NATURAL LANGUAGE PROCESSING

DECEMBER 2025

NEWS CLASSIFICATION

Adiluly Sayat

Dataset

FEATURES:

CLASS INDEX	TITLE	DESCRIPTION
Numerical label of the news article. 1-world 2-sport 3-business 4-Sci/Tech	The title of the news article.	The main content or summary text of the news article.

Data Preprocessing

LOWERCASE & CLEAN

Convert text to lowercase and remove URLs, HTML, numbers, and punctuation.

TOKENIZE

Split the text into individual words (tokens).

REMOVE STOP WORDS

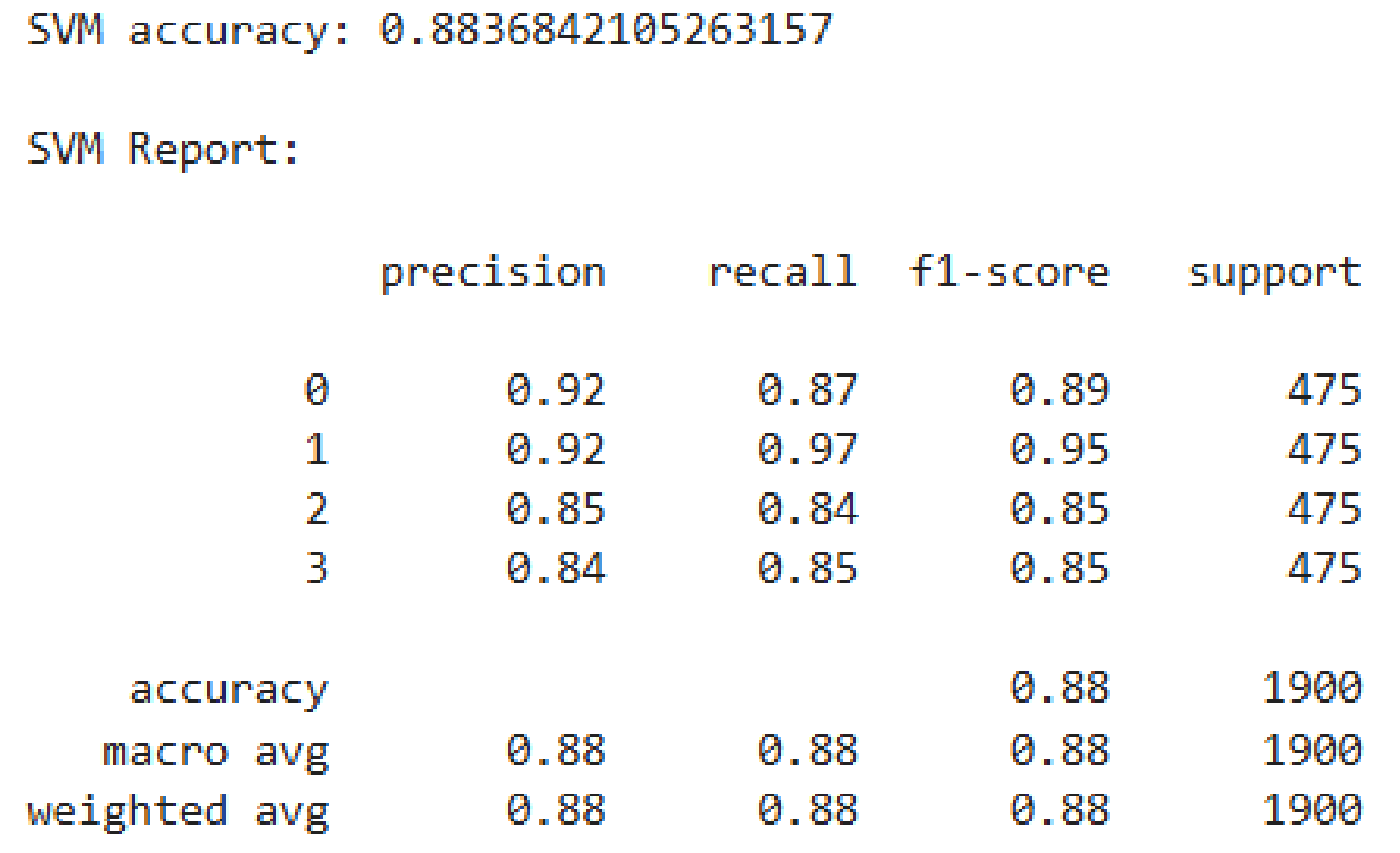
Filter out common but meaningless words.

