Learning from Explanations: Multi-aspect based Age-restricted Rating Prediction in Long Scripts

Anonymous ACL submission

Abstract

In the Motion Picture Association of America (MPAA), reviewers watch the entire film to determine the age-restricted category (MPAA rating) of the movie and provide the explanatory feedback for rating decision. As such human expert system is a time-consuming and non-scalable process, this paper proposes to develop a machine review system named MARS that automatically predicts the MPAA ratings of movie scripts. Specifically, in MARS, we first explore the use of the well-studied multi-aspect classification as machine-provided explanations, then leverage them to better learn the target rating prediction models. We demonstrate MARS outperforms various baselines by around 10 points in terms of F1 score, detecting severe contents with multi-aspect view.

1 Introduction

The age-restricted ratings for movies have a wide practical value (Gentile, 2008). For example, customers can rely on the ratings as a guideline when they plan to watch movies with their kids, while media service providers (*e.g.*, Netflix and Amazon) may use the ratings to enable age filters in parental controls. Filmmakers also edit their movie scripts based on the received ratings for lower ratings. The MPAA[1] provides five categories such as R and PG-13 for age-restricted ratings. In current movie rating systems, human experts determine the MPAA ratings based on multiple aspects (*e.g.*, violence and sex) in contents, and then provide feedback of ratings. However, as such manual rating decisions are very time-consuming through non-scalable review process, Shafaei et al. (2020) initiated the research of automatically predicting MPAA ratings based on neural models.

In spite of such recent neural revolution, there is much room for improving the machine review system. First, the predictive performance is relatively poor, as the target prediction task suffers from long texts in movie scripts. Although several techniques (Ding et al., 2020; Zhang et al., 2019; Huang et al., 2019) are developed for long text such as academic papers, our extreme case (*e.g.*, up to 21K tokens per movie) is not explored yet. Second, the automatic rating prediction process limits its impact for filmmakers as it does not provide any fine-grained explanation for received ratings, *i.e.*, "how do movies get their ratings?". If provided, the explanations can help producers to trim severe contents for a lower rating (*e.g.*, R→PG-13).

In this paper, we present Multi-Aspect Review System (MARS), a novel machine review system that predicts and explains the age-restricted ratings of movie scripts with multi-aspect view, *e.g.*, "this movie is too violent!" for a R-rated movie. Here, we argue that the well-studied multi-aspect classification (Martinez et al., 2019, 2020; Zhang et al., 2021) can guide not only human to understand the rating decisions but also machine to better predict age-restricted ratings. Specifically, MARS repurposes the learned multi-aspect classification model at movie level to an external machine-to-machine explainer at more fine-grained level (Liu et al., 2019a; Hase and Bansal, 2021) as follows:

- **Scene level**: By performing multi-aspect classification with each scene, we identify important scenes and concatenate them for summarization (Papalampidi et al., 2019, 2020a,b), which enables to replace the full-length scripts by a shorter input for better rating prediction.
- **Word level**: Multi-aspect classification can hierarchically identify attentive words from the important scenes. By this, we adopt attention supervision (Choi et al., 2020; Zou et al., 2018), which treats attention as output variables to robustly supervise rating prediction.

Our experiments demonstrate that only MARS

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

shows over 80% in F1 score (exactly, 85.68%), outperforming various baselines by around 10 points. To simulate when the original movie scripts should be revised/censored for a lower age rating, we also analyze how sensitively a change of input script leads to a change of output rating. As a result, we find that MARS is an effective and scalable tool to identify severe contents for minimal editing, *e.g.*, removing a few scene- and word-level explanation.

2 MARS: Proposed Movie Review System

To specify the suitability of movies for children, the MPAA provides five rating categories such as G, PG, PG-13, R, and NC-17 (see Appendix B). The objective of MARS is to improve predictive performance on age-restricted ratings as well as provide explanations for editing contents. To this end, given a movie script $\mathbf{x}$, a rating model learns to predict age-restricted rating $\mathbf{y}$ which is a one-hot categorical label. To counter class imbalance, inspired by (Martinez et al., 2019, 2020), we adopt three rating classes by using a median split: `LOW` (<PG-13), `MED` (=PG-13), `HIGH` (>PG-13).

