Detecting Objectionable Content in Online Media

by
Mahsa Shafaei

A dissertation submitted to the Department of Computer Science,

College of Natural Sciences and Mathematics

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in Computer Science

Chair of Committee: Thamar Solorio

Committee Member: Edgar Gabriel

Committee Member: Ioannis Kakadiaris

Committee Member: Fabio Gonzalez

University of Houston
December 2021

ACKNOWLEDGMENTS

I would like to express my heartfelt gratitude to my advisor, Dr. Thamar Solorio, for her constant support, patience, and supervision. Without her encouragement and supervision, this research would not have been possible. I would like to express my appreciation to Drs. Edgar Gabriel, Ioannis Kakadiaris, and Fabio Gonzalez for serving as members of my dissertation committee and offering helpful input to help me enhance my study.

This dissertation would not have been feasible without my husband's encouragement and assistance, as well as the unconditional love of my family specially my mother. I am eternally grateful to them all for providing the necessary support and motivation.

ABSTRACT

In this dissertation, we discuss methods for having a system detect automatically objectionable content in online media. Movies, animations, trailers, and video blogs are vastly accessible by younger audiences through the movie service providers (e.g., Amazon and Netflix), YouTube, movie theatres, and generally the web. The online content helps us learn and inspire societal changes. But it can also contain objectionable content that negatively affects viewers' behavior, especially children. For some media content (like movies, books and trailers), we do have a rating system. For example, the rating system for movies is adopted from the Motion Picture Association of America (MPAA), consists of manual inspection of movies to assign an age rating. However, there are some issues regarding this rating system. First, the current system announces a single rating for the whole content. Yet, suitability is partially related to the culture, people's background, emotional and cognitive skills of children. Thus, having a single rating is not always helpful, and more details are needed. Second, this manual process does not scale to an ever-increasing number of online videos available on the internet.

As the first step towards the main goal of this dissertation (detecting objectionable content), we design, implement, and evaluate a system that is capable to predict movies and trailers age suitability rating without a human observation to explore different models for the task. The system that we propose either employs only the script of the movies as the input, or it takes advantage of all modalities and combines all cues from acoustic, visual and textual information for detecting the objectionable content. The script-based system can be utilized at the early steps of the production when we only have the script. The multi-modal version, however, can be used after the production when a video is fully ready. Finally, we expand our multi-modal model to automatically generate the list of objectionable elements in any kind of video. In this dissertation, we focus exclusively on "Comic Mischief" elements, which no one has attempted previously.

Along with the system, we propose the biggest corpus of movie scripts that comes with metadata, poster images, and movie trailers that are rated by the MPAA institution. We also compile a dataset including a wide range of videos that are tagged with comic mischief elements in video scenes. Finally, we make the implementation and data resources available for further research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 Introduction

Recent studies on children's screen time reveal an alarming increase in the amount of time they spend watching movies or videos on electronic devices [10]. While some of the information in these films and videos is absolutely harmless, there is also some dangerous and inappropriate material that may have a detrimental effect on their conduct. For instance, watching some programs may encourage teenagers to engage in risky sexual relations and alcohol consumption [56, 52, 1] or instill dread and fear in children [63, 29].

1.1 Motivation

Identifying questionable components in media content serves a variety of practical purposes. For instance, parents can use this method to assess what type of content is appropriate for their children. Additionally, media service providers (for example, Amazon and Netflix) may leverage this system to enable age filters in parental controls. Possessing a list of questionable content is a potentially critical component for producers as well. They can then utilize this list to fine-tune the content for their intended audience.

To our knowledge, no automated technique for detecting questionable content in the media exists. However, we have mechanisms for rating media that are dependent on human judgment. In the United States, the Motion Picture Association of America maintains the current media rating system (MPAA). The MPAA is a film rating system that recommends the proper viewing age for films. Classification and Rating Administration (CARA), one of the subdivisions of MPAA organization, assigns the MPAA rating. CARA members watch the entire film in order to decide the film's age category and MPAA rating [43]. The first shortcoming of this approach is that the existing technique of ranking is a time-consuming and inefficient process. Not surprisingly, there are several films and videos online that lack any rating or guideline for sensitive content. As a result, we require a system capable of automatically processing information and identifying questionable aspects without human intervention.

Moreover, MPAA organization assigns a single rating to the entire film rather than a thorough list of questionable items. Questionable elements are those that may be deemed objectionable by a group of people. For instance, depictions of violence, alcohol consumption, and offensive language in spoken words are some examples of questionable content. A single rating may provide a problem due to the expansive definition of what constitutes appropriate content for youngsters. Appropriateness is determined by a variety of factors, including the content's context, the child's ability to absorb content, and cultural traditions. As a result, our ultimate goal is to provide granular details for a specific type of unacceptable behavior.

1.2 Characteristics and Challenges

For youngsters, objectionable content consists of a variety of components such as drugs, sex, violence, language, and sensitive subjects. Figure 1 depicts some scenes from movies with questionable content.

(a) (b) (c)

Figure 1: Examples of questionable content in movies

As a result, a simplistic approach such as using a list of bad words is insufficient to detect these sensitive aspects. Additionally, the influence of each component is context-dependent. For example, if the word f**k refers to a sexual context, with high probability, it leads to an "R" rating ("You can f**k me in the car" in the movie "To Rome with Love"). But, if it is used as a curse word, the movie can be rated as "PG-13". For instance, the sentence "that's a clear sign to back the f**k off!" is used in "Fast & Furious", and yet the movie is rated as PG-13. Additionally, while words like fat are frequently listed as offensive in social media comments, they are less likely to be

considered inappropriate in films. Swearing is also considered more offensive the greater the gap between speaker and listener in status. If an entry-level employee swore at their boss, that would be considered worse than if two students swore at one another. Thus, we must consider the context of the content.

Moreover, inappropriate content may appear in only one of the video's modalities, or the modalities may clash. For example, the depiction of vivid violent situations without dialogue or in conjunction with a piece of light jazz music. Figure 2 depicts a frame from "The King's Man", which features extremely violent visual content accompanied by jazz music and no dialogue. Thus, more complicated algorithms are required to incorporate data from all modalities of a movie in order to find questionable aspects.

Figure 2: Church fight scene in "The King's Man" movie

1.3 Research Objectives

In this dissertation, we aim to study the problem of detecting questionable content in movies and online video content. The general goal of this dissertation is the following:

Goal: *Design new algorithms to detect objectionable elements in online media content such as movies, trailers, and video clips.*

To achieve this objective, we must first define what constitutes questionable content. The

second step is to create a data resource that is consistent with the objective. This resource may include freely available online videos (conversation texts, audio clips, and image frames) with labels based on a questionable content definition or corresponding age-related ratings. Finally, we must develop algorithms capable of automatically detecting patterns of data indicative of various types of questionable content.

As a first step toward the eventual aim, we will collect publicly available resources from the internet and generate a dataset without the cost of any annotation. This dataset comprises film scripts and trailers that have been rated for their suitability for children by the MPAA organization. Then, we will develop a system for predicting the rating of movie scripts and trailers. To accomplish the project's primary objective (which is to develop a "content descriptor" rather than a "content rater"), we will create a dataset of various types of videos (not just movies) and will annotate them according to the definition of objectionable content discovered in the fields of psychology and mass media communication to detect a specific group of questionable content. Finally, we will develop a novel multi-modal system capable of detecting sensitive elements utilizing auditory, textual, and visual signals. Parents can be alerted about the content via the list of sensitive elements, and the choice about whether the content is appropriate or not will be made by the parents.

1.4 Contributions

We summarize below the contributions of this dissertation.

1. Create a resource for the task of detecting questionable content in media: a dataset of movies, trailers, and YouTube videos, along with the metadata, list of objectionable elements, and components of the rating (violence, sex and nudity, fear inducing material, alcohol and drug usage, and offensive language).

2. Design, develop, and evaluate a system to predict the movie rating at early stages of production by only using the script of the movie. This system will help us to find objectionable content in text-only media and make the prediction independent of having all modalities available.

3. Design, develop, and evaluate a multi-modal pipeline for utilizing the audio, video, and text modalities of online content. Because this system detects questionable elements by capturing all modalities, it can cover a wide range of objectionable content (for example, objectionable content related to speakers' tone, background music, graphical items, or speaker words).

4. Create a novel multi-modal system that employs a "hierarchical cross-attention" mechanism to detect "Comic Mischief" in videos by integrating all modalities in the video regardless of the video type; instead of predicting a single rating, we predict four subcategories of comic mischief: Mature Humor, Slapstick Humor, Gory Humor, and Sarcasm.

1.5 Structure of this Dissertation

In this section, we will provide a brief explanation of the chapters included in this dissertation as follows:

- **Chapter 2** provides a comprehensive overview of previous research done on the related topics.

- **Chapter 3** describes the methods we used to create three new datasets for the movie/trailer rating and video labeling tasks.

- **Chapter 4** introduces our proposed method for predicting MPAA ratings for movies based solely on scripts. Our end-to-end sequence labeling deep neural network architecture is designed to learn the pattern of movie scripts for each MPAA rating category.

- **Chapter 5** describes the multimodal methodology for predicting MPAA ratings for movie trailers. All three streams (audio, text, and imagery) from videos will be combined using well-known fusion models to provide a single model for categorizing a trailer into one of the MPAA ratings.

- **Chapter 6** discusses the novel approach we devised to label any type of video for comic mischief categories by combining textual, acoustic, and imagery modalities. To address this

issue, we propose a multi-task approach that uses an end-to-end neural network model to predict all four categories of comic mischief.

- **Chapter 5** summarizes the main contributions and the significant findings of this dissertation. We also mention possible future research in this area.

2 Related Work

This chapter discusses pertinent works that have been published on similar subjects to those covered in this dissertation.

To begin, we will discuss related text classification techniques. Second, we discuss related audio and video classification research, as we intend to incorporate audio and video classification modules into the proposed model to address the issue. In the fourth subsection, we describe multimodal data fusion methods that have been used in the related area. Finally, we review available datasets for the movie classification topic.

2.1 Text Classification

One of the papers that comes closest to our work is [39]. The authors of this study use scripts to determine whether a film is violent or not. They extract sentiment, semantic, and lexical features from movies and feed them into an RNN-based classifier to forecast violence. Our research is comparable to this because we also use dialogue between movie characters as an input for a portion of our work. However, rather than extracting features from text to feed the model, we mostly use raw data to avoid error propagation caused by feature extraction [45]. Additionally, models' outputs vary. We predict the MPAA rating in our study, not the level of violence (violence is one of the many aspects of the MPAA rating). Also, the authors of [62], in the field of movie text classification, proposed a deep learning architecture for predicting the genre of movies based on the plot synopsis. Their model employed an attention mechanism to determine the significance of features contained in each word in a synopsis. Similarly, in [17], the authors used a Bidirectional LSTM architecture to predict movie genres from plot summaries.

Several studies have been conducted on detecting offensive language in text based on a survey conducted on hate speech detection [53]. These works are relevant to our research because offensive dialogue can have an effect on a film's suitability for children. The researchers in [55, 66, 48, 40] adapted Convolutional and Recurrent Neural Networks to predict abusive language and hate speech

in Twitter data. Additionally, by extracting lexical, syntactic, sentiment, and semantic features and training traditional machine learning classifiers, [12] and [46] automatically predict hate speech in online content.

2.2 Audio Classification

Several types of handcrafted feature extraction techniques have been proposed for the audio modality in the past [13, 20, 47], with Mel-frequency cepstral coefficients (MFCCs) being the most frequently used in the literature. However, several approaches for performing audio classification have been proposed recently that combine audio features such as spectrogram information with deep learning architectures [47, 25, 34]. Numerous works, including [50, 23, 22], have examined audio as a modality for classifying film content. However, none of these methods have addressed the issue of determining the age suitability of videos. For instance, [22] works on predicting violence in videos via visual and audio features extracted from films.

2.3 Video Classification

Earlier approaches, such as [30] on video classification, investigated the use of several temporal fusion methods for combining data from multiple consecutive video frames using features extracted from CNN architectures. The authors in [16] introduced an end-to-end architecture that relies on a combination of CNNs for feature extraction from RGB frames. After that, the CNN features are forwarded to an LSTM layer that models frame temporal variation. A different approach is taken in [60], specifically 3D-CNN, where the authors propose the use of a CNN variant that takes into account both spatial and temporal domain convolutions in a video. [7] proposes an extension of the 3D-CNN approach, in which the authors propose a two-stream 3D-CNN architecture for video classification. Once again, two streams were used as input data: RGB frame data and optical flow images.

2.4 Multimodal Data Fusion

Multimodal data fusion is an attempt to integrate information from different resources to predict an output (class label or real value). Fusion methods can be categorized into three general groups [6]; early, late, and hybrid. In early fusion, low-level features are combined and fed to the prediction model. However, in the late fusion, we merge different modalities in the decision level using various rules (e.g., majority voting, averaging). Hybrid methods have the advantages of both methods. In [19] the authors introduced a late fusion scheme based on SVM classifiers and handcrafted features to perform movie genre classification using information from plot synopsis and movie posters. In [4], the authors introduced a deep learning model based on gated multi-modal units, that can be categorized as a hybrid method to predict movie genres by fusing movie image posters and movie textual plot information. This model is inspired by recurrent neural network. In RNN models, the recurrent units determine how much current and previous information is used to build the current state. In GMU, we measure how much different modalities trigger the activation function, which is used to build the output.

