Joint Estimation and Analysis of Risk Behavior Ratings in Movie Scripts

Victor R Martinez
University of Southern California
Los Angeles, CA
victorrm@usc.edu

Krishna Somandepalli
University of Southern California
Los Angeles, CA

Yalda T. Uhls
University of California, Los Angeles
Los Angeles, CA

Shrikanth Narayanan
University of Southern California
Los Angeles, CA
shri@sipi.usc.edu

Abstract

Exposure to violent, sexual, or substance-abuse content in media increases the willingness of children and adolescents to imitate similar behaviors. Computational methods that identify portrayals of risk behaviors from audio-visual cues are limited in their applicability to films in post-production, where modifications might be prohibitively expensive. To address this limitation, we propose a model that estimates content ratings based on the language use in movie scripts, making our solution available at the earlier stages of creative production. Our model significantly improves the state-of-the-art by adapting novel techniques to learn better movie representations from the semantic and sentiment aspects of a character's language use, and by leveraging the co-occurrence of risk behaviors, following a multi-task approach. Additionally, we show how this approach can be useful to learn novel insights on the joint portrayal of these behaviors, and on the subtleties that film-makers may otherwise not pick up on.

1 Introduction

In one of the longest running movie franchises in history, fictional British Secret Service agent James Bond is more often than not portrayed as an extremely charming gentleman, a cold-blooded killer, a smoker, and a severe alcoholic (Wilson et al., 2018). This is not a unique character trait, as other critically acclaimed films—such as, The Exorcist (Friedkin, 1973), Pulp Fiction (Tarantino, 1994), and A Clockwork Orange (Kubrick, 1972)—follow narratives where the main characters engage in a similar collection of risk behaviors. The portrayals of these risk behaviors typically include acts of violence, sexual and substance-abusive behaviors in scenes of fighting, bloodshed, gunplay; intercourse and nudity; and alcohol, smoking and drug use, respectively. While these tend to attract audiences (Barranco et al., 2017) and facilitate a movie's global market reach (Sparks et al., 2005), they have long sparked concerns about the potential side effects of repeated exposure. Particularly, in the case of at-risk populations, such as children and adolescents, where this exposure has been linked to increased risk for engaging in violence (Anderson and Bushman, 2001; Bushman and Huesmann, 2001), smoke and alcohol consumption (Sargent et al., 2005; Dal Cin et al., 2008), and earlier sexual initiation (Brown et al., 2006).

Although various automated tools have been designed to recognize risk behaviors portrayals (e.g., (Chen et al., 2011; Liu et al., 2008)), many rely on cinematic principles from film theory such as illumination, rapid shot transitions or musical score selection (Brezeale and Cook, 2008). This limits their practical impact to an almost-final edition of the content, specifically where visual and sound effects have been added in, making it too late or expensive to implement any modifications. Hence, there is an opportunity on being able to identify these depictions from an earlier stage of content creation as to offer additional useful insights for film-makers and movie producers during the complex creative process.

To this end, our work leverages on two key insights: first, that while all of these works focus on a specific behavior, risk behaviors frequently co-occur with one another both in real-life (Brener and Collins, 1998) and in entertainment media (Bleakley et al., 2017, 2014; Thompson and Yokota, 2004). Second, that the language use in movie scripts can characterize portrayals of risk behaviors at the earliest form of content creation—even before production begins. For example, by identifying when Mr. Bond orders his usual alcoholic drink, Pulp Fiction's main characters plotting to kill someone, or the evil incarnated in The Exorcist cursing in a sexually explicit manner.

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 4780–4790,
November 16–20, 2020.

The present work, to the best of our knowledge, is the first to model the co-occurrence of risk behaviors from linguistic cues found in movie scripts. Our proposed model is a multi-task approach that predicts a movie script's violent, sexual and substance-abusive content from vectorial representations of the character's utterances. We hypothesize that this multi-task approach will help improve violent content classification, as well as in providing insights on their relation to other dimensions of risk behaviors depicted in film media.

Specifically, the contributions of this work are:

1. A multi-task model that significantly improves the state-of-the-art for violent content rating prediction by leveraging the co-occurrence of sexual and substance-abusive content

2. *MovieBERT*[1]: A domain-specific fine-tuned BERT model (Devlin et al., 2019) pre-trained over a large collection of film and TV scripts. We use this model to obtain better representations for the semantics of a character's language

3. A novel large-scale analysis on the joint portrayals, and their relation to other ratings, of violence, sex, and substance abuse in film.

2 Related Work

To understand the prevalence of risk behaviors in film and TV, social scientists have often relied on relatively small human annotated data sets (typically under a 100). This includes a study of portrayals of violence in 74 to 77 films from the last decade (Yokota and Thompson, 2000; Webb et al., 2007), as well as portrayals of teenage sex in 90 of the top-grossing films (Callister et al., 2011). Among other findings, these studies provide evidence that MPAA[2] ratings (the primary rating system used for films in the U.S.) are overly sensitive to sexual content, and less effective at identifying other types of risk behaviors (Tickle et al., 2009; Thompson and Yokota, 2004). However, most of these works are limited to the study of a particular behavior, even though risk behaviors frequently co-occur with one another in media (Bleakley et al., 2017, 2014; Thompson and Yokota, 2004).

