**Movie Parental Guide Predictive Profiling from Raw Movie Script using AI for Automated Viewability Classification**

Sujeeth

Department of Psychology, The George Washington University

PSYC 3170: Clinical Psychology

Dr. Tia M. Benedetto

April 19, 2023

Contents

[1. Project Background & Overview 2](#_Toc133163394)

[2. Problem Statement 3](#_Toc133163395)

[3. Research Questions being Answered 4](#_Toc133163396)

[4. Literature Review 4](#_Toc133163397)

[4. Machine Learning Techniques used for Current Work 4](#_Toc133163398)

[5. Methodology 5](#_Toc133163399)

[5.1 Overview 5](#_Toc133163400)

[5.2 Data Source 6](#_Toc133163401)

[Movies Script 6](#_Toc133163402)

[Movies Metadata 7](#_Toc133163403)

[5.3 Data Cleaning and Transformation 7](#_Toc133163404)

[Scraping Movie Script from Internet 7](#_Toc133163405)

[Data Cleaning & Transformation (Script Preprocessing) 9](#_Toc133163406)

[Feature Selection and Engineering 11](#_Toc133163407)

[7. Model Selection and Evaluation 11](#_Toc133163408)

[Violence Prediction – Parameter Tuning & Bias Analysis 11](#_Toc133163409)

[Profanity Prediction – Parameter Tuning & Bias Analysis 13](#_Toc133163410)

[Intense Prediction – Parameter Tuning & Bias Analysis 14](#_Toc133163411)

[Sex Prediction – Parameter Tuning & Bias Analysis 15](#_Toc133163412)

[8. Model Results and Discussion 16](#_Toc133163413)

[Descriptive Analysis 16](#_Toc133163414)

[Violence 16](#_Toc133163415)

[Profanity 17](#_Toc133163416)

[Intense 18](#_Toc133163417)

[Sex 18](#_Toc133163418)

[Model Evaluation 19](#_Toc133163419)

[Movie Violence – Model Evaluation Report: 19](#_Toc133163420)

[Movie Profanity – Model Evaluation Report: 20](#_Toc133163421)

[Movie Intense – Model Evaluation Report: 21](#_Toc133163422)

[Movie Sex – Model Evaluation Report: 21](#_Toc133163423)

[9. Discussion on Key Findings 22](#_Toc133163424)

[10. Conclusion 23](#_Toc133163425)

[11. Reference 25](#_Toc133163426)

# 1. Project Background & Overview

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 [1, 2], children in specific. Government regulatory bodies like Classification & Ratings Administration (CARA)3 by Motion Picture Association of America (MPAA)4, 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). Hence a solution for auto parental guide prediction is very critical.

# 2. Problem Statement

The entertainment industry is rapidly expanding due to globalization, with content crossing regional and language barriers. While there are many benefits to this growth, there are also negative implications, particularly the impact of movies on children. Improper regulation and monitoring of content can have serious consequences on society. The Classification & Ratings Administration (CARA) by Motion Picture Association of America (MPAA) in the USA provides appropriate classification of parental guidance for movies based on potential violence, drug usage, and nudity. However, the manual rating system suffers from a lack of standardization and can be opinion-driven. Therefore, an automated solution for parental guide prediction is critical.

# 3. Research Questions being Answered

In order to address this problem of slow, manual and subjective approach we are proposing an 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.

# 4. Literature Review

“Age Suitability Rating: Predicting the MPAA Rating Based on Movie Dialogues” by Mahsa Shafaei, Niloofar Safi Samghabadi5 and “Machine Learning Models for Content Classification in Film Censorship and Rating” by Syma Afsha6 are few of the paper discussing the use of Movie Subtitle/Dialogues/Scripts for classifying movie ratings. Afore mentioned papers describes the use of both Machine Learning and Deep Learning techniques one of the papers saw 81.6% but doesn’t give specific detail if the reported F1 score is for test of train dataset.

In our current approach we propose the use of Advanced deep learning models (RNN) for evaluating movie scripts on Violence, Profanity, Intense and Sex.

# 4. Machine Learning Techniques used for Current Work

In the current project we propose the use of RNNs for predicting movie parental guide rating. Recurrent Neural Networks (RNNs)[7, 8] are a type of artificial neural network used in sequential data processing, such as natural language processing and time series analysis. Unlike other neural networks, RNNs have a "memory" element that allows them to process sequences of inputs and outputs, taking into account the order and context of the information.

RNNs work by feeding the output of the previous step back into the model as input for the current step, allowing the network to remember information from previous steps. This feedback loop creates a "hidden state" that captures the network's understanding of the input sequence so far.

One of the main advantages of RNNs is their ability to process sequences of variable length, making them well-suited for tasks such as language modeling, speech recognition, and machine translation.

# 5. Methodology

## 5.1 Overview

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.

* 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
* Word Vectorization: Continuous bag of words and TF-IDF approach will be test for converting text to word vectors for model development
* 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 built for each of sex, violence, profanity, drugs and intense dimensions
* 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.
* 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)

## 5.2 Data Source

Data for current project has been procured from two different sources.

