Requirements Document - Project 13

Team: AegiFi

Akshit Sinha - 2021101109

Adyansh Kakran - 2021111020

Project Outline

This project aims to predict movie genres based on their plot summaries, by exploring various machine learning techniques. The primary methods employed include Naive Bayes, XGBoost with Word2Vec embeddings, and Gated Recurrent Units (GRU) neural networks. The project involves text classification and multi-label genre tagging.

We plan to extend this project by implementation the following ideas on top of the models suggested by the paper -

- Different ways to generate Word2Vec embeddings
- Generating synthetic data to normalise the skew in the original data
- Using Doc2Vec instead of Word2Vec for richer embeddings to feed to XGBoost and GRUs

Deliverables

In this section we outline the deliverables for each checkpoint.

Checkpoint 2

- Data preprocessing
- Implementation of Binary Naive Bayes
- Implementation of Multinomial Naive Bayes
- Implementation of XGBoost with Word2Vec

Checkpoint 3

- Implementation of GRU
- Generation of synthetic data for normalisation
- Generation of Word2Vec with different ways
- Generation of embeddings using Doc2Vec
- Analysis and comparison of all the models using evaluation parameters given below

Final Deliverables

Through this project, we aim to accomplish multi label classification of movie plots using different kinds of ML models, comparing and analysing the results and getting the best approach for this task. This paper implemented these models and compared their results but we plan to extend this approach using multiple methods, some from the future works section of the paper itself. This might provide a better approach to the task as well.

This task requires us to have a basic understanding of NLP as well as Machine Learning and will make us more well versed with both.

The final deliverables thus include all the models implemented in the paper, as well as some future works suggested by the paper.

Analysis

As this is a multilabel classification problem, we evaluate the models on the following evaluation metrics:

- Jaccard Index
- hamming Distance
- Accuracy
- Precision
- Recall
- F1-score
- Hit Ratio

In addition to this, we analyse whether the various improvements suggested by us improve the performance and by how much.

We also compare the given models based on power consumption and time taken during retraining as well as inferencing.