Bidirectional transformers, such as BERT, have been used in a variety of text-related tasks (for example, text classification) and have significantly improved the results in many cases. Some works used bidirectional transformers to fuse different modalities and were able to outperform the state-of-the-art in various tasks. By modifying the BERT architecture, the authors of [38] proposed a multi-modal two-stream model (known as ViLBERT). The model incorporates two parallel streams of visual and linguistic processing, which interact via co-attention transformer layers. The intermediate representation of the input token, which is built using the encoder layer, is used in the main architecture of the BERT to create three vectors: key, query, and value vectors. These three vectors are then fed to the multihead attention layer. The key and value of one modality, as well as the query of the other modality, are sent to the multihead attention module in the ViLBERT model. The idea behind this architecture is to create attention-pooled features of one modality based on the other. For example, we can find the most important words in the text based on image features and vice versa. In another study [32], the authors use BERT in a novel way to

fuse modalities, dubbed the multimodal bitransformer (MMBT). In this work, the modalities are combined at the input level. Each modality is obtained from a pre-trained model and serves as an input segment in the BERT architecture, which is distinguished from other modality by the [CLS] and [SEP] tags. The advantages of this model are as follows: 1) it works for any number of modality, 2) it works when one modality is unavailable, and 3) it is simple to implement. However, as previously stated, each modality is transferred to a feature vector using a pre-trained model. As a result, during self-supervised pre-training, the model is unable to fully leverage multimodal information.

2.5 Existing Movie Classification Datasets

Numerous datasets for movie classification have been proposed in the past. The authors in [14] introduced MediaEval 2013 Violent Scene Detection, which provided annotations for detecting violent scenes in motion pictures. The authors of [11] proposed an evaluation framework for detecting violent scenes in Hollywood and YouTube videos, as well as a dataset (VSD96) containing over 96 hours of video from a variety of genres. They provide annotations at various levels of detail (e.g., shot- and segment-level), as well as annotations of intermediate concepts (e.g., blood, fire). The authors of [9] introduced Moviescope, a dataset for categorizing film genres. This dataset contains 5,043 movie records, as well as their associated metadata, video trailers, and text plots. None of the existing datasets include labels indicating the age suitability of films or movie trailers.

3 Datasets for Questionable Content Prediction

To our knowledge, there is no previous film dataset that includes the age suitability rating for films in addition to the scripts or trailers. Thus far in this research, we've introduced three different types of corpora in order to tackle the questionable content prediction task. The first version includes the film's script, MPAA ratings, and metadata. The second edition is devoted to film trailers along with MPAA ratings and metadata. Both versions are devoid of human annotation, as the labels are based on the MPAA organization's age rating schema. The final dataset is structured in such a way that it supports our research's ultimate objective. We gathered a variety of video types and annotated them with elements of a particular type of questionable content ontology.

3.1 Movie Scripts Dataset with MPAA Ratings

To compile the dataset, we gathered movie scripts from a publicly accessible source (springfields website). It's worth noting that the scripts contain only dialogue between movie characters (without the description of the scenes). We crawl the IMDB website to extract movie metadata. The metadata that we collect includes the MPAA rating, the directors and actors' names, the film's genre, a link to a downloadable poster, and the film's runtime. Additionally, we download poster images for all films in the database via poster links. From the approximately 15k films that we initially collected, the MPAA rating is available for approximately 7k films.

The MPAA classifies films into five categories (G, PG, PG-13, R, and NC-17) based on their suitability for children. G stands for general group; it denotes that all ages are welcome. PG indicates that the film contains some material that parents should review. The PG-13 rating indicates that the film contains some content that is deemed inappropriate for children under the age of 13. R stands for "restricted", which means that anyone under the age of 17 should watch the film with a parent. NC-17 indicates that the film is not recommended for anyone under the age of 17. These ratings are defined in detail at (https://www.mpaa.org). According to MPAA documentation, prior to 1996, this rating used different group names and slightly different definitions for each group

[31, 44, 36]. As a result, we exclude films released prior to 1996 (approximately 1,500 films in the corpus) to avoid inconsistency in the definition of the ratings.

According to the IMDB website, there are approximately 70 films with NC-17 rating, most likely due to the niche market for these films. [1] As a result, we exclude films from this category from our experiments. Table 1 contains the final version of our dataset's statistics.

Table 1: Movie script dataset statistics

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

The MPAA assigns ratings based on the following criteria: (i) Violence, (ii) Language, (iii) Substance Abuse, (iv) Nudity, and (v) Sexual Content, but the MPAA organization does not provide a method for quantifying the amount of content associated with these components. However, the IMDB website contains objectionable content for MPAA-compliant films (Parental Guide): 1) Blood and Gore, 2) Sex and Nudity, 3) Scary and Intense Scenes, 4) Profanity, and 5) Alcohol, Drugs, and Smoking. The IMDB website allows users to rate the severity of the aforementioned types of content using the following labels: *None*, *Mild*, *Moderate*, and *Severe*. We track the number of votes received for each label assigned to each component. Due to the fact that these tags are based on user votes, not all films with an MPAA rating include these components. The data statistics for these components are shown in Table 2. Each cell in the table represents the number of films assigned a severity level (None to Severe) for each component (violence, profanity, etc.) within each MPAA category (G, PG, PG-13, R, NC-17). For example, there are four films in category G that are *None* for *violence*. And the reason is that the majority of voters chose *None* (maybe not all of them).

In Table 3, we display the distribution of films by genre. Between different groups, there is a class imbalance (in terms of genre), as certain genres are more popular in the film industry. We leave this issue alone in order to maintain the dataset's representativeness of the real-world situation.

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

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

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

Table 3: Distribution of movies in each genre.

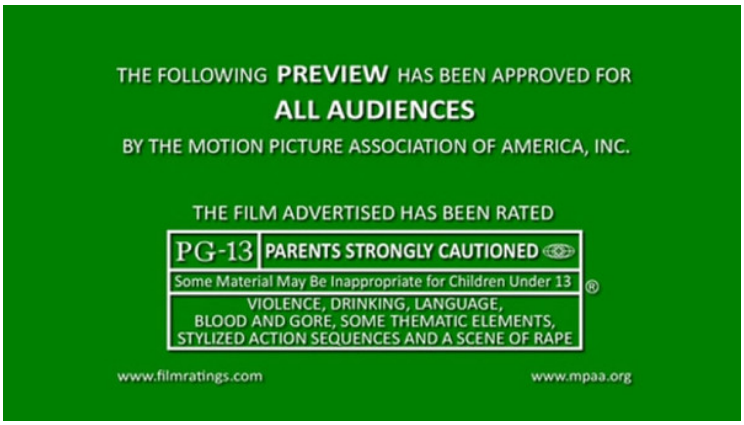
Genre	**#**	**Genre**	**#**
Science-Fiction	619	Action	1,277
Horror	800	Animation	296
Crime	1,000	Adventure	806
Romance	1,082	History	216
News	8	Western	78
Comedy	1,999	War	214
Thriller	1,785	Short	14
Mystery	618	Biography	374
Musical	297	Drama	2,965
Documentary	189	Family	552
Sport	220	Fantasy	558

3.2 Movie Trailer Dataset with MPAA Ratings

The Multimodal Movie Trailer (MM-Trailer) dataset was created by collecting rated trailers from the IMDB and YouTube websites. Typically, trailers promote upcoming films and are shown in theaters prior to the start of the film. Trailers are rated differently than films. As previously stated, there are five distinct MPAA categories for films: NC-17, R, PG-13, PG, and G, which are listed in order from most restricted audience (NC-17) to most open audience (G). However, in trailers, the rating is indicated by a colored band (red, yellow, or green) and a message at the beginning of the trailer. It's worth noting that the trailer rating is unrelated to the film's rating, as trailers can be sanitized to remove sensitive material. Figure 3 illustrates two different types of trailer bands.

The rating bands' color and message have changed over time. Prior to April 2009, the green band at the start of trailers implied that they were appropriate for all audiences, including children. After April 2009, the green band's message was changed to "appropriate audience". In other words, the trailers' rating corresponds to the film's theatrical release rating. For example, if the film playing in the theater is rated NC-17 (no one under the age of 17 is recommended to see this film), the

(a) Red band sample



(b) Green band sample

Figure 3: Two examples of trailer bands in green and red colors.

green band trailer advertised prior to the film may not be appropriate for children, regardless of the color green. The yellow band is intended for trailers advertised on the internet and indicates that the trailer is appropriate for "age-appropriate internet users", as site visitors are predominantly adults. The final group of trailers are those with a red band; the red color indicates that the content is only appropriate for a "mature" or "restricted" audience.

Given that our objective is to develop an automated system capable of predicting which movie trailers are not suitable for children, the dataset contains only two types of trailers:

1. **Green-band trailers:** this category includes (i) trailers with the message "all audiences", and (ii) green band trailers with "appropriate audience" for films with a G or PG rating.

2. **Red-band trailers:** all red-band trailers, these include restricted and mature audiences (not appropriate for children).

Table 4: Dataset statistics

Green Band Trailers	Red Band Trailers	Total Trailers
1,040	403	1,443

We also extracted separate audio files and subtitles for the trailer. Certain YouTube trailers incorporate the video's subtitle. In these instances, we pre-process the subtitles by omitting timestamps and retaining only the text. For trailers that do not include a subtitle file, we automatically generate the subtitle from the audio using the Python speech recognition tool [65]. Our dataset contains 11G of audio streams. The audio file for each trailer is a combination of background music and vocals, so the duration of the audio track is identical to the duration of the trailer. The total number of words in all trailer scripts is 1,478,139 (on average, there are 576 words per trailer). Take note that 20,783 of the vocabulary set's words are unique.

Additionally, we provide metadata for the films themselves, which includes the film's IMDB ID, title, genre, cast and director credits, and a link to the poster image. Our dataset's statistics are summarized in Table 5.

3.3 YouTube Videos with Comic Mischief Labels

Content that is questionable may have a variety of interpretations across cultures. Moreover, people's tolerance levels may vary according to their age, life experience, and cognitive skills. Thus, rather than assigning a single rating to a video, we aim to provide information to potential viewers about the presence of specific questionable content. A system of labeling questionable content is a more flexible alternative to the most widely used age-based rating systems. Based on the questionable labels, the viewer or parent/guardian can then determine whether the video satisfies their criteria for what constitutes acceptable content. For instance, someone may be excessively sensitive to violence but not to profanity.

To develop this technology, we will need a large repository of consistently and exhaustively labeled video content, which did not exist previously. To create this data resource, we completed these steps: 1) Compile a taxonomy of questionable content; 2) Compile a diverse collection of English-language online videos; 3) Enhance the annotation tool; 4) Establish an incremental annotation strategy.

As a first step, we must enrich the taxonomy of questionable content and establish guidelines for annotators. To create a guideline for annotation, we did the following: 1) Compiled desiderata for annotation of questionable content and developed annotation guidelines; 2) Compiled a list of diverse online videos for use in pilot annotations; 3) To revise the questionable content ontology and annotation process, we conducted four workshops with participants from the fields of psychology, artificial intelligence ethics, computer vision, and natural language processing; 4) Finally, fine-tuned the guidelines in light of pilot annotation experience and discussion.

Figure 4 depicts the ontology of questionable content. The taxonomy of all questionable content is extremely diverse, and it is impractical to annotate a video for every possible group within the questionable taxonomy (the error of annotation will be high). We will focus exclusively on "Comic Mischief" in this research, which no one has attempted previously. Each label may use any or all of the three video modalities: Sound, Dialogue, Soundtrack, and Video to represent itself. Definition of modalities are:

1. Dialogue: Audio and text transcription of spoken dialogue between characters.
2. Sound: Sound effects and ambient sounds (e.g., explosions, shootings, dismemberment, moaning).
3. Soundtrack: Music of the video (including transcripts of lyrics).
4. Video: Pixel information from video frames.

We have four types of content under the comic mischief category. Each of these categories is defined as follows:

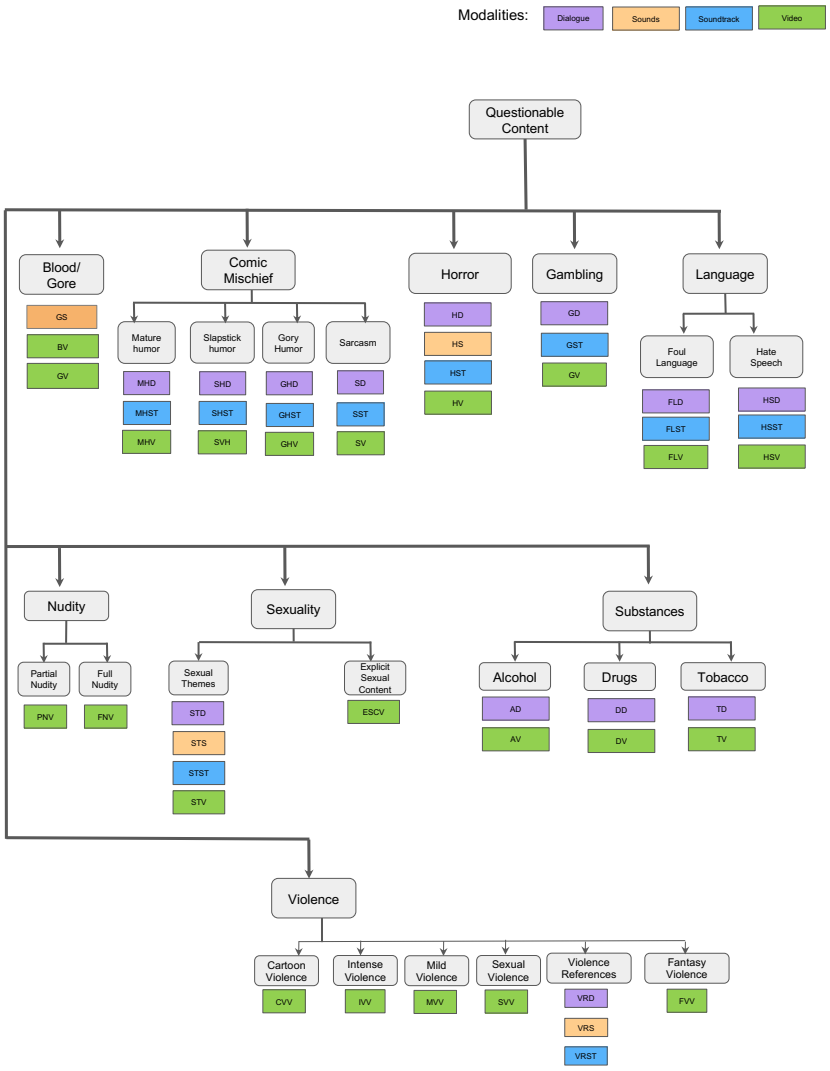


Figure 4: Ontology of Questionable Content

1. Gory Humor: Gory is a term that refers to something that involves a great deal of bloodshed and violence, such as a film in which victims are axed to death by a psychopath. Gory Humor is a term that refers to situations in which a gory scene is juxtaposed with humorous references.

