**Project Proposal**

**<Candidate Name>: <Mobile Number>**

**University of Memphis**

**Purnakumar Thota**

**Title: Movie parental guide predictive profiling from raw movie script using AI for automated viewability classification**

**Introduction:** Entertainment industry is one of the fastest growing industries in the current age of globalization where content is no longer restricted to regional or language boundaries. Though there are many positive sides to the growth there few negative implications as well of which one of the major concerns is the influence of Movies on audience, especially children. Improper regulation/monitor/classification of content can have very critical impact on society, children in specific. Government regulatory bodies like *Classification & Ratings Administration (CARA)*[1] by *Motion Picture Association of America (MPAA)*[2], in USA, regulates the content by provided appropriate classification of Parental Guide for Movies aired in USA. The rating provided by these agencies are generally based on potential violence, drug usage, and nudity etc., where a committee of 2-3 people manually watch a movie and provide ratings based on their observation. Though this approach of manual rating a movie may seem accurate it severely suffers from standardization as it can sometime be opinion driven (those there are certain guidelines for rating generation). In order to address this problem of slow, manual and subjective approach we are proposing a AI driven solution where a deep learning model can be trained to predict intensity of critical factors like violence, foul language, drug/alcohol usage etc., from movie script, which can later be used to provide overall rating for movie view-ability for children. In the current solution proposed we will be rating a movie on the following dimensions sex, violence, profanity, drugs and intense

**Proposed Solution Approach:** Data procurement for this project is a bit challenging as movie script information in not readily available. Movie scripts for a movie will be scrapped from Web and scrapped movie script is used for analysis. Post movie script procurement, script is tagged to its meta data (download from Kaggle/IMDB) along with parental rating. Post creation of analysis ready dataset with movie script and metadata following step will be followed for model development.

1. Data Pre-processing: This includes removing unwanted words in scripts like speaker names, removing stopwords, text normalization, performing word stem, lemmetization etc., Post data pre-processing cleaned text without unwanted words will be generated
2. Word Vectorization: Continuous bag of words and TF-IDF approach will be test for converting text to word vectors for model development
3. Model Training: Since we are dealing with text data with a good sample of movies for model training deep learning models works well under these setting hence we will be using RNN model for predicting ratings. Here single RNN mode will be build for each of sex, violence, profanity, drugs and intense dimensions
4. Model Evaluation: Since the problem we are trying to solve is a multi-class classification problem Accuracy will be used as primary measure for model evaluation. Along with accuracy other metrics like confusion matrix etc., will also be considered for model selection.
5. Model Save: Final model performing well on test, train and validation dataset will be save for future reference (for serving as input for frontend application)
6. Front-end Application: A streamlit application will be developed which takes script as user input in realtime and give rating prediction

**Data Source:**

Source 1

Dataset: Movies meta data long with parental ratings

Source: Kaggle/IMDB (https://www.kaggle.com/datasets/barryhaworth/imdb-parental-guide)

Method of Extraction: Direct download from Kaggle.com

Source 2

Dataset: Movie Scripts

Source: IMSDB (https://imsdb.com/)

Method of Extraction: Webscraping from imsdb.com

**Tools:**

* Programming Languages:
  1. Python: For programming entire modeling pipeline
  2. UI development: Streamlit, Javascript, HTML and CSS
* Libraries/Frameworks:
  1. Tensorflow & Keras: For building deep learning models
  2. Steamlit: For application development
* IDE:
  1. Model Development: Jupyter Notebook
  2. App Development: Streamlit

**References:**

1. <https://www.carafilmratings.com/>
2. <https://www.bbfc.co.uk/about-classification/classification-guidelines>