### Movies Script

Internet Movie Script Database (IMSDB)[9] website has large collection of movie scripts in HTML format. Which allows users to read and download scripts with no cost. Data scrapped contains movie scripts in the HTML format. Script can be linked to corresponding metadata using movie title (the id which is used to scrape scripts)

**Method Procurement:** Web Scrapping

**Source Link:** <https://imsdb.com/>

### Movies Metadata

Data contains selected movies from IMDB[10] along with parental guide information scrapped from web. Data captures other metadata of the movie like Title, Start Year, Run Time etc., Data has been extracted from Kaggle.com. Following is the source description as per Wiki “*Kaggle, a subsidiary of Google LLC, is an online community of data scientists and machine learning practitioners. Kaggle allows users to find and publish data sets, explore and build models*”

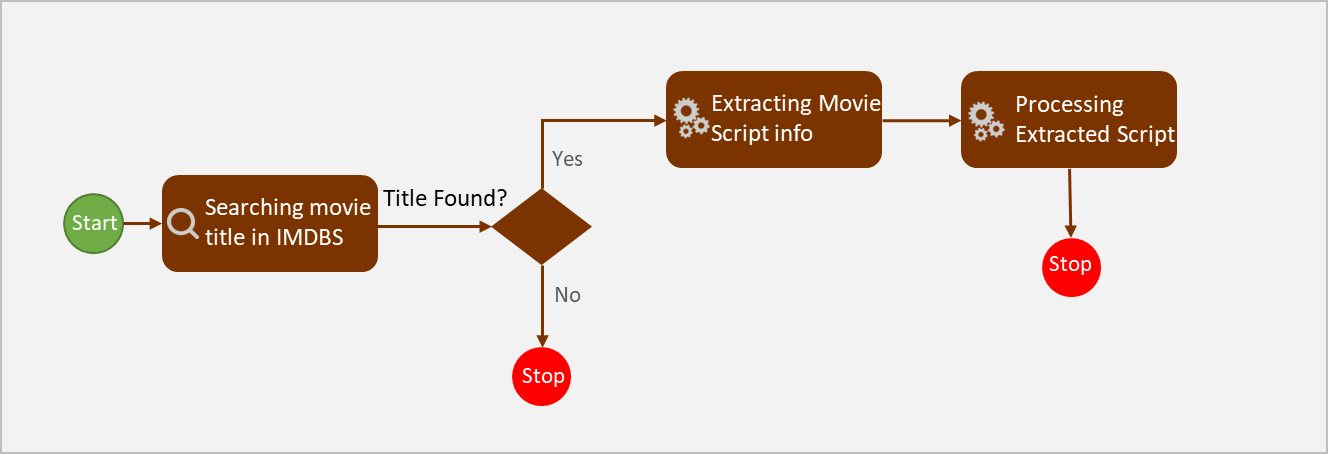
**Method Procurement:** Direct Download

**Source Link:** <https://www.kaggle.com/datasets/barryhaworth/imdb-parental-guide>

## 5.3 Data Cleaning and Transformation

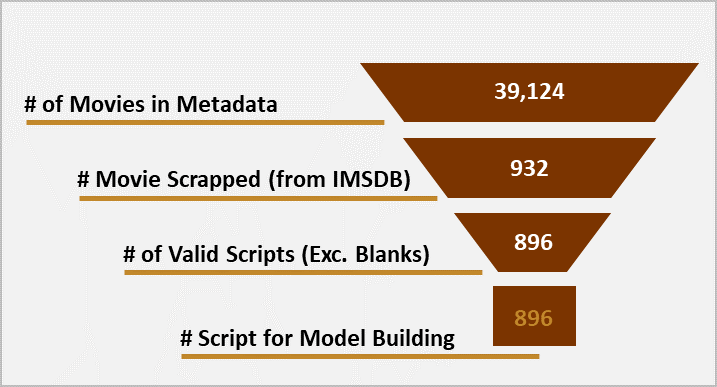
### Scraping Movie Script from Internet

* Movie tile name from metadata is searched for in IMSDB using beautiful-soup [11] package
* Availability of the movie in IMSDB can be take care using try except block
  + In case if movie is available script information is extract from the HTML tags [Refer to Figure 1 below]



***Figure 1:***Movie Script Extraction and Processing

* Extracted movie scripts are processed for removing unwanted HTML tags
  + 756 movies scripts have been finally scrapped from IMSDB. [Refer to Figure 2 below]



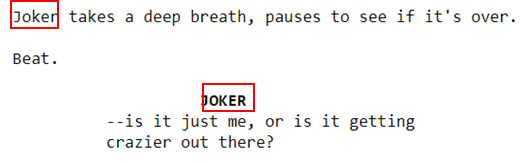
***Figure 2:*** *Funnel Chart Showing Extraction of Scripts from Metadata till Final Number of Script*

### Data Cleaning & Transformation (Script Preprocessing)

As majority of the data used for current project is textual information, it needs to be first converted into numerical format using text vectorization. Before vectorization data has to be treated. Following pre-processing steps have been performed on the data

**Text Normalization:**

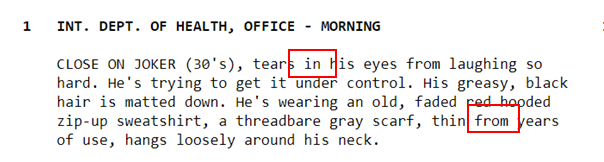
* Text normalization [12] is the process of converting text with mixed cases (lower/upper) cases to eighter lower or upper case [Refer to Figure 3 for example]
* Text normalization helps in make data richer by treating the words with cases as one. For example, the words “Script” and “script” have same meaning but are in different cases



***Figure 3:*** *Figure showing word Joker exiting in two different cases in Joke Movie script*

**Stop words and special character removal:**

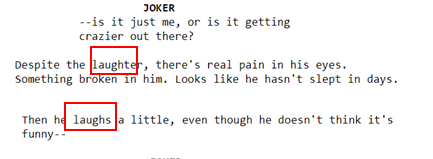
* Words like “I”, “am”, “is”, “are” etc., are generally useful for making sentences grammatically correct but doesn't carry any information hence these can be excluded before modelling [Refer to Figure 4 for example]



