Neighbour	metric: euclidean, n_neighbors: 30, weights: distance	0.1276
Naive Bayes	alpha: 0.1, class_prior: None, fit_prior: True	0.05148
Neural Networks	activation: relu, alpha: 0.0001, batch_size: 32, hidden_layer_sizes: (20, 20), learning_rate: constant, max_iter: 20	0.9882
Logistic Regression	C: 10, max_iter: 500, multi_class: ovr, solver: lbfgs	0.08862
Random Forest Classifier	max_depth: None, n_estimators: 1000	0.2410
Adaboost Classifier	algorithm: SAMME.R, estimator__max_depth: 5, learning_rate: 0.01, n_estimators: 100	0.2464
Support Vector Machine	max_iter: 1000, kernel: linear, C: 1.0	0.0442

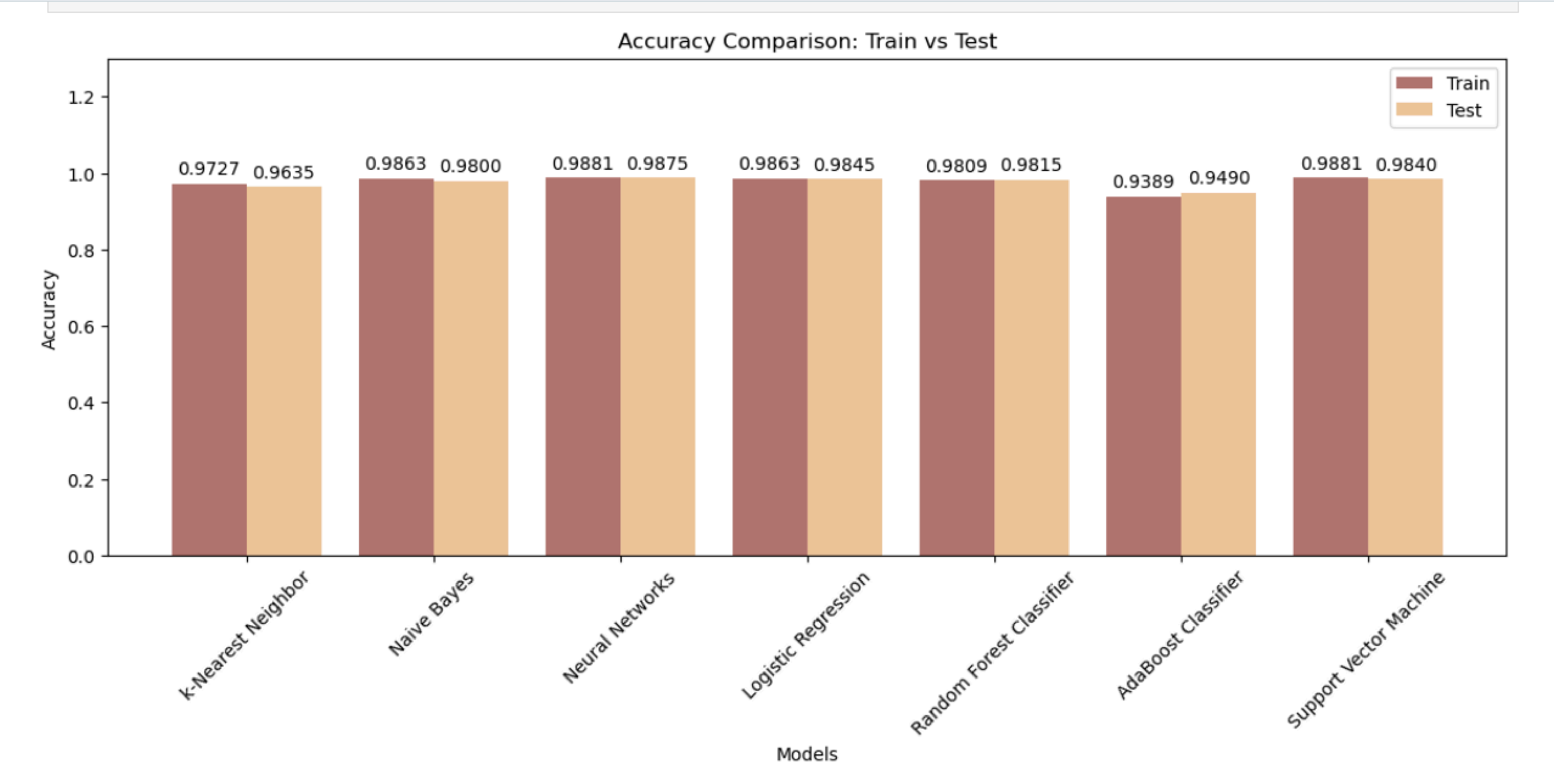
EVALUATION

- Model performance on unseen data.