2.1 Training a Multi-aspect Explainer

In current movie rating system, reviewers provide not only age-restricted ratings but also explanations based on multiple aspects. To mimic this process, we first train an explainer, which is a multi-aspect classification model (shortly, aspect model) that learns to predict the existence of five aspects in contents: Violence, Sex, Substance abuse, Profanity, Frightening. Specifically, as we treat each aspect as an individual binary classification task, the aspect model is trained by multi-task learning.

Given a script $\mathbf{x} = \{w_1, ..., w_T\}$ with T words, we first use a shared encoder (*e.g.*, Bi-LSTM layer) to obtain the representation h_t of each word w_t. Then, we add individual attention and predictor layers for different aspects. Formally, the existence of aspect a is predicted as $\hat{\mathbf{y}}^a = \texttt{Softmax}(W^{a\top}\mathbf{h}^a + b^a)$ where $\mathbf{h}^a = \sum_{t=1}^{T} h_t \hat{\alpha}_t^a$ with attention weights $\{\hat{\alpha}_1^a, ..., \hat{\alpha}_T^a\}$ indicating a probability distribution over the hidden representations. Note that these aspect-specific parameters are not shared among different aspects, which contributes to capturing what word is important for a particular type of aspects. The aspect model is trained by minimizing the cross-entropy loss L_{CE} between the (binary) ground truth label $\mathbf{y}^a$ and its predicted one $\hat{\mathbf{y}}^a$ for all the five aspects, *i.e.*, $\sum_{\forall \mathbf{x}} \sum_{\forall a} L_{CE}(\mathbf{y}^a, \hat{\mathbf{y}}^a)$.

2.2 Generating Fine-grained Explanations

Once the aspect model is learned, we leverage it as an external explainer that provides aspect-aware explanations for a given movie script. Specifically, by utilizing attention and prediction scores from the model, plausible explanations for age-restricted ratings can be captured at word and scene level. For example, in a `HIGH`-rated movie Deadpool, one scene with script "*... Shit! I forgot my ammo bag! Shall we turn back? No time. Fuck it. ...*" gets high attention score with bad words such as '*shit*' and '*fuck*', while another scene with script "*... If your left leg is Thanksgiving, and your right leg is Christmas, can I visit you between the holiday? ...*" gets high prediction score (but low attention score) for a specific aspect (*e.g.*, Sex) with metaphoric expression such as '*leg*'.

As a full-length movie script may include redundant parts (Ding et al., 2020) and have sparse attention distributions (Choi et al., 2020), we first segment a script $\mathbf{x}$ into n scenes with uniform length (we empirically set n as 200). Then, we individually perform the aspect classifications with each scene s_i to compute $score(s_i) = \sum_{\forall a}(\hat{\mathbf{y}}^a(s_i) + \sum_{w_j \in s_i} \hat{\alpha}_j^a)$. Based on these scores, we identify the scene-level explanation of $\mathbf{x}$ by $\mathbf{E}_{scene} = \{s_1, ..., s_m\}$ $(m < n)$ with the m-highest scores. That is, the most informative scenes are sampled with respect to multiple aspects. For more fine-grained explanations, we also extract highly attended words, and define the word-level explanation of script $\mathbf{x}$ by $\mathbf{E}_{word} = \{w_1, ..., w_k\}$ where the sum of attention weights of w_i for all aspects, *i.e.*, $\sum_{\forall a} \hat{\alpha}_i^a$, is the top-k in every scene.

2.3 Learning from Explanations

In what follows we leverage aspect-aware explanations into two directions for model to robustly predict age-restricted ratings (Hase and Bansal, 2021).

