Violence Rating Prediction from Movie Scripts

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

Krishna Somandepalli
University of Southern California
Los Angeles, CA
somandep@usc.edu

Karan Singla
University of Southern California
Los Angeles, CA
singlak@usc.edu

Anil Ramakrishna
University of Southern California
Los Angeles, CA
akramakr@usc.edu

Yalda T. Uhls
University of California Los Angeles
Los Angeles, CA
yaldatuhls@gmail.com

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

Abstract

Violent content in movies can influence viewers' perception of the society. For example, frequent depictions of certain demographics as perpetrators or victims of abuse can shape stereotyped attitudes. In this work, we propose to characterize aspects of violent content in movies solely from the language used in the scripts. This makes our method applicable to a movie in the earlier stages of content creation even before it is produced. This is complementary to previous works which rely on audio or video post production. Our approach is based on a broad range of features designed to capture lexical, semantic, sentiment and abusive language characteristics. We use these features to learn a vector representation for (1) complete movie, and (2) for an act in the movie. The former representation is used to train a movie-level classification model, and the latter, to train deep-learning sequence classifiers that make use of context. We tested our models on a dataset of 732 Hollywood scripts annotated by experts for violent content. Our performance evaluation suggests that linguistic features are a good indicator for violent content. Furthermore, our ablation studies show that semantic and sentiment features are the most important predictors of violence in this data. To date, we are the first to show the language used in movie scripts is a strong indicator of violent content. This offers novel computational tools to assist in creating awareness of storytelling.

Introduction

Violence is an important narrative tool, despite some of its ill effects. It is used to enhance a viewer's experience, boost movie profits (Barranco, Rader, and Smith 2017; Thompson and Yokota 2004), and facilitate global market reach (Sparks, Sherry, and Lubsen 2005). Including violent content may modify a viewer's perception of how exciting a movie is by intensifying the sense of relief when a plotline is resolved favorably (Topel 2007; Sparks, Sherry, and Lubsen 2005).

There is a *sweet-spot* of how much violent content filmmakers should include to maximize their gains. Too much violence may lead to a movie being rated as NC-17[1] which severely restricts both the promotion of the film as well as the viewership. This is sometimes considered a "commercial death sentence" (Cornish and Block 2012; Susman 2013). This usually forces filmmakers to trim the violent content to receive a rating of R or lower. As shown by (Thompson and Yokota 2004), MPAA ratings have been to be inconsistent. Hence it is crucial to develop an objective measure of violence in media content.

Using violent content is often a trade-off between the economic advantage and social responsibility of the filmmakers. The impact of portrayed violence on the society, especially children and young adults, has been long studied (See detailed review (American Academy of Pediatrics and Others 2001)). Violent content has been implicated in evoking aggressive behavior in real life (Anderson and Bushman 2001) and cultivating the perception of the world as a dangerous place (Anderson and Dill 2000). But this type of content does not appear to increase severe forms of violence (e.g., homicide, aggravated assault) at the societal level (Markey, French, and Markey 2015), and its impact on an individual is highly dependent on personality predispositions (Alia-Klein et al. 2014). Casting people of certain demographics as perpetrators more frequently than others may contribute to the creation of negative stereotypes (Potter et al. 1995; Smith et al. 1998), which may put these populations at a higher risk of misrepresentation (Eyal and Rubin 2003; Igartua 2010). Thus, it is important to study violence in movies at scale.

There is a demand for scalable tools to identify violent content with the increase in movie production (777 movies released in 2017; up 8% from 2016 (Motion Picture Association of America 2017)). Most efforts in detecting violence in movies have used audio and video-based classifiers (e.g., (Ali and Senan 2018; Dai et al. 2015)). No study to date has explored language use from subtitles or scripts. This limits their application to after the visual and sound effects have been added to a movie.

There are many reasons why identifying violent content from movie scripts is useful: 1) Such a measure can provide

[1]Motion Picture Association of America's rating system classifies media into 5 categories, ranging from suitable for all audiences (G) to Adults only (NC-17). NC-17 films do not admit anyone under 17. In contrast, restricted rating (R) films may contain adult material but still admit childrens accompanied by their parents.

filmmakers with an objective assessment of how violent a movie is 2) It can help identify subtleties that producers may otherwise not pick up on when violence is traditionally measured through action, and not language 3) it could suggest appropriate changes to a movie script even before production begins 4) It has the potential to analyze large scale portrayals of violence from a *who-is-doing-what* perspective. This could provide social scientists with additional empirical evidence for creating awareness of negative stereotypes in film.

Our objective in this work is to understand the relation between language used in movie scripts and the portrayed violence. In order to study this relationship, we present experiments to computationally model violence using features that capture lexical, semantic, sentiment and abusive language characteristics. In the following sections we discuss what is the prevalence of violence in media, and related work on how language can be used to identify violence. We next describe our computational framework followed by a battery of experiments to validate our approach.