EVALUATION

- Model performance on unseen data.



EVALUATION

Most important features for the best 3 models.

1. Naive Bayes Model

education
entertainment
news
technology
film

2. Neural Networks Model

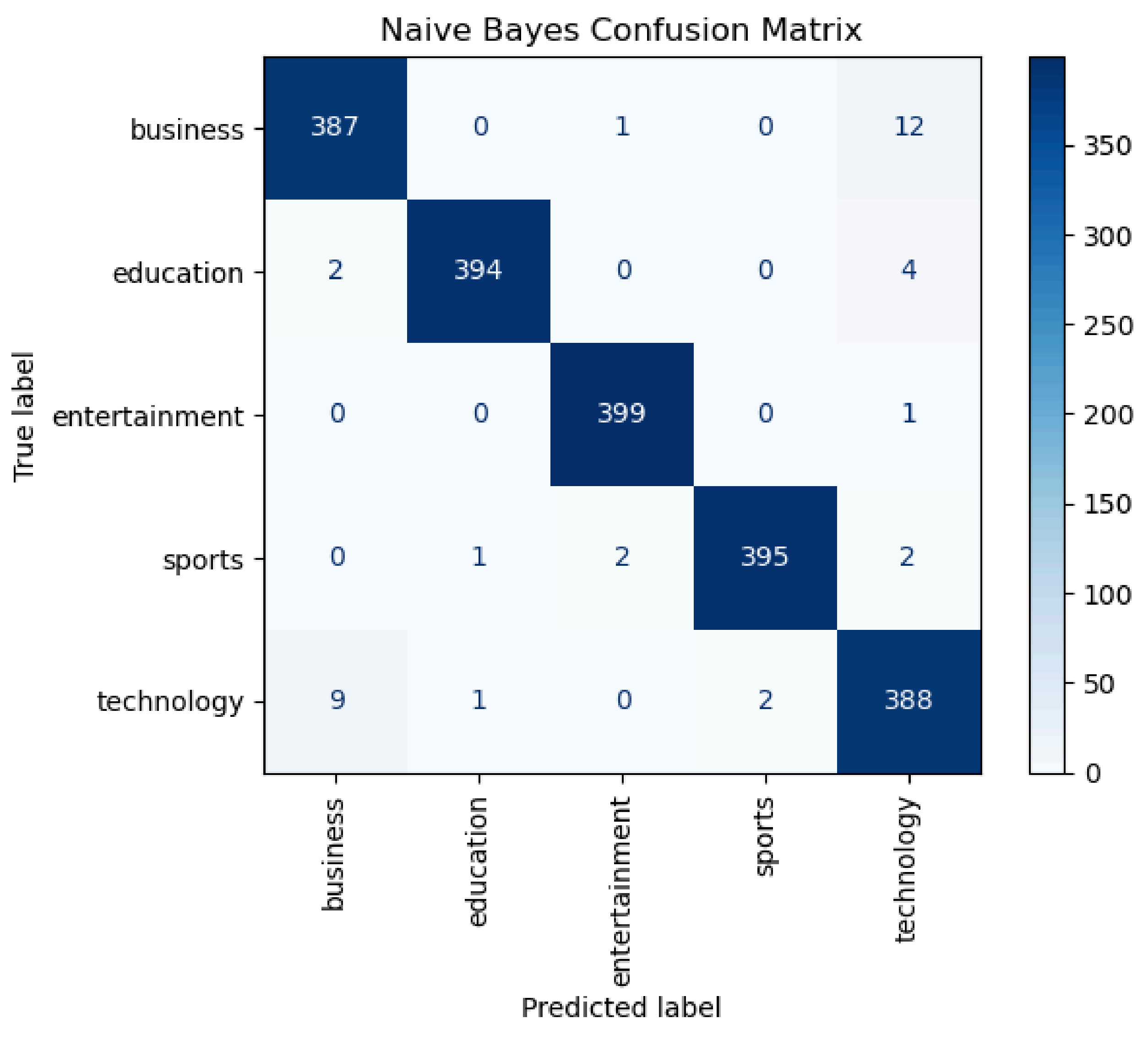
education
business
technology
student
entertainment

