Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 1327–1335
Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

Age Suitability Rating: Predicting the MPAA Rating Based on Movie Dialogues

Mahsa Shafaei, Niloofar Safi Samghabadi, Sudipta Kar and Thamar Solorio
Department of Computer Science, University of Houston
{mshafaei, nsafisamghabadi, skar3, tsolorio}@uh.edu

Abstract

Movies help us learn and inspire societal change. But they can also contain objectionable content that negatively affects viewers' behavior, especially children. In this paper, our goal is to predict the suitability of movie content for children and young adults based on scripts. The criterion that we use to measure suitability is the MPAA rating that is specifically designed for this purpose. We create a corpus for movie MPAA ratings and propose an RNN-based architecture with attention that jointly models the genre and the emotions in the script to predict the MPAA rating. We achieve 81% weighted F1-score for the classification model that outperforms the traditional machine learning method by 7%.

Keywords: Text Classification, MPAA Rating, Story Analysis

1. Introduction

The latest reports on screen time among children show an alarming trend in the amount of time they spend watching movies or videos on electronic devices (Chen and Adler, 2019). Some of the content in these movies and videos is completely innocuous, but there is also harmful and inappropriate content that can negatively affect their behavior. For example, watching specific programs may encourage irresponsible sexual behavior and alcohol usage in teenagers (Strasburger, 1989; Sargent et al., 2006; AAP, 2001) or instill anxiety and fear among children (Wilson, 2008; Johnson et al., 2002).

The Motion Picture Association of America (MPAA)[1] is a film rating system that establishes the appropriate age for movie viewers. MPAA ratings have a wide practical value. For example, parents can rely on them as a guideline to determine what movies are appropriate for their children. Also, media service providers (e.g., Amazon and Netflix) may use these ratings to enable age filters in parental controls. Having the MPAA rating is an important element for producers too. Although films can be shown without a rating, certain theaters refuse to show non-rated movies[2] which in turn negatively affects the potential popularity of the movie as well as its gross revenue.

The MPAA rating is determined by CARA[3] (one of the subdivisions of MPAA organization). Members of CARA watch the entire film to determine the age category and the MPAA rating of the movie (MPAA, 2010). The current rating method is a time-consuming and non-scalable process. Not surprisingly, there are many movies available that are missing an MPAA rating. Also, rating happens post production, when making changes in movies can cost a lot of money.

To solve the aforementioned obstacles, we explore predicting the MPAA rating by only using movie scripts. Our method is a practical solution because it can predict the rating at early steps of the production when we only have the script of the movie. As a result, producers can use the predicted rating to make adjustments in the script based on the desired audience. Furthermore, this system is an efficient way to rate movies that have been released without a rating and is an initial step toward rating online video content (like YouTube videos). Our promising results show that movie scripts include a reasonable amount of information for MPAA prediction.

In this paper, we provide a corpus and a method for predicting the MPAA rating. We build a deep neural model to jointly model conversations between characters, genres of the movie, and emotions conveyed within the conversations to predict MPAA ratings. We also explore the notion of similar movies to improve the performance of the model; this model is only applicable to movies that are already released but do not have the MPAA rating. To summarize, this paper presents the following main contributions:

- Propose a novel task: automatic prediction of the suitability of movies for children using the MPAA rating scheme.
- Provide the first corpus of movie scripts with their associated MPAA rating, values of MPAA components, MPAA rating for similar movies, and poster images of movies.
- Establish a strong benchmark for the task; we achieve 81.6% weighted F1-score performance that works 7% better than the traditional machine learning method.

2. Related Work

To the best of our knowledge, there is no previous work on predicting the MPAA rating for movies. However, some research has been conducted on detecting violent content in movies or predicting abusive language and hate speech in online texts.

The closest paper to our work is Martinez et al. (2019). In this study, the authors try to predict if a movie is violent or not using scripts. They extract sentiment, semantic, and lexical features and feed them to an RNN-based classifier to predict violence in movies. Our research is similar to this work because we also use dialogues among movie characters as the input. But, instead of extracting features from

[1] https://www.mpaa.org/film-ratings/

[2] https://en.wikipedia.org/wiki/Motion_Picture_Association_of_America_film_rating_system

[3] Classification and Ratings Administration

text to feed the model, we mostly use raw data to avoid error propagation due to feature extraction (Ning et al., 2019). Also, the outputs of models are different. In this paper, we predict the MPAA rating, not violence (violence is one of the many aspects of the MPAA rating). The dataset introduced by the aforementioned work is not publicly available, while we make our corpus available[4] to enable research in this direction.

There are some other works for violence detection in movies as well. Giannakopoulos et al. (2010) work on this topic by extracting visual and audio features from movies. Authors in (Gninkoun and Soleymani, 2011) build upon (Giannakopoulos et al., 2010) and add textual features in order to capture the ratio of swear words for violence detection. Using video and audio make the system unsuited for prediction before movie production. So, in this paper, we only rely on the script of the movie to do the prediction, and we use a bad word list based system as a baseline to compare with our proposed model.