2. Slapstick Humor: Slapstick is a style of comedy that is characterized by practical jokes, collisions, clumsiness, and embarrassing events. An example of slapstick is a comedy performed by the USA television characters called the Three Stooges, where people get poked in the eye or pies in the face.

3. Sarcasm: Sarcasm employs words that are typically used to mock or annoy someone, or for humorous effect. Sarcasm may employ ambiguity, but it is not always ironic. Sarcasm is most noticeable in spoken word and is largely context dependent. It is primarily distinguished by the inflection with which it is spoken.

4. Mature Humor: Sexual references are included in depictions or dialogue that contain "adult" humor. Strong Language, Alcohol/Drug Consumption, and Sexual References are all part of adult humor.

3.3.1 Video Collection Process

Our objective is to create a diverse collection of English-language videos. We chose YouTube as a data source because it has several advantages: it is freely and easily accessible to anyone with internet access. Also, it enables users to upload videos easily, ensuring a diverse range of video content. To discover new videos, we defined a simple algorithm that takes advantage of YouTube's recommendation system:

1. We begin by manually searching YouTube for videos that contain the type of content we're looking for. These are the videos that will serve as our seed material.

2. For each seed video, we compile a list of YouTube's recommended videos to watch next. YouTube recommends videos to users based on their content. As a result, we can anticipate some content similarity. These suggested videos become the next set of seed videos, and this process is repeated to collect additional videos.

3. To avoid large concept drift, we iterate three times through this process. However, as we progress deeper into the exploration process, we collect fewer videos per seed node.

3.3.2 Annotation Process and Quality

To ensure the annotation process is of high quality, we collect three-way annotations for videos. The annotation process was carried out by 6 annotators (4 graduate students and 2 bachelor students).

The annotation was done using an html-based annotation tool that included the capability to play the video and a list of all possible labels that annotators could assign to the video. Each label can have a value of 1 or 0, indicating whether it corresponds to the video content or not. It should be noted that each label can be displayed in any of the video's modalities (Dialogue, Sound, Soundtrack, Video). As a result, we have distinct labels for each modality (e.g., Gory Humor - Dialogue, Gory Humor - Video). The final value assigned to each label is determined by a majority vote of all annotators. Thus, if at least two annotators agree on a label, we will assign it to the video. To quantify the annotation's quality, we compute the Inter-Annotator Agreement (IAA) by calculating Cohen's kappa on the annotations elicited from each annotator and the majority voting on all annotations. Based on the computed IAA values, there is a substantial agreement (IAA = 0.70).

3.3.3 Dataset Statistics

Table **??** summarizes the dataset's statistics. As can be seen from this table, some of the videos lack dialogue.

Table 5: Dataset statistics

	Max	Min	Avg	Median
# Script Words	266	0	114.34	122
Video/Audio Length	71.97	0.13	57.42	60.39
# Frames	2157	1	622.115	469

Table 6: Dataset statistics for comic mischief class

Class 1	Max	Min	Avg	Median
# Script Words	266	0	117.83	125
Video/Audio Length	71.97	9.44	58.57	60.46
# Frames	2157	108	657.69	478

Table 7: Dataset statistics for class without any comic mischief labels

Class 0	Max	Min	Avg	Median
# Script Words	259	0	106.08	111
Video/Audio Length	64.9	0.13	54.73	60.07
# Frames	1836	1	538.4	460.5

Table 7 and Table 6 show the statistics for each class separately. Class 1 indicates that at least one of the comic mischief categories is present in the video; Class 0 indicates that none of the categories are present. According to the numbers, the length of videos and the average number of words per video are comparable for both classes, indicating that both have videos with sufficient length to include meaningful content.

4 Predicting the MPAA Rating Based on Movie Dialogues

In this chapter of our research, we will attempt to classify films according to their MPAA rating based on their script content. Using the scripts to predict the MPAA rating is not a simple task. This rating is based on a number of content elements, including those relating to drugs, sex, violence, language, and sensitive subjects. As a result, a simplistic approach such as using a list of bad words is insufficient to predict the MPAA rating. To our knowledge, no prior work has been conducted on predicting the MPAA rating. Existing work, such as [39, 2, 49], has a more narrow focus on detecting violence in films, and according to studies such as [28] and [61], violence prediction alone is insufficient to predict the MPAA rating.

4.1 Methodology

To address the challenges presented by this task, we incorporated a variety of different types of data into our model, including conversational data, emotional dynamics between characters, film genre, and similar films to the target film. The majority of these resources, such as the film's script or metadata information, are available from the start of production. However, similar films will not be available until the film is released. Thus, depending on the time period for which the prediction is required, we can use a variety of models. For instance, when predicting the MPAA rating of an unrated released film, we can use a model that leverages similar films to make a more accurate prediction.

Figure 5 depicts the proposed model's overall architecture. The model is composed of the following layers: 1) an embedding layer that converts the words to vector representations, 2) a long short-term memory (LSTM) layer that learns the temporal dependency of the words, 3) an attention layer that determines the importance of each word in the sequence, 4) emotion, genre, and similarity vectors that provide contextual information to the model, and finally 5) a prediction layer. The following sections will delve into the model's details.

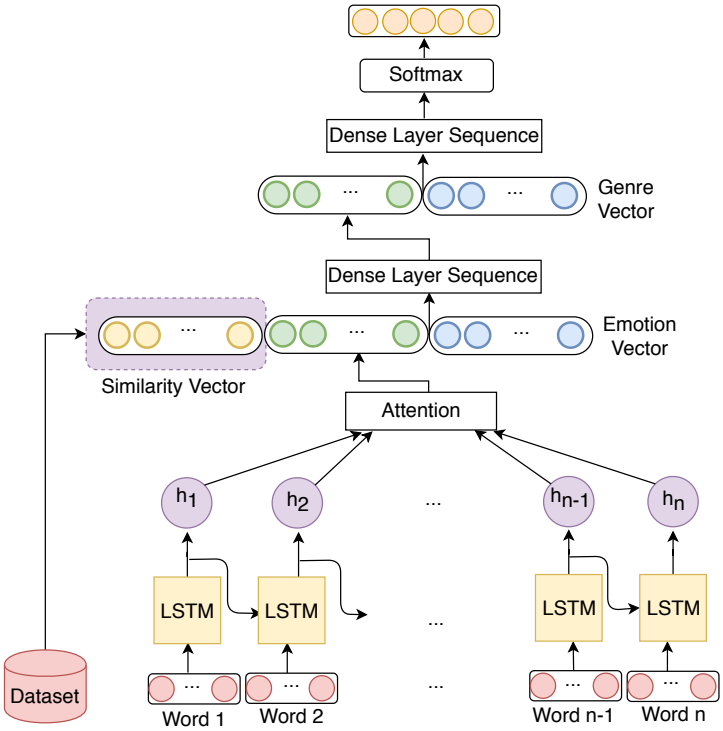


Figure 5: Overall architecture of the proposed model (the similarity vector is appended in case of late prediction; when similar movies are available for the target movie).

4.1.1 Embedding Layer

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

4.1.2 LSTM Layer with Attention

We used the LSTM layer to extract the sequential information from the scripts in order to capture the context of each word. This layer converts an array of embedded vectors to an array of hidden

[2] `https://nlp.stanford.edu/projects/glove/`

vectors. Then we provided the attention mechanism with the resulting hidden representation. We used the same attention model as [5]. This layer computes the weighted sum r as $\sum_i \alpha_i h_i$ to aggregate hidden layers of LSTM to a single vector. The model can learn the relative importance of hidden states (h_i) by learning the α_i. We compute α_i as follows:

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

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

4.1.3 Emotion Vector

Emotional content in film dialogues can aid the model in contextualizing conversations between characters and also in discriminating between films with varying rankings. For instance, we anticipated G films (the most appropriate films for children) to contain less content about "fear" or "disgust" and more about "joy" and "happiness". We used the NRC emotion lexicon [42] to extract emotion from the text. This dictionary associates words with binary values representing eight distinct emotions (anger, anticipation, joy, trust, disgust, sadness, surprise, and fear) and two sentiments (positive and negative). We calculated the normalized count of words used to express each emotion throughout the film. As a result, we have a vector $[e_1, e_2, ..., e_{10}]$ for each movie, where e_i is the percentage of words corresponding to emotion i. To demonstrate the validity of our emotion hypothesis, we illustrated the average emotion scores for each class (for movies with the same MPAA rating, we averaged all the scores per emotion). According to Figure 6, certain emotions are more prevalent in a particular class of films. For example, the values of negative emotions like *disgust*, *anger*, and *fear* are higher for movies rated for older audiences. While, *anticipation* and *surprise* have higher rates in the G category than in other classes.

We also demonstrated the validity of our claim using several sample movies from the training set. We randomly selected a G-rated film ("College Road Trip") and an R-rated film ("Guernica"). Then, we classified sentences from these films according to their average emotion score over the words (we ignored sentences with less than four words). Table 8 demonstrates that the R-rated film

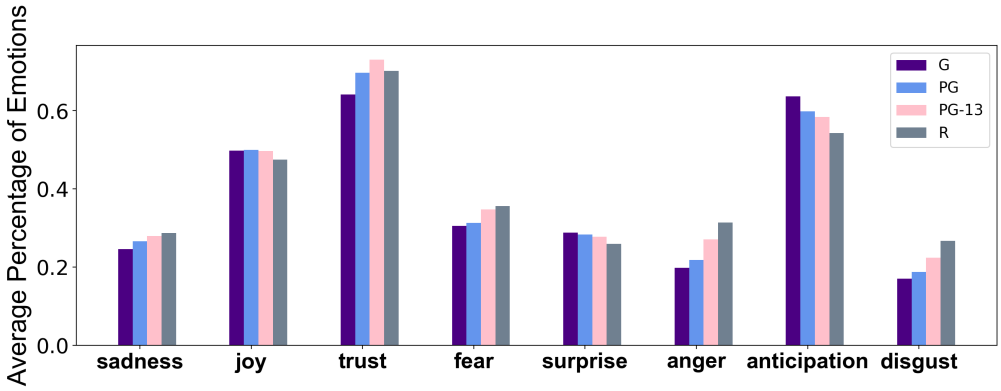


Figure 6: The distribution of MPAA categories across various emotions in the training set. The y-axis shows the average value for the emotion scores for all movies in each rating.

contains more intense sentences for *anger* and *fear*, while the G-rated film contains more intense sentences for *joy* and *surprise*.

Table 8: Sample of top rated sentences for some emotions(joy, anger, fear, surprise) for a G-rated ("College Road Trip") and an R-rated ("Guernica") movie.

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

These trends support our hypothesis that emotion vectors could aid in task improvement. As a result, we concatenated the emotion vector with the attention output to incorporate emotion information into the model.

4.1.4 Genre Vector

The genre can provide information about the theme of the movie. For instance, certain dialogues are considered violent in one genre but are considered harmless in another (an action film vs. a sport

film) [39]. As a result, we leveraged this information by augmenting the model with a genre vector. Our corpus contains a total of 24 genres, but some films are assigned to multiple genres. Thus, we created a binary multi-hot vector to represent the film's genres. Each cell in the vector represents a genre, with a value of one if the genre corresponds to one of the film's genres and zero otherwise.

To illustrate the effect of genre, we showed the distribution of various MPAA ratings for the training data across various genres (Figure 7). Certain genres are more appropriate for children than others, as indicated by this figure. For instance, the MPAA ratings for *Animation*, *Adventure* and *Family* indicate that these genres are more appropriate for children than *Drama*, *Horror*, and *Crime*.

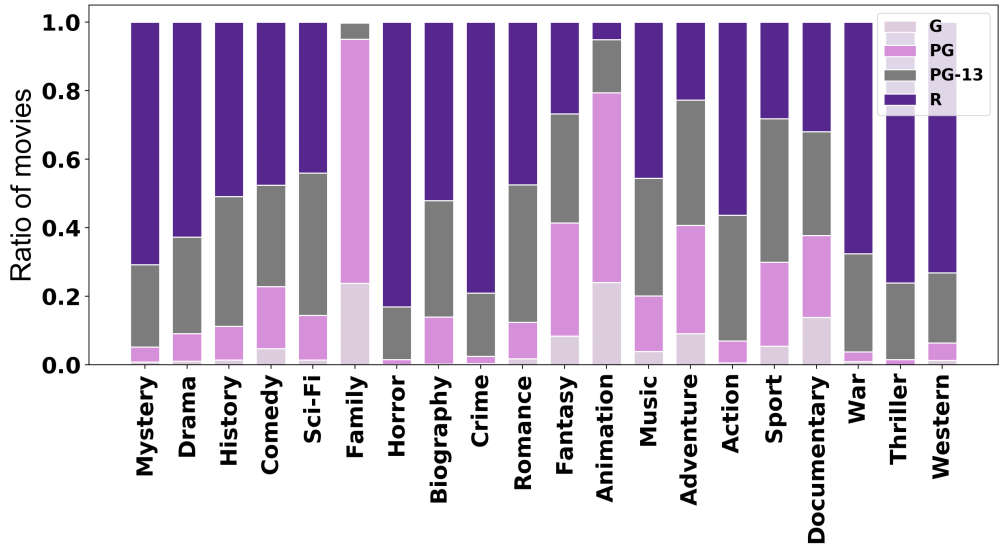


Figure 7: MPAA ratings distribution per genre in the training set; number of movies for each genre is normalized.

