NLP (COMM061)-PROJECT PLAN

Group 07: Gowrisaranyan Ganesan, Mouly Ravilla Hemachandran, Sharath Kumar Muthu Anand Kumar, Jisa Mithun, Rohini Raghukumar

Topic: Movie Genres based on plot summary – A Multi Label Classification

Problem Definition:
The multimedia streaming services such as Netflix, Amazon Prime Videos, etc. generate personalized feed for the members that help them to find content that they would enjoy. Movie genre classification is a critical component of such recommendation systems which use the reviews, movie posters, actors, genres to recommend the movie to the users. As part of this project, we will build a multi-label classification model to predict the movie genres based on the plot description.

Dataset:
CMU Movie Summary Corpus is a multi-label data set with 42,306 movie plot summaries extracted from Wikipedia + Freebase, including: Movie box office revenue, genre, release date, runtime, and language. There are 363 unique genres in the dataset. This dataset has been taken from http://www.cs.cmu.edu/~ark/personas/.

Planned Activities:
We plan to perform several experiments and machine learning techniques to classify the genres based on the movie plot information from the dataset. The general approach to execute the project is given below:

Tasks	Sub-Tasks	Owner	Estimated Duration
Data Understanding and Visualization		Team	1 week
Data Cleaning and Pre-processing	Tokenization	Team	2 weeks
	Normalization		
	Remove Stop Words		
	Lemmatizing or Stemming		
	Text Featurization/Transformation		
Data Modelling	Naive Bayes model	Jisa Mithun	1 week
	Support Vector Machine	Rohini	
	KNN	Sharath Kumar	
	LSTM	Gowrisaranyan	
	BERT	Mouly	
Evaluation	Evaluate model on testing data to find accuracy of the model.	Team	1 week
Deployment	Development and deployment of ML-models using Flask	Team	2 weeks

Table 1: Project Plan

We will experiment various techniques for Text Featurization and Transformation like Bag of words (BoW), n-grams, Term Frequency-Inverse Document Frequency (TF-IDF), POS Tagging, Parsing (Dependency & Constituency), Word Sense Disambiguation, Named Entity Recognition, Topic Modelling with team members working on at least 2 topics. We will perform Data modelling and hyper-parameter tuning for the models defined in Table 1: Project Plan

Development Environment & Code Repository:
We plan to use Jupyter or Google Colab as Python development environment for this project and work individually on the defined tasks and collaborate our work using Git and Github repository.

Python Libraries: numpy, re, pandas, NLTK, genism, spacy, seaborn, Matplotlib, sklearn, pickle, pytorch, Flask, umap-learn.

Note: Installation of any additional libraries will be explored during the experiments

Deployment
Towards the end, we will build a CI/CD workflow for an end-to-end ML pipeline to automate the execution of entire process from loading the dataset all through the production.