***Figure 4:*** *Figure showing presence of stop-words in Joker movie script*

**Root word extraction:**

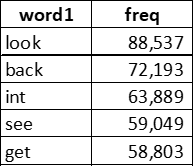
* Words like “script” and “scripted” convey same meaning with base word being “work”. Hence word like “scripted” can be converted to root word for building better generalized models [Refer to Figure 5 for example]
* Root word extraction technique **lemmatization [13]** has been used in the current work



***Figure 5:*** *Figure showing word laugh existing in two different forms*

**High Frequency Word Exclusion:**

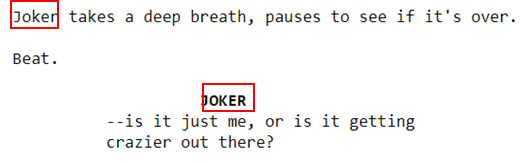
* Words like “look”, “day” etc., are not part of stopwords and they do not carry any major information which can help in movie rating profiling and hence can be excluded post manual inspection [Refer to Table 1 for example]



***Table 1:*** *Sample Frequent Words in the Script*

**Character Names and Script Related Tags Exclusion:**

* Words like “I”, “am”, “is”, “are” etc., are generally useful for making sentences grammatically correct but doesn't carry any information hence these can be excluded before modelling [Refer to Figure 6 for example]



***Figure 6:***Figure showing word “Joker” being repeated in Joker Movie

# Feature Selection and Engineering

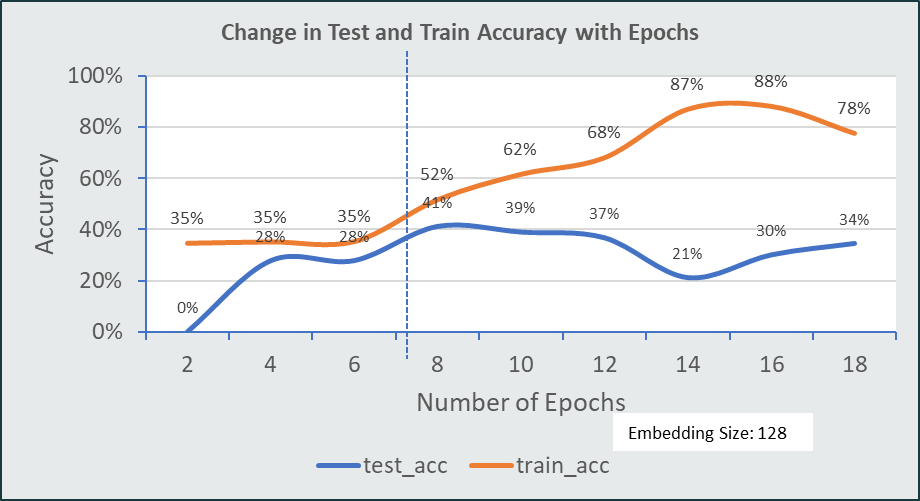
* Count vectorization has been used for converting text to word vectors
  1. Other forms of word vector creation like TD-TDF [14] etc., are not used as we are using advanced neural networds (RNN)
* Tokenizer and texts\_to\_sequences functionality from Keras [15] library has been used for word vector creation which then are used for feeding into RNNs
* Same feature engineering steps have been performed on both test and train datasets

# 7. Model Selection and Evaluation

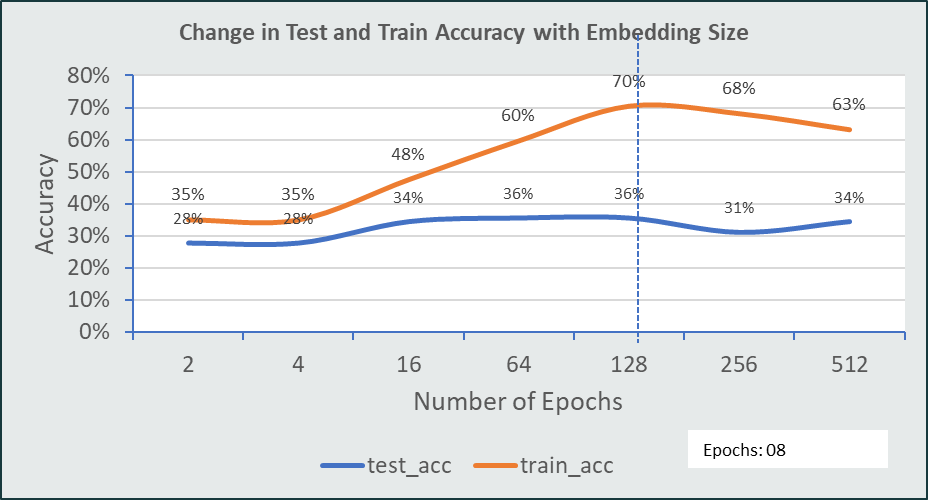
Multiple have been built for predicting the extent of Violence, Intense, Profanity and Sex in the movie. For each of the rating prediction multiple iterations were performed for improving prediction accuracy. Following is the selection report for each rating prediction

## Violence Prediction – Parameter Tuning & Bias Analysis

* Multiple RNN models have been developed on movies script for violence prediction
  + 10% of the total scripts are used as test data while rest 90% is used for model training
* Multiple iteration performed by changing depth of neural nets, embedding length and number of epochs
* Initially number of embedding and depth have been selected by keeping epochs constant, post selecting embedding and depth best performing model is iterated by changing epochs [Refer Figure 7]
* Train accuracy of the model for Violence prediction is increasing with increase in number of epochs while test accuracy is decreasing after 8th epoch [Refer Figure 8]
* So far model with 8 epochs has shown best train and test accuracy at overall level
  + Train accuracy: ~52%
  + Test accuracy: ~41%