3. Support Vector Machine Model

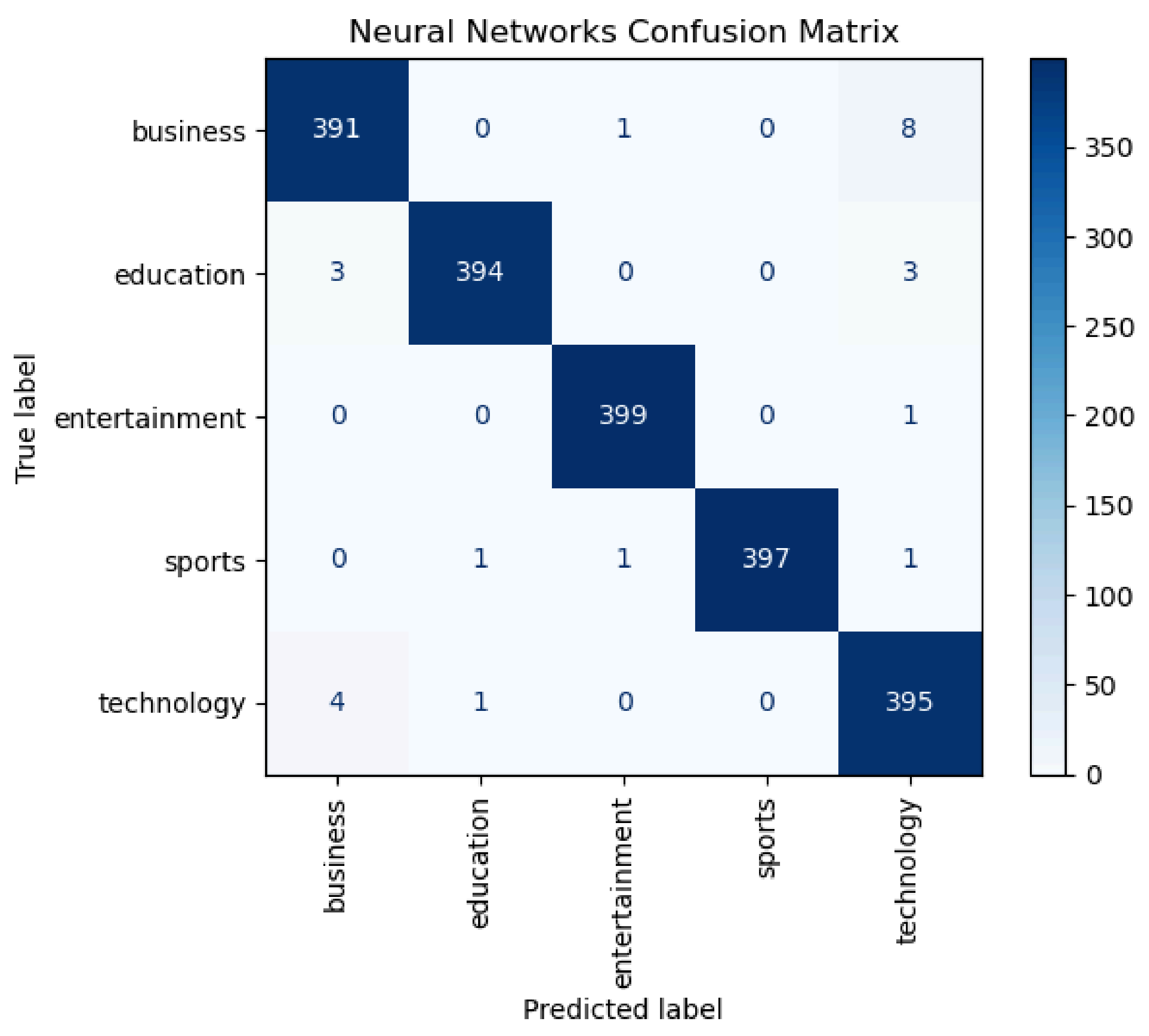
film
student
technology
education
business