The task of identifying risk behaviors from language is perhaps closely related to that of recognizing *Abusive Language* (AL; Waseem et al. 2017). AL is an umbrella term that includes offensive language, including sexist and racist language, and hate-speech. AL computational models are usually designed using popular document classification techniques (Mironczuk and Protasiewicz, 2018), based on features such as n-grams (Nobata et al., 2016); affective language (Wiegand et al., 2018) and distributed semantic representations (Wulczyn et al., 2017). Recent efforts (e.g., Mozafari et al., 2019) explore a supervised fine-tuning approach that start from pre-trained models of highly-contextualized word representations from transformers (Devlin et al., 2019).

Most similar to our work are efforts in predicting a single movie-level rating from language either in movie scripts (Martinez et al., 2019; Shafaei et al., 2019) or in transcripts (Mohamed and Ha, 2020). These works explore the use of recurrent neural networks (RNN) over sequences of vector representations, each composed by the concatenation of lexical, semantic and sentiment features, to learn a movie representation from which the target rating is predicted. There are two notable differences between these and our proposed model. First, our model incorporates additional information in the form of other prediction targets (i.e., multi-task paradigm) and multiple attention layers (Vaswani et al., 2017). The former is motivated by the previously mentioned notion that characters tend to engage in joint portrayals of risk behaviors (Bleakley et al., 2017); the latter allows the model to jointly attend to information from different representation sub-spaces. Second, these previous works explore an early-fusion method where linguistic features are concatenated and fed to a self-attention mechanism on top of the RNN layer. This assumes that in an effort to construct a meaningful interpretation of the features, the attention layer will be powerful enough to disentangle different aspects of language, such as semantic and sentiment. Instead, we use a late-fusion approach where we separate semantics from sentiment, and direct them through different pathways in our model—all the way up to independent attention layers. Thus, our attention layers have the relatively easier task of identifying what is of importance for a particular view of language in a particular task. This allows our model to attend to what is being said (semantic) and, independently, how it is being said (sentiment). We expect this

[1] https://github.com/usc-sail/mica-riskybehavior-identification

[2] Motion Picture Association of America

to be more informative about the content of each utterance, leading to a better representation construction.

3 Method

Our model learns to map sequences of character utterances' representations to overall movie-level ratings. Each representation is composed by two parts: one representing its semantics, and one for its sentiment. These representations are obtained from models trained on larger out-of-domain corpora but have been validated on related tasks in domains similar to those we study in this work (e.g., classification of movie review sentiment (Pagliardini et al., 2018)). Our decision to start from character utterance representation (as opposed to word representations) comes from the limited number of labeled expert curated content ratings in our dataset (see Section 4).

3.1 Semantic representations

The unique aspect of this work is the use of highly-contextualized vector representations for the particular domain of movie scripts to predict content ratings. These techniques have shown remarkable success on a variety of NLP tasks such as sentiment classification (Devlin et al., 2019) and identifying AL in social media (Mozafari et al., 2019).

A. Sentence embeddings: We obtain 700-dimensional Sent2Vec (Pagliardini et al., 2018) (a sentence-level extension of word2vec (Mikolov et al., 2013)) representations from either of two pre-trained sources: (a) BookCorpus (Zhu et al., 2015), and (b) our own collection of $6,000$ movie and TV scripts (see Sec. 4).

B. Highly-contextualized representations: Bidirectional Encoder Representations from Transformers (BERT; Devlin et al. 2019) is a novel language model that outperforms its predecessors due to an innovative architecture that incorporates information from both the left and right contexts. This is done through an interlacing of n fully-connected dense layers each with a multi-head attention layer (Vaswani et al., 2017). From BERT, we obtain vector representations for every utterance. These come from either of two pre-trained models: (a) *BERT-base* ($n = 12$; 768-dimensional), and (b) *BERT-large* ($n = 24$; 1024-dimension)—both trained on a large corpus of documents from Wikipedia and BookCorpus.

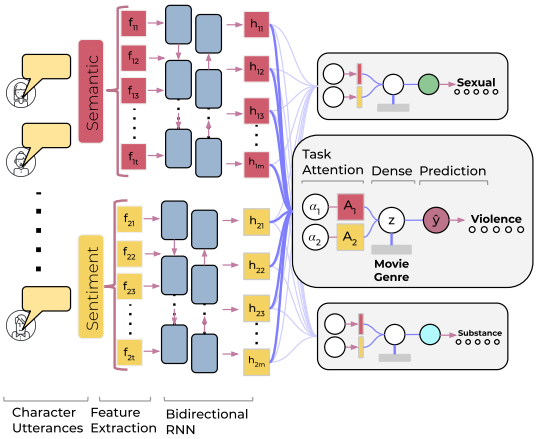

Figure 1: Multi-task model for content rating classification: Each utterance is represented by semantic and sentiment features, fed to independent RNN encoders. The sequence of hidden states from the encoders serve as input for task-specific layers (gray boxes).