Based on a survey done on hate speech detection (Schmidt and Wiegand, 2017), several works have been conducted on detecting the offensive language in the text. These works are relevant to our research since the offensive language in dialogues can affect the suitability of movies for children. Researchers in (Singh et al., 2018; Zhang et al., 2018; Park and Fung, 2017; Mathur et al., 2018) adapt Convolutional and Recurrent Neural Networks to predict abusive language and hate speech on Twitter data. (Davidson et al., 2017) and (Nobata et al., 2016) also automatically predict hate speech in online content by extracting lexical, syntactic, sentiment, and semantic features and training traditional machine learning classifiers. In one of our baselines, we also make use of lexical and sentiment features and feed them into an SVM classifier.

3. Dataset

We expand the movie script dataset collected by Shafaei et al. (2019) to include MPAA ratings. The original corpus provides scripts of movies as well as their metadata like name of actors, directors, genre, etc. It should be noted that the scripts only contain conversation between the characters (without the description of the scenes). From all the movies in the original dataset, the MPAA rating is only available for about 7k movies. There are five categories for the MPAA rating (G, PG, PG-13, R, NC-17) that specify the suitability of movies for children. G stands for the general group; it means all ages admitted. PG means that there is some content in the movie that parents should review. PG-13 indicates that the movie has some content deemed not appropriate for children under 13 years old. R stands for "restricted" and means people under 17 should watch the movie with a parent. NC-17 refers to no one under 17 is recommended to watch the movie. The exact definition of these ratings is available at (https://www.mpaa.org). Based on documentation on MPAA rating, this rating had different group names and slightly different meanings for each group before 1996 (Kennedy, 2013; New York, 1981; LIFE, 1969). Therefore, we do not consider movies before 1996 (about 1,500 movies of the corpus) to avoid inconsistency in the definition of the ratings.

Rating	G	PG	PG-13	R	NC-17	Total
#Movies	162	639	1,559	3,193	9	5,562

Table 1: Dataset statistics after considering production year and adding new G movies.

The first version of the corpus includes much fewer G-rated movies compared to other groups (PG, PG-13 and R), so we add 50 more movies in this category to our collection (limited number of scripts are available online). We also have a small number of movies for NC-17. Based on the IMDB website, in total, we have about 70 movies with this rating, probably since the market segment for these movies is limited.[5] Therefore, we ignore movies in this group in our experiments. Table 1 presents the statistics for the final version of our dataset.

The MPAA ratings are determined based on the following components (i) Violence, (ii) Language, (iii) Substance Abuse, (iv) Nudity and (v) Sexual Content, but MPAA does not provide a way to quantify the amount of content related to these components present in a film. However, the IMDB website provides objectionable content of movies compatible with MPAA components (Parental Guide): 1) Violence & Gore, 2) Sex & Nudity, 3) Frightening and Intense Scenes, 4) Profanity, and 5) Alcohol, Drugs & Smoking. The IMDB website also provides a way to rate the severity of the aforementioned types of content, through user votes using the following set of labels: *None*, *Mild*, *Moderate*, and *Severe*. We collect the number of votes for each label per component. Since these tags come from users' votes, not all of the movies with an MPAA rating have these components available. Table 2 shows the data statistics based on these components. Each cell in the table stands for the number of movies that are tagged with a severity-label (None to Severe) for each component (violence, profanity, etc) per MPAA category (G, PG, PG-13, R, NC-17). For instance, we have four movies in category G that are tagged as *None* for *violence*. And, the reason is the majority of the voters voted for *None* (maybe not all of them).

The IMDB website provides a field name "More Like This" for each movie. It includes the movies that share a similar aspect (like the genre) with the target movie. In our dataset, we have up to 12 most similar movies for each movie along with the corresponding MPAA rating for them. Using the IMDB website, we also add high-quality poster images for all movies in the corpus.

In Table 3, we show the distribution of movies across different genres. Class imbalance (in terms of genre) exists between different groups since some genres are more popular in the movie industry. We do not fix this issue to keep the dataset representative of the real-world situation.

4. Methodology

This paper aims at classifying movies into one of the MPAA ratings based on the content of the movie. Predicting the

[4] http://ritual.uh.edu/LREC2020/

[5] https://www.theguardian.com/film/1999/jul/25/2

	Violence				Profanity				Nudity				Frightening				Alcohol				Total
Rating	N	Mi	Mo	S	N	Mi	Mo	S	N	Mi	Mo	S	N	Mi	Mo	S	N	Mi	Mo	S	#
G	4	**24**	4	0	**20**	12	0	0	**22**	8	1	1	5	13	13	0	**19**	13	0	0	34
PG	100	**280**	74	11	164	**268**	31	7	**247**	210	14	8	101	**242**	91	16	201	**209**	28	16	502
PG-13	190	454	**539**	66	102	**620**	522	25	344	**687**	269	11	271	395	**451**	77	213	**790**	191	33	1,340
R	185	572	781	**863**	61	370	986	**997**	411	**900**	878	421	312	523	**772**	585	195	**1061**	724	294	2,681
NC-17	1	3	3	2	0	0	6	3	0	0	2	7	3	1	3	1	0	3	4	0	9

Table 2: Statistic of votes for each severity tag per MPAA category (N = None, Mi = Mild, Mo = Moderate, S = Severe). The most frequent severity-tag for each component are in bold

Genre	**#**	**Genre**	**#**
Science-Fiction	619	Action	1,277
Horror	800	Animation	296
Crime	1,000	Adventure	806
Romance	1,082	History	216
News	8	Western	78
Comedy	1,999	War	214
Thriller	1,785	Short	14
Mystery	618	Biography	374
Musical	297	Drama	2,965
Documentary	189	Family	552
Sport	220	Fantasy	558

Table 3: Distribution of movies in each genre.

