Emotions and Textual Feature Analysis for Motion Picture Association of America (MPAA) Rating Classification

Yaseen K. Abbas and Ahmed Al-Azawei
Department of Software, Colleage of Information Technology,University of Babylon, Al-Najaf Street, 51002 Hilla, Babylon, Iraq
yaseenkudhaira.sw@student.uobabylon.edu.iq, ahmedhabeeb@itnet.uobabylon.edu.iq

Keywords: Machine Learning, MPAA Rating, Movie Scripts, LightGBM, Multi-Class Classification, TF-IDF.

Abstract: Movies represent one of the most important media that combine narrative components with visual and auditory elements to entertain or inspire viewers. The Motion Picture Association of America (MPAA) is a prominent rating system that plays a great role for audience to select an appropriate film. This rating system is important because it helps parents filter the content to protect their children from unsuitable movies. It also assists audiences in their own choice. Traditionally, MPAA ratings are assigned by reviewers manually and this, in turn, leads to inconsistency, subjectivity, and time-consuming. This research introduces an automated method for predicting MPAA rating by using the scripts feature with the emotion feature that aiming to enhance classification accuracy and provide a more reliable assessment of age appropriateness. Scripts are preprocessed and converted into term frequency-invert document frequency (TF-IDF) vectors to capture significant linguistic patterns. Moreover, emotional features are extracted from movie scripts using transformer-based models due to their contextual understanding capabilities. These features are integrated to form a multi-class output. In this study, the LightGBM algorithm is applied as a gradient boosting technique. Experimental findings indicate that combining emotion features alongside textual representations enhances prediction accuracy in comparison to the use of scripts alone. The model achieves 84.2% and 84.6 for the weighted F1-score and the accuracy metric, respectively. This supports the effectiveness of the proposed model in predicting MPAA ratings from both movie scripts and emotional features.

1 INTRODUCTION

Movies are important means of cultural expression and artistic in people's life. They are not just a means of entertainment, but movies also a tool for social and intellectual influence. Film classification is an essential step in regulating the film industry and ensuring it reaches the appropriate audience. This enables viewers to choose films that suit their interests and preferences. It also contributes to directing audiences to age-appropriate content through age rating. Thus, movie classification is an effective way to protect children and sensitive groups from inappropriate content. In addition, it facilitates production and distribution companies to market films more accurately. This can also enhance the viewing experience by setting audience expectations before watching a film. Creators can rely on movie classification as it directs them towards a specific genre that reflects their artistic vision [1].

The age appropriateness of films has been recommended by censorship bodies through rating certificates. There are two main organizations specialized in performing this task. The former is the Motion Picture Association of America (MPAA) in the United States of America, whereas the latter is the British Board of Film Classification (BBFC) in the United Kingdom [2]. These two organizations build their ratings based on films' content. MPAA ratings have a wide practical value where parents can depend on them as a guideline to determine which movies are appropriate for children. Media service providers such as Amazon and Netflix may use these ratings to enable age-appropriateness filters. This mainly relies on movie scripts as a means to determine their content [3].

The MPAA uses five film ratings to guide audiences. The first one relates to the General type (G) that is suitable to all ages with no offensive content, nudity, or significant violence. The second one is Parental Guidance (PG) that may contain

unsuitable material for children with moderate violence and brief nudity. Parents Strongly Cautioned (PG-13) is the third type which includes warns content that may be inappropriate for viewer under-13 years with occasional use of substances, strong language, and minimal sexual content. Fourthly, Restricted (R) requires parent accompaniment for under-17 year due to adult content, including strong language, violence, sexuality, or substance use. Finally, No One Under 17 (NC-17) prohibits anyone under 17 from viewing due to explicit sexual scenes, extensive harsh language, or excessive violence, though it is not classified as obscene or pornographic [4].

A wide range of data sources can be utilized for the automatic classification of movies into age appropriateness rating. One of the important data sources available about movies is movie scripts, which contain only conversations between characters, which is considered the foundation of filmmaking. Movie scripts instruct how a scene should be filmed in addition to telling the audience what is happening. It is a screenplay-formatted document that authors, directors, and actors can use to help them prepare for their parts [5].

