CROSS TEST REPORT

SUMMARY

In this project, two models were trained and evaluated. Model A achieved an accuracy of 46% whereas Model B achieved 47% accuracy. Although both models showed relatively low performance, Model B performed better than Model A.

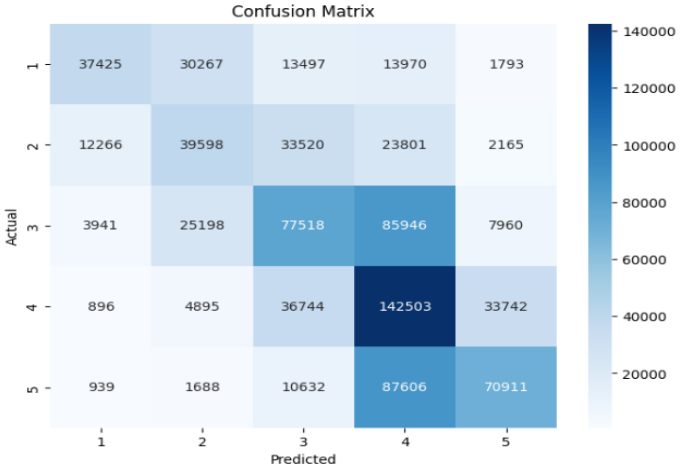
Model A tested using balanced dataset:

```
Accuracy: 0.4602768753885625

Classification Report:
              precision    recall  f1-score   support

           1       0.67      0.39      0.49     96952
           2       0.39      0.36      0.37    111350
           3       0.45      0.39      0.42    200563
           4       0.40      0.65      0.50    218780
           5       0.61      0.41      0.49    171776

    accuracy                           0.46    799421
   macro avg       0.51      0.44      0.45    799421
weighted avg       0.49      0.46      0.46    799421
```



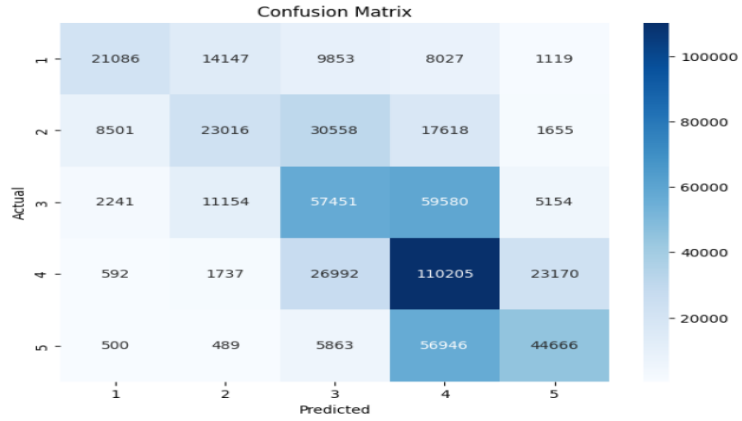
Model B tested using the imbalanced dataset:

```
Accuracy: 0.47282785071544475

Classification Report:
              precision    recall  f1-score   support

           1       0.64      0.39      0.48     54232
           2       0.46      0.28      0.35     81348
           3       0.44      0.42      0.43    135580
           4       0.44      0.68      0.53    162696
           5       0.59      0.41      0.48    108464

    accuracy                           0.47    542320
   macro avg       0.51      0.44      0.46    542320
weighted avg       0.49      0.47      0.46    542320
```



OBSERVATIONS.

- When trained on om=ne dataset and tested on a different dataset, the performance of both models dropped sharply.
- The drop indicates poor generalization across domains.

- Model B, even though it scored 52% before cross test, still struggled heavily on unfamiliar data.
- The model accuracy dropped because each dataset uses different vocabulary, style, and context, so the model learned patterns specific to the training domain.
- This shows the need for combined or merged datasets for better cross domain performance.

RECOMMENDATION: choosing the model

Based on initial evaluation, Model A should be deployed because it has better accuracy (44%) compared to Model B (44%) on its own test set. Even though both models struggle in cross testing. Between the two, Model B is recommended for deployment because it performed better overall with a higher accuracy (47%) and slightly better consistency in predictions.

Although the performance is still not ideal, Model A is the best choice among the options tested and can serve as a baseline model. It can be deployed with the intention of further improvement through hyperparameter tuning and balanced training data in future work.