MPAA rating using the scripts is not a trivial task. This rating is a combination of several content elements related to drugs, sex, violence, language, sensitive themes, etc. As a result, a naive method like employing a list of bad words is not sufficient to predict the MPAA rating (since this rating is more comprehensive and covers other aspects, not just offensive language). To the best of our knowledge, there is no previous work on predicting the MPAA rating. Existing work, such as (Martinez et al., 2019; Acar et al., 2013; Penet et al., 2011) has a more narrow focus on detecting violence in movies, and according to the studies from Jenkins et al. (2005) and Webb et al. (2007), violence prediction is not enough to predict the MPAA rating either.

To address the challenges in this task, we use different types of resources in our model, including conversational data, emotional dynamics between characters, genre of the movies, and similar movies to the target movie. Most of these resources, like the script of the movie or metadata information, are available since the early steps of the production. However, similar movies are not in hand until the movie is released. So, depending on the time that we need the prediction, we can use different models. For example, for predicting the MPAA rating of an unrated released movie, we can use the model that takes advantage of similar movies to have a better prediction.

Figure 1 illustrates the overall architecture of the proposed model. The model consists of: 1) an embedding layer to convert the words into the vector representation, 2) a long short-term memory (LSTM) layer to learn the spatial dependency of the words, 3) an attention layer to find the importance of each word in the sequence, 4) emotion, genre and similarity vectors to add contextual information to the model, and finally 5) a prediction layer. We will get into the details of the model in the following sections.

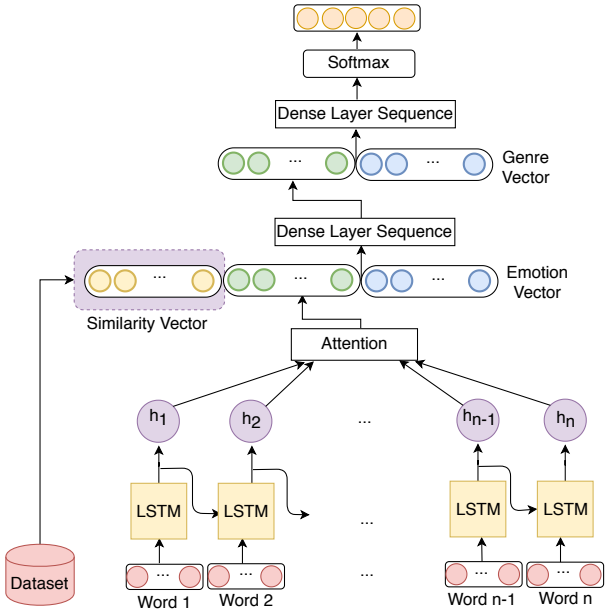


Figure 1: Overall architecture of the proposed model (the similarity vector is appended in case of late prediction; when similar movies are available for the target movie).

4.1. Embedding Layer

Word embedding is an effective representation learning method for text classification as it is capable to capture semantic information of the text. The input of the model is an embedding layer that gets a vector of word indexes $[I_1, I_2, ..., I_{10000}]$, and the output of the layer is a 2-D matrix $[[v_{1,1}, ..., v_{1,j}], [v_{2,1}, ..., v_{2,j}], ..., [v_{n,1}, ..., v_{n,j}]]$; each vector $[v_{i,1}, ..., v_{i,j}]$ is the embedding representation for the corresponding word i. We use 300 dimensional pre-trained Glove embedding to initialize this module.[6]

4.2. LSTM Layer with Attention

To capture the context of each word, we extract the sequential information from the scripts using the LSTM layer. This layer transforms a sequence of embedded vectors into a sequence of hidden vectors. Then we pass the resulting hidden representation to the attention mechanism. We use the same attention model as Bahdanau et al. (2014). This layer computes the weighted sum r as $\sum_i \alpha_i h_i$ to aggregate hidden layers of LSTM to a single vector. The model can learn the relative importance of hidden states (h_i) by learning the α_i. We compute α_i as follows:

$$\alpha_i = \text{softmax}(v^T \tanh(W_h h_i + b_h)) \quad (1)$$

[6] https://nlp.stanford.edu/projects/glove/

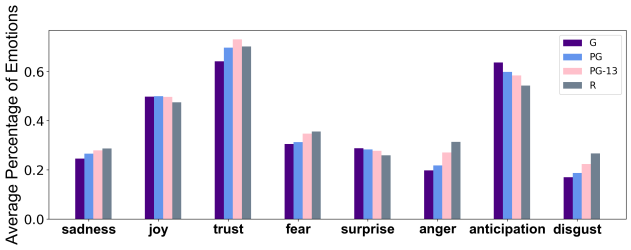


Figure 2: The distribution of MPAA categories across various emotions in the training set. The y-axis shows the average value for the emotion scores for all movies in each rating.

where W_h is the weight matrix, and v and b_h are the parameters of the network.