C. MovieBERT: A common approach to implement models that produce near state-of-the-art results is to fine-tune large pre-trained models (such as BERT) for a particular task. This aims to keep the generalization power of the original model while also adapting its vocabulary for the language use in a particular domain. Following this idea, here we fine-tune a BERT-base model by continuing its training over the $6,000$ movie scripts dataset. Our adapted model, *movieBERT*, consists of 12 transformer layers that learn a 768-dimensional representation of a movie script. We train this model over a $85\% - 15\%$ train-test data split and, as in (Devlin et al., 2019), we optimize the model for two tasks: next-sentence prediction and masked language modeling. In the former, the model has to predict the sentence that follows a given sentence; in the latter, a random word in a sentence is masked with a token, and the model has to recover the original word. We initialize the weights of our model with those from the pre-trained BERT-base model, and continue training for $10,000$ steps, using the base model's parameters: learning rate of 2×10^{-5}, batch size of 32, and sequences length of 128. *MovieBERT* achieves 96.5% accuracy on the next sentence prediction task, and a 65.9% accuracy on the masked language model—an absolute improvement from the BERT-base model of 24.5% and 12.43%, respectively. To obtain sentence-level representations, we concatenate and then average-pooled the output of the last 2 layers.

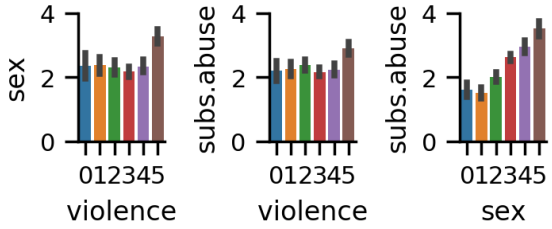

Figure 2: Risk behavior rating co-occurrence: on average, when one risk-behavior rating increases so does the others. Error bars denote 95% confidence intervals.

3.2 Sentiment representations

Previous works show the benefits of including lexical features that capture the expressed sentiment characteristics from language for media content prediction tasks (Martinez et al., 2019; Shafaei et al., 2019). However, most approaches to sentiment analysis on movie scripts rely on manually-constructed sentiment lexica (e.g., Gorinski and Lapata 2018, 2015). These lexica have a limited vocabulary, which is costly to scale or adapt to new domains. In contrast, here we explore neural-network-based sentiment models that learn representations from language used in the related task of movie reviews (Socher et al., 2013). While we are aware of the possible mismatch between the language use in movie reviews and that of movie scripts, our work relies on the assumption that these reviews provide a good initial step towards capturing sentiment expressed in movie scripts. These models not only learn how words are used from a larger vocabulary but also consider the relations between these words which may allow them to generalize better for unseen data. In this work, we experiment with two neural-based models: bidirectional long short-term memory models (Bi-LSTM; Tai et al. 2015), and bidirectional encoder representations from transformers (Devlin et al., 2019). We chose these models because they provide a good trade-off between the number of parameters and the performance on the sentiment prediction task (Barnes et al., 2017), and due to their outstanding performance in NLP tasks. Our sentiment representations are obtained from the last hidden state of the Bi-LSTM, and the previous to last layer of the BERT transformer.

3.3 Role of Movie Genre

Movie genres relate the elements of a story, plot, setting and characters to a specific category. Categorizing a movie indirectly assists in shaping the characters and the story of the movie, and determines the plot and best setting to use. Thus, movie genre contains information on the type of content one could expect in a movie (especially for the case of violent content (Martinez et al., 2019)). Thus, our models include movie genre as an additional feature. Genres for each movie were obtained from IMDb[3] and transformed into a multi-hot encoding.

3.4 Ratings Prediction Model

Our model (see Fig. 1) takes a sequence of utterance representations as input, and outputs predictions for target content ratings. Formally, let K be the number of content ratings to output (number of tasks), and $\{u_t\}_{t=1}^{N}$ be a sequence of N character utterances. For each u_t, we obtain features, f_{1t} and f_{2t} corresponding to the semantic and sentiment aspects of language respectively. These representations are input to separate bi-directional RNN layers. To improve model generalization, a dropout layer (probability p) was added after the feature extraction layer. Each RNN takes a sequence of representations and outputs a sequence of m hidden vectors $\{h_{j1}, \ldots, h_{jm}\}$; $h_{jl} \in \mathbb{R}^d$ where $j = 1, 2$ corresponds to semantic and sentiment features respectively. Each hidden vector represents a state of *conversational context*—i.e., what is being said in relation to what has been previously said. This context is important as it follows from the fact that most utterances are not independent of one another, but follow a conversation thread.

Both hidden-vector sequences $\{h_{1i}\}_{i=1}^{m}$ and $\{h_{2i}\}_{i=1}^{m}$ go through $k \in \{1 \ldots, K\}$ task-specific units, represented as gray boxes in Fig. 1. Each task-specific unit is composed of a sequence of four layers: (i) two separate self-attention mechanisms; (ii) a concatenation layer; (iii) a z-dimensional dense layer, and (iv) a softmax prediction layer. Self-attention (Bahdanau et al., 2014) aggregates the sequence of hidden vectors into a representation of what characters say during the movie. These attention layers, denoted by $\{\alpha_{kj} \in \mathbb{R}^m : j = 1, 2\}$, are not shared between the tasks to allow them to focus on what is important for their particular type of content. We chose this approach as it showed improved performance over our initial experiments with multi-head attention (Vaswani et al., 2017). Attention outputs corresponds to a weighted sum of the hidden states and the α_{kj} weights, $A_{kj} = \sum_{i=1}^{m} \alpha_{kji} \cdot h_{ji}$. In the concate-

