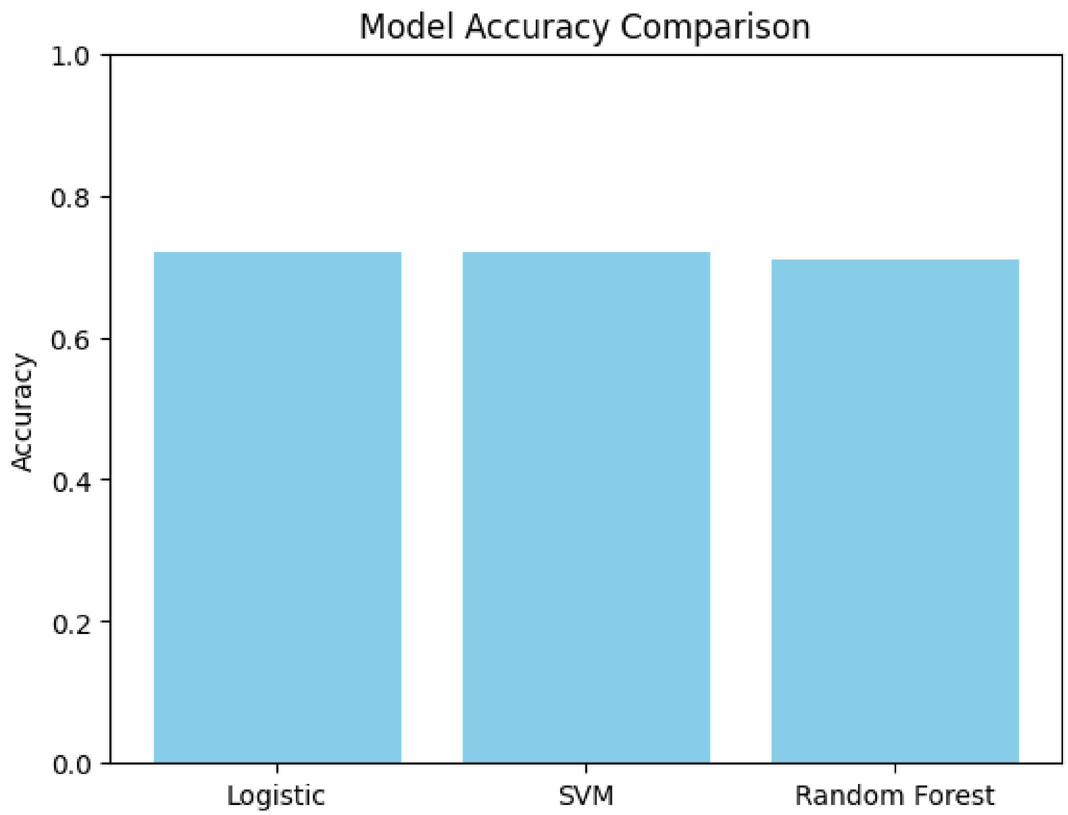
1. Recap of Current Model Performance

Three models were evaluated using accuracy, F1 score, and confusion matrices. Below are the key metrics:

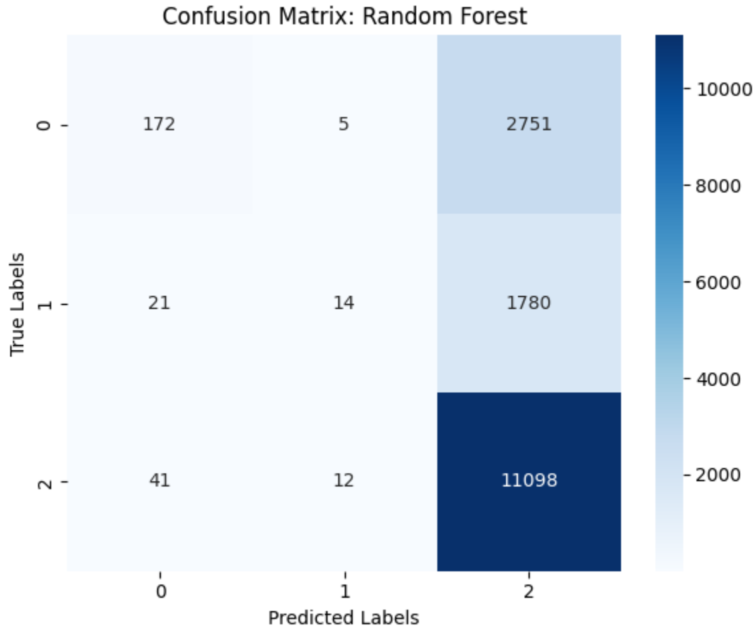
Model	Accuracy	F1 Score
Logistic Regression	0.7202	0.6451
SVM	0.7196	0.6343
Random Forest	0.7099	0.6033

