Project Log

Week 1

1. Getting the data from the FTP link - Adyansh
2. Understanding the general structure of the data - Akshit
3. Splitting the data and combining the data into one csv. - Akshit

Week 2

1. Implementation of binary naive bayes - Akshit
2. Implementation of multinomial naive bayes - Adyansh
3. Calculation of metrics for each - Akshit

Week 3

1. Generating Word 2 Vec Embeddings using google news corpus - Adyansh
2. Implementation of XGBoost algorithm - Akshit

Week 4

1. Implementation of Binary GRU - Adyansh
2. Implementation of Rank GRU - Akshit
3. Manual hyperparameter tuning of the parameters - Adyansh & Akshit

Week 5

1. Implementation of Multinomial GRU - Adyansh
2. Manual hyperparameter tuning of the parameters - Akshit
3. Giving structure to the code to be executed through a main file - Akshit

Week 6

1. Generation of TF-IDF weighted word2vec embeddings. - Adyansh
2. Generation of Doc2Vec Embeddings - Adyansh

3. Application of these embeddings into the different models. - Akshit

Week 7

1. Building streamlit frontend for training and inference on the IMDB data. - Akshit
2. Connecting to the backend for response collection and showing metric graphs. - Akshit
3. Implementing Basic Transformer class. - Adyansh

Week 8

1. Building streamlit inference - Akshit
2. Analysis and comparison of different ML Models, affect of different word embeddings on the models - Akshit & Adyansh

Overall Summary

Through this project, we accomplished multi label classification of movie plots using different kinds of ML models, compared and analysed the results and presented multiple approaches to this problem. Overall, we successfully and consistently recorded better results than the original paper.

We then further extended the paper by implementing features like TF-IDF weighted Word2Vec embeddings, and a Transformer to overcome the limitations stated by the paper. Along with these improvements, we plan to open source the code we have built, and we provide an interface built on Streamlit where anyone can use these models and can test on their own plot summaries, which we plan to publish through Streamlit.