Task 1 submission: Team Aivengers

Nohan, Deepak, and Shriya

Approaches:

1. We started by cleaning the data, applying the TFIDF vectorizer, and then using Logistic Regressor in a OnevsRest classifier (since its a multilabel classification)
2. Later, we tried out other ML models in sklearn and found that MultiLayerPerceptron performed better. We also tried a Voting Classifier with different models, but MLP was still found to be the best.
3. We tried finetuning Bert-based models pre-trained on movie datasets. But these did not perform as expected.
4. Later, the performance increased when we improved the data cleaning step of the MLP-based model by removing stopwords and doing stemming and lemmatization.
5. Finally, we concatenated final layer embeddings from a Bert-based model(nickmuchi/distilroberta-base-movie-genre-prediction) with the TFIDF vectors (with above data cleaning steps) and performed PCA to reduce dimensionality and used these features as input to MLP. This model gave a major boost in the f1 score,
6. The model that gave best public score is the tuned version of above model, where we tuned the probability threshold and the % variance captured in PCA. This model gave a public score of 0.17980.