4.3. Emotion Vector

The emotion in movie dialogues can help the model to better contextualize conversations among characters, and also better discriminate movies belonging to different rankings. For example, we expect G movies (the most suitable movies for children) to contain less content related to "fear" or "disgust" and instead include more "joy" and "happiness". To extract emotion from the text, we use NRC emotion lexicon (Mohammad, 2011). This dictionary maps words to eight different emotions (anger, anticipation, joy, trust, disgust, sadness, surprise, and fear) and two sentiments (positive and negative) with binary values. We calculate the normalized count of words per emotion over the whole movie. As a result, we have a vector $[e_1, e_2, ..., e_{10}]$ for each movie, where e_i is the percentage of words corresponding to emotion i. To show the truthfulness of our hypothesis about the emotion, we illustrate the average emotion scores per class (for movies with the same MPAA rating, we average all the scores per emotion). According to Figure 2, some of the emotions are more dominant in a specific class of movies. For example, the values of negative emotions like *disgust*, *anger*, and *fear* are higher for movies rated for older audiences. While, *anticipation* and *surprise* show higher rates in the G category compared to other classes.

Moreover, we show the validity of our claim through some sample movies from the training set. We select a random sample of G-rated movies ("College Road Trip") and an R-rated one ("Gernika"). Then, we sort sentences of these movies based on average emotion score over the words (we ignore sentences with less than four words). Table 4 shows that the R-rated movie has more intense sentences for *anger* and *fear*, while the G-rated movie has stronger sentences for *joy* and *surprise*.

These trends bode well with our assumption that emotion vectors could help to improve the task. Therefore, to integrate emotion information into the model, we concatenate the emotion vector with the attention output.

4.4. Genre Vector

The genre can provide information about the theme of the movie. For example, some dialogues are considered violent in a specific genre, but they are harmless in another genre (an action movie vs. a sport movie) (Martinez et al., 2019).

Emotion	Rating	Sentence	Score
Joy	R	I'll be glad to	0.062
	G	I love you, beautiful	0.090
Anger	R	Hit him, hit him	0.110
	G	Mom, this is crazy!	0.052
Fear	R	He was about to cross enemy lines	0.085
	G	We got a police emergency	0.076
Surprise	R	I'll deal with it	0.055
	G	Road trip, road trip!	0.952

Table 4: Sample of top rated sentences for some emotions(joy, anger, fear, surprise) for a G-rated ("College Road Trip") and an R-rated ("Gernika") movie.

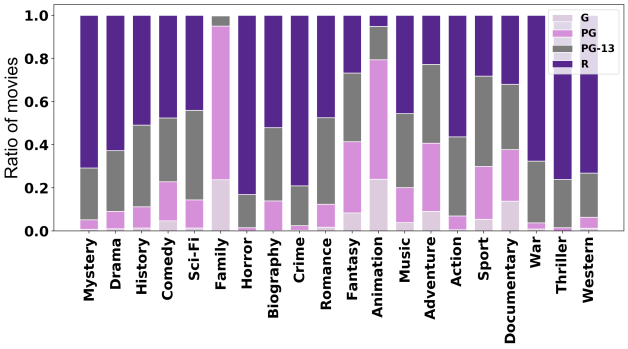


Figure 3: MPAA ratings distribution per genre in the training set; number of movies for each genre is normalized.

So, we exploit this information by adding a genre vector to the model. We have a total of 24 genres across our corpus, but some of the movies are assigned to several genres. Thus, we form a binary multi-hot vector for modeling the genres of a movie. Each cell of the vector represents a genre, and its value is one if that genre is one of the movie's genres; otherwise, it is zero.

To shed some light on the effect of genre, we show the distribution of various MPAA ratings across different genres for the training data (Figure 3). According to this figure, some genres are more appropriate for children compared to others. For example, MPAA ratings of *Animation*, *Adventure* and *Family* show that movies in these genres are more suitable for children compared to *Drama*, *Horror*, and *Crime*.

4.5. Similar Movies Vector

As we mentioned in Section 3, IMDB introduces a list of similar movies for each movie. The similarity is calculated by IMDB based on several factors that include genre, country of origin, and actors. Intuitively, we can say two similar movies may have a close MPAA rating as well. We can thus leverage this information when available. For this purpose, we generate a five-dimensional vector for each movie. The vector is $[p_G, ..., p_{NC-17}]$ and p_i shows the percentage of similar movies with the MPAA rating equal to $i \in \{G, PG, PG-13, R, NC-17\}$.

4.6. Dense Layer and Output Layers

We further use two dense layers to fine-tune the information after concatenating the vectors. We use batch normalization and dropout rate after the hidden layer to avoid over-

fitting. Since we have a multi-class classification, in the final step, we use the softmax activation function to calculate the probability of each group (G, PG, PG-13, R).

5. Experiments

As our data is imbalanced, we use the random stratified sampling and split data to 80% training, 10% development, and 10% test set (for all experiments we use the same train, validation, and test data). To smooth the imbalance problem, we employ class weights in the loss function. The metric we use to report the performance of models is the weighted F1-score.

5.1. Baseline Systems

In this section, we define several baselines to evaluate the performance of our proposed model.