4.1.5 Similar Movies Vector

As mentioned previously in Chapter 3, IMDB provides a list of similar movies for each movie. IMDB determines the similarity based on several factors, including genre, country of origin, and actors. Intuitively, we can deduce that two similar films may also have a similar MPAA rating. As a result, we can take advantage of this information when it becomes available. To accomplish this, we created a five-dimensional vector for each film. The vector is $[p_G, ..., p_{NC-17}]$ and p_i shows the percentage

of similar movies with the MPAA rating equal to $i \in \{G, PG, PG-13, R, NC-17\}$.

4.1.6 Dense Layer and Output Layers

After concatenating the feature vectors, we used two dense layers to fine-tune the information. To avoid over-fitting, we used batch normalization and a dropout rate after the hidden layer. Due to the fact that we have a multi-class classification, we used the softmax activation function to calculate the probability of each group in the final step (G, PG, PG-13, R).

4.2 Experiments

Due to the imbalance in our data, we used random stratified sampling and split the data into 80% training, 10% development, and 10% test sets (for all experiments we use the same train, validation, and test data). We used class weights in the loss function to smooth out the imbalance problem. The weighted F1-score is the metric we use to quantify model performance.

4.2.1 Baseline Systems

We define several baselines in this section to evaluate the performance of our proposed model.

Threshold Model: Our first baseline considers only offensive words that have been used in film scripts. To create a list of bad words, we compiled an online list[3] and supplemented it with words listed in [26]. Using this list, we calculated the percentage of bad words in each movie. Then, we found the best thresholds for the percentage of bad words among all threshold values {0.0001, 0.0002,...,0.05} by performing grid search on the validation set. The final model will consist of a list of thresholds; all films with bad words less than t_1 will be labeled G, all films with bad words between t_1 and t_2 will be labeled PG, and so forth. The purpose of this baseline is to demonstrate that simply having a list of bad words does not suffice to determine a film's suitability for children.

SVM Model: The second baseline is comparable to the model proposed by [54] for predicting movie success. The optimal combination of features for the model is a unigram, a bigram, a bag-of-genres,

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

and a bag-of-directors. Additionally, we included an emotion vector in the feature set to allow for a fair comparison to our deep learning model. We tuned hyper-parameter C of the SVM model, C, using grid search method $\in \{1, 10, 100, 1000\}$.

Martinez'19 [39]: As previously stated, this work achieved the state-of-the-art result for predicting violence in films using only scripts and metadata. Although the dataset used in this study is not publicly available, the code is.[4] We applied the same model to our dataset and used the result as an additional baseline.

SVM+Similarity: The final baselines are based on MPAA ratings of similar films. Due to the fact that the IMDB website does not explain all of the factors that contribute to the similarity metric, we may assume that one of the most significant factors is the MPAA rating. Thus, having these ratings for similar films in the model simplifies the problem. To demonstrate that this assumption is incorrect, we presented a baseline model that predicts the MPAA rating for the target film solely using the average MPAA rating of similar films. Additionally, we added this average rating to baseline 2 (SVM model) in order to compare the traditional and deep learning models fairly.

4.2.2 Experimental Setup

We used Pytorch to implement our model. To tune hyper-parameters, we ran experiments on validation set for the model with different learning rates {0.00001, 0.0001}, number of LSTM's hidden units {32, 64, 128, 256}, and dropout rates {0.3, 0.4, 0.5}. Additionally, to avoid over-fitting, we used L2 regularization in addition to the dropout. We calculated the loss between predicted and actual labels using the binary cross-entropy loss function and Adam as the optimizer [33]. Best performance obtained by the following set of parameters: dropout = 0.3, learning rate = 0.00001, LSTM hidden units = 256. We trained over 100 epochs and consider the model with the best weighted-F1 score on the validation set as the final model to apply on the test set.

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

4.3 Results

Quantitative Results: Table 9 summarizes the classification results for predicting the MPAA rating of films using the weighted-F1 score. To disentangle the effects of genre and emotion vectors on performance, we conducted experiments with our proposed LSTM with Attention architecture (L&A model) without incorporating genre or emotion information. Additionally, we investigated the contribution of each vector to the results by adding them separately to the model (L&A with emotion and L&A with genre).

Table 9: Classification results in terms of weighted F1-score for four-class classification.

Models	F1-Score
Baseline 1- Threshold model	65.89
Baseline 2- SVM	74.29
Baseline 3- [39]	75.06
LSTM with Attention layer (L&A)	78.30
L&A with genre	79.49
L&A with emotion	78.94
L&A with emotion+genre	**81.62**
Baseline 5- SVM (Only Similarity)	57.49
Baseline 6- Baseline 2+Similarity	77.70
L&A with emotion+similarity	83.26
L&A with similarity	80.53
L&A with genre+similarity	81.26
L&A with emotion+genre+similarity	**83.68**

Our proposed "LA with emotion+genre" model achieves the best result for early prediction (without relying on similar films information). This model has a weighted F1-score of 81.62%, which is 7.3% higher than the traditional machine learning model. Additionally, it outperforms the baselines "[39]" and "Threshold" by 6.56% and 15.73%, respectively. According to the findings, both genre and emotion vectors enhance the performance of a simple LSTM model with attention. These findings corroborate our hypothesis about the importance of emotion and genre modeling in predicting the MPAA rating. To aid in comprehension of the results, we included the confusion matrix for our best model in Table 10. Based on the matrix, our model is only capable of correctly predicting a few instances of the G-class, despite the fact that we added class weights to the loss function to smooth out the problem of imbalanced data (we only have 162 instances from class G). However, the model correctly predicts 88.5% of R movies. All of the errors in this category are

predicted to be PG-13, the closest classification to R.

Table 10: Confusion matrix of the best model (L&A with emotion+genre) for predicting MPAA ratings. Rows are target tags and columns are predicted ones.

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

In those instances where we have similar movies, we can create a more accurate model. While the similarity metric improves the SVM model's performance, it does not outperform the deep-learning model. By incorporating emotion, genre, and the average MPAA rating of similar films into the L&A model, the model achieves an F1-score of 83.68% and outperforms the corresponding SVM model by 5.98%.

While genre can aid the model in improving its performance, it is insufficient on its own to predict the MPAA rating. Table 11 shows the weighted F1-score for different genres using our best model. Based on this table, for genres like *Comedy* and *Drama*, which contain movies with different MPAA ratings, the performance is better than *Family*, *Western* and *War*, even though most of the movies in these genres (*Family*, *Western* and *War*) belong to a single rating. For genres like *Crime*, *Horror*, and *Thriller* (that for the most instances have one rating), the performance is better than other genres (93%, 87%, 88% respectively), but not that far from genres like *Romance*, and *Action* (with 84%, and 81% respectively) that contain movies with more varied ratings. Additionally, when only genre is used as an input, performance decreases across all genre categories, with no correlation between the amount of reduction and the single-rated status of a genre. Thus, while genre is useful in and of itself, it is not the most pertinent information for the MPAA rating.

Time effect: Another issue with this task is the time constraint. One assumption is that group definitions have evolved over time, which could have a significant effect on the model. For instance, a film that was rated R 20 years ago would now be classified as PG/PG-13 (because of social changes during the time). We demonstrated empirically in this section that, at least for the time period

Table 11: Weighted F1-score for different genres in the test set (Best= using the best model, OG= using only genre as the input).

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

considered, time has little effect on overall prediction performance.

The following experiments were conducted using our best model. As previously stated, our corpus spans the years 1996 to 2018. We divided the entire dataset into two time periods: prior to and subsequent to 2007. The second period includes additional films, but for the sake of fairness, we eliminated them (we kept equal number of each class for both periods).To have a reliable result, we employed a 2-fold cross-validation method to split data in each period ($q1$ and $q2$ for period one, $q3$ and $q4$ for the second period). Then, for each experiment, we reported the average result. We trained our model using a quarter of data (qi) from one of the periods and then test it using 1) the remaining quarter from that period and 2) two quarters from the other period. Due to the fact that we divided the data into four sections, we had a very small number of G movies in each group, and thus excluded the class G from this experiment.

According to Figure 8, there is no discernible pattern indicating a shift in MPAA rating definitions. If the definition changes significantly, we expect that if we train on recent data (e.g., movies released after 2007) and test on older data (e.g., movies released before 2007), we will see an increase in false predictions toward predicting (PG-13 as PG) (R as PG-13 or PG). Alternatively, if we train with older films and test with newer ones, we should expect to see a greater number of incorrect predictions toward predicting (PG as PG-13 or R) and predicting (PG-13 as R). However, the findings indicate that changes do not follow a predictable pattern. Thus, the dataset exhibits little evidence of temporal bias, at least empirically.

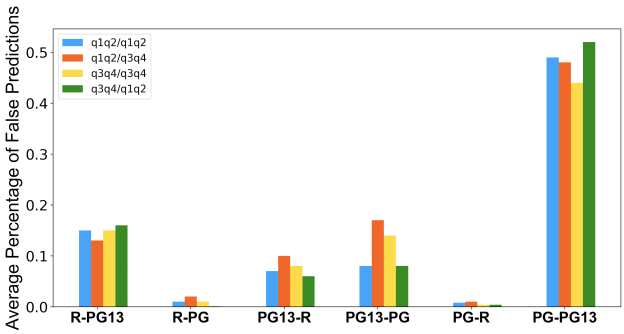


Figure 8: Average percentage of false predictions over all folds. The x-axis shows (true tag-predicted tag), e.g., R-PG13 = 1% means 1% of R movies are predicted as PG-13. Each bar is assigned to training on one quarter and test on other quarters. For example q1q2/q3q4 means we average over training on $q1$ and test on ($q3$ and $q4$) and training on $q2$ and test on ($q3$ and $q4$). q1q2/q1q2 means we average over training on $q1$ and test on $q2$ and training on $q2$ and test on $q1$.

4.4 Analysis

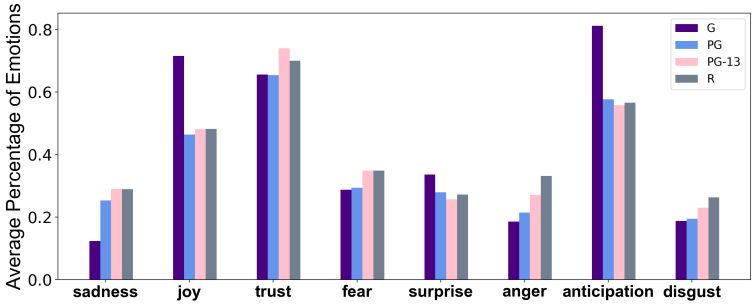
This section expands on our proposed model's analysis. We examined the effectiveness of emotion vectors first, and then the effect of the attention mechanism using some random movie samples. Finally, we examined the effect of offensive language on the MPAA rating.

4.4.1 Emotion Analysis

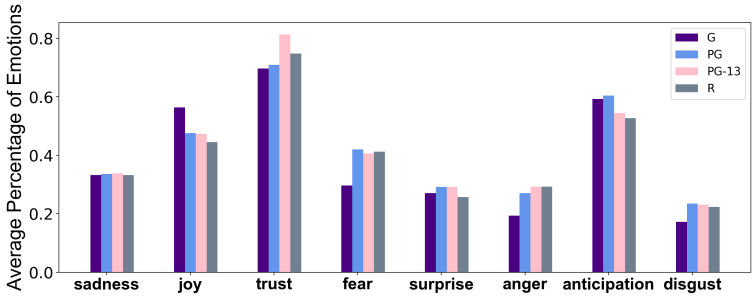
To examine the emotion vectors' effects further, we compared the histograms of emotion scores for correctly and incorrectly labeled instances.

Figure 9a illustrates a naturally occurring pattern for the average emotion score in correctly classified samples. Negative emotions such as *sadness*, *anger*, and *disgust* are expressed at the highest rates in class R. However, in the misclassified samples 9b, we observed no discernible trends in the expression of emotions across ratings. For example, the PG class places a higher premium on negative emotions such as *disgust*, *sadness*, and *fear* than the R class, despite the fact that PG is a more child-appropriate group.

To ascertain the reason for this observation, we examined some samples from the PG and R groups that contain a high proportion of words associated with *disgust*. The results indicate that while *disgust* is associated with terms such as *robbery*, *murder*, and *asshole* in R-rated films, it is



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



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

Figure 9: Average emotion score of correctly and incorrectly classified movies in the test set per class.

associated with *fool*, *sick*, and *painful* in PG-rated films. Although some PG-rated films contain a high number of words associated with negative emotions, the degree of negativity, or the strength of the emotion, appears to be lower than in R-rated films (Table 12). As a result, the model has room to improve in order to account for these more subtle differences.

Table 12: Sample sentences of R and PG rated movies that contain "disgust" emotion in the conversation.

Rate	Sentence
R	redEach victim was killed by a punctuate wound at the skull redGoddamn fucking asshole. redYour wife was murdered.
PG	blueLorenz! are you sick? blueDo you think the memory become less painful then! blueHow terribly awful it all is!