Traditionally, MPAA ratings are assigned manually by human reviewers. This approach is often considered time-consuming, subjective, and inconsistent. It may also affect the rating decision due to personal bias or differing interpretations of content. Consequently, there is a growing need for an automated and reliable method to predict MPAA ratings objectively. Developing such a system can enhance consistency, reduce human effort, and support filmmakers, distributors, and audiences in understanding content suitability more efficiently.

The present study addresses this problem by automatically predicting MPAA ratings for movies based on their textual content. The task is formulated as a multi-class classification problem. The input parameters include both textual features and other features that are extracted via the Transformer model. The classification process is based on the LightGBM algorithm, which predicts the most appropriate MPAA class for each movie. To evaluate the model performance, the accuracy and weighted F1-score are used as main criteria for assessment.

The key aims of this research are twofold. First, it develops an automatic method to predict MPAA movie ratings using textual script data with emotions that are extracted from the entire script. The distilled version of RoBERTa (DistilRoBERTa) [6]as a Transformer model is employed for feature extraction. The extracted features are combined with Term Frequency–Inverse Document Frequency (TF-IDF) features and they are fed into the LightGBM[7] model for multi-class prediction purposes. Second, the research addresses the limitations of the sequence length constraints existing in Transformer architectures. To achieve such aims, it seeks to answer the following research questions:

1) How can emotion features extracted from a Transformer-based model improve the performance of MPAA rating prediction in comparison to traditional text-based representations alone?
2) Which method can effectively handle the input sequence length limitation of Transformer models when processing long movie scripts for emotion feature extraction?
3) To what extent the combination of different features can contribute to achieving higher accuracy and interpretability in automated MPAA rating prediction?

Based on such aims and research questions, this research achieves the following contributions:

1) Proposing an automatic model for predicting MPAA rating based on fusion term weighting features with emotion features extracted from a Transformer-based model.
2) Presenting a method for handling the length size limitation of the input sequence processing that exists in the Transformer model during the emotion feature selection process.
3) Achieving a new performance benchmark for this task with 84.2% for the weighted F1-score and accuracy was 84.6%.

The remainder of this paper is organized as follows. Section 2 reviews the relevant literature and related work. Section 3 presents the dataset and methods used. Section 4 reports the proposed model phases. Section 5 presents and analyzes the experimental results. Finally, Section 6 provides conclusions and outlines potential future research directions.

2 LITERATURE REVIEW

This study attempts to address the gap related to extracting emotion features by employing a Transformer-based model that is capable of capturing contextual and fine-grained affective cues within movie scripts. Unlike lexicon-based methods such as the Name Entity Recognition (NRC) emotion lexicon [1][8], which assigns static emotion

values to individual words without considering their contextual meaning, the proposed approach utilizes the DistilRoBERTa-base model ("j hartmann/emotion-english-distilroberta-base"). It was trained on the GoEmotions dataset. This model enables the extraction of context-dependent emotional representations that better reflect the underlying sentiment and intensity of movie narratives. This study also introduces a method to overcome the input length limitation in Transformer-based emotion extraction by segmenting and aggregating emotional representations across script chunks.

The movie scripts are typically long and complex narratives, often exceeding the maximum input length (512 tokens) that Transformer-based models can process. This constraint causes loss of contextual and emotional information, leading to incomplete emotion representation. This method can preserve emotional continuity and ensures that long movie scripts are fully analyzed, capturing both local and global emotional contexts. By integrating these emotion features with statistical linguistic representations derived from TF-IDF, the proposed framework improves the accuracy and reliability of automated MPAA rating prediction.

Many studies were accomplished on age rating classification. These studies varied in using either traditional machine learning techniques or artificial neural network models. It was found that this literature utilized a single data type such as text data, images, or based on multiple form to achieve better accuracy results.

To our knowledge, the first published research about predicting MPAA rating was accomplished by Shafaei et al. [8]. The research works on predicting the suitability of movies for children and young adults based on movie scripts, relying on the MPAA rating system as a benchmark. A deep learning model was proposed by using a Recurrent Neural Network (RNN) with an attention layer. This incorporates movie genres and emotions in conversations to predict the five MPAA ratings. The experiment shows that the model achieves a weighted F1-score of 78% which outperforms traditional machine learning methods by 6%. The study highlighted that bad word lists alone are insufficient for MPAA rating prediction. Therefore, including multiple factors such as violence, language, and thematic material can enhance the prediction. The model performance was improved when genre and emotion vectors were included.