Threshold Model: Our first baseline only considers bad words that have been used in the movie scripts. To create a bad word list, we compiled an online list[7] and combined it with words listed in (Hosseinmardi et al., 2014). Using this list, we calculate the percentage of bad words in each movie. Then, we find the best thresholds for the percentage of bad words among all threshold values {0.0001, 0.0002,...,0.05} by performing grid search on the validation set. The final model will be a list of thresholds; all movies with bad words less than t_1 are labeled as G, all movies with bad words between t_1 and t_2 are labeled as PG, etc. The intention behind this baseline is to show that having a bad word list is not enough to decide about the suitability of movies for children.

SVM Model: The second baseline is similar to the model proposed by Shafaei et al. (2019) for movie success prediction. The best set of features for the model is a combination of unigram, bigram, bag-of-genres and bag-of-directors. We also add the emotion vector to the feature set to have a fair comparison with our deep learning model. We tune hyper-parameter C of the SVM model, C, using grid search method $\in \{1, 10, 100, 1000\}$.

Martinez'19 (Martinez et al., 2019): As we mentioned earlier, this work has the state-of-the-art result for violence prediction in movies using only scripts and metadata. The dataset in this research is not publicly available, however the code is published.[8] We apply the same model to our dataset and report the result as another baseline.

SVM+Similarity: The last baselines use MPAA ratings of similar movies. Since the IMDB website does not explain all the factors for similarity metric, we may assume that one of the dominant factors is MPAA rating. And, having these ratings of similar movies in the model makes the problem trivial. To show that this assumption is not correct, we present a baseline model that only uses the average MPAA rating of similar movies to predict the MPAA rating for the target movie. Also, we add this average rating to baseline 2 (SVM model) to have a fair comparison between the traditional and the deep learning model.

[7] https://code.google.com/p/badwordslist/downloads/detail?name=badwords.txt

[8] https://github.com/usc-sail/mica-violence-ratings/tree/master/experiments

5.2. Experimental Setup

We use Pytorch to implement our model. To tune hyper-parameters, we run experiments on validation set for the model with different learning rates {0.00001, 0.0001}, number of LSTM's hidden units {32, 64, 128, 256}, and dropout rates {0.3, 0.4, 0.5}. Also, to avoid over-fitting, besides the dropout, we use L2 regularization. We use binary cross-entropy loss function in order to calculate the loss between predicted and actual labels and employ Adam as the optimizer (Kingma and Ba, 2014). Best performance obtained by the following set of parameters: dropout = 0.3, learning rate = 0.00001, LSTM hidden units = 256. We train over 100 epochs and consider the model with the best weighted-F1 score on the validation set as the final model to apply on the test set.

6. Results

Quantitative Results: Table 5 shows the classification results for predicting the MPAA rating of movies in terms of weighted-F1 score. To disentangle the contributions of genre and emotion vectors to the performance, we experiment with our proposed LSTM with Attention architecture (L&A model) without using genre and emotion information. We also investigate the contribution of each vector to the results by separately adding them to the model (L&A with emotion and L&A with genre).

Models	F1-Score
Baseline 1- Threshold model	65.89
Baseline 2- SVM	74.29
Baseline 3- (Martinez et al., 2019)	75.06
LSTM with Attention layer (L&A)	78.30
L&A with genre	79.49
L&A with emotion	78.94
L&A with emotion+genre	**81.62**
Baseline 5- SVM (Only Similarity)	57.49
Baseline 6- Baseline 2+Similarity	77.70
L&A with emotion+similarity	83.26
L&A with similarity	80.53
L&A with genre+similarity	81.26
L&A with emotion+genre+similarity	**83.68**

Table 5: classification results in terms of weighted F1-score for four-class classification.

The best result for early prediction (without using similar movies information) is achieved by our proposed "L&A with emotion+genre" model. The weighted F1-score for this model is 81.62% which is 7.3% higher than the traditional machine learning model. It also outperforms "(Martinez et al., 2019)" and "Threshold" baselines by 6.56% and 15.73% respectively. Based on the results, both genre and emotion vectors improve the performance of the plain LSTM model with attention. These results support our assumption on the relevance of emotion and genre modeling for the task of predicting the MPAA rating. In order to have a better understanding of the results, we show the confusion matrix of our best model in Table 6. Based on the matrix, our model is able to predict only a few instances of G-class correctly, even though we add class weights to the loss function to smooth imbalanced data problem (we only

Tags	R	PG-13	PG	G
R	**282**	37	0	0
PG-13	14	**128**	14	0
PG	1	29	**34**	4
G	1	1	10	**4**

Table 6: Confusion matrix of the best model (L&A with emotion+genre) for predicting MPAA ratings. Rows are target tags and columns are predicted ones.

Genre	Best	OG	Genre	Best	OG
Science-Fiction	74.87	73.0	Action	81.64	62.0
Family	74.01	70.1	Animation	62.75	47.3
Crime	93.15	76.39	Biography	73.12	44.8
Romance	84.59	44.12	Sport	76.62	47.61
Comedy	83.52	60.86	Fantasy	75.08	58.38
War	73.07	33.33	History	76.00	39.69
Horror	87.11	76.39	Documentary	66.39	29.47
Adventure	69.33	55.23	Mystery	86.73	66.94
Musical	74.68	62.83	Drama	80.56	56.44
Thriller	88.70	76.69	Western	63.88	48.07

Table 7: Weighted F1-score for different genres in the test set (Best= using the best model, OG= using only genre as the input).

have 162 instances from class G). However, the model predicts 88.5% of R movies correctly. And all of the mistakes in this category are predicted as PG-13, which is the closest group to R.