Related Work

Violence in movies

Most studies that have examined the effect and prevalence of violence in film have generally used trained human annotators to identify violent content. This annotation approach limits the studies to a small set of movies, typically under a 100. For example, (Yokota and Thompson 2000) studied depictions of violence in 74 G-rated animated movies released between 1937 and 2000, and (Webb et al. 2007) examined 77 of the top-grossing PG-13 films of 1999 and 2000. Both studies show that in these datasets, a majority of the movies contain violent acts, in-spite of their ratings.

Although, to the best of our knowledge, no work has presented a large-scale analysis of violent content in movie scripts. Several works have studied movie scripts for actor's portrayals and negative stereotyping of certain demographics. For example, (Sap et al. 2017) studied 772 movie scripts with respect to actors' gender, and (Ramakrishna et al. 2017) explored 954 movie scripts with respect to actors' age, gender and race.

Violent content and Language

The task of analyzing violent content in movie scripts is closely related to detecting abusive language. Abusive Language (Waseem et al. 2017) is an umbrella term generally used to group together offensive language, hate speech, cyber-bulling, and trolling (Burnap and Williams 2015; Nobata et al. 2016). Abusive language typically target underrepresented groups (Burnap and Williams 2015), ethnic groups (Kwok and Wang 2013) or particular demographics (Dixon et al. 2017; Park and Fung 2017).

There is a good body of work dealing with identifying various types of abusive language; for a in-depth review please refer to (Schmidt and Wiegand 2017). Most of the computational approaches for abusive language detection use linguistic features commonly used in document classification tasks (Mironczuk and Protasiewicz 2018). For example, word or character n-grams (Nobata et al. 2016; Mehdad and Tetreault 2016; Park and Fung 2017), and distributed semantic representations (Djuric et al. 2015; Wulczyn, Thain, and Dixon 2017; Pavlopoulos, Malakasiotis, and Androutsopoulos 2017). Under the assumption that abusive messages contain specific negative words (e.g., slurs, insults, etc.) many works have constructed lexical resources to capture these type of words (for example (Wiegand et al. 2018; Davidson et al. 2017)). In addition to linguistic features, works have explored the use of social network meta-data (Pavlopoulos, Malakasiotis, and Androutsopoulos 2017) with the effectiveness of this approach still up for debate (Zhong et al. 2016). With respect to the computational models used, most studies employ either traditional machine learning approaches or deep-learning methods. Examples of the former are Support Vector Machines (Nobata et al. 2016) or Logistic Regression classifiers (Wulczyn, Thain, and Dixon 2017; Djuric et al. 2015). On the latter, studies have successfully used convolutional neural networks (Park and Fung 2017; Zhang, Robinson, and Tepper 2018) and recurrent neural networks (e.g., (Pavlopoulos, Malakasiotis, and Androutsopoulos 2017; Founta et al. 2018)).

Dataset

Movie screenplay dataset We use the movie screenplays collected by (Ramakrishna et al. 2017), an extension to Movie-DiC (Banchs 2012). It contains 945 Hollywood movies, from 12 different genres (1920—2016). Unlike other datasets of movie summaries or scripts, this corpus is larger and readily provides actors' utterances extracted from the scripts (see Table 1).

The number of utterances per movie script in our dataset varied widely. It ranged between 159 and 4141 ($\mu = 1481.70, \sigma = 499.05$). The median number of utterances of the scripts was $M = 1413.5$. Assuming that movies follow a 3-act structure, and that each act has the same number of utterances, this means that each act is made by about 500 utterances. In all sequence modeling experiments we focus only on the last act because filmmakers often include more violent content towards the climax of movie. This is supported by excitation-transfer theory (Zillmann 1971). It suggests that a viewer experiences a sense of relief intensified by the transfer of excitation from violence when the movie plot is resolved favorably. We also evaluate our sequence models using the introductory segment of the movie to assess this assumption (See **Sensitivity Analysis**).

Violence ratings In order to measure the amount of violent content in a movie, we used expert ratings obtained from Common Sense Media (CSM). CSM is a non-profit organization that provides education and advocacy to families to promote safe technology and media for children[2]. Expert raters, trained by CSM, review books, movies, TV shows, video games, apps, music and websites in terms of age-appropriate educational content, violence, sex, profanity and more to help parents make media choices for their kids. CSM experts watch movies to rate its violent content

[2] https://www.commonsensemedia.org/about-us

Number of movies	945
# Genres	12
# Characters	6,907
# Utterances	530,608

Table 1: Description of the Movie Screenplay Dataset.

	0	1	2	3	4	5	Total
no.	40	48	83	261	135	165	732
%	5.46	6.56	11.34	35.66	18.44	22.54	100.0

Table 2: Raw frequency (no.) and percentage (perc.) distribution of violence ratings

from 0 (lowest) to 5 (highest). Ratings include a brief rationale (e.g., fighting scenes, gunplay). Each rating is manually checked by the Executive Editor to ensure consistency across raters. These ratings can be accessed directly from the CSM website. From the total 945 movie scripts in the dataset, we found 732 movies (76.64%) for which CSM had ratings. The distribution of ratings labels is given in Table 2. To balance the negative skewedness of the rating distribution, we selected two cut-offs and encoded violent content as a three-level categorical variable (LOW$<$ 3, MED$=$ 3, HIGH$>$ 3). The induced distribution can be seen in Table 3.