Another research introduced in [1] presents a task of predicting the MPAA rating of movies using their scripts to assess age suitability for children and young adults. The Long Short-Term Memory (LSTM) model was leveraged here with the attention mechanism and it incorporated both emotion and genre vectors for predicting MPAA ratings. The pre-trained Global Vectors for Word Representation (Glove) was used as a word embedding method which can preserve the semantic information. Based on model performance, the weighted F1-score was 81.6% which outperforms traditional machine learning methods by 7%.

In [9] a multi-modal method was proposed based on deep learning methods to automate the age-suitability rating of movie trailers. A new dataset, the Multi-modal Movie Trailer (MM-Trailer) that contains 1,443 trailers and their corresponding age-suitability ratings. The research combines audio, video, and text modalities to improve prediction accuracy, demonstrating that multi-modal approaches outperform single models. A specialized neural network architectures were assigned for each modality. RNNs were for text and audio, and CNN-LSTM were for video. The results indicated that the Gated Multi-modal Unit (GMU) fusion method yields the highest performance, significantly improving upon the best single modality model. Based on the experimental results of the GMU fusion variant, it achieved a weighted F1 score of 86.06% that outperforms other fusion methods and single modality models. It should be noted that the output result in this study is a binary classification, where the film is classified as suitable for all audiences or restricted and specific to a specific audience.

Another research study [3] introduced a multimodal approach for predicting the film age appropriateness with MPAA and BBFC. It combined textual and visual features. The subtitles and synopses were used in textual data, while the movie poster was used as the visual data. Several deep learning and machine learning models were utilized for textual and visual data.

ResNet, Inception and NASNet were used for visual feature extraction, while the TF-IDF method was used for textual feature selection. The study demonstrated that textual features provide a strong predictor of MPAA ratings, while images provide complementary information. The experimental result shows that combining both modalities leads to a higher accuracy than using either alone, which highlights the value of multimodal fusion in movie rating research. The best accuracy result was accomplished using xgboost model with only three

class output (PG, PG-13, R), with 81.1% for MPAA and 66.8% for BBFC.

The research study introduced in [4] proposed using bidirectional LSTM models that were integrated with an attention mechanism to introduce MPAA multi-class movie rating prediction. The study was based on 12000 textual script data and presents four models to evaluate the best one using different feature extractors as Bidirectional Encoder Representations from Transformers (BERT) and A Robustly Optimized BERT Pretraining Approach (RoBERTa). According to the research results, the bad word frequency model achieves the best output with an accuracy metric of 56% classification of movie ratings.

3 MATERIALS AND METHODS

3.1 Movie Dataset

The dataset contains a movie corpus with a collection of 5.562 movie script files, and movie poster files. Furthermore, there is a metadata file contains information about movies, such as IMDB_ID, Title, Genre, MPAA_rating, Actors, Poster source, Director, and year [1]. The movie scripts file contains the whole conversations among characters which was saved in a separated file. There are five MPAA rating types in this dataset that are G, PG, PG-13, NC-17 and R. Figure 1 shows the distribution of these ratings.

3.2 The Emotional Extraction Model

The DistilRoBERTa-base model is a faster and smaller version of the RoBERTa [10] transformer model that is designed to retain most capabilities of the RoBERTa model, while reducing computational complexity and inference time. The distillation technique consists of training a smaller model to replicate the functioning of a larger model, allowing for effective deployment in resource-constrained settings. The DistilRoBERTa-base emotion model is fine-tuned on annotated emotion recognition datasets such as GoEmotions [11]. It contains 27 emotion categories, or six emotion types plus a neutral class. The fine-tuning enables the model to categorize input text into specific emotions by leveraging the contextual embeddings generated by the DistilRoBERTa architecture. This model balances between performance and efficiency that make it suitable for large-scale emotion detection or real-time tasks in applications, such as movie script analysis. Combining emotion features with other data into multimodal frameworks enhances the understanding of narrative tone in the used data to improve the performance in tasks such as MPAA rating prediction.