[3] https://www.imdb.com/

	LOW(< 3)	MED($= 3$)	HIGH(> 3)
violence	304 (*30.7%*)	329 (*33.3%*)	356 (*36%*)
sexual	446 (*45.1%*)	329 (*33.3%*)	214 (*21.6%*)
substance	469 (*47.4%*)	225 (*39.6%*)	129 (*13.0%*)

Table 1: Movie content rating counts and percentage distribution. Median split was induced on all ratings to balance class distribution.

nation layer, these aggregated representations are coupled with movie-genre $v_k = [A_{k1}; A_{k2}; g]$, and serve as inputs for a z-dimensional dense layer. This yields $s_k = \phi(W_k * v_k + b_k)$ where ϕ is a ReLu function, and W_k, b_k are the weight and bias matrix to be learned. We predict the ratings through a prediction layer as $\hat{y_k} = softmax(s_k)$. The complete model is trained by minimizing the aggregated loss $L = \sum_k l_k(y_k, \hat{y_k})$ where l_k is the cross-entropy loss associated with the k-th task.

4 Data

We collected a large number of movie scripts from three publicly available sources. The first source was related works who shared their movie scripts datasets (Gorinski and Lapata, 2018; Ramakrishna et al., 2017); the second source was online collections of produced scripts[4], and the final source was online communities where non-produced scripts are shared[5]. In total we collected $12,706$ scripts, some of which correspond to produced films or TV episodes. To improve the quality of this dataset, we clean it by extracting text, limiting to files with more than $1,000$ lines, and replacing non-ascii characters. In case of any error, we remove the file from the collection. This procedure resulted in $6,057$ movie scripts spanning 23 genres with an average of 1450.6 utterances per movie ($\sigma = 456.11, M = 1447.0$). We use this collection to fine-tune movieBERT.

To evaluate the performance of our model, and directly compare it to previous work, we manually align a subset of 989 movie scripts from our dataset to the content ratings found in (Martinez et al., 2019). These ratings come from Common Sense Media (CSM), a non-profit organization that promotes safe technology and media for children[6]. CSM experts rate movies from 0 (lowest) to 5 (highest) with each rating manually checked by the executive editor to ensure consistency across raters. A manual inspection of the dataset revealed that the movies with the least scores across all risk behaviors correspond to the romantic genre, whereas the movies with the most risky content were in the horror genre. Additionally, we investigate if CSM expert raters capture the co-occurrence of risk behavior portrayals. Figure 2 shows that, on average, when one risk-behavior rating increases so does the others. This was corroborated by significant positive Spearman's correlations between violence and sexual content ($r_s = 0.161$, $p < 0.001$); violence and substance-abuse ($r_s = 0.129$, $p < 0.001$), and sexual content and substance-abuse ($r_s = 0.467$, $p < 0.001$).

4.1 Preprocessing

We follow a procedure similar to that described in Martinez et al. (2019), which discards scene headers, actions and transitions to represent a movie script as a sequence of actors speaking one after another. This leads to a natural formulation of a sequence learning model for capturing the dialog narrative using recurrent neural networks. Additionally, we transformed the five-point ratings to three categories using a median split on each rating to counter class imbalance and to be consistent with previous work. The distribution of the ratings is shown in Table 1.

5 Experimental Setup

In this section we discuss the model implementation, parameter selection, baseline models and sensitivity analysis setup.

5.1 Model Implementation

Our model was implemented in Keras[7]. Although not common in most deep-learning approaches, we performed 10-fold cross-validation (CV) to obtain a more reliable estimation for our model's performance. In each fold, the model was trained until convergence (i.e. loss in consecutive epochs was less than 10^{-8} difference). To prevent over-fitting, we used Adam optimizer with a small learning rate (0.001), batch size of 16, and high dropout probability ($p = 0.5$). For the RNN layer, we used Gated Recurrent Units (GRU; Cho et al. 2014). For the sentiment models, Bi-LSTM parameters were informed by the work of Tai et al. (2015): 50-dimensional hidden representation, dropout ($p = 0.1$), trained with Adam optimizer on a batch size of 25 and a L_2 penalty of 10^{-4}. To allow for a fair comparison, all the BERT pre-trained models and *movieBERT*

[4] imsdb.com and scriptdrive.org
[5] reddit.com/r/Screenwriting
[6] http://www.commonsensemedia.org
[7] https://keras.io