4.4.2 Weight Analysis

We represented the sentences with the highest attention weights in some random sample movies with a different MPAA rating in this section (Table 13). According to these samples, the most intense sentences in R and PG-13 films are more intense than those in PG and G films. Second, depending on the genre of the film, the sentence structure varies within the same rating group. For example, while samples 5 and 6 are both rated R, sample 5 is a *Adventure* film, while sample 6 is a *Crime-Drama* film. As can be seen, sample 6 contains harsher language than sample 5 (sample 5 has inappropriate words, but in a non-aggressive manner). Additionally, the MPAA association provides a concise explanation for rated films in order to justify the film's rating (available at our dataset). We indicated the reasons for these samples in the table, and we can see that the sentences with the highest weights correspond to the reasons for these samples.

Table 13: Sentences with the highest attention weights in some sample movies.

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

4.4.3 Bad Word Ratio

We conducted another analysis on the same list of bad words used in our "Threshold" baseline to determine why these words are insufficient to predict the MPAA rating of the films. We combined all the scripts in our corpus for each class and calculated the frequency of bad words within that class.

Table 14 displays the top five negative words for each data class. As expected, the ratio of most

frequently occurring bad words varies across classes, but so does the intensity of the words, which cannot be captured by a bad word list in which all words are assumed to be equal in strength. Thus, while bad words can have an effect on the MPAA rating, a threshold is insufficient to accurately predict the rating. We must analyze these words in context in order to determine their impact.

Table 14: Top 5 bad words in each class. The numbers inside the parenthesis show the ratio of the word across all the scripts of the class.

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

5 Predicting the Age-Suitability Rating of Movie Trailers Based on Textual, Visual, and Acoustic Cues

The purpose of this chapter of our research is to predict the age-suitability rating of movie trailers using the MPAA's trailer rating guidelines. We formulate this problem as a binary classification task in which trailers are classified as either all-audiences-appropriate (green-band trailers) or restricted-audiences-appropriate (red-band trailers).

5.1 Methodology

Our objective is to predict the age suitability rating for movie trailers using the MPAA's trailer rating guidelines. The problem is formulated as a binary classification task in which trailers are classified as either appropriate for all audiences (green-band trailers) or appropriate to restricted audiences (red-band trailers). The Multi-modal Movie Trailer Rating (MMTR) system is proposed to accomplish this goal. The trailers are modeled in this system as a fusion of three modalities: subtitles, audio, and video. We train separate streams of Recurrent Neural Networks (RNNs) for subtitles and audio, and a combination of Convolutional Neural Networks (CNNs) and LSTM for video, to extract a representation for each modality. Then, we use a fusion module to combine all stream representations in order to take advantage of cues from various modalities. The overall design for the system architecture is depicted in Figure 10.

Our approach is based on independently identifying the best individual modality model and then combining data from all three monomodal models (subtitle, audio, and video) using one of the three fusion methods described below: (i) Gated Multi-modal Unit (GMU) [4], (ii) Feature Concatenation Fusion, or (iii) Late Fusion. The following sections detail each module of the system.

5.1.1 Text Stream

Trailers' subtitles contain a wealth of information. They can assist us in determining the video's subject matter. Additionally, the presence of specific words in the dialogue can be a strong indicator

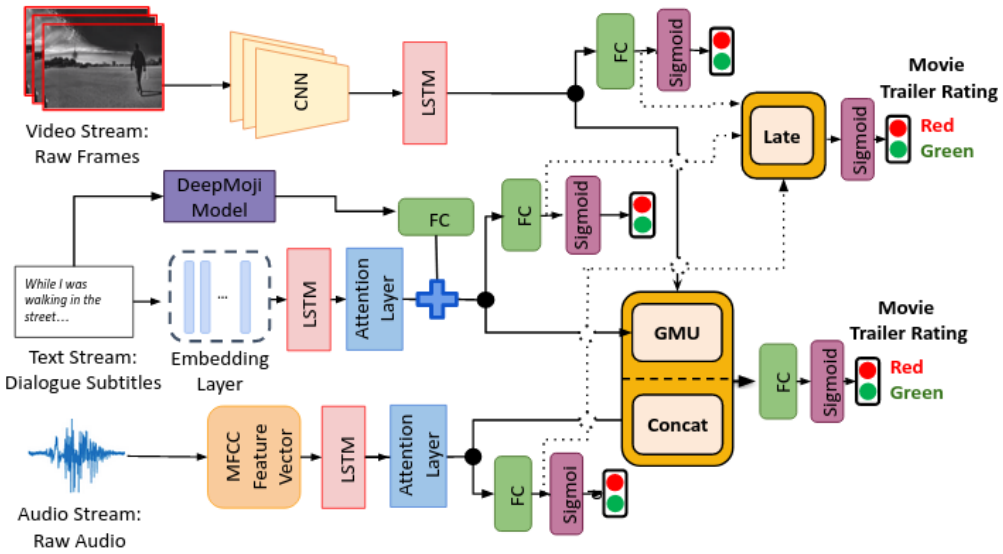


Figure 10: Overview of the deep learning movie rating system and comparison of fusion methods: (i) A video subtitle is transformed into a vector representation using an Embedding Layer and then forwarded to an LSTM network with an attention layer. We concatenate the output of the attention layer with the feature vector from the DeepMoji Model. (ii) A video volume is passed through a CNN-LSTM model that is used as a feature extractor, in order to obtain a single vector representation of the entire video. (iii) Raw audio signal from the video is represented as a sequence of MFCC feature vectors, passed to an LSTM layer. (iv) Lastly, information from all modalities is combined using one of the following fusion methods, namely Gated Multi-modal Unit (GMU), Late Fusion and Feature Concatenation Fusion, before labeling the age-suitability of each trailer. FC in the diagram stands for a fully connected layer.

of certain types of sensitive content, while analyzing the entire transcript can yield more subtle cues. We fed the following modules the information from the subtitles in order to model it.

BERT + Long Short-Term Memory (LSTM) with Attention: We leveraged the well-known power of transformer-based word representations by using BERT [15] as an embedding mechanism. Then, we passed the word vectors to an LSTM layer, which models the word sequence in order to extract the text's semantic information. Following that, we fed the LSTM's hidden representations to the attention mechanism [5] in order to determine the significance of each word in the dialogue. While BERT has improved over time (RoBERTa [37], ALBERT [35]), the purpose of this experiment is to demonstrate that a multi-modal approach can solve this task with acceptable performance.

Emotion Vector: We anticipated observing a stronger correlation between strong negative

emotions (fear, anger, and sadness) and red band trailers. Likewise, we anticipated that positive emotions, such as joy, will be more closely associated with green band trailers.

We used the DeepMoji model to extract emotion from text [18]. This model was trained on 1.2 billion tweets containing emojis in order to gain a better understanding of how language is used to convey emotions. Recent work in abusive language detection has demonstrated promising results when using DeepMoji [51], and thus it appears reasonable to expect similar results for this task. To incorporate this model into our system, we used the pretrained model's final hidden layer representation to convert the text to emotional feature vectors. Finally, we concatenated the emotion vector with the attention output and pass the resulting vector to a fully connected layer to fine-tune the joint representation further.

5.1.2 Video Stream

The video modality provides a wealth of visual and temporal cues that aid in the analysis of multimedia content. Specifically, video can assist in modeling objectionable content such as depictions of nakedness or bloody scenes, as well as suggestive elements. To accomplish this, a CNN-LSTM model based on the work of [16] and [64] was used to learn spatiotemporal video features. Each video was subsampled to a fixed number of frames distributed evenly throughout its duration to create a visual temporal sequence. The raw RGB frames were fed into a CNN model as input. This CNN model generates a feature representation for each frame's spatial information. The final pooling layer of the CNN output was passed to an LSTM, which models temporal dependencies between frames.

5.1.3 Audio Stream

The trailer's audio can assist the model in determining the film's genre and theme, and thus serves as a powerful tool for distinguishing red-band trailers from green-band trailers. For instance, *horror* and *thriller* films (which often feature suspenseful music) are less likely to be appropriate for children. Along with the music score, the tone and pitch of the speakers can provide relevant cues

for rating the trailer. It's worth noting that our model incorporates the entire audio (the music and dialogue combined). We extracted the Mel Frequency Cepstral Coefficients (MFCC) from the audio stream in order to model it. MFCCs are one of the most frequently used feature representations for audio classification [3] and speech recognition tasks [59]. We began by segmenting the audio into to n chunks, $n \in \{10, 20, 50, 100\}$, and then extracting the MFCC feature vector for each chuck. Additionally, by averaging the MFCC vectors within each chunk, we can obtain a fixed-length representation for the entire audio, regardless of its duration. Finally, we passed the vector to an LSTM module that models the MFCC changes over the course of the video; a fully connected layer following the LSTM aids in fine-tuning the model for the task.

5.1.4 Fusion

The fusion module's objective is to train the system to predict the trailer's rating by integrating evidence from the video, audio, and text modalities. Three well-established fusion methods are evaluated in order to create a unified representation for each trailer.

Gated Multi-modal Unit (GMU): The GMU enables the model to learn an intermediate representation by combining various modalities, with the gate neurons learning to determine how each modality contributes to the intermediate representation. A significant advantage of the GMU model is its ability to tailor activation from each modality to the particular instance. This method is inspired by recurrent architectures' control flow. The recurrent units in RNN models determine how much current and previous evidence is used to construct the current state. In GMUs, the activation function for building the output using different modalities is measured, in order to form a unified intermediate representation for all modalities.

To predict the movie genre, the original GMU was successfully applied to a movie dataset consisting of plot synopses and movie posters. The authors of the original paper implemented a bimodal system. We used the straightforward approach discussed in their paper to extend the model to include three modalities. The exact formulation is shown in Equation 2; where W_i, Y_i are learnable parameters, x_i is the feature vector for modality i and $[.,.]$ stands for concatenation.

$$
\begin{aligned}
h_1 &= tanh(W_1.x_1) \\
h_2 &= tanh(W_2.x_2) \\
h_3 &= tanh(W_3.x_3) \\
z1 &= \sigma(Y_1.[x_1, x_2, x_3]) \\
z2 &= \sigma(Y_2.[x_1, x_2, x_3]) \\
z3 &= \sigma(Y_3.[x_1, x_2, x_3]) \\
h &= z_1 * h_1 + z_2 * h_2 + z_3 * h_3
\end{aligned}
\tag{2}
$$

Feature Concatenation Fusion: One popular method for fusion is feature concatenation [6] , in which the representation vectors for each modality are concatenated and the resulting unified representation is routed via multiple hidden layers or utilized directly for prediction.

Late Fusion: Late fusion is another widely used fusion technique [19]. Using various rules (e.g., majority voting, averaging), different modalities are merged at the decision level in late fusion [6]. The average of all outputs from all modalities is calculated and used as the final output in this case.

Prior to performing feature concatenation or GMU-based fusion, each modality's information was represented by a feature vector extracted from pretrained models, which served as a modality stream. Then, we used the GMU or concatenation module to combine the vectors from all modalities into a single vector. Finally, we connected the fused representation to a fully connected layer, resulting in a two-dimensional vector (we have two classes). The two-dimensional vector was then labeled using the sigmoid function. For late fusion, we computed the average of the output of each single modality model prior to the sigmoid function (vectors of size two). Finally, we applied a sigmoid function to the single representation to perform classification.

5.2 Experiments

This section will demonstrate that a multi-modal approach is an effective method for solving the task. As a result, we compared the prediction performance of single modality models to that of

the system's multimodal variations. Additionally, we compared the performance of our proposed system's multimodal variations (MMTR system) to one another.

As mentioned previously, the MM-Trailer dataset is skewed. Thus, to ensure the reliability of the results, 5-fold cross-validation was chosen as the evaluation method. We chose 10% of the train set as the validation set in each fold to obtain the best model. It should be noted that the dataset was stratified in order to ensure that each set contains an equal number of examples from each class. The performance was measured using the weighted F1 score, which was averaged across all five folds in each experiment.

5.2.1 Baseline Methods

Most Frequent Baseline: The first baseline is a naive attempt to demonstrate how difficult it is to solve the problem. All instances in the validation and test sets were assigned to the most frequent class in this model, and the F1 score was calculated using the ground truth label.

Text Baseline - Traditional Machine Learning: We provided a traditional machine learning method with hand-crafted features for the text baseline model. We extracted unigram and bigram features from subtitles and weighted them using the term frequency-inverse document frequency (TF-IDF) formula. The feature vectors were then classified using an SVM model. We chose an SVM model because it performed well on a comparable task of detecting violence [39].

Text Baseline - BERT + Attention + NRC: The NRC emotion lexicon is a frequently used resource for extracting emotion from text [42]. This dictionary associates words with eight distinct emotions (anger, anticipation, joy, trust, disgust, sadness, surprise, and fear) as well as with two sentiments (positive and negative). We computed the normalized count of words per emotion across the entire subtitle and created a vector of size 10 for each trailer using this dictionary. This vector was used in place of the DeepMoji vector in the model.

Text Baseline - DeepMoji + Fully connected layer: To demonstrate how much emotion can contribute to rating prediction on its own, we used only the DeepMoji vector as an input and pass it through a fully connected layer and sigmoid classifier for prediction.

Video Baseline: Our video baseline is based on the deep three-dimensional convolutional network (3D CNN) architecture [60]. Instead of images, the 3D-CNN architecture performs 3D convolution and 3D pooling operations on video volumes. Each video was subsampled to a set of 18 evenly distributed frames that serve as the model's input. The training was conducted over 50 epochs with a 0.5 dropout rate, a learning rate of 10^{-5}, and an eight-sample batch size.

Audio Baseline: CNNs have demonstrated promising results in the classification of audio [25]. To accomplish this, we extracted the log-Mel spectrogram from the audio for each full video using the LibROSA python library [41] and then used it as input to a CNN architecture. For the log-Mel spectrograms, 128 Mel-spaced frequency bins were used, while Inception V3 was used as the CNN model for this baseline. The CNN model was trained for 100 epochs using a 64-sample batch size and a 10^{-5} learning rate. During training, an early stop policy was implemented to avoid over-fitting.