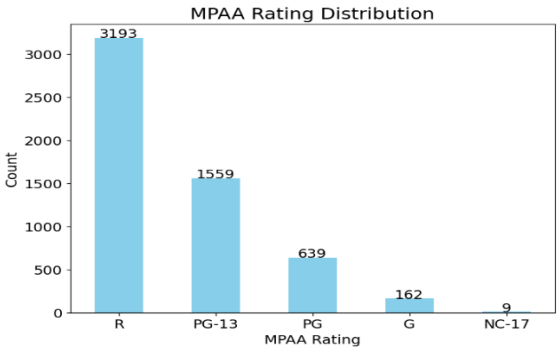

Figure 1: MPAA_rating distribution over movies.

3.3 Textual Feature Extraction

This research uses simple method to select script features by using TF-IDF technique.TF-IDF is the most known feature extraction technique that converts text data, such as summaries or scripts, into numerical vectors suitable for machine learning models. TF-IDF highlights terms that are most informative [12] for distinguishing between movie rating. For instance, words such as "kill" or "blood" may indicate a strong signal for R rating of films. Other words such as "school" or "friendship" are more indicative of G or PG rating. The general terms such as movie or story are assigned less weight because they provide little information regarding age suitability.

The ability of capturing specific vocabulary related to various rating types makes TF-IDF an effective representation for machine learning classifiers such as SVM, Random Forest, and ensemble models, thereby improving predictive accuracy in multi-label movie genre classification tasks. To compute TF-IDF as Equation (3), two variables should be calculated: TF based on (1) and IDF from (2) as in [13].

$$TF_{t,d} = \frac{f_{t,d}}{\sum_{t \in d} f_{t,d}}, \quad (1)$$

where, $f_{t,d}$ is the number of present of term t in document d, $\sum_{t \in d} f_{t,d}$ indicates the overall count of terms t within document d

$$IDF_{(t,D)} = log\ \frac{N}{|\{d \in D : t \in d\}|}, \quad (2)$$

where N = total number of documents d in corpus D and $|\{d \in D : t \in d\}|$ denotes the count of documents that include the term t.

$$\text{TF-IDF} = \text{TF} * \text{IDF}. \quad (3)$$

3.4 LightGBM Model

LightGBM [7] stands for Light Gradient Boosting Machine, which is a type of gradient boosting framework developed by Microsoft. LightGBM has two significant techniques that make it outperforms the traditional XGBoost model in the memory consumption and processing speed. The first technique is Gradient-based One-Side Sampling (GOSS) that emphasizes selecting samples using their gradients, giving priority to those with greater gradients during training. Exclusive Feature Bundling (EFB) is the second one that enhances memory utilization by grouping unique features, enabling more effective processing and quicker training, while maintaining performance levels substantially. It effectively handles a substantial number of data instances and a large variety of features. Therefore, these two features make the LightGBM model much quicker and more scalable than traditional GBM.

4 THE PROPOSED MODEL

This study presents a multi-class classification model for predicting four types of MPAA rating categories. This is based on fusion extracted emotional and script features using a powerful machine learning model. Before that, the data must pass through a preprocessing stage to introduce clear data for use in the suggested model.

4.1 The Preprocessing Stage

Preprocessing is a crucial step in any Natural Language Processing (NLP) tasks that affect a model's performance [14]. To use the movie script for subsequent analysis, text normalization and cleaning procedures are needed. Many cleaning procedures were utilized in this research, such as is lowercasing, stopwords removal and alphabetic filtering. The lowercasing process converts the entire input script into lowercase alphabetical to ensure consistency in letter casing, thereby minimizing sparsity in the token space. Another preprocessing procedure is the elimination of stopwords, which discards common English stopwords such as “and”, “is”, and “the”. This method was implemented using the NLTK library. Lastly, an alphabetic filtering process was applied to keep only tokens made up of purely alphabetic characters. This helps remove numbers, punctuation, and other non-letter symbols that usually do not add meaningful semantic value.
There is another preprocessing step related to the imbalance state of the class label. This involves identifying and removing the rare labels. This can potentially improve the performance of the learning model by reducing noise and preventing overfitting on the underrepresented class. Based on Figure 1, which illustrates the distribution of five MPAA rating labels, the NC-17 rating class was removed since it had low occurrence frequency over other rating.