Support Vector Machine

Trains a Classifier (SVM): The model learns patterns between preprocessed texts (TF-IDF vectors) and their correct categories (labels).

Tests the Model: The trained model makes predictions on new, unseen data (the test set).

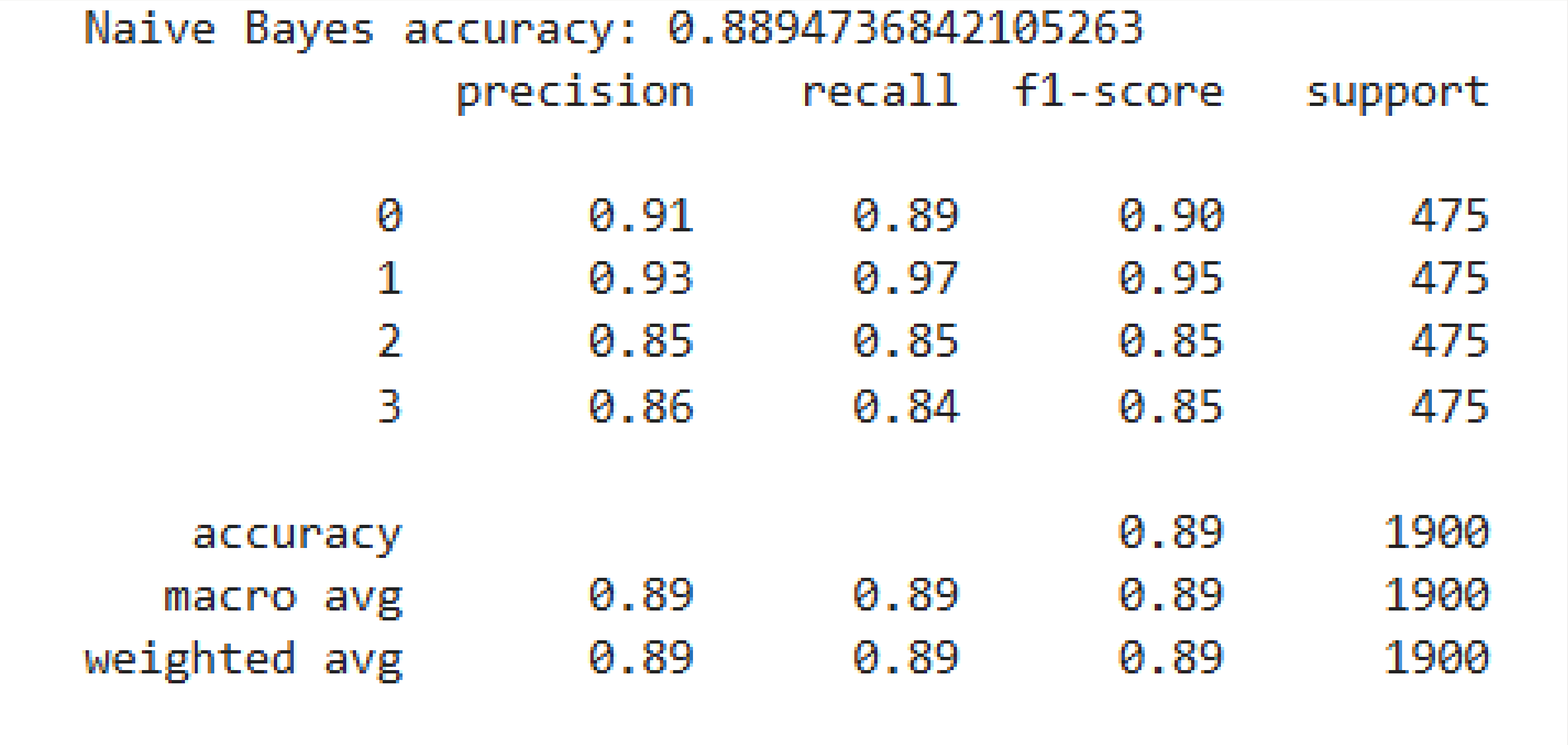
Outputs Metrics: Automatically calculates key performance indicators (accuracy, precision, recall, F1-score) to evaluate how well the model performs.



```
SVM accuracy: 0.8836842105263157

SVM Report:

              precision    recall  f1-score   support

           0       0.92      0.87      0.89       475
           1       0.92      0.97      0.95       475
           2       0.85      0.84      0.85       475
           3       0.84      0.85      0.85       475

    accuracy                           0.88      1900
   macro avg       0.88      0.88      0.88      1900
weighted avg       0.88      0.88      0.88      1900
```