	Features			Violence			Sex			Subs. Abuse		
	Semantic	Sentiment	Genre	P	R	F_1	P	R	F_1	P	R	F_1
Single-Task Baselines												
Adhikari et al. (2019)	BERT (*base*)	–	No	57.4	55.7	56.1	39.2	34.0	29.2	30.4	35.1	31.9
Nobata et al. (2016)	Abusive Lang.	–	No	52.4	52.4	52.3	44.3	44.3	44.2	42.8	42.4	42.6
Martinez et al. (2019)	AL + word2vec	*Lexical*	*Yes*	*60.1*	*61.1*	*60.4*	–	–	–	–	–	–
No Multi-Task												
Bi-GRU (16)	Sent2Vec (*BookCorpus*)	Bi-LSTM	Yes	64.7	65.6	64.9	45.2	43.8	43.2	52.5	45.1	46.1
	Sent2Vec (*adapted*)			64.5	65.6	64.8	47.2	43.3	42.4	51.7	46.4	47.8
	BERT (*base*)			64.1	64.5	64.2	46.5	44.3	43.5	50.3	46.0	47.3
	BERT (*large*)			63.0	63.7	63.2	44.2	42.1	40.2	52.8	44.2	45.0
	movieBERT			66.9	67.3	67.0	47.6	47.4	47.3	51.1	47.2	48.5
Proposed: Multi-Task & Task-specific Attention												
Bi-GRU (16)	Sent2Vec (*BookCorpus*)	**Bi-LSTM**	**Yes**	66.3	67.2	66.5	17.7	18.6	17.7	17.2	16.9	16.9
	Sent2Vec (*adapted*)			64.0	64.8	64.0	45.0	43.9	43.6	49.9	47.0	47.9
	BERT (*base*)			67.4	67.8	67.5	49.5	47.0	46.8	53.5	47.6	49.1
	movieBERT			**67.6**	**68.3**	**67.7**	**49.8**	**47.9**	**47.9**	51.7	48.7	**49.6**
	BERT (*large*)			64.3	65.0	64.5	46.1	44.5	43.8	**53.6**	46.9	48.6
	movieBERT	BERT (base)		66.2	66.5	66.3	48.7	46.1	46.2	50.8	**48.8**	**49.6**

Table 2: 10-fold cross validation multi-task classification performance. Precision (P), recall (R) and F1 macro average scores reported (percentages). Models trained independently for each task are denoted by double-line. The best model (shown in bold) performs significantly better than baseline for violence (perm. test $n = 10^5$, $p = 0.002$) and substance-abuse ($n = 10^5$, $p = 0.006$).

had the same set of parameters as the BERT-base model: 12 layers, 768 dimensions, learning rate of 2×10^{-5}, sequence length of 128 and batch size of 32. For the initial experiments, we set the model parameters to hidden dimension size of $d = 16$, to help prevent overfitting, and the sequence length $m = 500$, which is approximately the duration of one movie act (i.e., one third). This selection was informed by previous works (Martinez et al., 2019; Shafaei et al., 2019).

5.2 Experiments

In our first set of experiments, we compare the predictive power of each of the proposed features for predicting risk behavior content. In a second set, we explore how varying the number of dimensions ($d \in \{8, 16, 32, 64\}$) and the utterance sequence length ($m \in \{100, 300, 500, 1000\}$) impacts the performance of our model. Additionally, we explore the individual contribution of each feature to the overall prediction task using ablation studies. For all experiments, we report macro-average precision, recall and F-score (F_1) estimated through 10-fold cross validation.

5.3 Baselines

As baselines, we compare against: (i) AL classification (Nobata et al., 2016), since AL likely includes sexual and drug-related terms; (ii) the state-of-the-art for violence rating prediction from movie scripts (Martinez et al., 2019), and (iii) BERT-only document classification systems (Adhikari et al., 2019). Additionally, to measure whether the performance improves with the inclusion of co-occurring risk behaviors, we compare our model against the same architecture without the multi-task approach.

6 Results

6.1 Classification Results

Table 2 presents the classification performance for the baselines and our proposed model. In line with previous results (Martinez et al., 2019; Shafaei et al., 2019), we observe that including sentiment features (either in the form of lexica or neural network representations) greatly improves the model performance. Even without the multi-task framework, our model architecture shows significant improvement over the baselines (permutation test, $n = 10^5$, all $p < 0.05$). This is likely due to our design choice of reducing the model complexity by focusing just on the informative features (i.e., semantic, sentiment and genre) instead of dealing with redundant features (e.g., n-grams, word2vec, AL lexica). By including the co-occurrence information in the form of additional tasks, our proposed multi-task model with task-specific attention gained an average $F_1 = 1.22\%$ points. It also results in the best model (*movieBERT + sentiment + movie-genre*) with an $F_1 = 67.7\%$ for ($d = 16, m = 500$), performing significantly better than the previous state-of-the-art model for violent content rating prediction (perm. test $n = 10^5$, $p = 0.002$) as well as the AL baselines for violence (perm. test $n = 10^5$, $p = 0.005$) and substance-abuse content (perm. test $n = 10^5$,

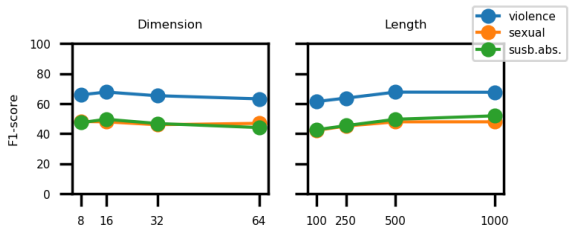

Figure 3: 10-fold cross validation multi-task classification performance based on GRU dimension (d) and sequence length (m).

Semantic	Sentiment	Genre	Violence	Sex	Subs. Abuse	Avg.
✓	✓	✓	67.6 (0.0)	47.9 (0.0)	49.6 (0.0)	0.0
–	✓	✓	60.8 (-6.8)	42.6 (-5.3)	38.2 (-11.4)	-7.83
✓	–	✓	65.2 (-2.4)	46.9 (-0.1)	49.0 (-0.6)	-0.96
✓	✓	–	64.5 (-3.1)	47.0 (-0.9)	50.0 (+0.4)	-1.2

Table 3: 10-fold CV ablation experiments using Bi-GRU (16). F1 macro average score (percentage) reported. In parenthesis: difference between full model and the individual ablation.