4.2 The Classification Stage

The proposed model is illustrated in Figure 2. It is based on three stage to achieve the goal of predicting MPAA rating. The first one is related to feature extraction from TF-IDF after cleaning the textual script during the preprocessing stage. The spaCy English language library, which is considered a natural language processing was used in the text cleaning procedure to standardize the movie script before the feature extraction stage. The cleaning process converts all script texts into lowercase, removes unnecessary symbols, and tokenizes the content into linguistically meaningful units. The tokens are then rejoined to form coherent sentences, while correcting spacing around punctuation marks to preserve natural text structure. This procedure ensures that the scripts are linguistically consistent, noise-free, and properly formatted for TF-IDF and emotion feature extraction.

In this study, the number of TF-IDF features was set to 12,000, aiming to capture as many as possible representative and informative terms from the movie scripts. This dimensionality was chosen to balance between computational efficiency and feature richness, ensuring that both common and distinctive

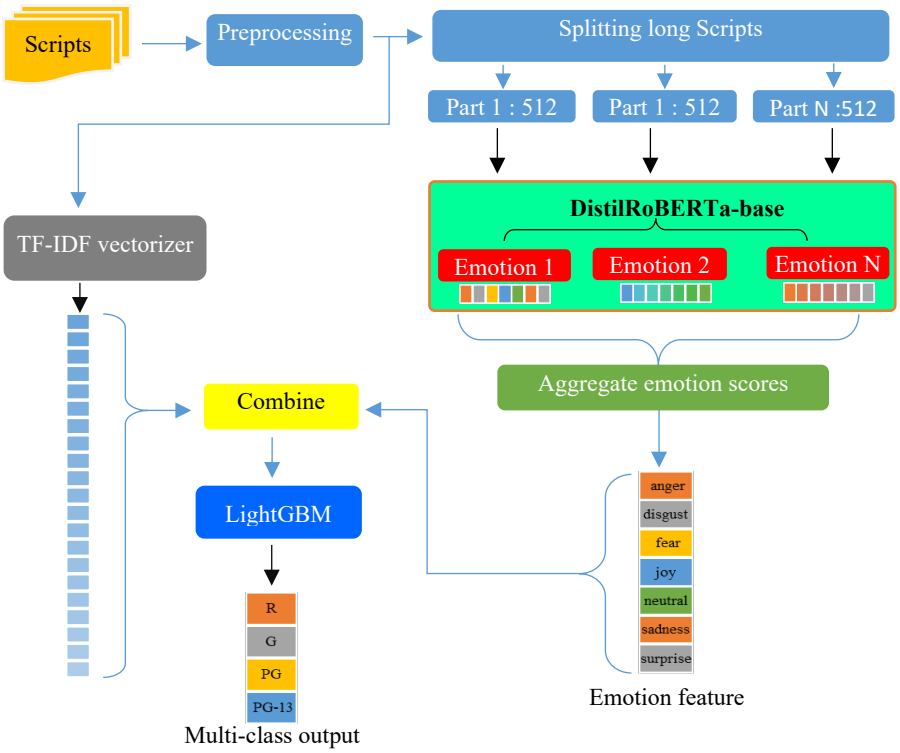

Figure 2: The proposed MPAA rating classification model.

linguistic patterns contribute to MPAA rating prediction. The second one is associated with the emotional feature extraction, that was accomplished using a trained DistilRoBERTa-base model. It was published on the Hugging Face website under the "j-hartmann/emotion-english-distilroberta-base" title. The model was trained over the GoEmotions dataset to predict six emotions namely, "anger", "disgust", "fear", "joy", "sadness" and "surprise", and "neutral" type.

The Transformer-based emotion recognition is designed to capture contextual emotional nuances in text by leveraging deep semantic representations rather than relying on keyword-based emotion lexicons. The maximum input sequence that the DistilRoBERTa model can process is 512 tokens, where the medium token length in the used script exceeds 2000 tokens based on the calculation sequence size.