For those cases that we have similar movies, we can make a more accurate model. The similarity metric improves the performance of the SVM model, yet it does not work better than the deep-learning model. Using emotion, genre, and the average value of the MPAA rating of similar movies in L&A model, the model achieves 83.68% weighted F1-score and outperforms the corresponding SVM model by 5.98%.

Although genre can help the model to improve the performance, it is not enough as a single source to predict the MPAA rating. Table 7 shows the weighted F1-score for different genres using our best model. Based on this table, for genres like *Comedy* and *Drama*, which contain movies with different MPAA ratings, the performance is better than *Family*, *Western* and *War*, even though most of the movies in these genres (*Family*, *Western* and *War*) belong to a single rating. For genres like *Crime*, *Horror*, and *Thriller* (that for the most instances have one rating), the performance is better than other genres (93%, 87%, 88% respectively), but not that far from genres like *Romance*, and *Action* (with 84%, and 81% respectively) that contain movies with more varied ratings. Also, if we only use genre as the input, the performance decreases in all genre categories, and there is no relation between the amount of reduction and how single-rated a genre is. So, genre by itself is helpful but not the most relevant information to the MPAA rating.

Time effect: Another concern about this task is time. One assumption is that the definition of groups has changed over time, so it might have a significant impact on the model. For example, a movie that was categorized as R 20 years ago, would be considered as PG/PG-13 nowadays (because

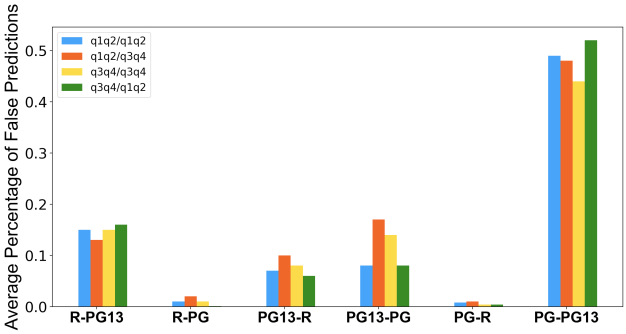


Figure 4: Average percentage of false predictions over all folds. The x-axis shows (true tag-predicted tag), e.g., R-PG13 = 1% means 1% of R movies are predicted as PG-13. Each bar is assigned to training on one quarter and test on other quarters. For example q1q2/q3q4 means we average over training on $q1$ and test on ($q3$ and $q4$) and training on $q2$ and test on ($q3$ and $q4$). q1q2/q1q2 means we average over training on $q1$ and test on $q2$ and training on $q2$ and test on $q1$.

of social changes during the time). In this section, we empirically show that at least in the time span that we have considered here, time has little effect on the overall prediction performance.

We conduct the following experiments using our best model. As we mentioned earlier, our corpus contains movies from 1996 to 2018. We divide the whole dataset into two periods, before and after 2007. The second period includes more movies, but for the sake of fairness, we delete extra movies (we keep equal number of each class for both periods). To have a reliable result, we employ a 2-fold cross-validation method to split data in each period ($q1$ and $q2$ for period one, $q3$ and $q4$ for the second period). Then we report the average result for each experiment. We train our model based on a quarter of data (qi) in one of the periods and test the model based on 1) the other quarter in the same period 2) two quarters in the other period. Since we split the data into four sections, we have very few numbers of G movies in each group, so we do not include the class G in this experiment.

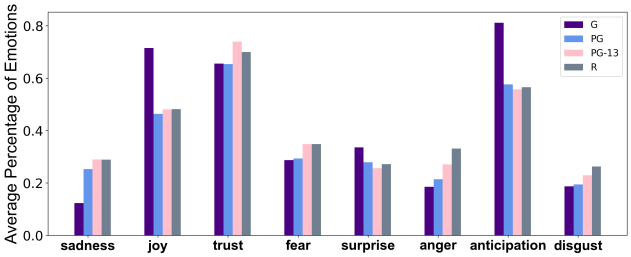
Based on Figure 4, there is no pattern that shows a certain shift in MPAA rating definitions. In case of having significant changes in the definition, we expect that if we train with recent data (e.g., movies after 2007) and test with older movies (e.g., movies before 2007), we see more false prediction towards predicting (PG-13 as PG) and (R as PG-13 or PG). Or the other way around, if we train with older movies and test with recent ones, we expect to see more false predictions toward predicting (PG as PG-13 or R) and predicting (PG-13 as R). But the results show that changes do not have a specific pattern. So, at least empirically, the dataset shows little signs of temporal bias.