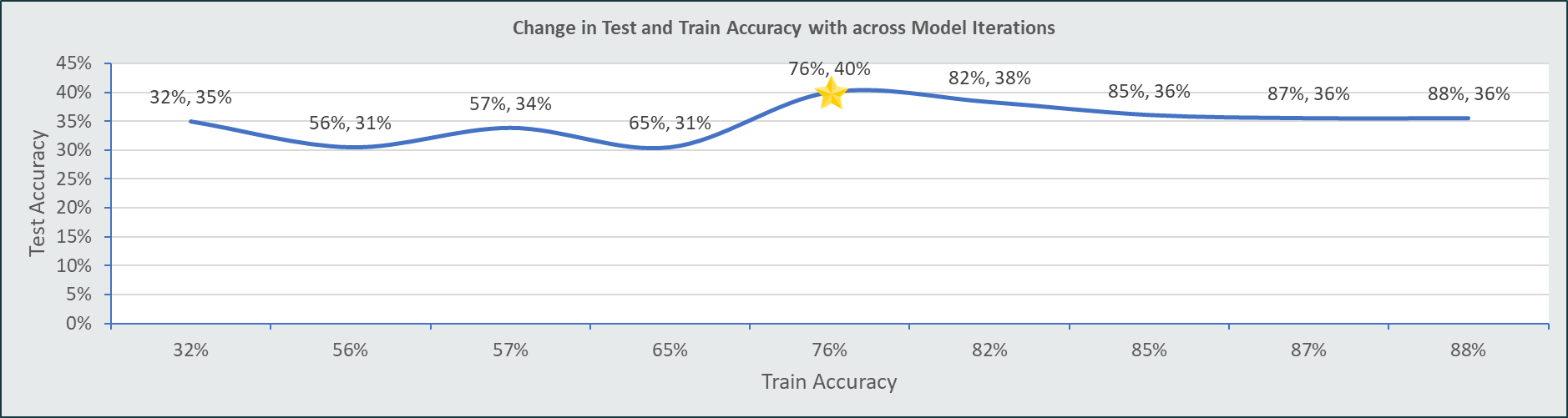
***Figure 7:***Figure showing change in Test and Train Accuracy with **Epochs**

**

***Figure 8:***Figure showing change in Test and Train Accuracy with **Embedding Size**

## Profanity Prediction – Parameter Tuning & Bias Analysis

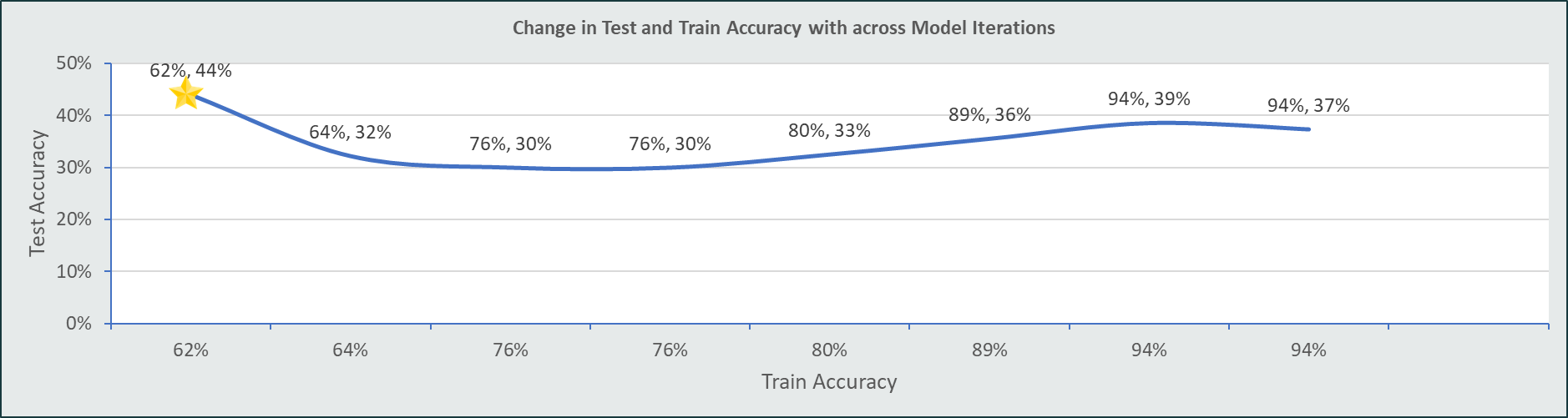
* Multiple RNN models have been developed on movies script for profanity prediction
  + 20% of the total scripts are used as test data while rest 80% is used for model training
* **Multiple iteration** performed by changing **depth of neural nets, embedding length, number of epochs, activation function and optimizer**
* Initially number of embedding and depth have been selected by keeping epochs constant, post selecting embedding and depth best performing model is iterated by changing epochs
* Balanced and best train/test accuracy is obtained for model with 1RNN layer (25Units) and 1 Hidden layer (50Units)
  + Final model iteration selected has a **train accuracy of 76%** and **test accuracy of 40%.** Though this model has **35% variation (suggesting overfitting)** between train and test accuracy, this has been selected as final model due to best test accuracy [Refer Figure 9]