Explanation as Input. To enhance predictive performance, we repurpose scene-level explanation as model input instead of original movie scripts. Specifically, we summarize the full-length script $\mathbf{x}$ into a shorter input $\mathbf{x}'$ by concatenating the selected text scenes in $\mathbf{E}_{scene}$. As studied in many recent studies (Du et al., 2020; Huang et al., 2019), such a summary can contribute to skimming the irrelevant and redundant parts due to linguistic redundancy effects (Yu et al., 2017), not reading the whole original text. To preserve the plot context of the script, we maintain the original relative ordering in $\mathbf{x}$.

Model	Precision	Recall	F1
Shafaei et al. (2020)	79.36	73.15	75.38
AttLSTM	78.00	74.00	75.36
AttLSTM-H	84.87	78.87	80.82
AttLSTM-S	76.11	78.12	76.66
MTL	72.51	76.87	73.76
HMTL	71.97	73.96	72.81
BERT	61.21	58.81	59.21
CogLTX	77.69	72.66	74.58
HAN	80.93	78.01	79.17
MARS (Ours)	**87.80**	**85.24**	**85.68**

Table 1: Age-restricted rating prediction performance.

Explanation as Target. To better transfer knowledge from the aspect model, we leverage attention supervision (Choi et al., 2020) from the word-level explanation $\mathbf{E}_{word}$ which guides to more focus on aspect-aware important words. We provide word-level annotation α_i, 1 if w_i in $\mathbf{E}_{word}$ and 0 otherwise, then the loss $L_{KL}(\alpha, \hat{\alpha})$ for attention supervision is computed based on the Kullback-Leibler divergence following Sood et al. (2020).

Based on these explanations, we learn the rating model $f(\mathbf{x}'; \theta) : \mathbf{x}' \rightarrow (\mathbf{y}, \alpha)$ parameterized by θ. Using Attentive LSTM (shortly, AttLSTM) as our base model architecture, the rating model is trained by minimizing the aggregated loss as $L_{CE}(\mathbf{y}, \hat{\mathbf{y}}) + \lambda \cdot L_{KL}(\alpha, \hat{\alpha})$ where two terms are from scene- and word-level explanations, and λ is a preference weight (we empirically set as 0.1).

3 Experiments

3.1 Experimental Settings

Datasets. We conduct experiments on a dataset introduced by Shafaei et al. (2020). The dataset consists of 3,639 movie scripts, accompanied with age-restricted rating and multi-aspect labels. Compared to conventional text classification tasks, this dataset contains much longer documents, where the average and maximum number of words per script is 4,653 and 21,981, respectively (see Appendix C for data statistics). We split and stratify the dataset by 80/10/10 ratio for training, development, and test set for each rating. We report macro average precision, recall, and F1 score on the test set.

Baselines. As baselines, we compare against: (i) Shafaei et al. (2020), which initiate the study of predicting age-restricted ratings by leveraging not only movie scripts but also genres, emotions, and similar movies; (ii) AttLSTM and its variants AttLSTM-H and AttLSTM-S, where hard label (*i.e.*, ground-truth aspect label) and soft label (*i.e.*, prediction from the aspect model) are added as features for rating model, respectively; (iii) Multi-task learning (MTL) (Liu et al., 2019b) and Hierarchical multi-task learning (HMTL) (Sanh et al., 2019), where the aspect and rating models are jointly learned; (iv) BERT (Devlin et al., 2019) and its variant CogLTX (Ding et al., 2020) for applying BERT to long texts; (v) HAN (Yang et al., 2016), the hierarchical attention network for long texts.