Methodology

Movie scripts often contain both the dialogues of an actor (or utterances) and scene descriptions. We preprocessed our data to keep only the actors' utterances. We discarded scene descriptions (e.g., camera panning, explosion, interior, exterior) for two reasons: 1) Our objective is study the relation between violence and what a person said. As such, we do not want to bias our models with descriptive references to the setup. Additionally, these descriptions vary widely in style and are not consistent in the depth of detail (in publicly available scripts) 2) This enables us to express a movie script as a sequence of actors speaking one after another using models such as recurrent neural networks (RNN).

Following this preprocessing, we collected language features from all the utterances. Our features can be divided into five categories: N-grams, Linguistic and Lexical, Sentiment, Abusive Language and Distributional Semantics. These features were obtained at two units of analysis: 1) *Utterance-level*: text in each utterance is considered independently to be used in sequence models, and 2) *Movie-level*: where all the utterances are treated as a single document for classification models. Because movie genres are generally related to the amount of violence in a movie (e.g., romance vs. horror), we evaluated all our models by including movie-genre as an one-hot encoded feature vector. We used the primary

	LOW	MED	HIGH	Total
no.	171	261	300	732
%	23.36	35.65	40.98	100.0

Table 3: Frequency (no.) and percentage (%) distribution of violence ratings after encoding as a categorical variable

	Utterance-level	Movie-level
N-grams	TF (IDF)	TF (IDF)
Linguistic	TF	TF
Sentiment	Scores	Functionals
Abusive Language	TF	TF
Semantic	Average	Average

Table 4: Summary of feature representation at utterance and movie level. **TF** - Term frequency, **IDF** - Inverse document frequency, **Functionals** - Mean, Variance, Maximum, Minimum, and Range

genre since a movie can belong to multiple genres. See Table 4 for a summary of the feature extraction methods at the two levels. We now describe each feature category in detail.

Features

N-grams: We included unigrams and bigrams to capture the relation of the words to violent content. Because screenwriters often portray violence using offensive words and use their censored versions, we included 3, 4 and 5 character n-grams as additional features. This window of size of 3–5 is consistent with (Nobata et al. 2016) who showed this to be effective to model offensive word bastardization (e.g., fudge) or censoring (e.g., *f**k*). These features were then transformed using term frequencies (TF) or TF-IDF (Sparck Jones 1972). We setup additional experiments to assess the choice of transformation and the vocabulary size (See section **Sensitivity Analysis**).

Linguistic and Lexical Features. Violent content may be used to modify the viewers' perception of how exciting a movie is (Sparks, Sherry, and Lubsen 2005). Examples of how excitation can be communicated through writing is by repeating exclamation marks to add emphasis, or by indicating yelling/loudness by capitalizing all letters. Hence, we include: number of punctuation marks (periods, quotes, question marks), number of repetitions, and the number of capitalized letters. These features were also proposed in the context of abusive language by (Nobata et al. 2016).

Psycho-lingustic studies traditionally represent text documents by the percentage of words that belong to a set of categories using predetermined dictionaries. These dictionaries that map words to categories are often manually crafted based on existing theory. In our work, we obtained word percentages across 192 lexical categories using Empath (Fast, Chen, and Bernstein 2016). Empath is similar to other popular tools found in Psychology studies such as Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al. 2015) and General Inquirer (GI) (Stone, Dunphy, and Smith 1966). We chose Empath because, it analyzes text on a wider range of lexical categories including those in LIWC or GI.

Sentiment Features. We include a set of features from sentiment classification tasks because it is likely that violent utterances contain words with negative sentiment. We used two sentiment analysis tools that are commonly used to process text.

AFINN-111 (Nielsen 2011): For a particular sentence, it produces a single score (-5 to +5) by summing the valence (a dimensional measure of positive or negative sentiment) ratings for all words in the sentence. AFINN-111 has also been effectively used for movie summarization (Gorinski and Lapata 2018).

VADER (Gilbert 2014): valence aware dictionary and sentiment reasoner (VADER) is a lexicon and rule-based sentiment analyzer that produces a score (-1 to +1) for a document.

For the two measures described above, we estimated a movie-level sentiment score using the statistical functionals (mean, variance, max, min and range) across all the utterances in the script. Formally, let $U = \{u_1, u_2, \ldots, u_k\}$ be a sequence of utterances with associated sentiment measures given by $S_U = \{s_1, s_2, \ldots, s_k\}$. We obtain a representation of movie-level sentiment $S_M \in \mathbb{R}^5$:

$$S_M(U) = \begin{pmatrix} \mu(S_U) \\ \sigma^2(S_U) \\ \max(S_U) \\ \min(S_U) \\ \max(S_U) - \min(S_U) \end{pmatrix}$$

Where $\mu(S_U) = \frac{1}{k}\sum_{i=0}^{k} s_i$ and $\sigma^2(S_U) = \frac{1}{k-1}\sum_{i=0}^{k}(s_i - \mu(S_U))^2$.