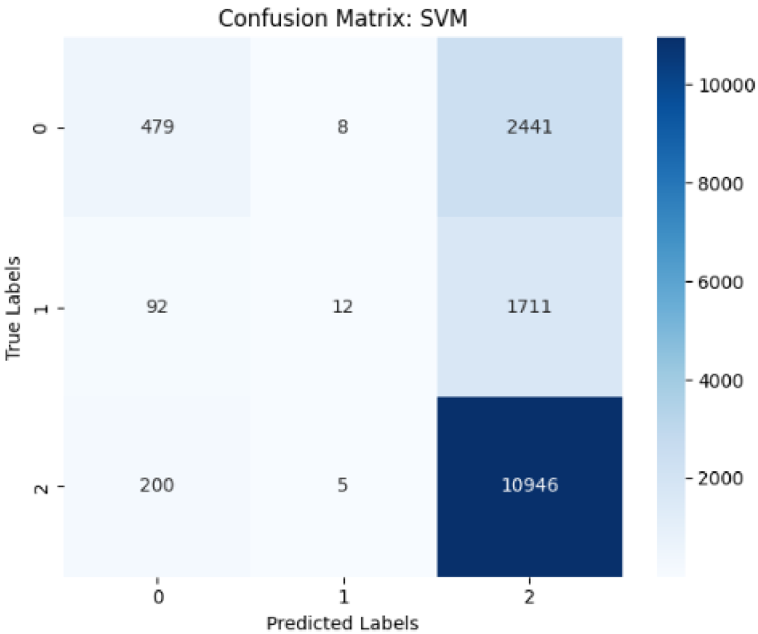


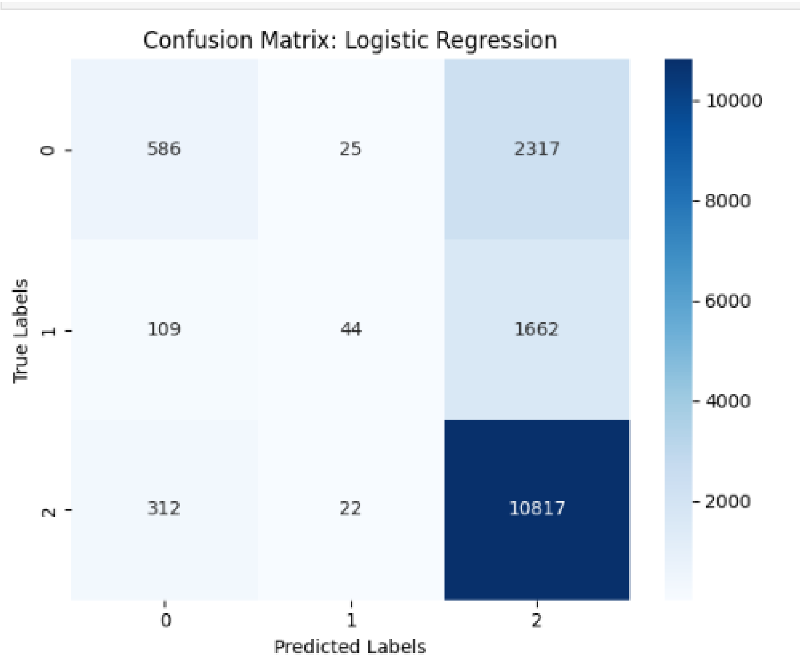
Observations:

- Logistic Regression performed best overall, both in terms of accuracy and F1 score.
- SVM showed competitive performance but had slightly lower F1 due to class imbalance.
- Random Forest demonstrated noticeable bias toward the majority class, especially struggling with under-represented classes.

Performance graphs and confusion matrices visually confirmed these conclusions, especially highlighting poor classification for class 0 (negative reviews).


Confusion Matrix: Random Forest
True Labels
Predicted Labels
0
1
2
172
5
2751
21
14
1780
41
12
11098
10000
8000
6000
4000
2000


Confusion Matrix: SVM
True Labels
Predicted Labels
0
1
2
479
8
2441
92
12
1711
200
5
10946
10000
8000
6000
4000
2000



2. Reflections on Issues Encountered

a. Data Imbalance

- The dataset contains significantly more positive reviews compared to neutral and negative ones.
- This led to models, especially Random Forest, favoring the majority class.
- Result: Poor recall for negative and neutral classes.

Action: Use SMOTE or other oversampling techniques to address this imbalance in future iterations.

b. Training Time and Computational Constraints

- SVM and Random Forest required longer training times, particularly on larger feature sets.
- Logistic Regression was more efficient and scalable.

Action: Consider dimensionality reduction or feature selection to reduce computational burden.

c. Model Interpretability

- Logistic Regression provides clear insight into feature contributions, making it a valuable baseline for explainability.
- Random Forest offers feature importance metrics, which can be explored in further depth.
- SVM lacks interpretability, which is a limitation when justifying model decisions.

3. Key Learnings

- Addressing class imbalance is critical to achieving balanced performance across all sentiment classes.
- Accuracy alone is insufficient for imbalanced data; F1 score and per-class metrics offer more reliable insight.
- Simple models like Logistic Regression can outperform more complex ones when well-optimized.
- Interpretability is a vital aspect, particularly when the model is intended for stakeholders or end-users.

4. Areas for Improvement

- Implement SMOTE or similar sampling strategies to balance the dataset.
- Apply dimensionality reduction (e.g., PCA) or select top features using statistical methods to reduce model complexity.
- Experiment with class weighting in models to counteract label imbalance.
- Explore the use of n-grams in TF-IDF to better capture contextual information.

5. Week 6 Roadmap

a. Introduction of Deep Learning Models

- Develop and test a Long Short-Term Memory (LSTM) model for sequence modeling of text.
- Incorporate BERT for contextual embeddings and sentence-level understanding.

b. Hyperparameter Tuning

- Use GridSearchCV or RandomizedSearchCV to fine-tune parameters for all classical models:
    - Logistic Regression: Regularization strength (C)
    - SVM: Kernel, C, gamma
    - Random Forest: Number of trees, max depth, min samples split
- Explore class_weight adjustments to handle imbalanced classes.

c. Data Enhancements

- Integrate additional metadata such as:
    - Book genres

    - Reviewer profiles
    - Star ratings
- These could provide richer context and improve classification performance.

d. Advanced Evaluation Metrics

- Plot ROC and Precision-Recall curves for each class.
- Use macro and weighted averages for metrics to capture model performance more comprehensively.