3.2 Overall Performance

Table 1 presents the age-restricted rating prediction performance of baselines and MARS. In line with all baselines using the same base architecture (*i.e.*, Shafaei et al. (2020) and variants of AttLSTM), we observe that a *learning-from-explanations* scheme in MARS greatly improves the model performance, while conventional feature engineering (*e.g.*, genres) brings marginal performance improvements. Especially, the performance gap with AttLSTM-S (or AttLSTM-H) suggests that learning the target prediction task is strongly affected by how aspect and rating labels are jointly used together. As another ways to leverage both aspect and rating labels, we consider MTL and HMTL as baselines, but they show even worse performance than learning with only rating labels (*e.g.*, AttLSTM). This indicates the difficulty and necessity of finding a way to properly model the relation between aspect and rating information, where MARS present the best remedy outperforming other alternatives.

MARS also outperforms other baselines, including CogLTX and HAN designed for dealing with long documents. Here, we argue that movie script is a practical testbed to evaluate the model capability of long text processing. For example, CogLTX is evaluated on 20NewsGroups (Lang, 1995) in its paper, but the average text lengths in the dataset are barely 256 tokens, which are far shorter than the average 4654 tokens of our movie script dataset. We find that, in such an extreme case, our simple and lightweight model works better than the complex and expensive models based on BERT.

3.3 Analysis of MARS

We further investigate how MARS works based on aspect-aware explanations on the test set.

Ablation Study. We first conduct ablation study on MARS to measure how the use of word- and scene-level explanations affects the prediction per-

Movie Title	Explanations
Deadpool	... *If your left leg is Thanksgiving, and your right leg is Christmas, can I visit you between holidays?* ... ... *Big chrome* ***cockgobbler*** *! That's nice. Really got* ***fuck*** *?* ...
American Virgin	... *Take look camera say dirty whores. I'm dirty* ***slut*** *. Go crazy see. Go artai* ***butts*** ... ... *Ask something? Fantazii ever? Know* ***sexual*** *fantasies? Course believe human. Tell one. Come.* ...
No Country for Old Men	... *Charlie grabs gun shoot* ***damn*** *thing head swaying trashing glanceshot ricochets around comes* ... ... *Huntsville back. Arrest testimony.* ***Killed*** *fourteen year old girl* ...

Table 2: Scene- and word-level explanations in MARS. Highlighted words represent word-level explanation.

Model	Precision	Recall	F1
MARS (Ours)	**87.80**	**85.24**	**85.68**
MARS w/o $\mathbf{E}_{word}$	84.90	77.50	80.40
AttLSTM w/ SUMMER	47.04	43.38	35.16
MARS w/o $\mathbf{E}_{scene}$	83.39	79.22	80.47
AttLSTM w/ Lexicon	74.00	78.00	75.66

Table 3: Effects of various explanation methods.

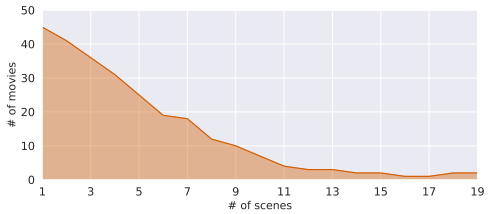

Figure 1: The number of movies whose predictions are changed by varying the number of scenes removed.

formance in MARS. In Table 3, we observe that both explanations significantly contribute to performance gains, while removing either of word- or scene-level explanation leads to the performance drop of 5.21 and 5.48, respectively, in F1 score.

Scene-level Explanation. To validate the effectiveness of scene-level explanations, we further compare MARS with AttLSTM using SUMMER (Papalampidi et al., 2020a), which summarizes movie scripts as shorter inputs for AttLSTM, as a result, Table 3 shows that using SUMMER achieves poor performance. It is because the conventional summarization methods such as SUMMER aim to capture turning points of the movie storyline such as changes of plans, major setback, climax, which are not effective for our target prediction task. In contrast, as shown in Table 2, MARS takes important task-relevant scenes, *e.g.*, sexual contents such as "*If your left leg is Thanksgiving, and your right leg is Christmas, can I visit you between holidays?*".