EVALUATION

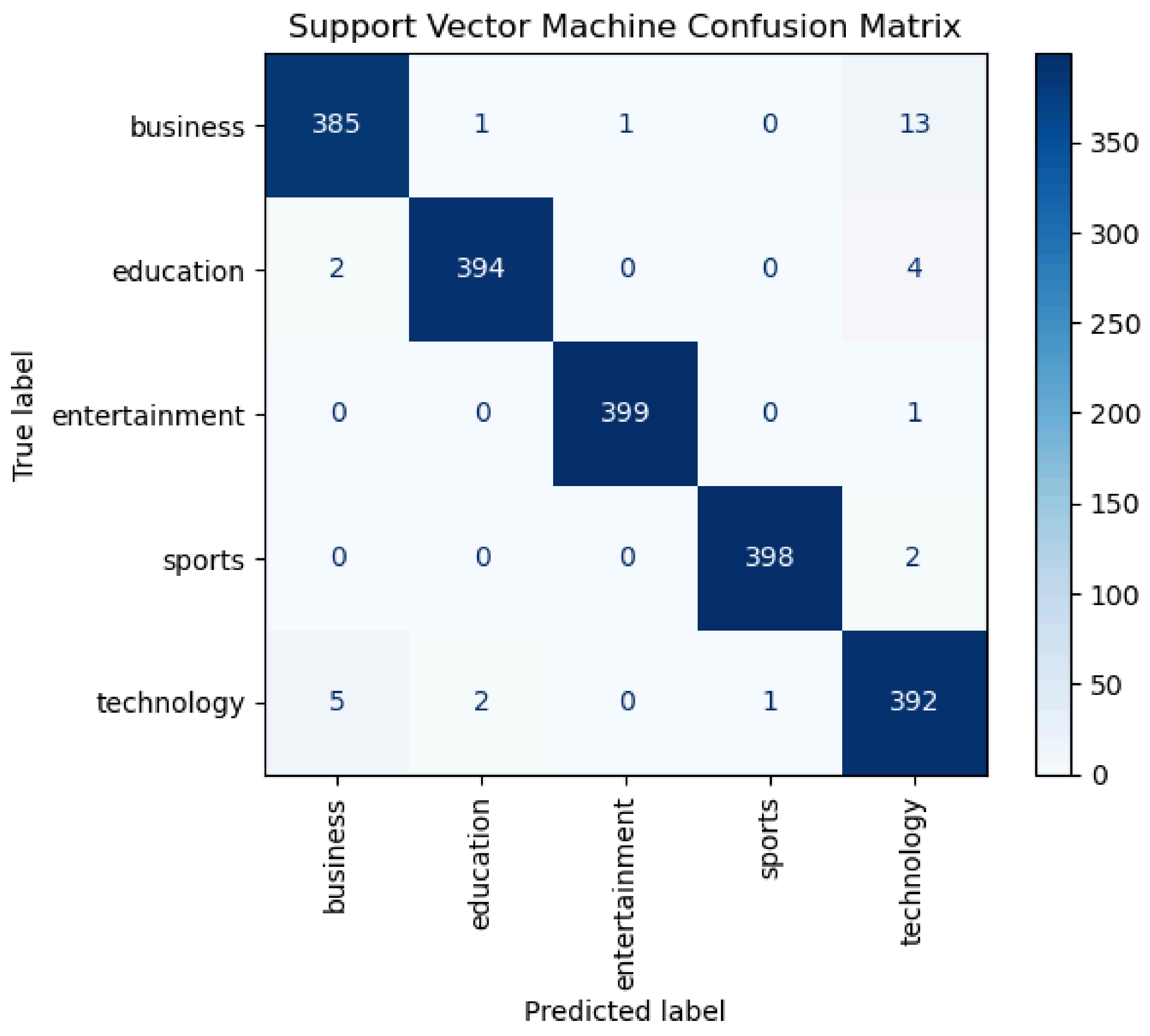
Best 3 models


Naive Bayes Confusion Matrix

True label \ Predicted label	business	education	entertainment	sports	technology
business	387	0	1	0	12
education	2	394	0	0	4
entertainment	0	0	399	0	1
sports	0	1	2	395	2
technology	9	1	0	2	388


Neural Networks Confusion Matrix

True label \ Predicted label	business	education	entertainment	sports	technology
business	391	0	1	0	8
education	3	394	0	0	3
entertainment	0	0	399	0	1
sports	0	1	1	397	1
technology	4	1	0	0	395


Support Vector Machine Confusion Matrix

True label \ Predicted label	business	education	entertainment	sports	technology
business	385	1	1	0	13
education	2	394	0	0	4
entertainment	0	0	399	0	1
sports	0	0	0	398	2
technology	5	2	0	1	392

- F1 SCORE: 0.9800
- LOG LOSS: 0.0668

- F1 SCORE: 0.9875
- LOG LOSS: 0.0512

- F1 SCORE: 0.9840
- LOG LOSS: 0.0556

APP DEMONSTRATION

Streamlit

- News article predictor app

CONCLUSION

- Predictive model to identify the category of a given news article.
- 7 models were developed.
- 3 models outperformed other models.
- Predict the category with a minimum accuracy of 98%.
- Likely to misclassify business articles as technology.
- Models were deployed on Streamlit app.