We also obtain the percentage of words in the lexical categories of positive and negative emotions from Empath. Finally we concatenate the $S_M(U)$ from AFINN-111 and VADER with the two measures from Empath to obtain a 12-dimensional movie-level sentiment feature.

Distributed Semantics. By including pre-trained word embeddings, models can leverage semantic similarities between words. This helps with generalization as it allows our models to adapt to words not previously seen in training data. In our feature set we include a 300-dimensional word2vec word representation trained on a large news corpus (Mikolov et al. 2013). We obtained *utterance-level embeddings* by averaging the word representations in an utterance. Similarly, we obtained *movie-level embeddings* by averaging all the utterance-level embeddings. This hierarchical procedure was suggested by (Schmidt and Wiegand 2017). Other approaches have suggested that paragraph2vec (Le and Mikolov 2014) provides a better representation than averaging word embeddings (Djuric et al. 2015). Thus, in our experiments we also evaluated the use of paragraph2vec for utterance representation (See Sec. **Sensitivity Analysis**).

Abusive Language Features. As described before, violence in movies is related to abusive language. *Explicit* abusive language can often be identified by specific keywords, making lexicon-based approaches well suited for identifying this type of language (Schmidt and Wiegand 2017; Davidson et al. 2017). We collected the following lexicon-based features: (i) number of insults and hate blacklist words from Hatebase[3] (Nobata et al. 2016; Davidson et al. 2017); (ii) cross-domain lexicon of abusive words from (Wiegand et al. 2018), and (iii) human annotated hate-speech terms, collected by (Davidson et al. 2017).

Implicit abusive language, however is not trivial to capture as it is heavily dependent on context and the domain (e.g, twitter vs. movies). Although we do not model this directly, the features which we have included, e.g., N-grams (Schmidt and Wiegand 2017) and word embeddings (Wulczyn, Thain, and Dixon 2017) have been shown to be effective to identify implicit abusive language—albeit, in social media. Additionally, we use sequence modeling which keeps track of the context that can capture this implicit nature of the abusive language. A detailed study of the relation between the domain, context and implicit abusive language is part of our future work.

[3]www.hatebase.org

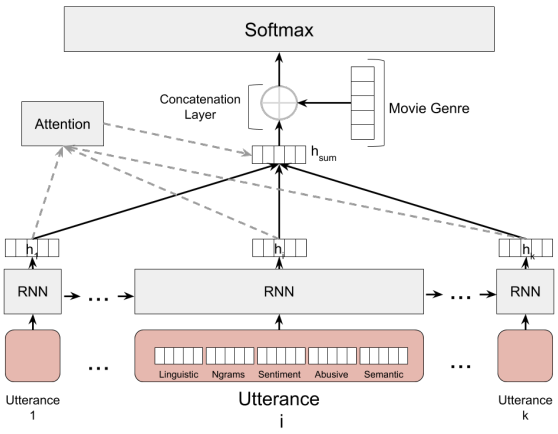

Figure 1: Recurrent Neural Network with attention: Each utterance is represented as a vector of concatenated feature-types. A sequence of k utterances is fed to a RNN with attention, resulting in a H-dimensional representation. This vector is then concatenated with genre representation and fed to the softmax layer for classification.

Models

SVM. We train a Linear Support Vector Classifier (LinearSVC) using the movie-level features to classify a movie script into one of three categories of violent content (i.e., LOW/MED/HIGH). We chose LinearSVC as they were shown to outperform deep-learning methods when trained on a similar set of features (Nobata et al. 2016).

RNN. We investigate if context can improve to predict violence similar to previous works e.g., (Founta et al. 2018) for related tasks. In this work, we consider two main forms of context: *conversational context*, and *movie genre*. The former refers to what is being said in relation to what has been previously said. This follows from the fact that most utterances are not independent from one another, but rather follow a thread of conversation. The latter takes into account that utterances in a movie follow a particular theme set by the movie's genre (e.g., action, sci-fi). Our proposed architecture (See Figure 1) captures both forms of context. The

conversational context is captured by the RNN layer. It takes all past utterances as input to update the representation for the utterance-of-interest. Movie genre is encoded as a one-hot representation concatenated to the output of the attention layer. This allows our model to learn that some utterances that are violent for a particular genre may not be considered violent in other genres.

Formally, let $U = \{u_1, u_2, \ldots, u_k\}$ be a sequence of utterances each representated by a fixed-length vector $x_t \in \mathbb{R}^D$. Let M_G be a one-hot representation of a movie's genre. The RNN layer transforms the sequence $\{x_1, x_2, \ldots, x_k\}$ into a sequence of hidden vectors $\{h_1, h_2, \ldots, h_k\}$. These hidden vectors are to be aggregated by an attention mechanism (Bahdanau, Cho, and Bengio 2014). Attention mechanism outputs a weighed sum of hidden states $h_{sum} = \sum_{i=0}^{k} \alpha_i h_i$, where each weight α_i is obtained by training a dense layer over the sequence of hidden vectors. The aggregated hidden state h_{sum} is concatenated to M_G (See Figure1) and input to a dense layer for classification.