$p = 0.006$).

While the proposed model also improves sexual content rating prediction, this improvement is non-significant ($p > 0.05$). As previously mentioned, this could be attributed to the fact that MPAA's ratings are particularly sensitive to sexual content (Thompson and Yokota, 2004). In fact, filmmakers are advised to avoid the repeated usage of sexually-derived words—either as an expletive or in a sexual context—as to avoid a non family-friendly rating (Myers, 2018). Thus, they might refer to sexual acts through the use of euphemisms or innuendos, which the model seems unable to pick up on. Our experiments in using BERT for sentiment representations (last row in Table 2) did not significantly improve performance any further ($p > 0.05$). Future work will explore further fine-tuning to better capture affective language.

6.2 Performance Analysis

Parameter Selection: We evaluate model performance under different selections of parameters, namely the number of hidden dimensions in the GRU layer (d) and the length of the character utterance sequences (m). The model performance for different dimensions is presented in the left section of Fig. 3. For all tasks, we notice an improvement in performance for $d = 16$, which drops for higher dimensions. This suggests that the larger models are overfiting the data. There is a slight improvement for sexual content estimation for $d = 8$ ($F_1 = 48.1$), but its performance is not significantly different from the original model (perm. test $p > 0.05$).

With respect to m, the right section of Fig. 3 presents the F_1 performance of the multi-task model. Overall, we see that longer sequences improve the model's performance. However, there was no significant difference between the performance of $m = 500$ and that of $m = 1000$ (perm.test, $p > 0.05$). Although we did not test sequences longer than 1000 utterances, the smaller performance gains between increments of m lead us to believe that the model is saturated, which suggests that any longer sequence length will not provide any significant performance gains.

Ablation studies: Table 3 shows the individual contributions of each of the three representations. We find that semantic representations are the most important source of information. Removing this feature results in an average performance drop of $-7.83 F_1$. This difference in performance was significant for violence (perm.test $n = 10^5$, $p = 0.003$) and substance-abuse (perm.test $n = 10^5$, $p < 0.0001$) tasks. The second most informative feature was genre, closely followed by sentiment with average performance drops of -1.2 and -0.96 respectively. These results suggest that, while useful, our sentiment features still have scope for improvement. In particular, we note that a potential limiting factor might be the possible mismatch between the language used in movie reviews and that of the movie scripts. A study on how to bridge this possible mismatch will be part of our future work.

Attention Analysis: Finally, we verify our assumption that the attention layers are correctly identifying the important aspects of language with respect to each behavior. We do so by exploring how the attention weights are distributed across the movies scripts. Each of the 6 attention layers (two per task: one for semantic and one for sentiment) learns a m-dimensional weight vector, where each entry corresponds to a particular utterance in the sequence. The higher the weight, the more *importance* the model assigns to that particular utterance. For example, for the violent behavior task, we would expect utterances assigned a higher attention weight to be more reflective violent expressions than utterances with lower attention weights. To verify that each attention layer is correctly focusing on the behavior we are in-

terested in, we set up a hypothesis test where we compare the maximum weight of each attention layer for movies rated HIGH against movies rated LOW on each behavior. Our null hypothesis is that there will be no difference in the way attention concentrates weights for different levels of the behavior. We reject this null hypothesis for the case of the semantics of the violence task (Mann-Whitney $U = 59377.5$, $n_1 = 356$, $n_2 = 304$, $p = 0.015$), and for the sentiment in the sexual content task (Mann-Whitney $U = 52937.5$, $n_1 = 214$, $n_2 = 446$, $p = 0.011$). These results suggest that our model picks up on violence by focusing on the content of the words, whereas identification of sexual behaviors is dependent on the emotional aspects of the language.

7 Co-Occurrence Analysis

In this section, we focus on some of the insights that our proposed model may provide film-makers and producers during the creative process. In particular, our analyses centers on three insights: first, on understanding how joint portrayals of risk behaviors appear on screen; second, in identifying temporal patterns that arise from these joint portrayals, and finally, in showcasing the relation between risk behaviors and MPAA ratings. For this analysis, we re-trained the best performing model over the complete movie script dataset ($n = 989$).

On the relation between joint portrayals of risk behaviors. We find a strong association between predictions of substance-abuse and sexual content: the odds for a movie script to be rated high on sexual content are twice as high when it has a high rating in substance-abuse compared to when it has a low rating (95% Confidence Interval [CI] 2.01 to 34.05). Moreover, we find that the odds of rating high on all three risk behaviors simultaneously are inversely proportional to the predicted violence rating (95% CI, HIGH:0.11 to 0.82 and MED:0.12 to 0.88). Hence, this suggests that *film-makers compensate low levels of violence with joint portrayals of sexual and substance-abuse behaviors.*