In this state, this model was unable to extract the emotion features from the entire script. Therefore, we proposed a method to address this limitation by using the chunking method. It splits the long-cleaned script into small chunks with 512 tokens, where every chunk is passed to the emotional model to extract the emotion vector for that chunk. This process is repeated to complete the remaining chunks. The individual emotion features per chunk are aggregation together to produce the final vector that represents the emotion tone in the script.

This emotion feature vector was combined with the TF-IDF vector to form the final feature vector that is fed into the LightGBM classifier. The data was split into 80% for training and 20% for testing to ensure that the model was trained on the majority of the samples, while preserving a separate portion for unbiased evaluation. Since the dataset had an imbalanced class distribution, as shown in Figure 1, creating the necessity to address this issue, which can be model-based or data-based. This study utilizes a class weights strategy, which is a type of model-based approach to tackling class imbalance. Typically, models tend to be biased towards the majority classes since the loss function influences heavily by the frequent labels. When employing the class weights technique, the loss for minority classes increases and this, in turn, leading the model to focusing more on rare categories during the training stage. This modification enhances the capability of the model to classify rare labels accurately and results in more balanced performance across all classes, especially in application such as MPAA rating

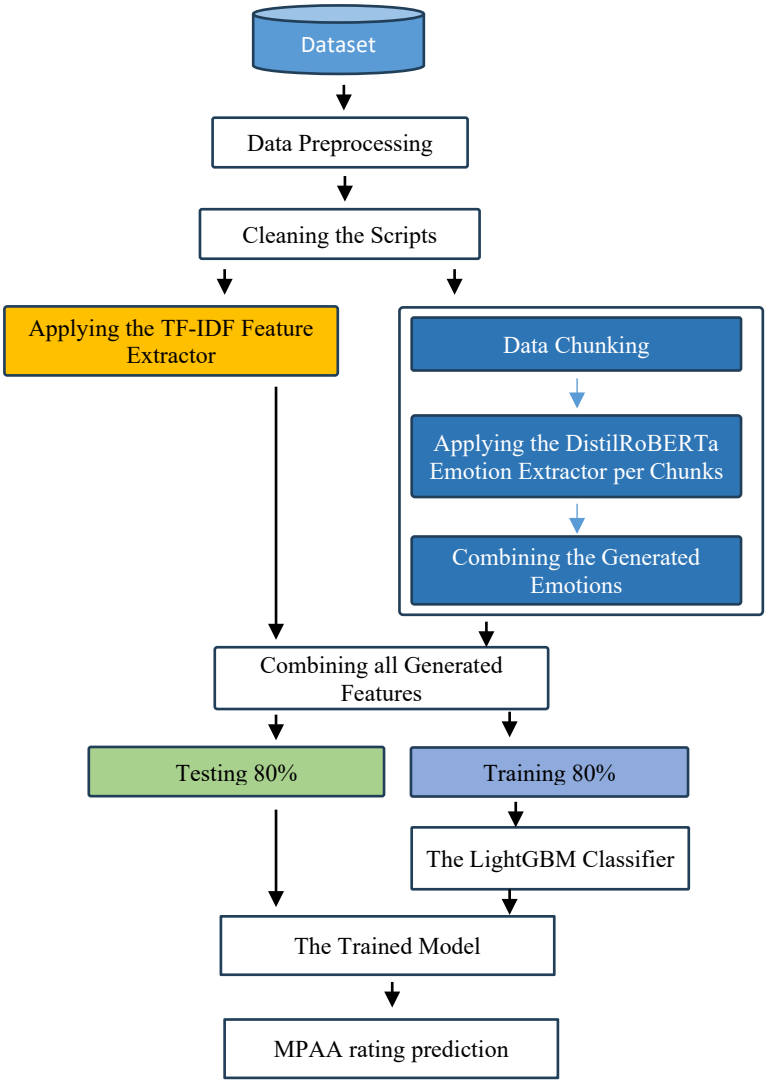

Figure 3: The entire model flowchart.

prediction. In this system, some categories may be significantly less common than others. After completing the training process with four multi-class labels, the LightGBM model can predict the MPAA rating of a given script by applying the SoftMax activation function on the output probabilities. The class with the highest probability is selected as the final predicted rating. This can be either G, PG, PG-13, or R. The overall workflow of the proposed MPAA rating prediction model is illustrated in Figure 3. It presents the main processing stages that begin from data preprocessing and chunking, until TF-IDF feature extraction and emotion representation using the DistilRoBERTa model. The extracted features are then fused and passed to the LightGBM classifier to predict the final MPAA categories.