Experiments

In this section we discuss the model implementation, hyper-parameter selection, baseline models and sensitivity analysis setup. The code to replicate all experiments is publicly available[4].

Model Implementation. Linear SVC was implemented using scikit-learn (Pedregosa et al. 2011). Features were centered and scaled using sklearn's robust scaler. We estimated model's performance and optimal penalty parameter $C \in [0.01, 1, 10, 100, 1000]$ through nested 5-fold cross validation (CV).

RNN models were implemented in Keras (Chollet and others 2015). We used the Adam optimizer with mini-batch size of 16 and learning rate of 0.001. To prevent over-fitting, we use drop-out of 0.5, and train until convergence (i.e., consecutive loss with less than 10^{-8} difference). For the RNN layer, we evaluated Gated Recurrent Units (Cho et al. 2014) and Long Short-Term Memory cells (Hochreiter and Schmidhuber 1997). Both models were trained with number of hidden units $H \in [4, 8, 16, 32]$. Albeit uncommon in most deep-learning approaches, we opted for 5-fold CV to estimate our model's performance. We chose this approach to be certain that the model does not over-fit the data.

Baselines For baseline classifiers, we consider both explicit and implicit abusive language classifiers. For explicit, we trained SVCs using lexicon based-approaches. The lexicon considered were wordlist from Hatebase, manually curated n-gram list from (Davidson et al. 2017), and cross-domain lexicon from (Wiegand et al. 2018). Additionally, we compare against implementations of two state-of-the-art models for implicit abusive language classification: (Nobata et al. 2016), a LinearSVC trained on Linguistic, N-gram, Semantic features plus Hatebase lexicon, and (Pavlopoulos, Malakasiotis, and Androutsopoulos 2017), an RNN with deep-attention. Unlike our approach, deep-attention learns the attention weights using more than one dense layer. In addition to these baseline models, we also compare against RNN models trained using only Semantic features (i.e., only word embeddings).

[4]https://github.com/usc-sail/mica-violence-ratings-predictions-from-movie-scripts

	Prec	Rec	F-score
Abusive Language Classifiers			
Hatebase	29.8	37.4	30.5
(Davidson et al. 2017)	13.6	33.1	19.3
(Wiegand et al. 2018)	28.3	34.8	26.8
(Nobata et al. 2016)	55.4	54.5	54.8
(Pavlopoulos et al. 2017)	53.3	52.0	52.5
Semantic-only (word2vec)			
Linear SVC	56.3	55.8	56.0
GRU (16)	53.6	52.3	51.5
LSTM (16)	52.7	54.0	52.5
Movie-level features			
Linear SVC	60.5	58.4	59.1
Utterance-level features			
GRU (4)	52.4	49.5	49.5
GRU (8)	58.2	58.2	58.2
GRU (16)	**60.9**	**60.0**	**60.4**
GRU (32)	58.8	58.4	58.4
LSTM (4)	54.1	54.2	52.1
LSTM (8)	56.6	57.6	57.0
LSTM (16)	57.4	57.2	57.2
LSTM (32)	56.4	56.2	55.9

Table 5: Classification results: 5-fold CV precision (Prec), recall (Rec) and F1 macro average scores for each classifier. In parenthesis are the number of units in each hidden layer.

Sensitivity analysis We present model performance under different selections of initial feature extraction parameters. First, we evaluate the impact of limiting vocabulary size to the most frequent $|\mathcal{V}|$ word n-grams and character n-grams. For this, we explored $|\mathcal{V}| \in [500, 2000, 5000]$. Additionally, we evaluated whether TF or TF-IDF word n-gram transformation was better. We also assessed the use of word embeddings (word2vec) against using embeddings trained on both words and paragraphs (paragraph2vec). We do so because previous works suggested that paragraph2vec creates a better representation than averaging word embeddings (Djuric et al. 2015). Finally, sequence models were evaluated on each one of the three segments obtained from the three-act segmentation heuristic.

Results

Classification results

Table 5 shows the macro-averaged classification performance of baseline models and our proposed models. Precision, recall and F-score (F1) for all models was estimated using 5-fold CV. Consistent with previous works, lexicon-based approaches resulted in a higher number of false positives, leading to a high recall but low precision (Schmidt and Wiegand 2017); in contrast, both implicit abusive language

	Prec	Rec	F-score	δ
All	60.9	**60.0**	**60.4**	0.0
		Ablations		
-Genre	59.9	59.0	59.4	-1.0
-N-grams	59.9	59.2	59.5	-0.9
-Linguistic	**62.1**	59.0	60.1	-0.3
-Sentiment	59.6	58.3	58.8	-1.6
-Abusive	60.6	58.9	59.6	-0.8
-Semantic	59.8	58.3	58.9	-1.5

Table 6: 5-fold CV ablation experiments using GRU-16. δ column shows the difference between original model and individual ablations. '-' indicates removing a certain feature

classifiers and our methods achieve a better balance between precision and recall. In line with previous work (Nobata et al. 2016), for the feature set we selected traditional machine learning approaches perform better than deep-learning methods trained on word2vec vectors only.

