Age Suitability Prediction System using MPAA Rating

Christeen S Joseph[1] and Dr. Jayashree R[2]

[1]Student, Department of Computer Science and Engineering, PES University, Bangalore

[2]Professor, Department of Computer Science and Engineering, PES University, Bangalore

ABSTRACT

This paper describes the architectural design and implementation of a system to predict the MPAA rating of a set of movies, thereby illuminating viewers on the suitability of these movies for various age groups. The model is designed on the basis of dialogues in movies, the bad word ratio and the context in which said words are used. By establishing the suitability factor, young adults and children can be restricted from being influenced by age inappropriate movies. The MPAA rating system classifies movies into G (General Audience), PG (Parental Guidance Suggested), PG-13 (Parents Strongly Cautioned), R (Restricted) and NC-17 (Adults Only).

Keywords: MPAA Rating, Age Suitability, Natural Language Processing, Naive Bayes, Support Vector Machine

1. INTRODUCTION

With an increase in work pressure, comes the added necessity to spend time in leisure. Today's world sees ample evidence of the influence of the media on a youngster’s mind. It becomes imperative to ensure that the youth and kids watch age appropriate movies. To ensure youngsters watch appropriate content, an MPAA rating scheme is determined based on dialogues in the movie. This project aims at trying to determine the age groups suitable for watching a movie.

This paper describes the architectural design and implementation of a model to predict the MPAA rating of a set of movies, to illuminate viewers on the suitability of these movies for various age groups. This model attempts to predict the suitability of movie content based on dialogues in scripts using the MPAA rating. The usage of popular machine learning algorithms such as Naive Bayes, Random Forest, Support Vector Machines and Logistic Regression simplified the training process. The model uses bad word ratio and the context in which it is used to predict the overall suitability.

The scope of this project extends from movie theaters to home theater systems. Ensuring that young viewers are not affected or influenced by age inappropriate movies has become of vital importance. It becomes a mandatory task to rate movies based on content to warn and restrict children and young adults from such content.

Knowledge of the five major categories involved in the MPAA rating is crucial to understand the methodology of this system. G represents the general group; which states that all ages can be admitted. PG signifies that certain content in the movie should be reviewed by parents. PG-13 indicates that the movie script constitutes certain dialogues that may be deemed as inappropriate for children under the age of 13. R, which stands for restricted, represents a tag that means people under 17 should ideally watch the movie with a parent. NC-17 refers to a restriction stating that no one under 17 is allowed to watch the movie under any circumstance.

2. Literature Survey

One of the earliest models to be developed, employed a wide range of language-based features required for improving readability, provided by Second Language Acquisition and Psycholinguistics research to differentiate between spoken languages that are aimed toward various age groups. The model [8] had managed to achieve an accuracy rate of around 96. It used the frequency database SUBTLEX-UK. The WEKA toolkit was used to perform certain classification related tasks and also evaluate the respective accuracy for every classified label. To build the model, Sequential Minimal Optimization, J48 Decision tree, Random Forest and Logistic Regression algorithms were implemented. The feature set that was internally used to create this model comprises of around 152 lexical and syntactic features that are primarily obtained from the research over complexity of text carried out in SLA and Psycholinguistics, namely Lexical richness features (LEX), Syntactic complexity features (SYNTAX), Psycholinguistic features (PSYCH) and Celex features (CELEX).

Another model developed earlier evaluated ratings based on scenes in movies using a machine learning based classification technique and Word2Vector with accuracy of 59 [1]. The procedures involved in creating the model were mainly three in number, namely text pre-processing, feature extraction, and classification. Pre-processing techniques involve removing non-symbol characters, converting words to lowercase, tokenizing words, reducing words to their word stems and removing person names. The process of feature extraction involves using WordVec to generate vector representations of words that were primarily based on their linguistic context. Classification uses classifiers such as SVM, NBC, Decision Trees and Cross Validation to select a model.

Creation of a CNN model [7] that is both effective and light-weight, and is quick in predicting output was later developed. The datasets used to train the model include Violence in Movies, Violent Scene Detection and Hockey Fights. The violence detection scheme uses three techniques. The movie is firstly split into a set of shots, from which a representative frame is selected. This choice is based on a parameter directed by the level of salience. The frames, which are passed one-by-one from a deep learning model, undergo a fine-tuning process which uses an approach called transfer learning (a pre-trained MobileNet model) to differentiate between violent and non-violent scenes in a movie script. These non-violent scenes are now combined sequentially to eventually produce a violence-free movie. Using Violence in Detection dataset, an accuracy of 99.5 was achieved, while the dataset involving Hockey Fights only managed 87 percent accuracy. Future enhancements involve improving the model by utilizing sequential learning parameters like LSTM with CNNs for effective and easy detection of violent scenes.