Word-level Explanation. To understand the advantage of word-level explanations from MARS, we compare MARS with AttLSTM using attention supervision with a pre-defined lexicon named *Google badword list*[2]. In Table 3, we find that MARS using only word-level explanation shows better performance over AttLSTM using the lexicon, mainly because MARS considers multi-aspect signals while the lexicon limits its coverage to a single aspect, *i.e.*, Profanity. Besides, MARS can capture important words best fit to each movie while the lexicon with a fixed word set can not. For example, as shown in Table 2, MARS highly attends '*dirty slut*', '*butts*', and '*sexual*' for American Virgin and '*damn*' and '*killed*' for No Country for Old Men, which contain sexual and violent contents respectively.

Counterfactual Explanation. Finally, we analyze counterfactual explanations by MARS. Specifically, we adopt *What-if* simulation, where we repeatedly remove scenes having highest scores, *i.e.*, $score(s_i)$, from the original movie scripts, then observe their counterfactual predictions until the prediction changes from `HIGH` to `MED`. Figure 1 reports the number of movies whose predictions are changed by varying the number of scenes removed. In this figure, we can find that more than 40 movies successfully change their age ratings by deleting only one scene from their original scripts, which may enable filmmakers to easily revise their contents with minimal edits for a lower rating.

4 Conclusion

In this paper, we present MARS as a scalable and effective tool for age-restricted rating prediction, significantly improving the predictive performance by aspect-aware fine-grained explanations. More thorough analysis of causality between the explanations and predictions is promising future work.

[2] https://code.google.com/archive/p/badwordslist

References

Raymond E Barranco, Nicole E Rader, and Anna Smith. 2017. Violence at the box office: Considering ratings, ticket sales, and content of movies. *Communication Research*, 44(1):77–95.

Seungtaek Choi, Haeju Park, Jinyoung Yeo, and Seungwon Hwang. 2020. Less is more: Attention supervision with counterfactuals for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6695–6704.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Ming Ding, Chang Zhou, Hongxia Yang, and Jie Tang. 2020. Cogltx: Applying bert to long texts. *Advances in Neural Information Processing Systems*, 33:12792–12804.

Jinhua Du, Yan Huang, and Karo Moilanen. 2020. Pointing to select: A fast pointer-lstm for long text classification. In *Proceedings of the 28th international conference on computational linguistics*, pages 6184–6193.

Johan Edstedt, Amanda Berg, Michael Felsberg, Johan Karlsson, Francisca Benavente, and Anette Novak. 2021. Is this harmful? learning to predict harmfulness ratings from video. *arXiv preprint arXiv:2106.08323*.

Douglas A Gentile. 2008. The rating systems for media products. *Handbook of children, media, and development*, pages 527–551.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18(5-6):602–610.

Peter Hase and Mohit Bansal. 2020. Evaluating explainable ai: Which algorithmic explanations help users predict model behavior? In *ACL*.

Peter Hase and Mohit Bansal. 2021. When can models learn from explanations? a formal framework for understanding the roles of explanation data. *arXiv preprint arXiv:2102.02201*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Ting Huang, Gehui Shen, and Zhihong Deng. 2019. Leap-lstm: Enhancing long short-term memory for text categorization. In *IJCAI*.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. *arXiv preprint arXiv:1902.10186*.

Sarthak Jain, Sarah Wiegreffe, Yuval Pinter, and Byron C. Wallace. 2020. Learning to faithfully rationalize by construction. In *ACL*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *ICLR*.

Advait Lad, Shivani Butala, and Pramod Bide. 2019. A comparative analysis of over-the-top platforms: Amazon prime video and netflix. In *International Conference on Communication and Intelligent Systems*, pages 283–299. Springer, Singapore.

Ken Lang. 1995. Newsweeder: Learning to filter netnews. In *ICML*.

Tao Lei, Regina Barzilay, and T. Jaakkola. 2016. Rationalizing neural predictions. In *EMNLP*.

Xiaoyue Li, Haonan Zhao, Zhuo Wang, and Zhezhou Yu. 2020. Research on movie rating prediction algorithms. In *2020 5th IEEE International Conference on Big Data Analytics (ICBDA)*, pages 121–125. IEEE.