***Figure 9:***Figure showing change in Test and Train for Profanity Accuracy across **Model Iterations**

## Intense Prediction – Parameter Tuning & Bias Analysis

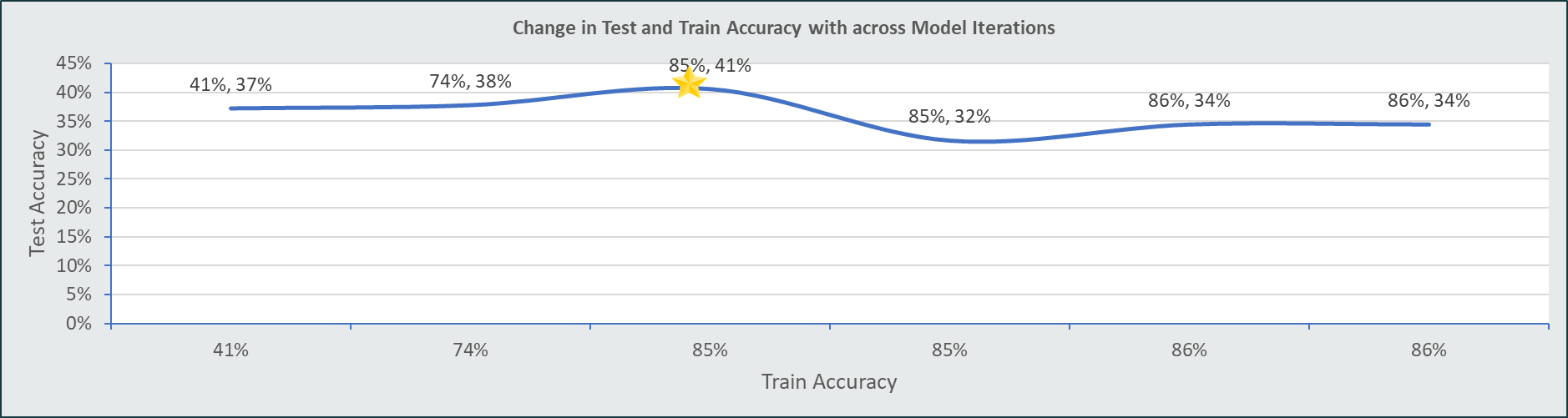
* Multiple RNN models have been developed on movies script for profanity prediction
  + 20% of the total scripts are used as test data while rest 80% is used for model training
* **Multiple iteration** performed by changing **depth of neural nets, embedding length, number of epochs, activation function and optimizer**
* Initially number of embedding and depth have been selected by keeping epochs constant, post selecting embedding and depth best performing model is iterated by changing epochs
* Balanced and best train/test accuracy is obtained for model with 1RNN layer (25Units) and 1 Hidden layer (50Units)
  + Final model iteration selected has a **train accuracy of 62%** and **test accuracy of 44%.** Though this model has **18% variation (suggesting very minor overfitting)** between train and test accuracy, this has been selected as final model due to best test accuracy [Refer Figure 10]



***Figure 10:***Figure showing change in Test and Train Accuracy of Intense Prediction across **Model Iterations**

## Sex Prediction – Parameter Tuning & Bias Analysis

* Multiple RNN models have been developed on movies script for profanity prediction
  + 20% of the total scripts are used as test data while rest 80% is used for model training
* **Multiple iteration** performed by changing **depth of neural nets, embedding length, number of epochs, activation function and optimizer**
* Initially number of embedding and depth have been selected by keeping epochs constant, post selecting embedding and depth best performing model is iterated by changing epochs
* Balanced and best train/test accuracy is obtained for model with 1RNN layer (25Units) and 2 Hidden layer (50Units, 10UNits)
  + Final model iteration selected has a **train accuracy of 85%** and **test accuracy of 41%.** Though this model has **30% variation (suggesting overfitting)** between train and test accuracy, this has been selected as final model due to best test accuracy [Refer Figure 11]



***Figure 11:***Figure showing change in Test and Train Accuracy of Sex Prediction across **Model Iterations**

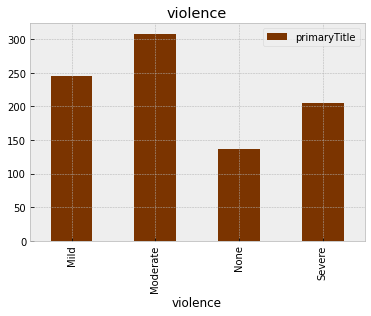
# 8. Model Results and Discussion

## Descriptive Analysis

An exhaustive descriptive analysis is performed on each of movie rating category. Word frequency analysis on Uni, Bi and Tri grams [16] have been performed. Following are the key observations

### Violence

* Violence has a good distribution of scripts across all four classes [Refer Figure 12]
* Density of the words showing violence like police station, machine gun, shooting etc., is increasing as we move from violence “None” to “Severe” – Bigram Analysis



***Figure 12:***Distribution of Class in Movie Violence Classification

### Profanity

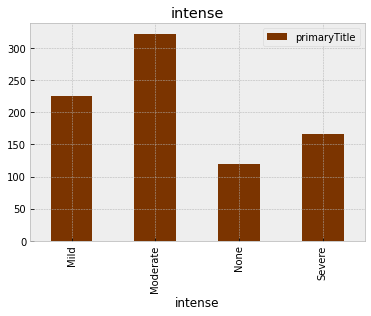
* Profanity has a good distribution of scripts across all four classes (Profanity category “None” has relatively lower number of scripts) [Refer Figure 13]
* Profanity category “None” has no or very less words conveying profanity while category “Severe” has words like f\*\*k, S\*\*t etc., - Observation from Word Cloud



