Documentation

Data Training:

- **Model Training:** The ***LSTM*** models were used to train using the data. The training process involved feeding the model with ***input sequences*** and their labels.
- **SMOTE:** ***SMOTE*** was further applied to the second training data to address ***class imbalance***.

Findings (Metrics):

- **Classification report:** ***Precision***, ***recall***, ***F1-score***, and ***support*** for each class were calculated to evaluate the model's performance.
- **Recall:** Reported for each class in the ***classification report***.
- **Class with least recall:** The class with the ***lowest recall*** was identified to pinpoint areas for improvement.
- **Confusion matrix:** A visualization of the model's predictions versus actual labels, showing the number of ***true positives***, ***true negatives***, ***false positives***, and ***false negatives*** for each class.
- **Further analysis:** ***Misclassified examples*** were examined for better understanding. ***Ratio of data points to total*** was also considered.

Data Cleaning:

- **Prediction on 'Others' class:** Titles initially classified as "***Others***" were further analyzed by model with ***confidence score***.
- **Confidence Threshold:** A confidence ***threshold*** was applied. Predictions above ***0.9*** were accepted, below **0.6** were classified as "***Others***," and those between ***0.6*** and ***0.9*** were marked for ***manual review***.

Accuracy, Recall:

- **Accuracy:** Reported on the test set. Model 1: ***97.5%***, Model 2: ***93.2%***
- **Total recall (macro average recall):** Calculated to provide a single score across all classes. Model 1: ***93%***, Model 2: ***77%***

Suggestions (mainly for the 2nd Model):

- **Hyperparameter Tuning:** Experimenting further with the ***number of epochs***, ***batch size***, and ***learning rate***. ***Early stopping*** can also be used to avoid *overfitting*.
- **Data Augmentation:** Exploring techniques for ***data augmentation*** of the ***underrepresented class*** (the one with the ***lowest recall***) to improve the model's performance on ***minority classes***.
- **More Advanced Models:** Experimenting with more ***advanced architectures*** like ***transformers*** (***BERT***) or other deep learning models.
- **Ensemble Methods:** Combining predictions from multiple models can help reduce ***bias*** and ***variance*** and may lead to better generalization.