MultinomialNB

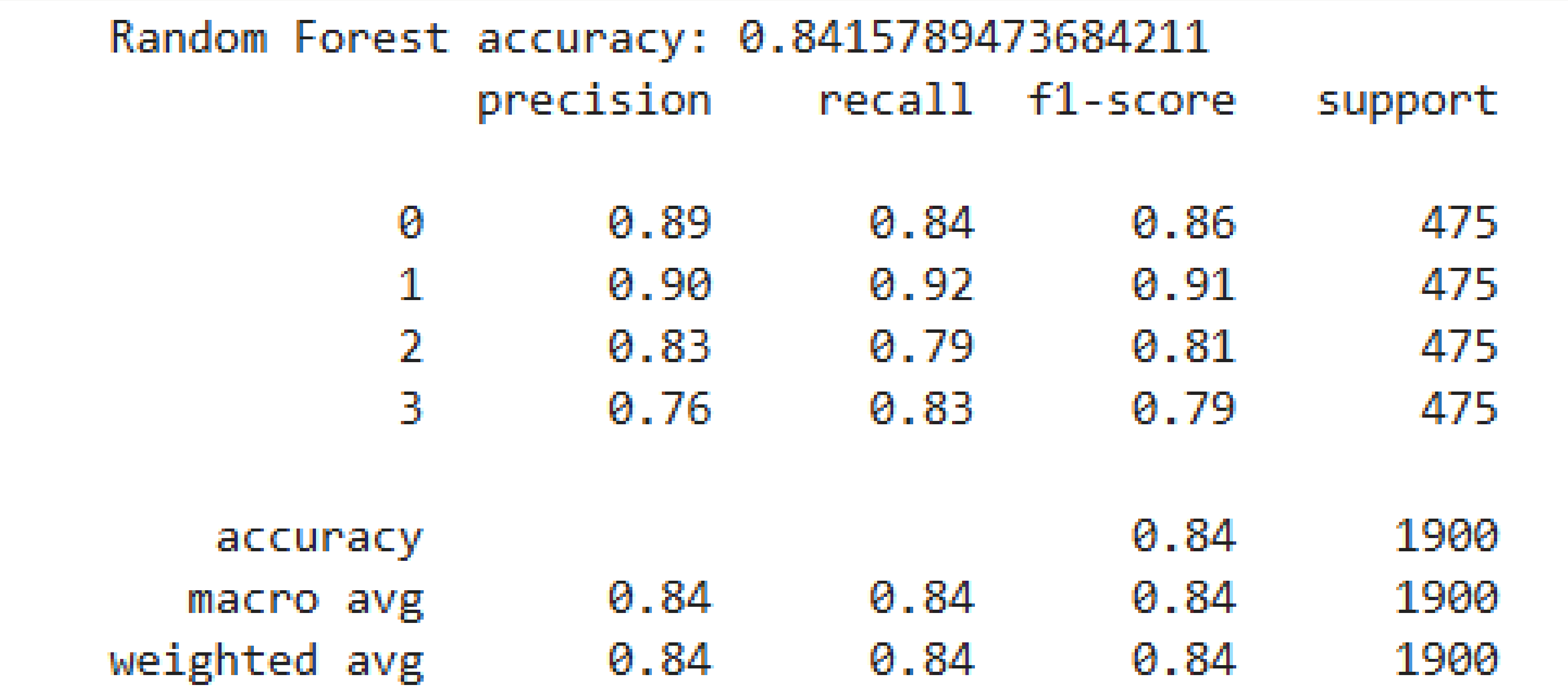
Trains a Naive Bayes classifier : This model is based on Bayes' theorem and is particularly effective for text classification. It learns from the prepared TF-IDF vectors and their labels.
Tests and Evaluates: The model makes predictions on the test data, and then key performance metrics are automatically printed.



```
Naive Bayes accuracy: 0.8894736842105263
              precision    recall  f1-score   support

           0       0.91      0.89      0.90       475
           1       0.93      0.97      0.95       475
           2       0.85      0.85      0.85       475
           3       0.86      0.84      0.85       475

    accuracy                           0.89      1900
   macro avg       0.89      0.89      0.89      1900
weighted avg       0.89      0.89      0.89      1900
```