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*.

Hui Liu, Qingyu Yin, and William Yang Wang. 2019a. Towards explainable nlp: A generative explanation framework for text classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5570–5581.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019b. Multi-task deep neural networks for natural language understanding. In *ACL*.

Victor Martinez, Krishna Somandepalli, Yalda Tehranian-Uhls, and Shrikanth Narayanan. 2020. Joint estimation and analysis of risk behavior ratings in movie scripts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4780–4790.

Victor R Martinez, Krishna Somandepalli, Karan Singla, Anil Ramakrishna, Yalda T Uhls, and Shrikanth Narayanan. 2019. Violence rating prediction from movie scripts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 671–678.

Emad Mohamed et al. 2020. A first dataset for film age appropriateness investigation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1311–1317.

Pinelopi Papalampidi, Frank Keller, Lea Frermann, and Mirella Lapata. 2020a. Screenplay summarization using latent narrative structure. In *ACL*.

Pinelopi Papalampidi, Frank Keller, and Mirella Lapata. 2019. Movie plot analysis via turning point identification. In *EMNLP*.

Pinelopi Papalampidi, Frank Keller, and Mirella Lapata. 2020b. Movie summarization via sparse graph construction. In *AAAI*.

Alexis Ross, Ana Marasović, and Matthew E. Peters. 2021. Explaining nlp models via minimal contrastive editing (mice). In *FINDINGS*.

Victor Sanh, Thomas Wolf, and Sebastian Ruder. 2019. A hierarchical multi-task approach for learning embeddings from semantic tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6949–6956.

Mahsa Shafaei, Niloofar Safi Samghabadi, Sudipta Kar, and Thamar Solorio. 2020. Age suitability rating: Predicting the mpaa rating based on movie dialogues. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1327–1335.

Roderik Smits and EW Nikdel. 2019. Beyond netflix and amazon: Mubi and the curation of on-demand film. *Studies in European Cinema*, 16(1):22–37.

Ekta Sood, Simon Tannert, Diego Frassinelli, Andreas Bulling, and Ngoc Thang Vu. 2020. Interpreting attention models with human visual attention in machine reading comprehension. *CoNLL*.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *EMNLP*.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard H. Hovy. 2016. Hierarchical attention networks for document classification. In *NAACL*.

Adams Wei Yu, Hongrae Lee, and Quoc V. Le. 2017. Learning to skim text. In *ACL*.

Mo Yu, Shiyu Chang, Yang Zhang, and Tommi S Jaakkola. 2019. Rethinking cooperative rationalization: Introspective extraction and complement control. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4094–4103.

Ruixuan Zhang, Zhuoyu Wei, Yu Shi, and Yining Chen. 2019. Bert-al: Bert for arbitrarily long document understanding.

Yigeng Zhang, Mahsa Shafaei, Fabio Gonzalez, and Thamar Solorio. 2021. From none to severe: Predicting severity in movie scripts. In *EMNLP*.

Yicheng Zou, Tao Gui, Qi Zhang, and Xuan-Jing Huang. 2018. A lexicon-based supervised attention model for neural sentiment analysis. In *Proceedings of the 27th international conference on computational linguistics*, pages 868–877.

A Overview

In the following sections, we provide more details on MARS. In Appendix B, we describe previous works for movie rating systems and explanations, respectively. Appendix C presents the overall statistics of the dataset used in this paper. Implementation details of MARS are in Appneidx D. We also compare MARS with random selection approach in Appendix E.

B Related Work

B.1 Movie Rating System

The age-restricted ratings for movies have a wide practical value (Gentile, 2008). For example, parents can rely on them as a guideline to determine what movies are appropriate for their children. Also, media service providers (*e.g.*, Amazon and Netflix) may use these ratings to enable age filters in parental controls (Lad et al., 2019; Smits and Nikdel, 2019). Having the rating is an important element for producers too (Barranco et al., 2017), as certain theaters refuse to show non-rated movies and it negatively affects the potential popularity of the movie as well as its gross revenue.