Our results suggest that models trained on the complete feature set performed better than other models. The difference in performance is significant (permutation tests, $n = 10^5$, all $p < 0.05$). This suggests that the additional language features contribute to the classification performance. As shown in the next section, this increase can be attributed mostly to Sentiment features. The best performance is obtained using a 16-unit GRU with attention (GRU-16), trained on all features. GRU-16 performed significantly better than the baselines (permutation test, smallest $|\Delta| = 0.056$, $n = 10^5$, all $p < 0.05$), and better than the RNN models trained on word2vec only ($|\Delta| = 0.079$, $n = 10^5, p < 0.05$). We were unable to find statistical differences in performance between LinearSVC trained on movie-level features and GRU-16 (perm test, $p > 0.05$).

Ablation studies

We explore how each feature contributes to the classification task through individual ablation tests. Difference in model performance was estimated using 5-fold CV, which are shown in Table 6. Overall, our results suggest that GRU-16 takes advantage of all feature types (all ablations below zero). Sentiment and word2vec generalizations (i.e., Semantic) contribute the most.

Our ablation studies showed no significant differences in classification performance (perm. test, $n = 10^5$, all $p > 0.05$). This could be because our non-parametric tests do not have enough statistical power, or that there is no real difference in how the model performs. A possible explanation of no real difference is that language features share redundant information. For instance, when N-grams are not present the classifier may be relaying more on Semantic features. This might also explain why Sentiment accounts for the highest ablation drop ($\delta = -1.6$), since word embeddings and n-grams ignore sentiment-related information (Tang et al. 2014). We found linguistic features to be the least informative features (ablation drop of $\delta = -0.3$) and when removed, the classifier achieves a higher precision score. An explanation of this behavior is that Linguistic features include general-domain lexicons which tend to produce low precision and high recall classifiers (Schmidt and Wiegand 2017). Hence, removing these features reduced recall and increased precision. Finally, removing genre resulted in a drop in performance (ablation drop of $\delta = -1.0$) suggesting the importance of genre for violent rating prediction.

Model	**Semantic**	**N-grams**	$\lvert V\rvert$	**F-score(%)**
Linear SVC	word2vec	TF	500	58.3
		TF	2000	59.6
		TF	5000	59.1
		TF-IDF	5000	58.9
	paragraph2vec	TF	5000	59.1
GRU-16	word2vec	TF	500	55.9
		TF	2000	59.1
		TF	5000	**60.4**
		TF-IDF	5000	56.9
	paragraph2vec	TF	5000	59.1

Table 7: 5-fold cross validation F scores (macro-average) for the two best models by varying model parameters. $|V|$= vocabulary size for word and character n-grams.

Sensitivity Analysis

We measured the performance of our best classifiers (LinearSVC and GRU-16) with respect to the choice of different parameters. Table 7 shows 5-fold CV macro-average estimates for precision, recall and F1 scores. Our results suggest that using the IDF transformation negatively impacts the performance of both classifiers. However these differences were not significant (perm. test, $|\Delta| = 0.035$, $n = 10^5, p > 0.05$).

We did not find any significant difference (perm. test, $|\Delta| \approx 0.00$, $n = 10^5, p > 0.05$) in performance when using paragraph2vec rather than averaging word embeddings. A possible explanation is that unlike previous approaches, which studied multi-line user comments, utterances are one or two short lines of text—typically.

Regarding vocabulary size, classifiers seem to be impacted differently. LinearSVC achieves a better score when 2000 word n-grams and character n-grams are used. In contrast, the bigger the vocabulary, the better GRU-16 performs.

Finally, scores suggest that GRU-16 performs better when trained on the final segment than when trained on the first segment ($|\Delta| = 0.067$) or the second segment ($|\Delta| = 0.017$). The difference in performance was significant when trained on the first segment (perm. test, $n = 10^5, p < 0.05$). This result seems to suggest that filmmakers include more violent content towards the end of a movie script.

Attention Analysis

By exploring the utterances with the highest and lowest attention weights, we can get an idea of what the model labels as violent. We would expect utterances assigned to a higher attention weight to be more violent than utterances with lower attention weights. To investigate if the attention scores highlight violent utterances, we obtained the weights from GRU-16 on a few movie scripts. These movie scripts were selected from a held-out batch of 16. We sorted utterances from each movie based on their attention weights. To