7. Analysis

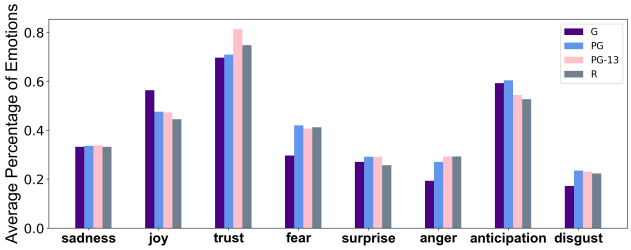
In this section, we provide further analysis of our proposed model. First, we look into the effectiveness of emotion vectors, then we investigate the impact of attention mechanism

ID	Rate	Sentence	Reason	Genre
1	G	what is this nonsense you insolent?	None	Family
2	PG	You want to end up like those bozos?	Mild action, rude humor, some thematic elements and brief scary images	Animation, Adventure, Comedy
3	PG-13	Now here we have a Brazilian tapir; I have to say I've dated better-looking women.	Sexual innuendo and language	Comedy
4	PG-13	You're such a punk-ass bitch	Sexuality including references, drug content, violence and some strong language	Crime, Drama, Mystery, Thriller
5	R	Seismic researchers, my ass! They've excavated something.	Some nudity and language	Adventure, Fantasy, Horror
6	R	Fuck the fucking car	Language, some drug use and violence	Crime, Drama

Table 8: Sentences with the highest attention weights in some sample movies.



(a) Average emotion score per rating for correctly classified samples



(b) Average emotion score per rating for miss-classified samples

Figure 5: Average emotion score of correctly and incorrectly classified movies in the test set per each class.

through some random movie samples. Finally, we analyse the effect of bad words on the MPAA rating.

7.1. Emotion Analysis

To further investigate the effects of the emotion vectors, we compare the histogram of emotion scores for correctly and incorrectly labeled instances.

Figure 5a shows an intuitive pattern for average emotion score in samples that are classified correctly. The class R shows the highest rates for negative emotions like *sadness*, *anger* and *disgust*. But, in the miss-classified samples 5b, we do not observe any clear trends for emotions in different ratings. For example, the PG class shows a high value for negative emotions like *disgust*, *sadness*, and *fear* compared to R, while PG is a more appropriate group for children.

To understand the reason behind this observation, we investigate some samples in the groups PG and R that show high rates of words associated with *disgust*. The results indicate that those words in R films include terms like *robbery*, *murder*, and *asshole*, but in PG-rated movies the words associated with *disgust* include *fool*, *sick* and *painful*. Although

Rate	Sentence
R	Each victim was killed by a punctuate wound at the skull Goddamn fucking asshole. Your wife was murdered.
PG	Lorenz! are you sick? Do you think the memory become less painful then! How terribly awful it all is!

Table 9: Sample sentences of R and PG rated movies that contain "disgust" emotion in the conversation.

we have a high number of words associated with negative emotions in some PG-rated movies, the degree of negativity, or the strength of the emotion, seems to be lower than words in the rated R films (Table 9). Thus, the model has room to improve for these more subtle differences.

7.2. Weight Analysis

In this section, we represent the sentences with the highest attention weights in some random sample movies with a different MPAA rating (Table 8). In these samples, the highest weighted sentences in R and PG-13 movies are more intense compared to PG and G movies. Secondly, based on the genre of the movie, the sentence type is different in the same rating group. For instance, sample 5 and 6 both are "R" movies, but sample 5 is an *Adventure* movie, while sample 6 is a *Crime-Drama* movie. We can see that the words in sample 6 are harsher than sample 5 (sample 5 has inappropriate words, but in a non-aggressive manner). Also, the MPAA association provides a brief explanation for rated movies to justify the rating of a film (available at our dataset). We indicate reasons in the table, and we can see that the highest weighted sentences align with the reasons for these samples.

7.3. Bad Word Ratio

We conduct an analysis over the same bad word list we used in our "Threshold" baseline to further investigate why these words are not enough to predict the MPAA rating of the movies. For each class in our corpus, we merge all the scripts and calculate the frequency of bad words over the same class.

Table 10 shows the top 5 negative words for each class of data. As expected, the ratio of most frequent bad words is different across the classes, but also the intensity of the words is different, and this cannot be captured through a

G	PG	PG-13	R
bad (0.03%)	bad (0.04%)	hell (0.04%)	fucking (0.12%)
hate (0.01%)	die (0.01%)	bad (0.03%)	shit (0.09%)
stupid (0.01%)	kill (0.01%)	shit (0.03%)	fuck (0.09%)
kill (0.009%)	hate (0.01%)	kill (0.03%)	kill (0.04%)
die (0.009%)	stupid (0.01%)	ass (0.02%)	hell (0.04%)

Table 10: Top 5 bad words in each class. The numbers inside the parenthesis show the ratio of the word across all the scripts of the class.

bad word list where all words are assumed to have an equal strength. So, bad words can affect the MPAA rating, but a threshold is not enough to predict the rating with reasonable accuracy. We need to analyze these words in their context to be able to measure the impact. For example, if the word f**k refers to a sexual context, with high probability, it leads to an "R" rating (*"You can f**k me in the car"* in *"to Rome with love"*). But, if it is used as a curse word, the movie can be rated as PG-13. For instance, the sentence *"that's a clear sign to back the f**k off!"* is used in *"Fast & Furious"*, and yet the movie is rated as PG-13. Furthermore, words like *fat* are mostly listed as offensive words in social media comments, but they are less probable to be considered as inappropriate words in movies.