On the temporal patterns of the joint portrayals. If there is a temporal relation between the portrayals, when the model picks up a cue for a particular behavior at time t (i.e., a spike in the attention signal), we expect to see a corresponding spike in the attention signal of another task some time after t. To compute this relation, for each movie script we obtained the maximum correlation and its corresponding time lag ($\Delta \in [-m, m]$) by using sample cross correlation function (CCF) between the attention weights of each task. CCF is a measure of similarity between two time series as a function of the displacement of one relative to the other. As an example, Fig. 4 shows the co-evolution of attentions weights and the lags corresponding to their maximum correlation for two movies. On average, *attention to the sexual sentiment content precedes attention to violence semantics* by $\bar{\Delta} = 15.50$ utterances (95% CI, 10.88 to 17.4), with an average correlation coefficient of $r_z = 0.192 \pm 0.02$. This lag increases for movies with higher content ratings on both violence and sex ($\bar{\Delta} = 21.46$, $r_z = 0.202$), whereas movies with low sex and violent content have almost no temporal difference, and a significantly lower correlation coefficient ($\bar{\Delta} = 0.75$, $r_z = 0.172$, perm.test $n = 10^5, p = 0.034$). These results suggests, as Bleakley et al. (2014) points out, that characters engage in sexual and violent behaviors in a small time span from one another.

On the relation between risk behaviors and MPAA ratings. Finally, we measure the relation between the predicted risk behaviors and the movie's MPAA rating. We find that *as sexual content increases, the association between violent (or substance-abuse) content and MPAA rating decreases*. Specifically, movies with high sexual rating are more likely to be rated as R[8], irrespective of their violent or substance-abuse content (odds ratio $OR = 12.172$ (95% CI: 7.86 to 19.46)). In contrast, the MPAA rating of a movie with low sexual content is strongly associated with both their violent content rating ($\chi^2(6) = 18.595$, $p = 0.004$) and their substance-abuse content rating ($\chi^2(3) = 17.99$, $p < 0.001$). These results point out the overly sensitiveness of MPAA raters towards sexual content and corroborate previous findings from small manually-annotated samples of films (Tickle et al., 2009; Thompson and Yokota, 2004).

8 Conclusion

We designed a multi-task model to capture the co-occurrence of depictions of violent content as well as sexual and substance abuse risk behaviors in film through the language data available in scripts. Our proposed model achieves significant improvements

[8] R–Restricted: under 17 requires accompanying parent or adult guardian.

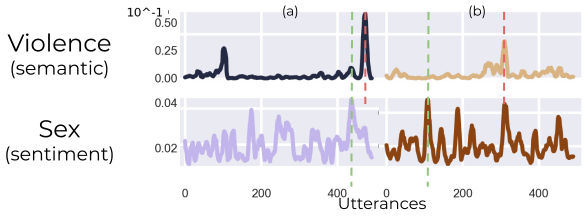

Figure 4: Attention weights for violence and sex for (a) The Exorcist (Friedkin, 1973), and (b) From Russia With Love (Young, 1963). Sex-sentiment (green) leads the violence-semantics (red) by 31 ($\rho = 0.23$) and 203 ($\rho = 0.29$) utterances respectively.

over previous state-of-the-art models for violent content rating prediction. While complementing audio-visual methods, our language-based models can be used to identify subtleties in the way risk behavior content is portrayed, before production begins, offering a valuable tool for content creators and decision makers in entertainment media.

8.1 Future Work

Our overarching goal is to identify when (and how often) are characters being portrayed as targets of risk behaviors—especially in the case where characters are women and minorities. The next step towards this goal would be to recognize when characters refer to one another, and how this contributes to the movie-level risk behavior rating. We hope this leads to tools that can be helpful during the creative process, rather than after the fact.

Acknowledgements

We thank the reviewers for their insights and guidance. This study was done at the Center for Computational Media Intelligence at University of Southern California (USC) which is supported by research awards from Google, and the U.S. Chamber of Commerce Foundation. We would also like to thank Sandeep Nallan Chakravarthula and Nikolaos Malandrakis for the helpful comments on the first draft of this work.

References

Ashutosh Adhikari, Achyudh Ram, Raphael Tang, and Jimmy Lin. 2019. Docbert: BERT for document classification. *CoRR*, abs/1904.08398.

Craig A Anderson and Brad J Bushman. 2001. Effects of violent video games on aggressive behavior, aggressive cognition, aggressive affect, physiological arousal, and prosocial behavior: A meta-analytic review of the scientific literature. *Psychological Science*, 12(5).

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *CoRR*, abs/1409.0473.

Jeremy Barnes, Roman Klinger, and Sabine Schulte im Walde. 2017. Assessing state-of-the-art sentiment models on state-of-the-art sentiment datasets. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*.

Raymond E. Barranco, Nicole E. Rader, and Anna Smith. 2017. Violence at the Box Office. *Communication Research*, 44(1).

Amy Bleakley, Morgan E. Ellithorpe, Michael Hennessy, Atika Khurana, Patrick Jamieson, and Ilana Weitz. 2017. Alcohol, sex, and screens: Modeling media influence on adolescent alcohol and sex co-occurrence. *The Journal of Sex Research*, 54(8):1026–1037.

Amy Bleakley, Daniel Romer, and Patrick E. Jamieson. 2014. Violent film characters' portrayal of alcohol, sex, and tobacco-related behaviors. *Pediatrics*, 133(1).

Nancy D Brener and Janet L Collins. 1998. Co-occurrence of health-risk behaviors among adolescents in the united states. *Journal of Adolescent Health*, 22(3):209–213.

Darin Brezeale and Diane J Cook. 2008. Automatic video classification: A survey of the literature. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(3).