A paper that provided great insight into the tactics of prediction of risk factor [9] proposed a solution to characterize certain features of violence-related content in movies, which were fetched from the language and dialogues found within the movie scripts. This approach was heavily reliant on a broad range of features and aspects that captured a number of lexical, semantic and sentiment characteristics. An extension to Movie-DiC dataset was used in order to train the model. Language features from all the dialogue utterances by actors were collected to train models like RNN and SVM. Features can be divided into five basic categories: N-grams, Linguistic and Lexical, Sentiment, Abusive Language and Distributional Semantics. Linear SVC was implemented using scikit-learn, along with a set of RNN models that were implemented in Keras. Subsequent work aims to explore the realm of sentiment analysis and its models, which are complementary to the existing lexicon-based approaches that were used.

The model created by Shafaei et al (2019) [6] predicts the suitability of movie content based on scripts using MPAA rating, which comprises five main categories, namely G, PG, PG-13, R, NC-17. An RNN-based design was created to predict rating using emotions and genre, internally using bad word ratio and the context in which it is used to predict suitability. It also provides the first document consisting of movie scripts along with their associated MPAA rating, values of certain MPAA components, MPAA rating for movies of similar genres, along with poster images for said movies. It achieved a 81.6 weighted F1-score performance that seemed to work better than the traditional models. The dataset is an expansion of the dataset that was manually collected by Shafaei et al (2019). Conversation related data, dynamics pertaining to emotions between characters, genre of the said movies, and movies that have a similar plotlines to the target movie are used to address challenges in the task.

Further enhanced systems started to learn how to better represent movies from semantic and sentiment aspects found in a character's use of language [10]. 10-fold cross-validation methods were used to fetch a very reliable estimation of the efficiency and performance of the model. For the RNN layer, Gated Recurrent Units were used along with Bi-LSTM parameters for implementation of sentiment models. An accuracy of 67 was achieved. Data was collected from numerous available online resources. The model was implemented in Keras. The multi-task model which was proposed, constitutes a task-specific attention model with an F1 = 67.7, an increase of 1.22 over previous models. Future enhancements involve how and when characters are being referenced in risk behavior scenarios.

Before creating a model, a dataset of about 17000 film transcripts together with their respective age-rated values to predict correct age classification was explored. The model created [2] involved Gradient boosting techniques for effectiveness, in comparison to deep learning design architectures. An accuracy close to 74 was achieved for US ratings and 65.3 for UK ratings. Three models were used: FastText, XGBoost and Hierarchical Attention. FastText can be described as a document classifier that is dependent on a language model that is built using the concept that words are represented as the combination of sub-word vectors. XGBoost, an efficient and effective implementation of gradient boosting machines, has proven to be a success in multiple real life scenarios. Hierarchical Attention constitutes a neural network architecture that depicts a text shown as a sequence of sentences, which can subsequently be represented in the form similar to a list of words, and uses relations between words to build a model. Data scripts were collected from https://www.springfieldspringfield.co.uk/. The main purpose of the Spacy NLP library was to tokenize the script into sentences. Two settings were used for evaluation of accuracy: strict and relaxed. Future enhancements involve consideration of time factors in the classification process.

A dissertation [3] over the subject stated a thesis to build a system that could allow users to determine whether or not a media text content is suitable to their age or preference. It constitutes the usage of various NLP techniques and machine learning algorithms. The Book Cave database is used as a dataset to train the model, which performs numerical analysis on text. The model represents the entire document as a bag-of-words, applies word embedding techniques and implements learning algorithms. The created model is actually able to reasonably detect semantic meaning in portions of text that are far smaller than the input texts when only labels for the input texts are present, i.e., even when ground-truth labels for the smaller portions of text are not available. The steps involved in the prediction process include tokenization, applying bag-of-words techniques and creation of word vectors. The learning algorithms used for bag-of-words ranges from K-Nearest Neighbors, Multinomial Naive Bayes, Logistic Regression, Random Forest, Multi-layer Perceptron and Support Vector Machine. The algorithms used for Word Vectors include the use of Self-attention and Hierarchical Attention Networks, Recurrent Neural Networks and Convolution Neural Networks.