8. Conclusion and Future Work

In this paper, we present the new task of automatic prediction of MPAA rating from movie scripts. We also present a new resource to support the design and benchmarking of machine learning approaches for the task. Lastly, we use a neural network architecture to provide initial results for state of the art comparisons. We model the conversations among characters of the movies, emotions behind the conversations, and genre of the movie in order to predict the film rating. Our best model improves the results compared to the traditional machine learning approach, by 7% weighted F1-score.

In the near future, we plan to explore the use of information from the video. Furthermore, we will design a multi-task model to predict MPAA rating as well as severity levels of relevant aspects for the rating (violence, profanity, nudity). We also plan to extend the approach to other types of online content that are easily accessible to children.

Acknowledgments

We would like to thank reviewers for their valuable feedback. We also thank Jason Ho for relevant comments on a previous draft of this paper.

9. Bibliographical References

AAP. (2001). Media violence. *Pediatrics*, 108(5):1222–1226.

Acar, E., Hopfgartner, F., and Albayrak, S. (2013). Violence detection in hollywood movies by the fusion of visual and mid-level audio cues. In *Proceedings of the 21st ACM International Conference on Multimedia*, pages 717–720.

Bahdanau, D., Cho, K., et al. (2014). Neural machine translation by jointly learning to align and translate. arxiv preprint arxiv: 1409.0473.

Chen, W. and Adler, J. L. (2019). Assessment of screen exposure in young children, 1997 to 2014. *JAMA pediatrics*, 173(4):391–393.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Eleventh International AAAI Conference on Web and Social Media*.

Giannakopoulos, T., Makris, A., Kosmopoulos, D., Perantonis, S., and Theodoridis, S. (2010). Audio-visual fusion for detecting violent scenes in videos. In *Hellenic Conference on Artificial Intelligence*, pages 91–100. Springer.

Gninkoun, G. and Soleymani, M. (2011). Automatic violence scenes detection: A multi-modal approach. *MediaEval 2011, Multimedia Benchmark Workshop*.

Hosseinmardi, H., Han, R., Lv, Q., Mishra, S., and Ghasemianlangroodi, A. (2014). Towards understanding cyberbullying behavior in a semi-anonymous social network. In *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*, pages 244–252. IEEE.

Jenkins, L., Webb, T., Browne, N., Afifi, A. A., and Kraus, J. (2005). An evaluation of the motion picture association of america's treatment of violence in pg-, pg-13–, and r-rated films. *Pediatrics*, 115(5):e512–e517.

Johnson, J. G., Cohen, P., Smailes, E. M., Kasen, S., and Brook, J. S. (2002). Television viewing and aggressive behavior during adolescence and adulthood. *Science*, 295(5564):2468–2471.

Kennedy, M. (2013). *Roadshow!: The Fall of Film Musicals in the 1960s*. Oxford University Press.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

LIFE, M. (1969). Life magazine.

Martinez, V. R., Somandepalli, K., Singla, K., Ramakrishna, A., Uhls, Y. T., and Narayanan, S. (2019). Violence rating prediction from movie scripts. In *Proceedings of the AAAI Conference*.

Mathur, P., Sawhney, R., Ayyar, M., and Shah, R. (2018). Did you offend me? classification of offensive tweets in hinglish language. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 138–148.

Mohammad, S. (2011). From once upon a time to happily ever after: Tracking emotions in novels and fairy tales. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 105–114. Association for Computational Linguistics.

MPAA. (2010). Classification and rating rules.

New York, M. (1981). New york media.

Ning, Q., Subramanian, S., and Roth, D. (2019). An improved neural baseline for temporal relation extraction. *arXiv preprint arXiv:1909.00429*.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., and Chang, Y. (2016). Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, pages 145–153. Inter-

national World Wide Web Conferences Steering Committee.

Park, J. H. and Fung, P. (2017). One-step and two-step classification for abusive language detection on twitter. *arXiv preprint arXiv:1706.01206*.

Penet, C., Demarty, C.-H., Gravier, G., and Gros, P. (2011). Technicolor and inria/irisa at mediaeval 2011: learning temporal modality integration with bayesian networks.

Sargent, J. D., Wills, T. A., Stoolmiller, M., Gibson, J., and Gibbons, F. X. (2006). Alcohol use in motion pictures and its relation with early-onset teen drinking. *Journal of Studies on Alcohol*, 67(1):54–65.

Schmidt, A. and Wiegand, M. (2017). A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10.

Shafaei, M., Lopez-Monroy, A. P., and Solorio, T. (2019). Exploiting textual, visual and product features for predicting the likeability of movies. In *The 32nd International FLAIRS Conference*.

Singh, V., Varshney, A., Akhtar, S. S., Vijay, D., and Shrivastava, M. (2018). Aggression detection on social media text using deep neural networks. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 43–50.

Strasburger, V. C. (1989). Adolescent sexuality and the media. *Pediatric Clinics of North America*, 36(3):747–773.

Webb, T., Jenkins, L., Browne, N., Afifi, A. A., and Kraus, J. (2007). Violent entertainment pitched to adolescents: an analysis of pg-13 films. *Pediatrics*, 119(6):e1219–e1229.

Wilson, B. J. (2008). Media and children's aggression, fear, and altruism. *The Future of Children*, 18(1):87–118.

Zhang, Z., Robinson, D., and Tepper, J. (2018). Detecting hate speech on twitter using a convolution-gru based deep neural network. In *European Semantic Web Conference*, pages 745–760. Springer.