Jane D Brown, Kelly Ladin L'Engle, Carol J Pardun, Guang Guo, Kristin Kenneavy, and Christine Jackson. 2006. Sexy media matter: exposure to sexual content in music, movies, television, and magazines predicts black and white adolescents' sexual behavior. *Pediatrics*, 117(4).

Brad J Bushman and L Rowell Huesmann. 2001. Effects of televised violence on aggression. *Handbook of children and the media*, pages 223–254.

Mark Callister, Lesa A. Stern, Sarah M. Coyne, Tom Robinson, and Emily Bennion. 2011. Evaluation of sexual content in teen-centered films from 1980 to 2007. *Mass Communication and Society*, 14(4).

Liang-Hua Chen, Hsi-Wen Hsu, Li-Yun Wang, and Chih-Wen Su. 2011. Violence detection in movies. In *2011 Eighth International Conference Computer Graphics, Imaging and Visualization*. IEEE.

KyungHyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. *CoRR*, abs/1409.1259.

Sonya Dal Cin, Keilah A Worth, Madeline A Dalton, and James D Sargent. 2008. Youth exposure to alcohol use and brand appearances in popular contemporary movies. *Addiction*, 103(12).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*. Association for Computational Linguistics.

W. Friedkin. 1973. The Exorcist. Warner Bros. Pictures.

Philip Gorinski and Mirella Lapata. 2015. Movie script summarization as graph-based scene extraction. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1066–1076.

Philip John Gorinski and Mirella Lapata. 2018. What's this movie about? A joint neural network architecture for movie content analysis. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Stanley Kubrick. 1972. A Clockwork Orange. Warner Bros. Pictures.

Anan Liu, Yongdong Zhang, Yan Song, Dongming Zhang, Jintao Li, and Zhaoxuan Yang. 2008. Human attention model for semantic scene analysis in movies. In *IEEE International Conference on Multimedia and Expo*. IEEE.

Victor Martinez, Krishna Somandepalli, Karan Singla, Anil Ramakrishna, Yalda Uhls, and Shrikanth Narayanan. 2019. Violence rating prediction from movie scripts. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*. AAAI press.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Marcin Mironczuk and Jaroslaw Protasiewicz. 2018. A recent overview of the state-of-the-art elements of text classification. *Expert Systems with Applications*, 106.

Emad Mohamed and Le An Ha. 2020. A first dataset for film age appropriateness investigation. In *Proceedings of The 12th Language Resources and Evaluation Conference*, Marseille, France. European Language Resources Association.

Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. 2019. A BERT-based transfer learning approach for hate speech detection in online social media. *CoRR*, abs/1910.12574.

Scott Myers. 2018. Reader question: Is there a rule as to how many "cuss words" can be used in a script? https://bit.ly/35hKwhY. Accessed: 04/09/2020.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*.

Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. 2018. Unsupervised Learning of Sentence Embeddings using Compositional n-Gram Features. In *NAACL 2018 - Conference of the North American Chapter of the Association for Computational Linguistics*.

Anil Ramakrishna, Victor R Martínez, Nikolaos Malandrakis, Karan Singla, and Shrikanth Narayanan. 2017. Linguistic analysis of differences in portrayal of movie characters. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017*.

James D Sargent, Michael L Beach, Anna M Adachi-Mejia, Jennifer J Gibson, Linda T Titus-Ernstoff, Charles P Carusi, Susan D Swain, Todd F Heatherton, and Madeline A Dalton. 2005. Exposure to movie smoking: its relation to smoking initiation among US adolescents. *Pediatrics*, 116(5).

Mahsa Shafaei, Niloofar Safi Samghabadi, Sudipta Kar, and Thamar Solorio. 2019. Rating for parents: Predicting children suitability rating for movies based on language of the movies. *CoRR*, abs/1908.07819.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013*, pages 1631–1642.

Glenn G Sparks, John Sherry, and Graig Lubsen. 2005. The appeal of media violence in a full-length motion picture: An experimental investigation. *Communication Reports*, 18(1-2).

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. *CoRR*, abs/1503.00075.

Quentin Tarantino. 1994. Pulp Fiction. Miramax.

Kimberly M Thompson and Fumie Yokota. 2004. Violence, sex, and profanity in films: correlation of movie ratings with content. *Medscape General Medicine*, 6(3).

Jennifer J Tickle, Michael L Beach, and Madeline A Dalton. 2009. Tobacco, alcohol, and other risk behaviors in film: how well do mpaa ratings distinguish content? *Journal of health communication*, 14(8).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In *Proceedings of the First Workshop on Abusive Language Online*.

Theresa Webb, Lucille Jenkins, Nickolas Browne, Abdelmonen A Afifi, and Jess Kraus. 2007. Violent entertainment pitched to adolescents: an analysis of PG-13 films. *Pediatrics*, 119(6):e1219–e1229.

Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. 2018. Inducing a Lexicon of Abusive Words–a Feature-Based Approach. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018*.

Nick Wilson, Anne Tucker, Deborah Heath, and Peter Scarborough. 2018. Licence to swill: James bond's drinking over six decades. *Medical journal of Australia*, 209(11):495–500.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*.

Fumie Yokota and Kimberly M. Thompson. 2000. Violence in G-rated animated films. *Journal of the American Medical Association*, 283(20).

Terence Young. 1963. From Russia with Love. United Artists.

Yukun Zhu, Ryan Kiros, Richard Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*.