Subsequently, a system [5], which involved creation of a multi-modal RNN model for subtitles and audio, along with an addition of CNN with LSTM for video, led to further research. It incorporates concepts including late fusion, feature concatenation fusion methodology and Gated Multi-modal Unit fusion. A novel task is introduced in multi-modal prediction, i.e. rating videos by using the MPAA rating for various movie trailers. The Multi-modal Movie Trailer Rating (MM-Trailer) dataset was created, which is composed of movie trailers along with MPAA tag values, audio files, subtitles of dialogues, and metadata of the movie. This model demonstrates how the combination of various modalities would lead to significant improvements. The selection of 5 fold cross-validation for evaluation proved to be effective as results obtained were fairly reliable. The weighted F1 score was chosen as a parameter to evaluate the efficiency and accuracy.

Further research led to the development of a model that involved creation of bi-modal dataset, adding both images and texts for around 17000 films, to identify appropriate age groups with the aim of improving the performance of age appropriateness [4]. Image feature extraction models that are commonly used in deep learning, like DENSENet, ResNet, Inception and NasNet were used. Accuracy of 68.4 and 56.7 was achieved using images, while accuracy of 81.1 and 66.8 was achieved using text. Results involved an analysis stating that for balanced data, it is easier to classify PG when compared to R and PG-13. Future work on the project would involve two main aspects, namely investigations related to using whole video and audio of the film in classification and analysis pertaining to the distribution of risk behavior.

3. DATASET

The dataset used in the creation of this model was collected by Shafaei et al (2019) [6]. It included a set of movie scripts and dialogues, along with some meta data pertaining to each movie.There are mainly two databases. One consists of a table that includes features like plot, title, binary class success, genre, year, etc. The other database, which consists of dialogue exchanges between characters in the movie, is found in the scripts folder.

4. PROPOSED METHODOLOGY

The main steps involved in the development of this project include pre-processing data, splitting up data into training and testing data, exploratory data analysis, feature extraction, implementing NLP algorithm, calculating bad word ratio, classifying suitability, training the model to predict context for bad words and the reclassifying suitability.

A methodology designed to implement the system involves designing a machine learning model to represent conversations between characters and their context within the movie. Models are designed considering bad words in the script and the overall context in which said words were used.

A number of algorithms, namely NBC, SVM, Random Forest and Logistic Regression are used to train the model to understand the context of statements in the scripts. Using the bad word ratio calculated from the script, the model analyzes the inappropriateness factor of the document and declares a rating for the movie script based on the percentage of encountered bad words. A further analysis over the script ensures the classification of the context in which these words were used. Based on the context, an overall rating is provided for the movie script. From this value, users can decide if the movie is appropriate for their age group.

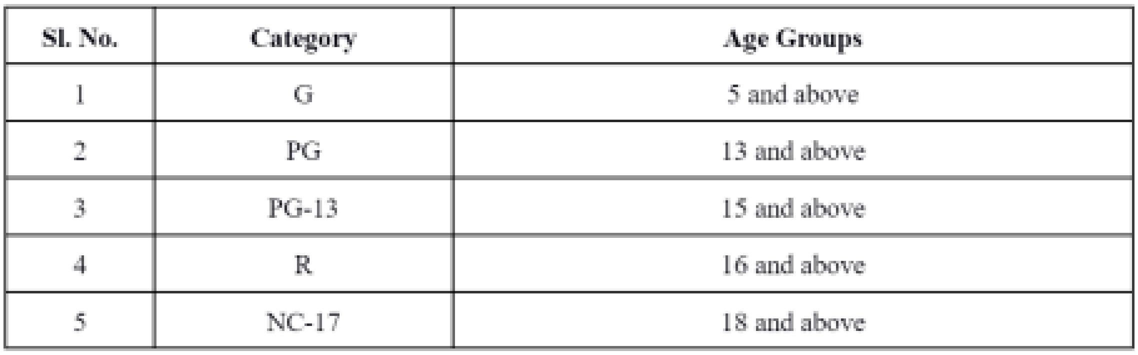
To build the model, the bad word ratio is taken into consideration. Using a bad word list that was created, the script is analyzed to find the most frequent words used and determine the frequency of bad words and their respective polarity. If said word is used multiple times or a larger portion of the text file consists of bad words, the movie is deemed inappropriate. Term-frequency inverse-document-frequency values were calculated to determine the count of words in the corpus and the probability of the occurrence of said words. Results show the bad word ratio based on these algorithms, along with the polarity of the scripts, thus determining if a movie is recommended or not. If the percentage of bad words exceeds a certain threshold, it becomes mandatory to flag the movie as containing inappropriate content. Based on the percentage values, a rating is given to the movie script. The bad word content can be used in three different contexts and based on the intensity of these contexts, the inappropriateness factor changes. For movies with a high bad word ratio, the context for the same is checked and based on the context, a reclassified value is determined.