5 RESULTS

The proposed model was tested with four machine learning models namely, Support Vector Machine (SVM), Naive Bayes (NB), Random Forest (RF) and XGBoost. Moreover, the findings of other studies were compared with the proposed model to demonstrate its efficiency. Both evaluation metrics, Accuracy and the Weighted F1-Score [15], were used to comprehensively evaluate the model's performance in the multi-class

classification task. Accuracy measures the overall proportion of correctly classified samples, while the Weighted F1-Score accounts for class imbalance by averaging the F1-scores of each class according to their support. Achieving high accuracy and a strong weighted F1-Score indicates reliable predictions across all classes. This demonstrates that the model not only performs well overall but also maintains balanced performance for both the majority and minority classes.

Some studies relied on the weighted F1–Score metric [16] This is based on the normal F1 – Score. This metric is presented in (4). It can provide more weight to classes that have more samples.

$$F1_{\,weighted} = \sum_{i=1}^{k} \frac{n_i}{N} . F1_i \,, \qquad (4)$$

n_i = number of true samples in class i , $F1_i$ = F1-score for class i , N = total number of samples.

Table 1 shows the model performance using various numbers of selected TF-IDF features. This is starting from 10000, 12000, and 15000 vector features. The model was evaluated using 5000 features, but it achieved a lower result than 10000 features. The LightGBM model outperformed other models. The weighted F1-Score, and accuracy were 84.2, 84.6, respectively.

Table 1: Evaluation of TF-IDF feature variants.

Method	10000 TF-IDF		12000 TF-IDF		15000 TF-IDF	
	weighted F1–Score	Accuracy	weighted F1–Score	Accuracy	weighted F1–Score	Accuracy
SVM	68.1	66.5	68.3	66.6	68.5	66.9
Naive Bayes	42.2	57.7	42.2	57.7	42.2	57.7
Random Forest	70	72.7	69.8	73.1	66.3	70.2
XGBoost	82.2	82.2	83.8	83.3	83.4	83.5
LightGBM	83.1	83.3	**84.2**	**84.6**	83.9	83.9

Table 2 illustrates the model performance in comparison with previous studies. The best result was achieved by the suggested model, except for the findings presented in [9]. However, it should be clear that the output of [9] is binary classification only. The accuracy value was 84.6 after integrating emotional features, while the weighted F1-Score metric was 84.2. The model was tested without using the emotional feature to identify the contribution of utilizing the emotion features. Its overall results were 83.7 and 83.8 for both weighted F1-Score and accuracy metric, respectively. This result was obtained when using the top 12000 features during the calculation TF-IDF feature, where it represents the best value that was selected based on experimental results.

Table 2: Performance comparison with literature.

Study	Method	weighted F1–Score	Accuracy
[8]	RNN	78	-
[1]	LSTM + Glove	81.6	-
[9]	CNN-LSTM + GMU	86.06	-
[3]	TF-IDF + ResNet, Inception	-	81.1
[4]	bidirectional LSTM	-	56
Our model	TF-IDF , LightGBM	**83.7**	**83.8**
Our model	TF-IDF + emotion , LightGBM	**84.2**	**84.6**

6 CONCLUSIONS

Movie ratings prediction plays a crucial role to ensure appropriate content for audience. It provides guidance for parents and audience by categorizing films according to their thematic content, language, and level of violence. This makes the importance of creating a model to classify movie ratings automatically.