***Figure 13:***Distribution of Class in Movie Profanity Classification

### Intense

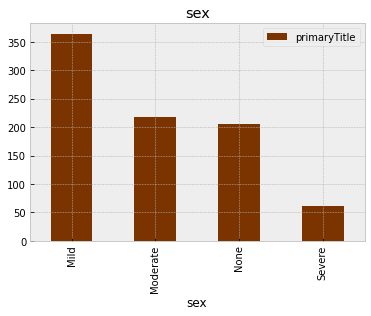
* Intenseness has a good distribution of scripts across all four classes [Refer Figure 14]
* Intenseness at a high level has similar word distribution as that of violence



***Figure 14:***Distribution of Class in Movie Intense Classification

### Sex

* Category SEX has a good distribution of scripts across all four classes (Sex category “None” has relatively lower number of scripts)



***Figure 15:***Distribution of Class in Movie Sex Classification

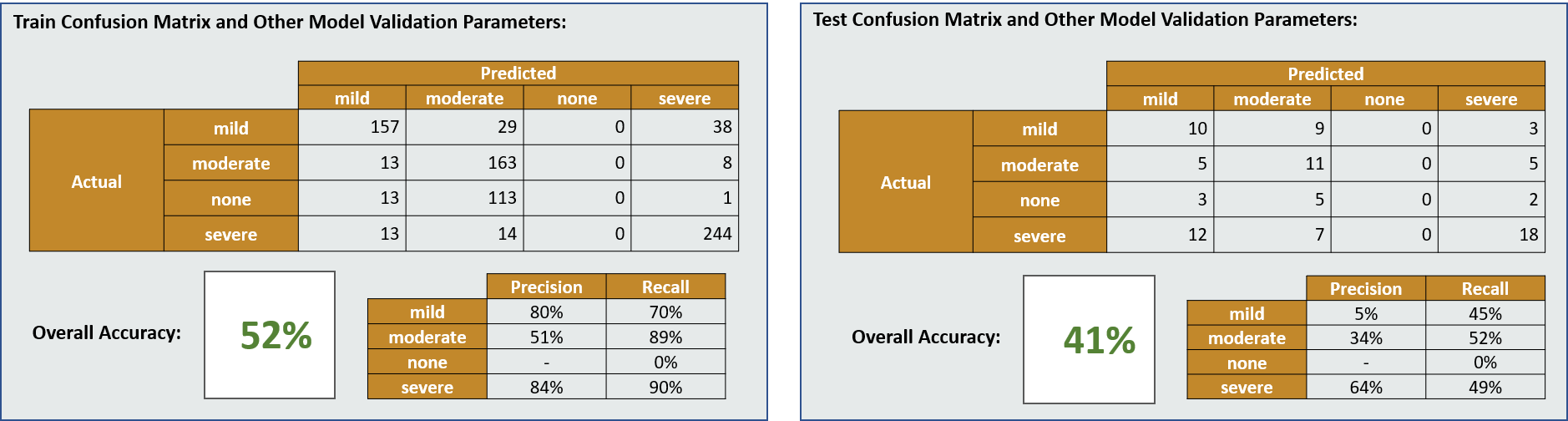
## Model Evaluation

Separate evaluation has been performed on each of Violence, Profanity, Intense and Sex rating predictions. Model accuracy in conjunction with precision and recall [17] has been used for model evaluation. To further ensure the consistency of model predictions evaluation has been performed on both Train and Test datasets. Following is the evaluation report for Violence, Profanity, Intense and Sex prediction.

### Movie Violence – Model Evaluation Report:

* Model performance on Violence prediction is moderate with Test and Train accuracy of 52% and 41% respectively
* Current best model iteration is doing good job for classifying “mild” and “severe” rating (good precision and recall observed) but failing in predicting “moderate” and “none” during train phase
* Though model performance has been exceptional in predicting “mild” ratings during training phase but failed a bit in testing phase
* Class “None” has zero predictions during Train and Test phases this can be due to relatively lower number of scripts with rating “None”

