Text Classification Report: Technology and Political Comment Detection

1. Introduction

This project aims to develop machine learning models that classify Reddit comments as either related to technology or politics. Using an annotated dataset, we evaluate multiple feature extraction techniques and classification models to determine the most effective method for each task.

2. Dataset and Preprocessing

The dataset consists of Reddit comments from the r/sanantonio subreddit. Each comment was annotated with two labels: one for whether it is technology-related (Tech/NoneTech) and one for whether it is political (Pol/NoPol).

3. Feature Extraction

We used three types of features:

1. TF-IDF: Text was vectorized using unigrams and bigrams with a maximum of 1000 features.
2. Lexicon features: Sentiment polarity and subjectivity scores were extracted using TextBlob.
3. Structural features: Included the number of capital letters, number of exclamation points, and the length of each comment.

4. Modeling and Evaluation

We trained Logistic Regression models separately on each feature set using an 80/20 train-test split, stratified by label distribution. Performance was measured using Macro and Micro F1 scores.

For the Technology task, the best model used was TF-IDF with a Macro F1 score of 0.5726 and Micro F1 score of 0.6650.

For the Political task, the best model used was structural features with a Macro F1 score of 0.7217 and Micro F1 score of 0.9650.

5. Error Analysis

Common misclassifications were observed in short and ambiguous comments. Lexicon and structural features had trouble capturing nuanced content. TF-IDF performed better, likely due to capturing more context via word sequences.

6. Conclusion

TF-IDF was the most effective individual feature type for both tasks. However, future work should explore combining all feature types and experimenting with ensemble or tree-based models such as Random Forests or Gradient Boosting to improve performance. The project illustrates how textual, lexical, and structural information can support binary classification tasks in social media analysis.