5.3 Results

The results of our experiments are summarized in Table 15. To assess the contribution of each modality to the rating task, we presented results for all single modality models: Audio Only Model (A-MFCC), Text Only Model with DeepMoji (T-BAD), and Video Only Model (V-CNN/LSTM). As expected, our experimental results confirm that utilizing all available modalities results in a superior outcome. As shown in Table 15, the GMU fusion variant of the MMTR model with all modalities achieves the highest weighted F1 score, 86.06 percent. This result increases the weighted F1-score of the best single modality model (T-BAD) by more than three percentage points ($P < 0.05$ based on the t-test).

We also reported results for various combinations of two modalities using all fusion methods to demonstrate the effect of engaging all modalities (T-BAD + A-MFCC, T-BAD + V-CNN/LSTM,

Table 15: Evaluation of the different variants of the MMTR system and other baselines using the MM-Trailer dataset. WF stands for weighted F1 score and results are averaged over 5 folds. A '*' indicates that the difference between the two classifiers' performance is shown to be statistically significant.

	Model	Val-WF	Test-WF
Single Modality Baselines	Most Frequent Baseline	60.82	60.37
	Text Baseline: Traditional Machine Learning	76.38	75.02
	Text Baseline: BERT+ Attention (T-BA)	84.91	81.99
	Text Baseline: BERT+ Attention+ NRC	85.45	81.67
	Text Baseline: DeepMoji+ fully connected layer (DeepMoji+FC)	73.19	68.23
	Video Baseline	81.82	75.33
	Audio Baseline	82.25	72.62
Single Modality Models	Audio- MFCC (A-MFCC)	77.56	73.86
	Text- BERT+ Attention+ DeepMoji (T-BAD)	86.41	82.67*
	Video- CNN/LSTM (V-CNN/LSTM)	87.06	79.41
Late (Fusion using two modalities)	T-BAD + A-MFCC	87.85	82.41
	T-BAD + V-CNN/LSTM	87.31	84.12
	A-MFCC + V-CNN/LSTM	84.63	79.68
Concatenation (Fusion using two modalities)	T-BAD + A-MFCC	87.15	82.17
	T-BAD + V-CNN/LSTM	89.25	82.80
	A-MFCC + V-CNN/LSTM	87.41	78.70
GMU (Fusion using two modalities)	T-BAD + A-MFCC	88.08	83.37
	T-BAD + V-CNN/LSTM	88.00	83.34
	A-MFCC + V-CNN/LSTM	85.63	80.35
Fusion using all Modalities (MMTR)	Late (T-BAD + A-MFCC + V-CNN/LSTM)	89.88	85.60
	Mid (Concat) (T-BAD + A-MFCC + V-CNN/LSTM)	89.97	82.75
	Mid (GMU) T-BAD + A-MFCC + V-CNN/LSTM	**91.05**	**86.06***

A-MFCC +V-CNN/LSTM, T-BAD + A-MFCC + V-CNN/LSTM, T-BAD + A-MFCC + V-CNN/LSTM). According to the findings, a combination of two modalities performs better than each modality alone, but not as well as a combination of all modalities.

When the various fusion approaches are compared, it is clear that GMU fusion outperforms concatenation fusion systems. The gains from GMU, we hypothesize, stem from the gated unit's ability to dynamically adjust the contribution of each modality to the intermediate representation. Using the t-test for statistical significance, we discovered a significant difference between GMU and feature concatenation fusion ($p - value < 0.05$). The test, however, does not demonstrate a statistically significant difference between late fusion and GMU. Thus, we can assert that late fusion can generalize as well as GMU fusion for the trailer age-suitability problem.

The results for T-BAD and T-BA indicate that DeepMoji is a useful feature for the rating task, as it aids the model in discriminating between red-band and green-band trailers. However, the DeepMoji+FC result demonstrates that the DeepMoji model alone is insufficient to solve the task.

To aid in comprehension of the fusion results, we also provided additional evaluation metrics

using the MMTR system variant in conjunction with GMU fusion (as GMU version is the winner approach based on the result table). According to the detailed result, the majority of the predicted instances are red-band trailers. The first possible explanation for this observation is that our training set contains fewer red-band trailers than green-band trailers. As a result, the model has a harder time capturing all patterns in this class. The second reason could be that the video content in red-band trailers is diverse. Bear in mind that this class includes any material that is not suitable for children. As a result, it is reasonable to assume that this class is more varied than the green band class.

Table 16: Performance of the MMTR system using alternative metrics by performing 5 fold cross-validation evaluation method. The results are averaged over 5 folds.

Model	**precision**	**recall**	**F1-score**
Green	87.4%	95.0 %	91.0 %
Red	83.6%	65.0 %	72.8 %
Macro avg	85.6%	79.8 %	82.0 %
Weighted avg	86.6 %	86.4 %	86.0 %

5.4 Discussion

To examine the MMTR system's weaknesses and strengths, we first investigated the incorrectly predicted cases using the system's most effective version (GMU fusion) on each fold of the data. When results are averaged across all folds, approximately 35% of incorrectly predicted cases with the MMTR system are also incorrectly predicted independently by each modality; thus, fusion is unlikely to help in this case. We discovered that in approximately 50% of cases where two modalities predict the incorrect rating, the MMTR system can trust the single correct modality. In approximately 93% of cases where only one modality is incorrect, the MMTR GMU Fusion variant correctly predicts the label using the other two modalities.

After averaging across all folds, we observed that the MMTR GMU Fusion variant system is unable to predict approximately 38 of 294 instances per fold. We analyzed 40 incorrectly classified trailers (selected across all folds) to determine why the model is unable to predict the label successfully. For each of the individual modalities, we proposed the following hypothesis:

1) *Text Modality:* One of the primary sources of text modality errors is the output of the speech recognition tool. To begin, the tool's free version is limited to short audio files. As a result, we divided the audio into 10-second segments. Thus, if the audio is cut off in the middle of a word, we may miss some words. Additionally, low-quality audio affects the accuracy of the automated speech recognition system, causing the model to miss specific bad words in the video or, conversely, generate bad words by mistake (detect "please" as "pussy"). However, in some cases, the trailer contains very little speech (less than ten words) or the language used contains no sensitive content. Unsurprisingly, the text modality is unable to function properly. Finally, we noticed that the trailer subtitles for some green band videos contain the words "gun" and "shot", which are predicted incorrectly by the text modality. The text model appears to be biased against the occurrence of these words, which are most likely associated with violent content.

2) *Video Modality:* One possible explanation for why the video modality model misses sensitive content is the video sampling rate. The inappropriate/violent scenes in these trailers vanish quickly or appear infrequently. As a result, we may miss them when sampling our model's frames. The second possible explanation is the trailers' quality. We recognized that some of the trailers were created in the past or are available in compressed formats, resulting in blurry frames and, in some cases, frames that are not very clear to the human viewer. Finally, we discovered that some green-band trailers contain brief sensitive content such as depictions of guns and blood, which our video modality model classifies as red-band. These are typically R-rated films that have been sanitized for the trailer. However, the film's theme is reflected in several frames. We can conclude that sometimes a single rating is insufficient to convey the nature of the content, and the next step of our research should include the prediction of a list of sensitive material in the video rather than a single label.

3) *Audio Modality:* Occasionally, the trailer's music is incompatible with the content. For instance, we came across musical films with a high level of violence but featuring smooth jazz music. As a result, the audio modality has a difficult time distinguishing between appropriate and inappropriate content. Additionally, in audio modality (as with video modality), we capture samples

from the continuous stream. As a result, if the intense audio (such as a scream or a gunshot) occurs quickly, our model may miss it.

Furthermore, we investigated the genre of trailers that were incorrectly predicted in one of the data folds. The interesting point is that 55% of incorrectly classified red-band trailers are classified as Thriller or Horror films, while 30% are classified as Comedy films (based on IMDB metadata). We do not include metadata in our model in order to make it applicable to any type of online content. This observation demonstrates that if we have available metadata for the film, the genre can be a potential feature for the model.

6 Labeling Comic Mischief Content in Online Videos

The primary objective of this chapter of our research is to develop a completely automated model for categorizing videos containing comic mischief material. For the first time in the literature, we focus on detecting comic mischief content in videos, which is a subset of questionable content based on the ontology of questionable content. In a comic mischief video, inappropriate content (violence, adult content, or sarcastic material) is combined with a humorous context. According to psychologists, when something like violence is presented in a serious context (like war), it has a less distributive effect than when it is presented in a pleasant and humorous context.

6.1 Methodology

Comic mischief is divided into four groups based on the ontology of questionable content: Mature Humor, Slapstick Humor, Gory Humor, and Sarcasm. We proposed two models to address the issue. We first defined the task as a binary classification model in the first model. If the video has content that falls under one of the comic mischief categories, the label will be 1, whereas if none of the categories apply, the label will be 0. In the second setting, we presented a multi-task multi-label model capable of simultaneously predicting all four subcategories.

We proposed a unified end-to-end model in this section of our study; hierarchical cross attention model with captions added (HICCAP). We first encoded each modality using a pre-trained model, and then sent the encoded vectors to RNN models to model the sequential information included within each modality. When the modality representations are complete, we transferred them to the cross attention network, which uses two other modalities to compute the attention of each modality. Figure 11 illustrates the system diagram.

Cross attention module produces a two-dimensional vector (length of sequence, #features). We employed the attention mechanism implemented in [5] to aggregate the data for the entire sequence and construct a single dimension vector (#features) for each modality based on the sequence's relevance. Finally, we integrated the output of all attention layers and fed it through a fully

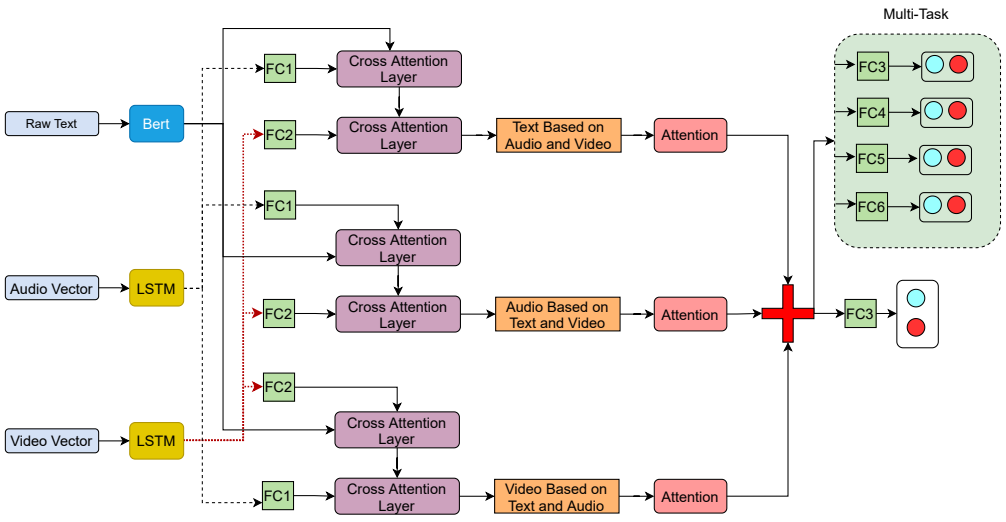


Figure 11: System architecture

connected network for classification purposes using the feature concatenation method. The next sections will detail each module.

6.2 Feature Encoding

The comic mischief content may appear in the speakers' dialogues (for example, sarcasm, mature humor jokes), in the image frames (for example, a large amount of graphic violence containing blood and gore in a humorous situation), or it may be accompanied by clues in the background audio, such as a blood splatter or laughing sound track. Thus, in order to find the most aligned pattern that is applicable to a wide variety of scenarios, we require a model that utilizes all accessible modalities (subtitles, acoustic, and visual information) to better grasp the context and collect all relevant data. The next sections will discuss each approach of encoding a modality:

6.2.1 Text Modality

Subtitles are an important resource for comprehending videos since they can specify the story or topic of a scene. Specifically, when it comes to comic mischief labeling task, some specific terms can help the model differentiate between various groups within the comic mischief category (blood, knife, gun vs sex, naked, and drink). Moreover, the sequence of words conveys information about

the context, such as in a humorous circumstance. To encode word vectors, we use the pre-trained BERT model [15] and gather the hidden states of the model's final layer, which results in embedded vectors that are the length of the words and feature vectors that are the size of the BERT hidden layer feature size (768). It's worth noting that we fine-tuned BERT for our task.

Automatic Video Captioning: Due to the fact that some of the videos lack dialogue, they do not include subtitles. To fill this void, we used a pre-trained model for video captioning to generate captions for these videos and use them instead of subtitles. We employed the most advanced algorithm available today, Dense Video Captioning with Bi-modal Transformer (BMT) [27]. Generally speaking, in Dense Video Captioning, the model locates interesting events in an untrimmed video and generates a unique textual description for each event.

6.2.2 Audio Modality

We hypothesize that audio is a valuable source of information for labeling videos containing comic mischief. Sound effects may convey information about the scene's context. For instance, blood splatter sound might serve as an acoustic cue for violent moments, or a laugh sound can serve as a cue for a humorous situation. Additionally, diverse background music can elicit a range of emotions among the audience. As a result, it is critical to incorporate audio information into the model. To extract audio features, we used the VGGish network [24] that was pre-trained on AudioSet [21]. More precisely, the VGGish model processes segments that are 0.96 seconds in length. The audio segments are then represented as log mel-scaled spectrograms using the Short-time Fourier Transform. VGGish's pre-classification layer generates a 128-dimensional embedding for each spectrogram. As a result, the audio track of a movie in the dataset is represented by a sequence of 128-d features with a length of the video divided by 0.96 (each feature in the stack corresponds to 0.96 seconds of the source audio). After generating the audio vectors, we modeled the sequence of audio segments using an LSTM model.