Refer to Table 2 below showing model performance on Train and Test datasets

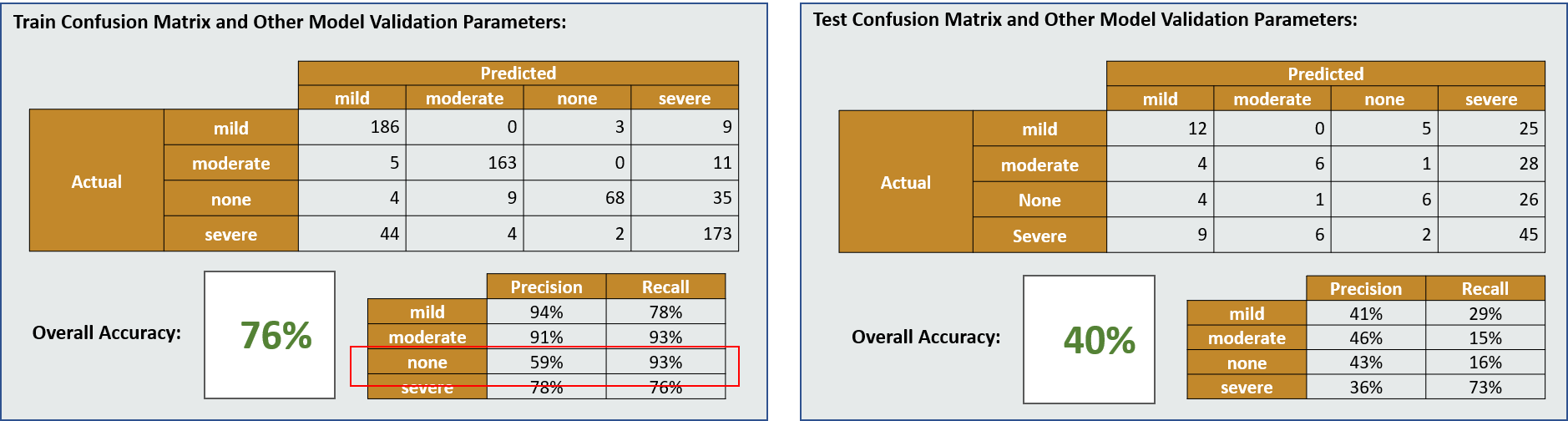


***Table 2:*** Violence Prediction – Train [Left] & Test [Right] Performance Evaluation

### Movie Profanity – Model Evaluation Report:

* Model performance on Profanity prediction is moderate with Test and Train accuracy of 40% and 76% respectively
* Difference in train and test accuracy of ~35% is suggesting overfitting which not being addressed by modelling technique to low size of dataset
* Majority of the scripts in test dataset are being classified as severe, high recall rate (73%) and low precision rate (36%) further confirms this
* Overall model for Profanity prediction is showing similar results to that of Violence prediction with test accuracy of 40%

Refer to Table 3 below showing model performance on Train and Test datasets

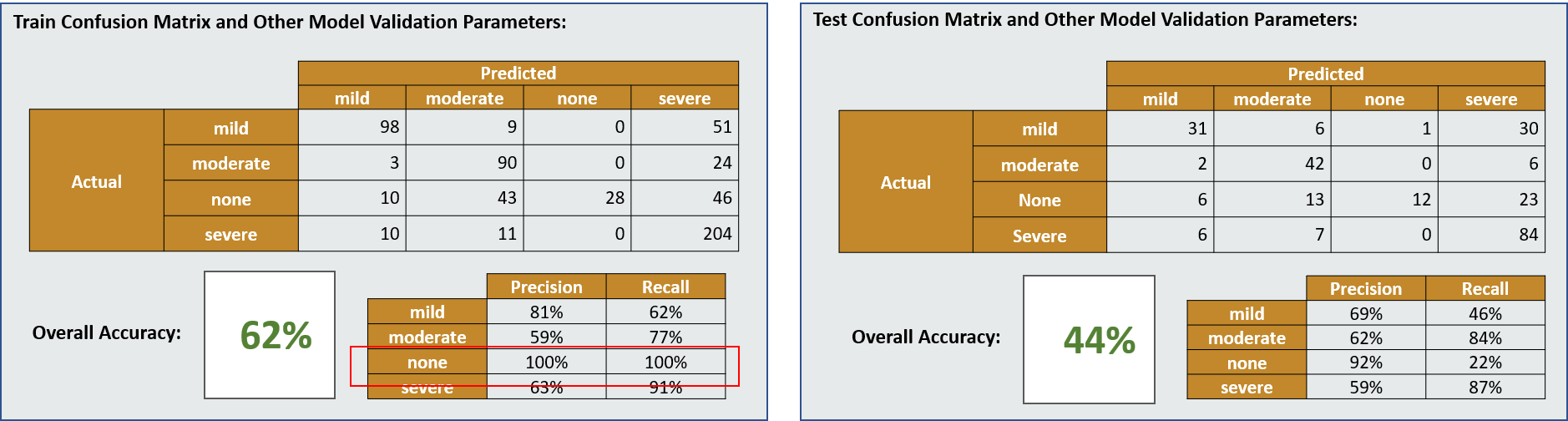


***Table 3:*** Profanity Prediction – Train [Left] & Test [Right] Performance Evaluation

### Movie Intense – Model Evaluation Report:

* Model performance on Intense prediction is moderate with **Test** and **Train** accuracy of **44%** and **62%** respectively
* Difference in train and test accuracy of ~18% is suggesting mild overfitting [18] which not being addressed by modelling technique to low size of dataset
  + Majority of the scripts in test dataset are being classified as severe or moderate (observation from precision rate)
  + Category “None” has excellent performance in train but has recall rate of only 22% in test
* Overall model for Intense classification performing well than that of Profanity & Violence prediction