5. RESULTS AND DISCUSSION

Results obtained from this project involve determining appropriateness of film for young adults and kids using MPAA rating. If the movie can be viewed by all, the category provided by the model is G, which represents General. The age group defined for the same ranges from 5 and above. If there are certain aspects of the movie rendered unsuitable for kids, it is marked as PG or PG-13, which implies parental guidance is recommended and strongly recommended respectively. For a movie marked PG, the ideal age group ranges from 13 and above, whereas for PG-13, the age group is defined as 15 and above. If the content is marked for a mature audience, it is marked as R, which stands for restricted audience, which implies ages 16 and above are allowed to watch such movies. If such content is severe, it is marked as NC-17, which implies that no one under the age of 17 is to watch this movie, thus age groups 18 and above are allowed to watch these sets of movies.

The cross validation values obtained from training the model using various algorithms ranges from approximately 65 to 90. Every model's accuracy is recorded to compare the best fit for training the model. This factor aids one in choosing the apt algorithm.

Table 1: Suitable age groups for different MPAA rating



Sl. No.	Category	Age Groups
1	G	5 and above
2	PG	13 and above
3	PG-13	15 and above
4	R	16 and above
5	NC-17	18 and above

6. Conclusion

Determining the age suitability of movies would benefit the young minds of today by ensuring they do not encounter any age inappropriate content. This would restrict the impact factor of a movie's content on their behavior and language, keeping the youth and children safe from negative influence.

This paper details a design to implement the system involving designing a machine learning model using NBC, SVM, Logistic Regression and Random Forest to represent conversational bits between characters and emotions experienced by characters in scenes within the movie. Bad words in the script and MPAA ratings of similar movies would be a basis for deciding the MPAA factor of a set of movies.

Scope for future enhancements involve using images to help in providing better context for understanding script dialogues. As the model is trained using above mentioned algorithms over a manually created dataset, there is a scope for improvement of accuracy of the model.

References

[1] Caner Balim, Ugur Gurel, “MPAA Rating Prediction based on Deep Learning”, 3rd International Symposium on Innovative Approaches in Scientific Study, 2019.

[2] Emad Mohamed, Le An Ha, “A First Dataset for Film Age Appropriateness Investigation”, Proceedings of the 12th Conference on Language Resources and Evaluation, 2020.

[3] Eric Robert Brewer, “Age-Suitability Prediction for Literature Using Deep Neural Networks”, Theses and Dissertations, 2020.

[4] Le An Haa, Emad Mohamed, “Combining Text and Images for Film Age Appropriateness Classification”, 5th International Conference on AI in Computational Linguistics, 2021.

[5] Mahsa Shafaei, Christos Smailis, Ioannis A. Kakadiaris, Thamar Solorio, “A Case Study of Deep Learning Based Multi-Modal Methods for Predicting the Age-Suitability Rating of Movie Trailers”, arXiv.org, 2021.

[6] Mahsa Shafaei, Niloofar Safi Samghabadi, Sudipta Kar and Thamar Solorio, “Age Suitability Rating: Predicting the MPAA Rating Based on Movie Dialogues”, Proceedings of the 12th Conference on Language Resources and Evaluation, 2020.

[7] Samee Ullah Khan, Ijaz Ul Haq, Seungmin Rho, Sung Wook Baik, Mi Young Lee, “Cover the Violence: A Novel Deep Learning Based Approach Towards Violence Detection in Movies”, MDPI Applied Science, 2019.

[8] Sowmya Vajjala, Detmar Meurers, “Exploring Measures of Readability for Spoken Language: Analyzing Linguistic Features of Subtitles to Identify Age-Specific TV Programs”, Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations, 2014.

[9] Victor R. Martinez, Krishna Somandepalli, Karan Singla, Anil Ramakrishna, Yalda T. Uhls, Shrikanth Narayanan, “Violence Rating Prediction from Movie Scripts”, The Thirty-Third AAAI Conference on Artificial Intelligence, 2019.

[10] Victor R Martinez, Krishna Somandepalli, Yalda T. Uhls, Shrikanth Narayanan, “Joint Estimation and Analysis of Risk Behavior Ratings in Movie Scripts”, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 2020.

AUTHOR

Christeen S Joseph received a B.E. degree in Computer Science Engineering from BNM Institute of Technology in 2018. From July 2018 to Jan 2021, she worked for Infrrd Pvt Ltd as a UI developer. She is currently pursuing her masters in Computer Science and Engineering from PES University.

Dr. Jayashree R is currently a professor at PES University. She has more than 25 years of experience teaching in colleges at Bangalore University, VTU and Mysore University and is a recognized supervisor of both VTU and PES universities.