This research presented a multi-class classification model for predicting four MPAA ratings based on a fusion approach. It integrated the script features extracted from TF-IDF and emotion features obtained from the DistilRoBERT model. This study demonstrated that Transformer-based emotion features enhance MPAA rating prediction beyond traditional method such as NER. Due to the limitation of sequence length processing, this present study introduced a new method based on a chunking method to address this constraint. It enabled the model to acquire the emotion tone from the entire script length. TF-IDF vectors, along with the emotion vector, were combined and passed into the LightGBM classifier to predict the height of movie ratings. Based on the results, the combination approach showed an enhancement in the prediction of MPAA ratings by achieving high accuracy and weighted F1-score.

The class weight technique was utilized to handle the imbalance state of labels, where some classes had far fewer samples than others. The presented results showed that integrating the

emotional features enhanced the performance of the LightGBM model. Furthermore, the suggested model result was compared with earlier literature based on the evaluation metric. This research findings outperformed previous studies with 84.2 and 84.6 for weighted F1–score and Accuracy, respectively.

Regardless of the research findings, some limitations should be addressed to open the door for future research. First, a multimodal approach can be applied to multiple data types. This may help generalize the proposed model. Moreover, applying more methods to address the class imbalance may enhance the model's performance. Finally, the proposed model should be applied on another rating way to improve its efficiency and applicability.

REFERENCES

[1] M. Shafaei, N. S. Samghabadi, S. Kar, and T. Solorio, "Age suitability rating: Predicting the MPAA rating based on movie dialogues," in LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings, 2020.

[2] L. Deni Setiawan and B. Bestari Puspita, "classification of children short films for mobile movie screening by bioscil," 2019. doi: 10.17501/24246778.2019.5101.

[3] L. A. Ha and E. Mohamed, "Combining Text and Images for Film Age Appropriateness Classification," in Procedia CIRP, 2021. doi: 10.1016/j.procs.2021.05.087.

[4] R. Jayashree and A. Nayan Varma, "MPAA Rating Prediction Using Script Analysis for Movies," in 2022 IEEE 7th International conference for Convergence in Technology, I2CT 2022, 2022. doi: 10.1109/I2CT54291.2022.9825434.

[5] V. R. Martinez, K. Somandepalli, K. Singla, A. Ramakrishna, Y. T. Uhls, and S. Narayanan, "Violence rating prediction from movie scripts," in 33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, 2019. doi: 10.1609/aaai.v33i01.3301671.

[6] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," arXiv preprint arXiv:1910.01108, 2019.

[7] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in Adv. Neural Inf. Process. Syst., vol. 30, 2017.

[8] M. Shafaei, N. S. Samghabadi, S. Kar, and T. Solorio, "Rating for Parents: Predicting Children Suitability Rating for Movies Based on Language of the Movies," arXiv preprint arXiv:1908.07819, Aug. 2019, [Online]. Available: http://arxiv.org/abs/1908.07819

[9] M. Shafaei, C. Smailis, I. A. Kakadiaris, and T. Solorio, "A Case Study of Deep Learning Based Multi-Modal Methods for Predicting the Age-Suitability Rating of Movie Trailers," arXiv preprint arXiv:2101.11704, Jan. 2021, [Online]. Available: http://arxiv.org/abs/2101.11704

[10] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv preprint* arXiv:1907.11692, 2019.

[11] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "GoEmotions: A dataset of fine-grained emotions," in Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.372.

[12] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," 1972. doi: 10.1108/eb026526.

[13] S. W. Kim and J. M. Gil, "Research paper classification systems based on TF-IDF and LDA schemes," Human-centric Computing and Information Sciences, vol. 9, no. 1, 2019, doi: 10.1186/s13673-019-0192-7.

[14] M. Siino, I. Tinnirello, and M. La Cascia, "Is text preprocessing still worth the time? A comparative survey on the influence of popular preprocessing methods on Transformers and traditional classifiers," Inf Syst, vol. 121, 2024, doi: 10.1016/j.is.2023.102342.

[15] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," Inf Process Manag, vol. 45, no. 4, pp. 427–437, Jul. 2009, doi: 10.1016/j.ipm.2009.03.002.

[16] C. Goutte and E. Gaussier, "A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation," in Lecture Notes in Computer Science, 2005. doi: 10.1007/978-3-540-31865-1_25.