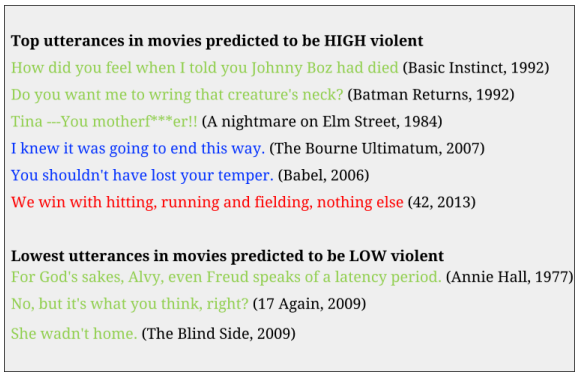

Figure 2: Examples of utterances with highest and lowest attention weights for a few movies. **green** - correctly identified, **blue** - depends on context (implicit), **red** - miss identified

illustrate a few examples (See Figure 2), we picked the top- or bottom-most utterances when a movie was rated HIGH or LOW respectively. From movies predicted as HIGH, the top utterances show themes related to killing or death. It also appears to pick up on more subtle indications of aggression such as "loosing one's temper". However, it assigned a high attention weight to an utterance about sports (marked in red). This suggest that movie genre, although helpful, does not disambiguate subtle contexts for violence. Understanding these contexts is a part of our future work.

Conclusion and Future Work

We present an approach to identify violence from the language used in movie scripts. This can prove beneficial for filmmakers to edit content and social scientists to understand representations in movies.

Our work is the first to study how linguistic features can be used to predict violence in movies both at utterance- and movie-level. This comes with certain limitations. For example, our approach does not account for modifications in post-production (e.g., an actor delivering a line with a threatening tone). We aim to analyze this in our future work with multimodal approaches using audio, video and text. Our results suggest that sentiment-related features were the most informative among those considered. In our future work, we would like to explore sentiment analysis models complementary to the lexicon-based approaches we used.

Acknowledgements

We thank the reviewers for their insights and guidance. We acknowledge the support from Google and our partners Common Sense Media. VRM is partially supported by Mexican Council of Science and Technology (CONACyT). We would also like to thank Naveen Kumar for the helpful feedback during this work.

References

Ali, A., and Senan, N. 2018. Violence Video Classification Performance Using Deep Neural Networks. In *International Conference on Soft Computing and Data Mining*, 225–233.

Alia-Klein, N.; Wang, G.-J.; Preston-Campbell, R. N.; Moeller, S. J.; Parvaz, M. A.; Zhu, W.; Jayne, M. C.; Wong, C.; Tomasi, D.; Goldstein, R. Z.; Fowler, J. S.; and Volkow, N. D. 2014. Reactions to Media Violence: Its in the Brain of the Beholder. *PLOS ONE* 9(9):1–10.

American Academy of Pediatrics and Others. 2001. Media violence. *Pediatrics* 108(5):1222–1226.

Anderson, C. A., and Bushman, B. J. 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):353–359.

Anderson, C. A., and Dill, K. E. 2000. Video games and aggressive thoughts, feelings, and behavior in the laboratory and in life. *Journal of Personality and Social Psychology* 78(4):772.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *CoRR* abs/1409.0473.

Banchs, R. E. 2012. Movie-DiC: A Movie Dialogue Corpus for Research and Development. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, ACL 2012*, 203–207.

Barranco, R. E.; Rader, N. E.; and Smith, A. 2017. Violence at the Box Office. *Communication Research* 44(1):77–95.

Burnap, P., and Williams, M. L. 2015. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet* 7(2):223–242.

Cho, K.; van Merrienboer, B.; Bahdanau, D.; and Bengio, Y. 2014. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. *CoRR* abs/1409.1259.

Chollet, F., et al. 2015. Keras. `https://keras.io`.

Cornish, A., and Block, M. 2012. NC-17 Rating Can Be A Death Sentence For Movies. [Radio broadcast episode] https://www.npr.org/2012/08/21/159586654/nc-17-rating-can-be-a-death-sentence-for-movies.

Dai, Q.; Zhao, R.; Wu, Z.; Wang, X.; Gu, Z.; Wu, W.; and Jiang, Y. 2015. Fudan-Huawei at MediaEval 2015: Detecting Violent Scenes and Affective Impact in Movies with Deep Learning. In *Working Notes Proceedings of the MediaEval 2015 Workshop*.

Davidson, T.; Warmsley, D.; Macy, M. W.; and Weber, I. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017*, 512–515.

Dixon, L.; Li, J.; Sorensen, J.; Thain, N.; and Vasserman, L. 2017. Measuring and Mitigating Unintended Bias in Text Classification. In *Proceedings of AAAI/ACM Conference on Artificial Intelligence, Ethics and Society*.

Djuric, N.; Zhou, J.; Morris, R.; Grbovic, M.; Radosavljevic, V.; and Bhamidipati, N. 2015. Hate Speech Detection with Comment Embeddings. In *Proceedings of the 24th International Conference on World Wide Web*, 29–30.

Eyal, K., and Rubin, A. M. 2003. Viewer aggression and homophily, identification, and parasocial relationships with television characters. *Journal of Broadcasting & Electronic Media* 47(1):77–98.

Fast, E.; Chen, B.; and Bernstein, M. S. 2016. Empath: Understanding topic signals in large-scale text. In *Proceedings of the Conference on Human Factors in Computing Systems, CHI 2016*, 4647–4657.

