Emotion Prediction Model Report

Objective:
The goal of this project was to develop a machine learning model to predict emotional reactions to New York Times articles based on their headlines and summaries. The model aims to enhance content personalization and improve user engagement by accurately classifying articles into multiple emotion categories.

Approach:

1. **Data Processing:**
    - Cleaned and tokenized text data from headlines and summaries.
    - Addressed missing values and removed columns that were not useful to analysis.
2. **Feature Engineering:**
    - Extracted additional features such as sentiment scores, text similarity metrics, and relevant bigrams.
    - Included terms that were most unique to particular emotions.
3. **Model Selection:**
    - Evaluated multiple models utilizing Random Forest, and XGBoost algorithms.
    - XGBoost was chosen based on its slight performance gain over the best random forest model.

Results:

- Micro-average ROC-AUC: **0.7765**
- Macro-average ROC-AUC: **0.6889**
- ROC-AUC scores for individual emotions ranged from **0.599 to 0.792**, indicating variability in model performance across different emotions.

Challenges:

- Significant class imbalance leading to lower recall for minority emotions.
- Overlap in textual content for different emotions causing misclassification.
- Model bias towards frequently occurring emotions.

Next Steps:

1. Explore transformer-based embeddings (e.g., BERT) for improved contextual understanding.
2. Further optimize hyperparameters and decision thresholds to balance precision and recall.
3. Investigate ensemble approaches combining deep learning with traditional models.
4. Utilizing a library like Hugging face to gain acces to pretrained embeddings.
5. Remove customs stopwords like Trump to see if there were would be a gain.
6. More N-gram refinement

Conclusion:
The current model provides a strong foundation for emotion prediction, offering valuable insights for content recommendation. However, further improvements in feature representation and model optimization are required to achieve higher accuracy and generalization across all emotion categories.