For this purpose, the Motion Picture Association of America (MPAA), a movie rating system, establishes five categories for the MPAA rating (G, PG, PG-13, R, NC-17) that specify the suitability of movies for children. G stands for the general group; it means all ages admitted. PG means that there is some content in the movie that parents should review. PG-13 indicates that the movie has some content deemed not appropriate for children under 13 years old. R stands for "restricted" and means people under 17 should watch the movie with a parent. NC-17 refers to no one under 17 is recommended to watch the movie. These MPAA ratings are determined based on the following aspects: (i) Violence, (ii) Language, (iii) Substance Abuse, (iv) Nudity (v) Sexual Content. On the other hand, the IMDB[3] website provides objectionable content of movies compatible with MPAA aspects (Parental Guide): (i) Violence & Gore, (ii) Sex & Nudity, (iii) Alcohol, Drugs & Smoking, (iv) Profanity (v) Frightening and Intense Scenes. The IMDB website also provides a way to rate the severity (None, Mild, Moderate, Severe) of the aforementioned aspects of content through user votes.

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

#Train	#Dev	#Test	Min. #w	Avg. #w	Max. #w	Min. #s	Avg. #s	Max. #s
2,910	364	365	483	4,782	21,181	20	944	3,022

Table 4: Data statistics: #w and #s denotes the minimum/average/maximum number of words and sentences per script, respectively.

	Vio.		Sex.		Pro.		Sub.		Fri.		Total
	Pos.	Neg.	Pos.	Neg.	Pos.	Neg.	Pos.	Neg.	Pos.	Neg.	
LOW	252	194	167	279	200	246	238	208	206	240	446
MED	533	606	730	409	580	559	398	741	763	376	1,139
HIGH	1,348	706	1,577	477	1,750	304	1,083	971	1,579	475	2,054
Total	2,133	1,506	2,474	1,165	2,530	1,109	1,719	1,920	2,548	1,091	-

Table 5: Distribution of ratings and aspects in the dataset.

In current movie rating systems, human experts watch the entire film and determine age-restricted rating for the movie. (Li et al., 2020), which is a time-consuming and non-scalable process. Although there are small efforts in automatically predicting age-restricted ratings (Mohamed et al., 2020; Shafaei et al., 2020), they are not able to explain "how do movies get their ratings?", which is crucial for its wider use in real-world. Moreover, rating happens post production, when making changes in movies can cost a lot of money (Edstedt et al., 2021). In this paper, our ultimate goal is to advance the movie rating system by improving predictive performance and providing explanations for predictions based on multiple aspects of movie contents. We leverage external multi-aspect model to explain the rating prediction based on multiple predefined aspects. Further, by predicting age-restricted ratings solely based on movie scripts, we benefits their application to wider range.

B.2 Explanation

In a similar vein, neural models dominate these days, but it remains difficult to know why such models make specific predictions for sequential text inputs (Jain and Wallace, 2019; Hase and Bansal, 2020; Ross et al., 2021). This problem has been exacerbated by the adoption of deep contextualized word representations, whose architectures permit arbitrary and interdependent interactions between all inputs, making it particularly difficult to know which inputs contributed to any specific prediction. Wiegreffe and Pinter (2019) argues for classifying model interpretability into two groups: faithfulness and plausibility. Lei et al. (2016) notes that a desirable set of criteria for explainable rationales is that they are sufficient, short, and coherent. Yu et al. (2019) extends these criteria by additionally arguing for comprehensiveness, which dictates that a explanation should contain all relevant and useful information. For the faithfulness, Jain et al. (2020) assumes that an explanation provided by a model is faithful if it reflects the information actually used by said model to come to a disposition.