Random Forest

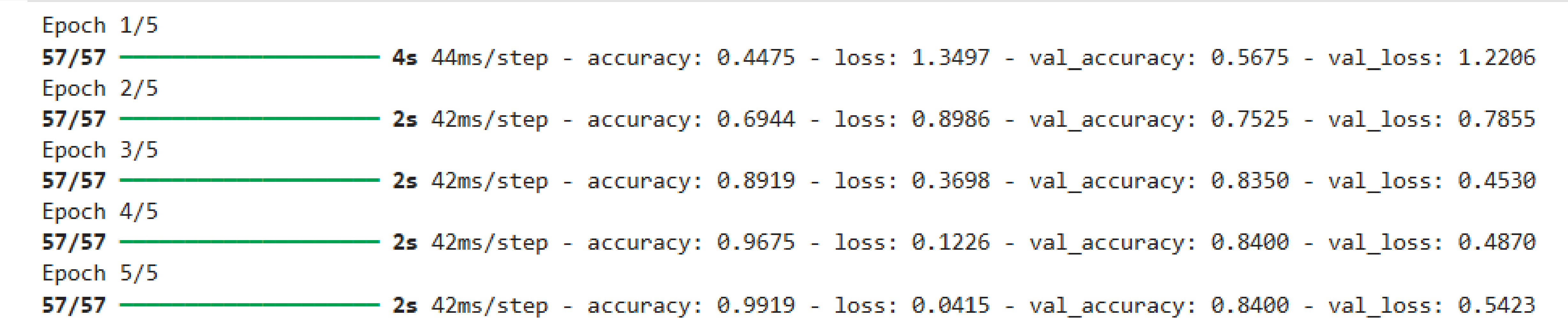
Trains a Random Forest model: Creates a group of 200 decision trees that learn from our data and vote together for the final prediction.
Tests and shows results: The model predicts categories for new texts, and the code automatically prints the percentage of correct answers (accuracy) and detailed statistics for each class.



```
Random Forest accuracy: 0.8415789473684211
              precision    recall  f1-score   support

           0       0.89      0.84      0.86       475
           1       0.90      0.92      0.91       475
           2       0.83      0.79      0.81       475
           3       0.76      0.83      0.79       475

    accuracy                           0.84      1900
   macro avg       0.84      0.84      0.84      1900
weighted avg       0.84      0.84      0.84      1900
```

GRU neural network

Prepares the text: Converts words into numbers (tokens) and pads all texts to the same length so the neural network can process them.
Creates and trains a neural network (GRU): Builds a model with an embedding layer, a GRU layer for sequence analysis, and trains it for 5 epochs, using a portion of the data to validate the quality.



```
Epoch 1/5
57/57 ━━━━━━━━━━━━━━━━━━━━ 4s 44ms/step - accuracy: 0.4475 - loss: 1.3497 - val_accuracy: 0.5675 - val_loss: 1.2206
Epoch 2/5
57/57 ━━━━━━━━━━━━━━━━━━━━ 2s 42ms/step - accuracy: 0.6944 - loss: 0.8986 - val_accuracy: 0.7525 - val_loss: 0.7855
Epoch 3/5
57/57 ━━━━━━━━━━━━━━━━━━━━ 2s 42ms/step - accuracy: 0.8919 - loss: 0.3698 - val_accuracy: 0.8350 - val_loss: 0.4530
Epoch 4/5
57/57 ━━━━━━━━━━━━━━━━━━━━ 2s 42ms/step - accuracy: 0.9675 - loss: 0.1226 - val_accuracy: 0.8400 - val_loss: 0.4870
Epoch 5/5
57/57 ━━━━━━━━━━━━━━━━━━━━ 2s 42ms/step - accuracy: 0.9919 - loss: 0.0415 - val_accuracy: 0.8400 - val_loss: 0.5423
```

References:

https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html#sklearn.svm.LinearSVC

https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html#sklearn.naive_bayes.MultinomialNB

https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html#sklearn.ensemble.RandomForestClassifier

https://www.tensorflow.org/api_docs/python/tf/keras/preprocessing/text/text_to_word_sequence

https://www.tensorflow.org/api_docs/python/tf/keras/layers

https://docs.python.org/3/library/re.html#text-munging

NLP

DECEMBER 2025

THANK YOU

220101169@stu.sdu.edu.kz