RECOMMENDATIONS

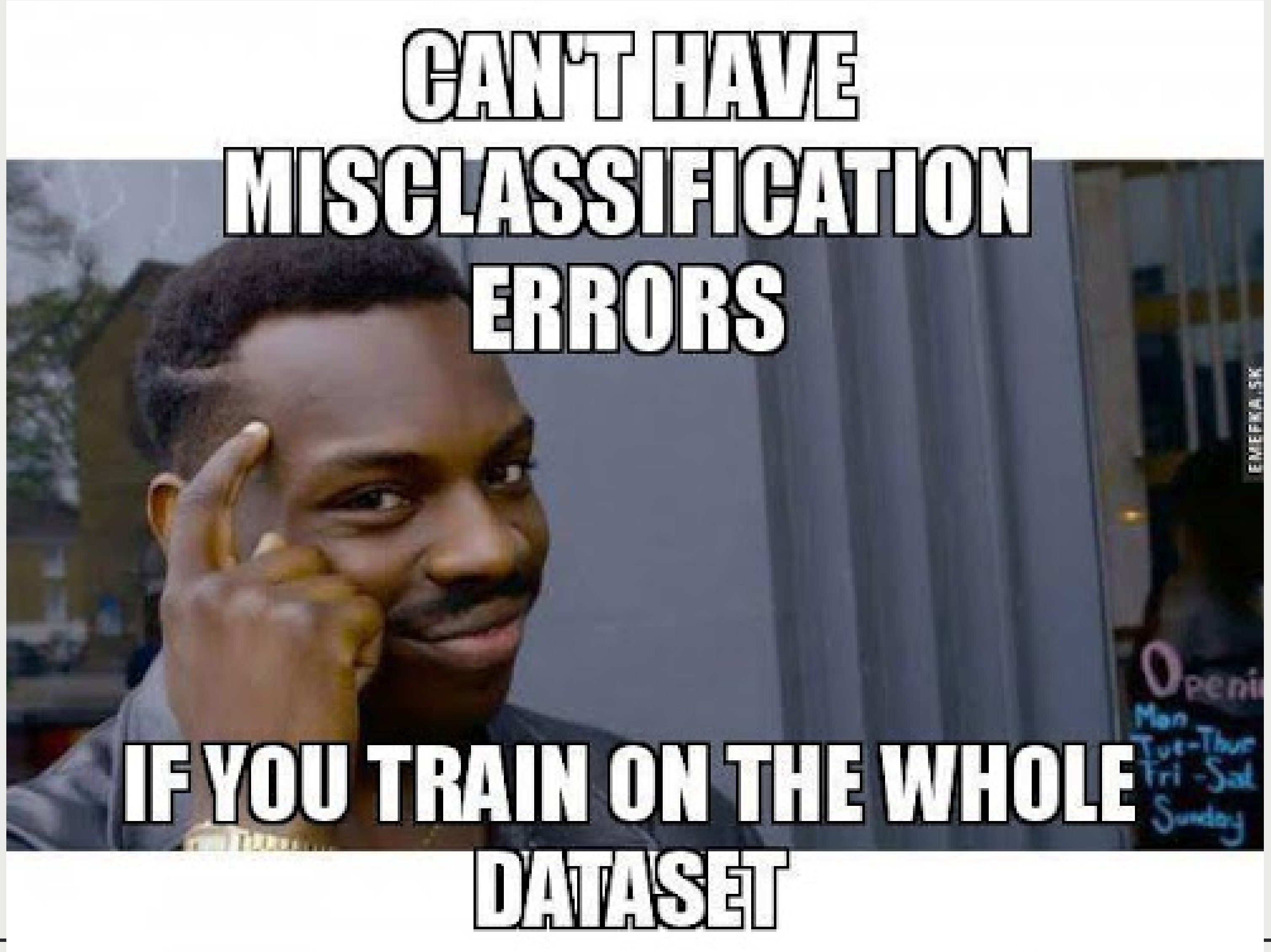
- Investigate model performance based on externally sourced data.
- Update the model to better predict articles where the url content does not contain the target category.
- Update the model to cast predictions with missing information.
- Update the model to identify additional categories.

RECOMMENDED

THURSDAY 06/02/2025

PRESENTED BY TEAM 3

THANK YOU


CAN'T HAVE MISCLASSIFICATION ERRORS
IF YOU TRAIN ON THE WHOLE DATASET
EMEFKA.SK
Openi
Mon
Tue-Thur
Fri -Sat
Sunday

THE Q&A SESSION BEGINS