6.2.3 Video Modality

Image frames are a critical resource for detecting comic mischief in videos. Objects and aspects in the scene, such as a knife and blood, might convey the gory nature of the situation. Naked bodies may indicate adult material. Certain facial expressions convey sarcasm, while others, such as eye poking, are slapstick. Furthermore, laughing faces can convey a sense of humor. To encode visual modality, we used a pre-trained I3D network [8] using the Kinetics dataset. This model's input is a row of 64 RGB and optical flow frames extracted at a specified frame rate. We followed the same procedure as the paper [27], in which the flow frames are extracted using PWCNet [57]. Both sets of frames are resized first, followed by cropping the core region. Following that, both frame stacks are transmitted over I3D's appropriate streams. The network generates a 1024-d representation for RGB and flow; we then summed the representations from both streams (same as the original paper of I3D). Thus, the visual track of a movie is represented as a sequence of 1024-d features of length L (L = (video length * frame rate)/64) each feature spanning (64/frame rate) seconds of the original video. As with audio, we passed video vectors through an LSTM module to extract sequential information.

6.2.4 Hierarchical Cross-attention and Fusion

We used the cross attention mechanism as a unit module to mix different modalities. Although this module is developed from the self-attention mechanism, it is intended to locate the attention of one modality based on another modality. We have query, key, and value vectors in the self-attention, and they are all identical to the input stream. However, with cross attention, the query is one modality and key and value vectors belong to the context modality. Because the cross attention mechanism is effective for two modalities, we extended it to compute the attention of multiple modalities (more than two). To accomplish this, we created a hierarchical cross-attention mechanism in which we first compute the attention of modality 1 based on modality 2, and then transmit the output of the first layer cross-attention to the second layer cross-attention together with modality 3. The output of this layer is modality 1's attention, which is determined by modality 2 and modality 3.

Mathematically hierarchical cross-attention can be represented by Equation 3 where K,Q, and V stand for key, query, and value respectively, and d_k is dimension of the key vector.

$$\begin{aligned} head_1 &= \text{softmax}(\frac{K_{m2}^T Q_{m1}}{\sqrt{d_k}})V_{m2} \\ head^{m1} &= \text{softmax}(\frac{K_{m3}^T Q_{head_1}}{\sqrt{d_k}})V_{m3} \end{aligned} \tag{3}$$

After calculating $head^{m1}$, $head^{m2}$, $head^{m3}$, we pass them separately through attention layer proposed by [5]. This layer computes the weighted sum r as $\sum_i \alpha_i head_i^{mj}$ where $j \in \{1, 2, 3\}$ to aggregate hidden layers of cross attention layer to a single vector. The model can learn the relative importance of hidden states ($head_i^{mj}$) by learning the α_i. We compute α_i as follows:

$$\alpha_i = \text{softmax}(v^T \tanh(W_h head_i^{mj} + b_h)) \tag{4}$$

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

Then we concatenated the output of all attentions and pass them through a fully connected network for the classification purpose.

6.2.5 Multi-task Model

When underlying principles or information are shared between tasks, multi-task learning improves performance. In our case, we want to classify a video into four distinct categories of comic mischief. These categories are likely to be correlated, as all tasks will require learning to detect items such as color and objects in image frames, ups and downs in audio stream, and context of the words in subtitles. To this end, the model is identical to the binary model, except that four distinct classification layers are defined for each task. The network's task-specific components all begin with the same base representation from the previous shared layer. Multi-task learning may enhance the generalizability of this representation by compelling the model to focus on the features that are useful across all tasks.

We minimize a linear combination of the loss functions associated with the individual tasks.

Each task will have its own unique loss function, denoted by the L_i. Hence, we weight each loss function in our multi-task model and minimize the sum of these weighted losses.

$$L_{total} = \sum_i w_i L_i \tag{5}$$

6.3 Experiments

The experiments' objective is to demonstrate that: 1) Combining all modalities results in the best model when compared to monomodal and bimodal models; 2) In comparison to other methods such as feature concatenation, late or GMU fusion, hierarchical cross-attention is an effective method for labeling videos for comic mischief labels. 3) Hierarchical cross-attention works better than bi-modal cross-attention. 4) With the multi-task model, we wish to demonstrate that tackling the problem in this manner is more effective and efficient than training a distinct model for each task. 5) And finally, to demonstrate how an automatic caption generating method based on image frames can be used to supplement the text modality in videos that lack it.

6.3.1 Evaluation

Data Partitions: We divided the dataset into three parts: 80% for training, 10% for validation, and 10% for testing. We have four aspects to the multi-task model (Mature Humor, Gory Humor, Slapstick Humor, and Sarcasm). To stratify the dataset based on all four classes, we created a new label consisting of the binary values for all four tasks (e.g., 0110), and then used this new label to stratify the dataset for training, testing, and validation. As a result, each aspect's distribution is quite close within each partition. Table 17 shows data distribution for each partition.

Table 17: Dataset partitions

	Mature Humor	**Slapstick Humor**	**Gory Humor**	**Sarcasm**	**None**	**All instances**
Train	222	166	86	374	307	1007
Validation	24	18	6	48	31	113
Test	35	19	11	41	30	113

Metric: Due to the data imbalanced, for the binary classification, we reported macro F1, weighted

F1, and average accuracy of each class. For multi-task model, we reported macro F1 for each task, the average and weighted average of macro F1 for all four tasks, hamming score (how close are the sequence of predicted labels and true labels), and accuracy score (exact match between two sequence of labels).

6.3.2 Baselines

Baselines are defined as the following categories:

Single modality models: As a baseline, we showed the result of the classification model using each modality separately. For text modality, the model is a pre-trained BERT + LSTM + attention + FC (fully connected network); for the audio and video modalities, we pass the input vectors through LSTM + attention + FC.

Concatenation models: In this set of baselines, we integrated modalities by concatenating feature vectors. First, we combined, each two modalities, to report the results for bi-modal models. Then, we presented the result of a three-modality combination. It should be noted the model will be trained in an end to end manner, and this method will be considered as an early fusion.

LXMERT: LXMERT [58] is a framework for learning the connections between vision and language. The authors of LXMERT developed a large-scale Transformer model composed of three encoders: an object relationship encoder, a language encoder, and a cross-modality encoder. To make this framework compatible with our project, we passed the video encoded vector (I3D network features) to the model rather than object vectors.

Bi-modal for cross-attention mechanism: We assert in this study that the hierarchical cross-attention model produces the best results. To substantiate this assertion, we demonstrated the outcome of an original cross attention experiment that is limited to two modalities.

Late fusion: Late fusion is a highly successful method of combining various modalities. Thus, in this strategy, we train the model independently for each modality and then combine all modalities at the decision level by averaging the probabilities for each class and selecting the class with the greatest probability.

Intermediate fusion with GMU: As with the MM-trailer model, which was developed for rating movie trailers, we trained the model for each modality separately, then saved the final layer before the classification layer, and ultimately used the GMU model to integrate the output from all modalities.

6.4 Results

Table 18: Binary model results on the test set. A-ACC stands for average accuracy of both classes (0 and 1), W stands for weighted

Models	Modalities	A -ACC	Micro F1	Macro F1	W F1
Single Modality Model	Text Modality	68.91	63.71	62	65.78
	Text Modality + Caption	62.31	71.68	62.88	71.36
	Audio Modality	66.00	67.25	63.08	68.9
	Video Modality	55.82	66.37	55.93	65.99
Concatenation Model	Text + Audio	67.71	61.94	60.44	64.05
	Text + Video	68.45	64.6	62.45	66.66
	Audio + Video	63.13	64.6	60.35	66.43
	Text + Audio + Video	66.54	53.98	53.89	54.84
Late Fusion	Text + Audio + Video	62.95	69.02	62.20	69.73
Intermediate Fusion with GMU	Text + Audio + Video	60.63	46.90	46.89	46.67
LXMERT	Image + Text	65.68	69.91	64.24	70.9
	Audio + Text	64.61	69.91	63.61	70.71
Bi-Modal Cross Attention	Image + Text	67.95	71.68	66.34	72.63
	Audio + Text	69.61	72.56	67.66	73.56
Hierarchical Cross-attention	Text + Audio + Video	72.16	77.87	71.93	77.99
Hierarchical Cross-attention + Caption (HICCAP)	Text + Audio + Video	**75.04**	**80.53**	**75.04**	**80.53**

In Table 18, we show the results for the binary model. According to the table, the best result is obtained from "Hierarchical Cross-attention + Caption (HICCAP)" Model. To evaluate the hierarchical cross-attention mechanism and fairly compare it to competing approaches, we report results for models without adding captions to the input data first. We improved the best monomodal model by 11.96% using hierarchical cross attention mechanism. The proposed hierarchical Cross-attention mechanism works better than the best concatenation model, late fusion and intermediate GMU fusion, and the best LXMERT model by 9.48%, 9.73%, 25.04%, 6.79% respectively based on average macro F1.

To demonstrate the effectiveness of captioning videos that lack subtitles, we added captions to

both text single modality model and hierarchical cross attention model. Results indicate that both models improve on various metrics; HICCAP outperforms hierarchical cross attention by 3.11% based on macro F1. It should be noted that HICCAP is the winner model based on all other metrics as well (average class accuracy, micro F1, and weighted F1).

Table 19: Multi-task model results on the test set. A-M, W-M, HS, ACC stand for average macro F1, weighted macro F1, hamming score, and multi-label accuracy score respectively.

		Mature Humor	Gory Humor	Slapstick Humor	Sarcasm				
		Macro	Macro	Macro	Macro	**A-M**	**W-M**	**HS**	**ACC**
Single Modality Models	Text Modality	80.49	52.54	49.38	61.93	61.05	64.83	70.79	17.69
	Text Modality + Caption	80.49	52.38	61.47	65.42	64.94	68.33	73	30.97
	Image Modality	57.43	66.58	53.34	66.61	60.9	61.19	66.37	25.66
	Audio Modality	66.88	57.94	65.15	71.14	65.27	67.28	70.57	19.46
Concatenation Model	Audio + Image	63.15	52.26	70.02	70.46	63.97	66.07	68.36	20.35
	Image + Text	79.48	46.37	58.35	65.38	62.39	66.802	69.02	20.35
	Audio + Text	84.08	52.97	45.93	64.31	61.82	66.36	69.69	20.35
	Audio + Image + Text	80.91	50.48	59.28	62.8	63.36	66.87	69.69	23.89
Bi-Modal Cross Attention	Text + Audio	84.08	68.36	48.9	68.08	67.35	69.95	75.44	30.97
	Text + Image	75.54	67.46	60.48	67.29	67.69	68.81	76.1	31.85
Hierarchical Cross-attention	Text + Audio + Video	84.11	67.46	58.63	69.64	69.96	72.21	**77.43**	33.62
Hierarchical Cross-attention + Caption (HICCAP)	Text + Audio + Video	77.23	66.13	62.74	75.33	**70.35**	**72.74**	75.88	**35.39**

Table 19 shows the result for the multi-task model. Same as the binary model HICCAP outperforms all the baselines based on the average weighted macro. To show that the multi-task approach is a reasonable approach for solving this problem, we compared the result of the multi-task hierarchical cross attention with the four single-task hierarchical cross attention models (Table 20). So, all the models are exactly the same, but we trained them for each task separately. Based on the weighted macro average of all tasks, the multi-task model works better overall. On the other hand, we want to show that in the single-task models, the hierarchical cross-attention mechanism is a reasonable approach. So, we compared the single-task models with late fusion and intermediate GMU fusion model. The hierarchical cross-attention model outperforms late fusion and GMU fusion by 10.16% and 23.36% respectively based on average macro metric.

In the multi-task approach, when we compare single modality models, audio modality is the

Table 20: Comparing a) multi-task model with single task models, b) hierarchical cross-attention layer for each task with late and GMU fusion. A-M and W-M, stand for average macro F1, and weighted macro F1 respectively.

	Mature	Gory	Slapstick	Sarcasm		
	Macro	Macro	Macro	Macro	**A-M**	**W-M**
Multi-task hierarchical cross-attention	84.11	67.46	58.63	69.64	69.96	72.21
Hierarchical cross-attention - per task	79.99	74.14	44.35	74.97	68.36	71.05
Late fusion - per task	67.56	42.59	52.34	70.34	58.20	63.31
Intermediate GMU fusion - per task	49.35	46.94	45.14	38.58	45.00	44.17

winner for sarcasm and slapstick humor. For the sarcasm unusually laugh sound accompanies the video and slapstick usually come with a specific background music (sometimes even no conversation). Gory humor is more of visual feature (e.g., depiction of blood), and mature humor is mostly conveyed through adult-type conversations.

6.5 Discussion

For the binary model, the model fails to predict 22 instances correctly (11 instances from comic mischief, 11 instances from the "none" group). The incorrectly classified comic mischief instances include 10 single-aspect cases (e.g., only include sarcasm); while in total we have 61 single-aspect instances. One video with more than one aspect is also predicted incorrectly out of 21. Thus, we can hypothesis that the more aspects the video has, it is easier to be detected as comic mischief. From the group "none" 11 out of 31 are predicted incorrectly, we have two assumptions for that: first, some videos include mature or violent content, but not in a humorous context; and second, we have a smaller number of none instances, making it more difficult for the model to find a clear pattern.

