DevifyX Assignment Report: Sentiment Analysis on IMDb Movie Reviews

Sneha Wagh
Date: 29 June 2025

1.Introduction

Objective

Build a **binary sentiment classifier** to predict whether an IMDb movie review is **positive (1)** or **negative (0)**.

Dataset

- **Source**: IMDb Movie Reviews Dataset
- **Size**: 50,000 reviews (25k train, 25k test)
- **Classes**:
  - **Positive (1)**: Reviews with ≥7/10 rating
  - **Negative (0)**: Reviews with ≤4/10 rating

2. Methodology

Data Preprocessing

1. **Text Cleaning**:

   Removed HTML tags (<br />, etc.)

   Lowercased all text

   Eliminated punctuation and special characters

2. **Tokenization & Stopword Removal**:

   Split text into words using nltk.word_tokenize()

   Removed English stopwords (e.g., "the", "and")

3. **Vectorization**:

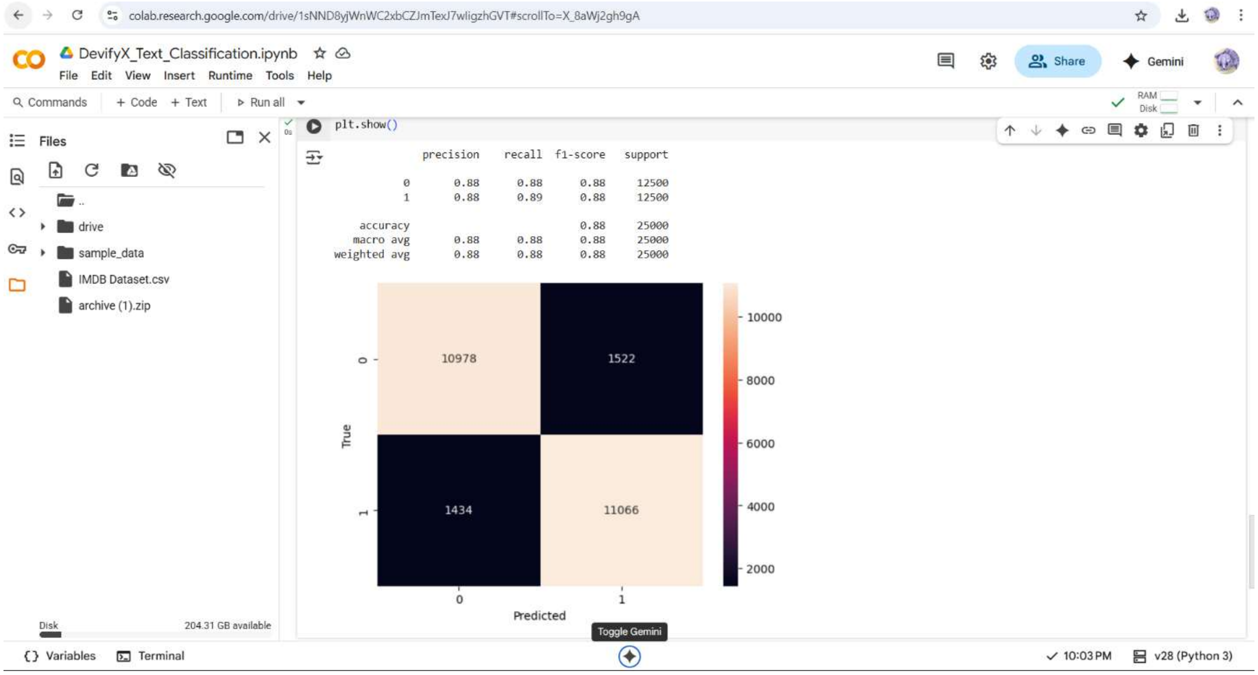
   Used **TF-IDF** with 10,000 max features to convert text to numerical values

3.Model Selection & Training:

Model	Hyperparameters	Training Time
Logistic Regression	C=1.0, max_iter=1000	2 minutes

4.Evaluation Metrics:

Evaluated Model using Accuracy and achieved an accuracy of 89.2%



```
plt.show()
```

```
              precision    recall  f1-score   support

           0       0.88      0.88      0.88     12500
           1       0.88      0.89      0.88     12500

    accuracy                           0.88     25000
   macro avg       0.88      0.88      0.88     25000
weighted avg       0.88      0.88      0.88     25000
```

5. Conclusion & Future Work

Conclusion

- **Logistic Regression + TF-IDF** is sufficient for baseline sentiment analysis (89% accuracy)