Founta, A.; Chatzakou, D.; Kourtellis, N.; Blackburn, J.; Vakali, A.; and Leontiadis, I. 2018. A Unified Deep Learning Architecture for Abuse Detection. *CoRR* abs/1802.00385.

Gilbert, C. H. E. 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media ICWSM 2014.*

Gorinski, P. J., and Lapata, M. 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, NAACL-HLT 2018*, 1770–1781.

Hochreiter, S., and Schmidhuber, J. 1997. Long Short-Term Memory. *Neural Computation* 9(8):1735–1780.

Igartua, J.-J. 2010. Identification with characters and narrative persuasion through fictional feature films. *Communications* 35(4):347–373.

Kwok, I., and Wang, Y. 2013. Locate the Hate: Detecting Tweets Against Blacks. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, 1621–1622.

Le, Q., and Mikolov, T. 2014. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, 1188–1196.

Markey, P. M.; French, J. E.; and Markey, C. N. 2015. Violent Movies and Severe Acts of Violence: Sensationalism Versus Science. *Human Communication Research* 41(2):155–173.

Mehdad, Y., and Tetreault, J. 2016. Do Characters Abuse More Than Words? In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 299–303.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*. 3111–3119.

Mironczuk, M., and Protasiewicz, J. 2018. A recent overview of the state-of-the-art elements of text classification. *Expert Systems with Applications* 106:36–54.

Motion Picture Association of America. 2017. Theme Report: A comprehensive analysis and survey of the theatrical and home entretainment market environment (THEME) for 2017. Technical report. [online] Accessed: 07/25/2018.

Nielsen, F. Å. 2011. A new ANEW: evaluation of a word list for sentiment analysis in microblogs. In *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages*, 93–98.

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*, 145–153.

Park, J. H., and Fung, P. 2017. One-step and Two-step Classification for Abusive Language Detection on Twitter. *CoRR* abs/1706.01206.

Pavlopoulos, J.; Malakasiotis, P.; and Androutsopoulos, I. 2017. Deeper Attention to Abusive User Content Moderation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017*, 1125–1135.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12(Oct):2825–2830.

Pennebaker, J. W.; Boyd, R. L.; Jordan, K.; and Blackburn, K. 2015. The development and psychometric properties of LIWC2015. Technical report.

Potter, W. J.; Vaughan, M. W.; Warren, R.; Howley, K.; Land, A.; and Hagemeyer, J. C. 1995. How real is the portrayal of aggression in television entertainment programming? *Journal of Broadcasting & Electronic Media* 39(4):496–516.

Ramakrishna, A.; Martínez, V. R.; Malandrakis, N.; Singla, K.; and Narayanan, S. 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*, 1669–1678.

Sap, M.; Prasettio, M. C.; Holtzman, A.; Rashkin, H.; and Choi, Y. 2017. Connotation Frames of Power and Agency in Modern Films. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017*, 2329–2334.

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*, 1–10.

Smith, S. L.; Wilson, B. J.; Kunkel, D.; Linz, D.; Potter, W. J.; Colvin, C. M.; and Donnerstein, E. 1998. Violence in television programming overall: University of California, Santa Barbara study. *National Television Violence Study* 3:5–220.

Sparck Jones, K. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation* 28(1):11–21.

Sparks, G. G.; Sherry, J.; and Lubsen, G. 2005. The appeal of media violence in a full-length motion picture: An experimental investigation. *Communication Reports* 18(1-2):21–30.

Stone, P. J.; Dunphy, D. C.; and Smith, M. S. 1966. *The general inquirer: A computer approach to content analysis.* MIT press.

Susman, G. 2013. Whatever happened to nc-17 movies? https://www.rollingstone.com/movies/movie-news/whatever-happened-to-nc-17-movies-172123/. Accessed: 2018-08-29.

Tang, D.; Wei, F.; Yang, N.; Zhou, M.; Liu, T.; and Qin, B. 2014. Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014*, 1555–1565.

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

Topel, F. 2007. TMNT's Censored Violence. Retrieved from http://www.canmag.com/nw/6673-kevin-munroe-tmnt-violence.

Waseem, Z.; Davidson, T.; Warmsley, D.; and Weber, I. 2017. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, 78–84.

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.

Wiegand, M.; Ruppenhofer, J.; Schmidt, A.; and Greenberg, C. 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*, 1046–1056.

Wulczyn, E.; Thain, N.; and Dixon, L. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, 1391–1399.

Yokota, F., and Thompson, K. M. 2000. Violence in G-rated animated films. *Journal of the American Medical Association* 283(20):2716–2720.

Zhang, Z.; Robinson, D.; and Tepper, J. 2018. Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network. In *The Semantic Web*, 745–760.

Zhong, H.; Li, H.; Squicciarini, A. C.; Rajtmajer, S. M.; Griffin, C.; Miller, D. J.; and Caragea, C. 2016. Content-Driven Detection of Cyberbullying on the Instagram Social Network. In *International Joint Conferences on Artificial Intelligence*, 3952–3958.

Zillmann, D. 1971. Excitation transfer in communication-mediated aggressive behavior. *Journal of Experimental Social Psychology* 7(4):419–434.