Refer to Table 4 below showing model performance on Train and Test datasets

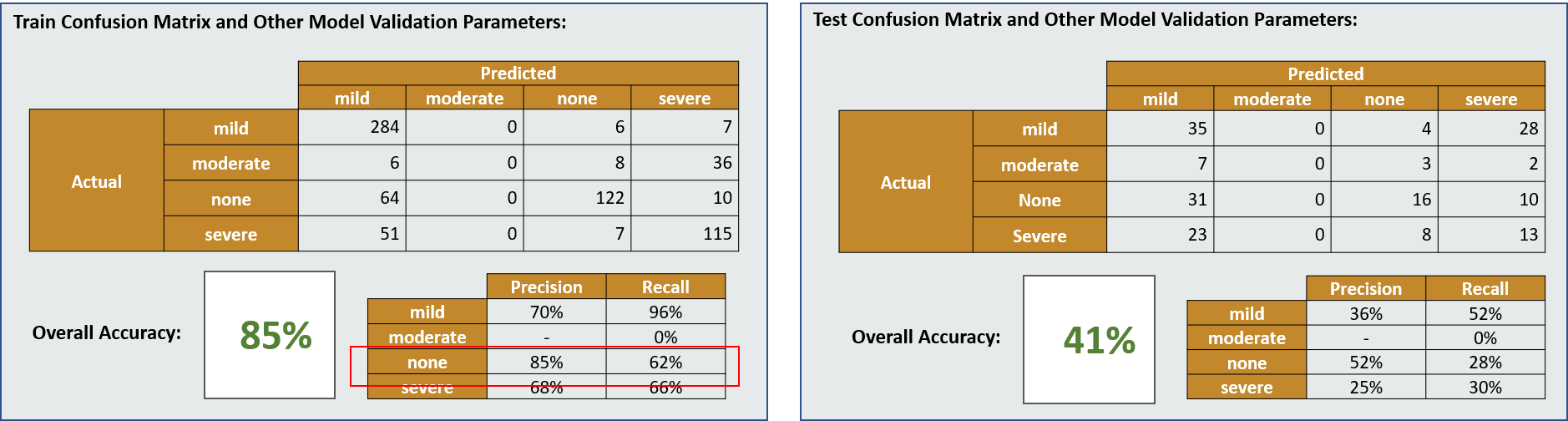


***Table 4:*** Intense Prediction – Train [Left] & Test [Right] Performance Evaluation

### Movie Sex – Model Evaluation Report:

* Model performance on Intense prediction is moderate with **Test** and **Train** accuracy of **41%** and **85%** respectively
* Difference in train and test accuracy of ~18% is suggesting mild overfitting which not being addressed by modelling technique to low size of dataset
  + Majority of the scripts in test dataset are being classified as severe or moderate (observation from precision rate)
  + Category “None” has excellent performance in train but has recall rate of only 22% in test
* Overall model for Intense classification performing well than that of Profanity & Violence prediction

Refer to Table 5 below showing model performance on Train and Test datasets



***Table 5:*** Sex Prediction – Train [Left] & Test [Right] Performance Evaluation

# 9. Discussion on Key Findings

“Age Suitability Rating: Predicting the MPAA Rating Based on Movie Dialogues” by Mahsa Shafaei, Niloofar Safi Samghabadi [1] and “Machine Learning Models for Content Classification in Film Censorship and Rating” by Syma Afsha are few of the paper discuss the use of Movie Subtitle/Dialogues/Scripts for classifying movie ratings. Afore mentioned papers describes the use of both Machine Learning and Deep Learning techniques one of the paper saw 81.6% but doesn’t give specific detail if the reported F1 score is for test of train dataset.

Our approach for rating classification across different segments is consistently giving us more than ~80% accuracy on train dataset and ~40% accuracy on test dataset. We suspect that the low-test accuracy is due

* 1. Low data volume
  2. Number of classes for predictions

As we are building a multiclass classification problem where we see very thin barrier from one class to other leading to higher error. Overall, our model is doing good job in profiling scripting except for classifying a script with “None” rating, which is due to relatively few numbers of scripts with class “None”.

# 10. Conclusion

The model performances for Intense, Profanity, and Violence predictions are moderate with varying degrees of accuracy on test and train datasets. Overfitting is observed in the models due to the small size of the dataset. The majority of the scripts in the test dataset are classified as severe or moderate. Category "None" has excellent performance in the train dataset but has low recall rates in the test dataset. The overall performance of the Intense prediction model is better than that of the Profanity and Violence prediction models.

The approach for rating classification is giving consistent train accuracy of over ~80% and test accuracy of ~40%, which could be due to low data volume and number of classes. The multiclass classification problem has a thin barrier between classes leading to higher errors. This issue of low accuracy and overfitting can be addressed models are trained on large of volumes scripts there by enable models to better generalize.

# 11. Reference

[1] Gentile, D. A., Lynch, P. J., Linder, J. R., & Walsh, D. A. (2004). The effects of violent video game habits on adolescent hostility, aggressive behaviors, and school performance. Journal of Adolescence, 27(1), 5-22.

[2] Carnagey, N. L., & Anderson, C. A. (2005). The effects of reward and punishment in violent video games on aggressive affect, cognition, and behavior. Psychological Science, 16(11), 882-889.

[3] https://www.healthychildren.org/English/family-life/Media/Pages/TV-Ratings-A-Guide-for-Parents.aspx

[4] <https://www.theschoolrun.com/parents-guide-movie-age-ratings>

[5] Shafaei, M., & Samghabadi, N. S. (2020). Age Suitability Rating: Predicting the MPAA Rating Based on Movie Dialogues. Proceedings of the 2020 12th Conference on Computer and Robot Vision, CRV 2020, 33-39. doi: 10.1109/CRV50468.2020.00012

[6] Syma Afsha. (2020). Machine Learning Models for Content Classification in Film Censorship and Rating. International Journal of Computer Applications, 179(43), 10-16.

[7] Graves, A. (2013). Generating sequences with recurrent neural networks. arXiv preprint arXiv:1308.0850.

[8] Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Advances in neural information processing systems (pp. 3104-3112).

[9] Web address for IMSDb (Internet Movie Script Database) [www.imsdb.com](http://www.imsdb.com)

[10] Web address for [www.imdb.com](http://www.imdb.com)

[11] Richardson, L. (2013). Beautiful Soup Documentation. Crummy Gummy. Available online: https://www.crummy.com/software/BeautifulSoup/bs4/doc/ (accessed on 23 April 2023).

[12] Jurafsky, D., & Martin, J. H. (2009). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Pearson Education.

[13] Miller, G. A. (1995). WordNet: A lexical database for English. Communications of the ACM, 38(11), 39-41.

[14] An introduction to TF-IDF by Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze (2008)

[15] Chollet, François. "Keras: Deep Learning for humans." 2015. Available online at <https://keras.io>.

[16] N. S. Kesavan and V. Subramaniyaswamy, "Text document clustering using n-grams", in Proceedings of the International Conference on Advances in Computing and Communications, 2013, pp. 23-26. doi: 10.1109/ICACC.2013.28.

[17] Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), 427-437. <https://doi.org/10.1016/j.ipm.2009.03.002>

[18] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction (2nd ed.). Springer.