We annotated videos for four modalities, as described in the dataset section. In this section, we examined the effect of modalities on the outcome and the efficacy of our best model in including various modalities into the prediction. We used binary HICAAP on the test set to identify cases that were predicted as "none" when they include comic mischief material. We noticed that only one of them (out of 11 incorrectly classified cases) is a single modality, meaning that only one modality has comic mischief, while the remaining modalities are none. We have 26 monomodal examples

out of 113 samples, with only two of them being active via video and 24 being active via dialogue. Additionally, five of the improperly classified cases are three-modality examples, with four of these cases containing all modalities except dialogue and only one containing the dialogue modality. As a result of this discovery, we can hypothesize that dialogue is a powerful modality among all modalities and that when it is active, the model has an easier time predicting properly. Is it, however, solely a matter of dialogue modality or is it also a matter of how models encode dialogue modality? To address this subject, we used the concatenation model and extracted some statistics from it. 15 out of 50 incorrectly classified occurrences (from the comic mischief class) are single-modality instances (all of them have dialogue as the active modality). It demonstrates that the HICAAP model, which employs a hierarchical cross attention paradigm, is capable of superior encoding modalities.

As previously stated, a multi-task model works best when targets are related. We used the Jaccard similarity score to compare labels. As shown in Table 21, slapstick and gory humor are the most similar. And sarcasm's similarity scores to other categories are low.

Table 21: Similarity scores between comic mischief aspects

	Gory Humor	Slapstick Humor	Sarcasm
Mature Humor	74.04	65.69	51.17
	Gory Humor	75.26	54.42
		Slapstick Humor	54.66

7 Conclusions and Future Work

This chapter provides an overview of the research conducted for this dissertation. First, we summarize our research's objectives and significant findings. Then we talk about potential future work in the field.

7.1 Conclusions

The purpose of this dissertation is to contribute to the literature by developing automated methods for computationally analyzing the context of the input video and detecting questionable content. The tasks presented in this dissertation are related to the interdisciplinary research space of Artificial Intelligence, Natural Language Processing, and Multi-Modal Systems. Having an automated system for detecting objectionable content is crucial as it contributes to a safer media environment for children and enables parents to make more informed decisions about screen permission for their children by utilizing the content descriptor for online content. In the long run, preventing young viewers from viewing harmful content may reduce their anxiety and fear, resulting in a healthier society.

The primary goal of this dissertation is to develop an automated model for detecting objectionable content in multimodal media content such as videos. To address this issue, we first created an automated model to predict the age rating for movies and trailers, and then leveraged those techniques to improve the model and label the video for a specific type of objectionable content. In the first project, we create a sequential model based on movie scripts to analyze the movie. The word sequence can describe the movie's topic and will aid the model in determining the pattern for different age groups. Furthermore, based on our experiments, emotion within the words and movie genre are two effective features that help the model predict age rating more accurately.

In the second project, we broaden the model to include all modalities, including audio, text, and image frames. As a result, the model can draw conclusions from a broader perspective. All modalities accompanying each other give the model the ability to understand the context more

accurately, as well as understand the content even when one of the modalities lacks information or is absent. We proposed a hybrid fusion method in this project by identifying the best individual modality model independently and then combining data from all three monomodal models (subtitle, audio, and video) using one of the fusion methods (feature concatenation, late fusion and gated module unit).

We concentrated on predicting comic mischief content, one of the categories of questionable content, in the video for the final project. According to psychology research, negative action is more distributive when combined with a positive context. One of these types of content is comic mischief, which occurs when violence, adult content, or offensive language is mixed with humor. For the first time, we propose in this project a model for labeling any type of video with a comic mischief label. First, we formulate the problem as a binary problem in which we predict whether or not there is comic mischief in the video. Second, we propose a multi-task model that uses a single end-to-end model to predict the four categories of comic mischief simultaneously. To label videos, the proposed system uses a combination of textual, acoustic, and imagery streams. Experiment results show that a model with all modalities produces the best results.

Furthermore, this dissertation contributes significantly by introducing new language and multimedia resources for the tasks of detecting questionable content in videos. We proposed the largest dataset for movie scripts, as well as metadata and poster images for all of the films in the dataset. Then, for those movies rated by the MPAA organization, we gathered age ratings. Furthermore, we expanded the dataset to a multimodal dataset by collecting trailers for movies that have a rating band at the beginning. Finally, we compiled the first dataset from a diverse set of videos and annotated them for four types of comic mischief content. Chapter 3 contains information on all datasets.

There are a variety of practical applications for detecting questionable content in videos. It can be used as a plugin in movie service providers, video streaming channels, or any social media platform to describe the video before people watch it. Furthermore, our proposed system can detect objectionable content in other types of media, such as books and music. As a result, we can provide

a safe environment for our children to absorb sanitized content in an age of rapid technological advancement.

7.2 Future Works

This work has a lot of potential. The ontology of questionable content covers a wide range of topics that we did not address in this dissertation. As a future work, one can work on training and fine-tuning models to detect other types of questionable content in videos. We also propose some possible solutions to our projects' limitations that can be explored as the future work:

1) One source of model prediction mistakes is due to ASR errors. To alleviate this issue, we propose two alternative solutions:

Confidence in ASR Results: ASR generates several possible replies and also provides a confidence score for each. In this research, we choose the most likely response. However, That answer is not 100 percent certain. As a result, we can infuse the model with the corresponding confidence score. Thus, the less confidence ASR has in the result, the less the sentence's words will alter the text representation.

Utilize ASR Alternative Outcomes: ASR generates several outputs for some audio inputs. We can identify the words that differ in distinct outputs and send them to the model together with the phrase confidence, which represents the probability of the existing words.

2) Occasionally, the proposed model is biased toward certain sensitive words or phrases (e.g., gun), but the context is not sensitive, resulting in an incorrect prediction. As a result, we can combine the audio signal for each word in order to capture the emotion conveyed by the words by recognizing the speaker's tone for the specific word.

3) In this dissertation, we mostly focused on using raw data from videos to solve the problem. Although employing raw data has its own advantages (feature vectors may propagate errors in the model), including some feature vectors in the model may help to detect a broader range of sensitive content with greater precision. For instance, we can use trained models to recognize nude persons, certain objects (gun and knife), or facial emotions in scenes.

4) When dealing with lengthy texts (such as a whole film script), hierarchical models might aid in comprehending more information. For instance, encoding sentences first, then combining sentences to encode the entire document.

5) Finally, another possible direction for this work is to focus on developing a more explainable model for detecting questionable content. The model could be designed in such a way that it points to the exact section of the video that contains questionable content.

.

Bibliography

[1] AAP. Media violence. *Pediatrics 108*, 5 (2001), 1222–1226.

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

[3] Andén, J., and Mallat, S. Multiscale scattering for audio classification. In *Proc. International Society for Music Information Retrieval Conference* (2011), Miami, FL, pp. 657–662.

[4] Arevalo, J., Solorio, T., Montes-y Gómez, M., and González, F. A. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992* (2017).

[5] Bahdanau, D., Cho, K., et al. Neural machine translation by jointly learning to align and translate. arxiv preprint arxiv: 1409.0473.

[6] Baltrušaitis, T., Ahuja, C., and Morency, L. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence 41*, 2 (2018), 423–443.

[7] Carreira, J., and Zisserman, A. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 4724–4733.

[8] Carreira, J., and Zisserman, A. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 6299–6308.

[9] Cascante-Bonilla, P., Sitaraman, K., Luo, M., and Ordonez, V. Moviescope: Large-scale analysis of movies using multiple modalities. *arXiv preprint arXiv:1908.03180* (2019).

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

[11] Constantin, M. G., Stefan, L. D., Ionescu, B., Demarty, C.-H., Sjoberg, M., Schedl, M., and Gravier, G. Affect in multimedia: Benchmarking violent scenes detection. *IEEE Transactions on Affective Computing* (2020).

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

[13] Davis, S., and Mermelstein, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing 28*, 4 (1980), 357–366.

[14] Demarty, C.-H., Ionescu, B., Jiang, Y.-G., Quang, V. L., Schedl, M., and Penet, C. Benchmarking violent scenes detection in movies. In *2014 12th International Workshop on Content-Based Multimedia Indexing (CBMI)* (2014), IEEE, pp. 1–6.

[15] Devlin, J., Chang, M., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[16] Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., and Darrell, T. Long-term recurrent convolutional networks for visual recognition and description. In *In Proc. IEEE Conference on Computer Vision and Pattern Recognition* (June 2015).

[17] Ertugrul, A. M., and Karagoz, P. Movie genre classification from plot summaries using bidirectional lstm. In *2018 IEEE 12th International Conference on Semantic Computing (ICSC)* (2018), IEEE, pp. 248–251.

[18] Felbo, B., Mislove, A., Søgaard, A., Rahwan, I., and Lehmann, S. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. *arXiv preprint arXiv:1708.00524* (2017).

[19] Fu, Z., Li, B., Li, J., and Wei, S. Fast film genres classification combining poster and synopsis. In *Intelligence Science and Big Data Engineering. Image and Video Data Engineering* (Cham, 2015), X. He, X. Gao, Y. Zhang, Z.-H. Zhou, Z.-Y. Liu, B. Fu, F. Hu, and Z. Zhang, Eds., Springer International Publishing, pp. 72–81.

[20] Geiger, J. T., Schuller, B., and Rigoll, G. Large-scale audio feature extraction and svm for acoustic scene classification. In *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (2013), IEEE, pp. 1–4.

[21] Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2017), IEEE, pp. 776–780.

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

[23] Hebbar, R., Somandepalli, K., and Narayanan, S. Improving gender identification in movie audio using cross-domain data. In *Interspeech* (2018), pp. 282–286.

[24] Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., et al. Cnn architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2017), IEEE, pp. 131–135.

[25] Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., Slaney, M., Weiss, R., and Wilson, K. Cnn architectures for large-scale audio classification. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017.

[26] Hosseinmardi, H., Han, R., Lv, Q., Mishra, S., and Ghasemianlangroodi, A. Towards understanding cyberbullying behavior in a semi-anonymous social network. In *Advances in*

Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on (2014), IEEE, pp. 244–252.

[27] Iashin, V., and Rahtu, E. A better use of audio-visual cues: Dense video captioning with bi-modal transformer. *arXiv preprint arXiv:2005.08271* (2020).

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

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

[30] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., and Fei-Fei, L. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2014), pp. 1725–1732.

[31] Kennedy, M. *Roadshow!: The Fall of Film Musicals in the 1960s*, reprint ed. Oxford University Press, 2014, p. 183.

[32] Kiela, D., Bhooshan, S., Firooz, H., Perez, E., and Testuggine, D. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950* (2019).

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

[34] Koutini, K., Eghbal-Zadeh, H., Dorfer, M., and Widmer, G. The receptive field as a regularizer in deep convolutional neural networks for acoustic scene classification. In *2019 27th European Signal Processing Conference (EUSIPCO)* (2019), IEEE, pp. 1–5.

[35] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. ALBERT: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942* (2019).

[36] LIFE, M. Life magazine. `https://books.google.com/books?id=a08EAAAAMBAJ&pg=PA55&dq=\%22X+Persons+Under+16+Not+Admitted\%22#v=onepage&q=\%22X\%20Persons\%20Under\%2016\%20Not\%20Admitted\%22&f=false`, 1969. Accessed: 2021-07-30.

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

[38] Lu, J., Batra, D., Parikh, D., and Lee, S. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265* (2019).

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

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

[41] McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., and Nieto, O. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th Python in Science Conference* (2015), vol. 8.

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

[43] MPAA. Classification and rating rules. `https://www.filmratings.com/Content/Downloads/rating_rules.pdf`, 2010. Accessed: 2021-11-29.

[44] New York, M. New york media. `"https://books.google.com/books?id=HOYCAAAAMBAJ&pg=PA64#v=onepage&q&f=false"`, 1981. Accessed: 2020-06-14.

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

[46] Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., and Chang, Y. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web* (2016), International World Wide Web Conferences Steering Committee, pp. 145–153.

[47] Papakostas, M., Spyrou, E., Giannakopoulos, T., Siantikos, G., Sgouropoulos, D., Mylonas, P., and Makedon, F. Deep visual attributes vs. hand-crafted audio features on multidomain speech emotion recognition. *Computation 5*, 2 (2017), 26.

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

[49] Penet, C., Demarty, C.-H., Gravier, G., and Gros, P. Technicolor and INRIA/IRISA at mediaeval 2011: learning temporal modality integration with Bayesian networks. In *MediaEval 2011, Multimedia Benchmark Workshop* (2011), vol. 807.

[50] Rasheed, Z., and Shah, M. Movie genre classification by exploiting audio-visual features of previews. In *Object Recognition Supported by User Interaction for Service Robots* (2002), vol. 2, IEEE, pp. 1086–1089.

[51] Safi Samghabadi, N., Hatami, A., Shafaei, M., Kar, S., and Solorio, T. Attending the emotions to detect online abusive language. *arXiv preprint arXiv:1909.03100* (2019).

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

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

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

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

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

[57] Sun, D., Yang, X., Liu, M.-Y., and Kautz, J. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 8934–8943.

[58] Tan, H., and Bansal, M. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490* (2019).

[59] Tiwari, V. Mfcc and its applications in speaker recognition. *International Journal on Emerging Technologies 1*, 1 (2010), 19–22.

[60] Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)* (USA, 2015), ICCV '15, IEEE Computer Society, p. 4489–4497.

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

[62] Wehrmann, J., Lopes, M. A., and Barros, R. C. Self-attention for synopsis-based multi-label movie genre classification. In *The Thirty-First International Flairs Conference* (2018).

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

[64] Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., and Toderici, G. Beyond short snippets: Deep networks for video classification. In *In Proc. IEEE Conference on Computer Vision and Pattern Recognition* (June 2015).

[65] Zhang, A. Speech recognition (version 3.8). `https://github.com/Uberi/speech_recognition`. Accessed: 2020-04-30.

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