C Data Statistics

Shafaei et al. (2020) initiates the research of predicting MPAA ratings from movie scripts by providing benchmark dataset. The dataset includes five categories for the MPAA rating (G, PG, PG-13, R, NC-17) and four severity level (None, Mild, Moderate, Severe) in five aspects (Violence, Sex, Substance abuse, Profanity, Frightening). Table 4 summarizes the statistics of the dataset. It is observed that the dataset contains much longer documents compared to conventional text classification datasets. To balance the negative skewness of the rating distribution, we encode ratings as a three-level categorical variable, namely LOW(<PG-13), MED(=PG-13), HIGH(>PG-13). Table 5 presents the distribution of ratings and aspects in the dataset.

D Implementation Details

For all components in MARS, *i.e.*, aspect model and rating model, we use the same Bi-directional LSTM (Graves and Schmidhuber, 2005; Hochreiter and Schmidhuber, 1997) with attention (Lin et al., 2017). The aspect model is trained by the multi-task learning scheme with the binary cross entropy loss. The rating model is trained with the movie-level rating label with the cross entropy loss and with the word-level annotations with the Kullback-

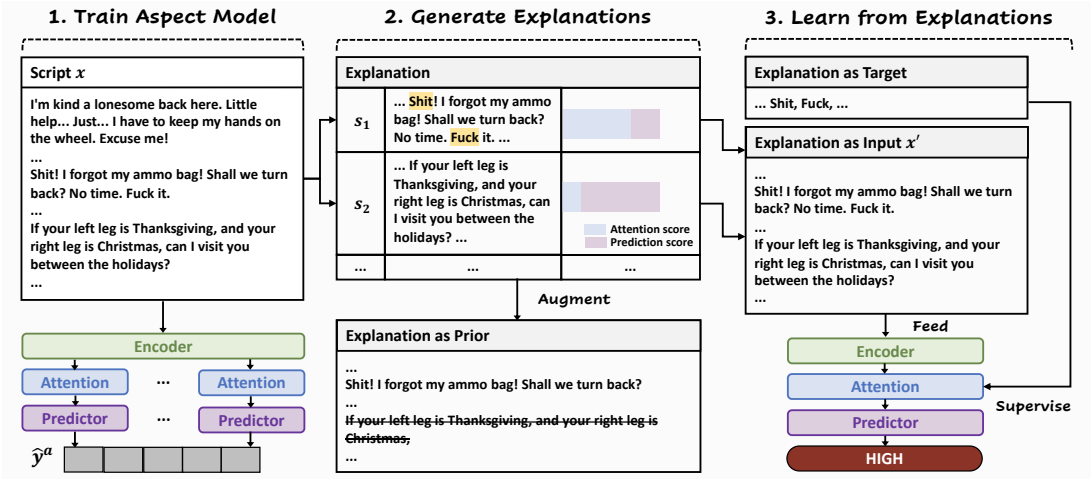

Figure 2: Illustration of MARS framework. The script is from the movie **Deadpool** with a rating of R (HIGH).

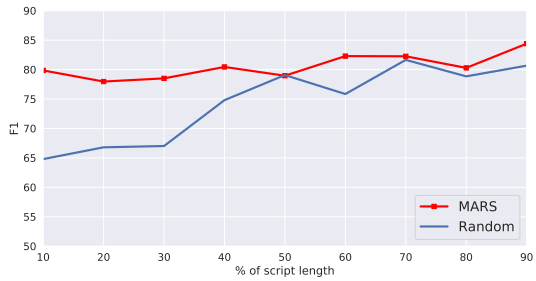

Figure 3: The performance of Random and MARS on rating prediction by varying the length of inputs (as % of script length).

Leibler divergence. Optimization is performed using Adam optimizer (Kingma and Ba, 2015) with a learning rate of 3e-4. To prevent overfitting, we adopt dropout rate of 0.2, weight decay with 5e-4, and early stop with 15 patience.

E Comparison with Random Selection

To verify explanations from MARS, we compare MARS with random selection approach, where MARS selects a subset of the scene-level explanations with their aspect scores (*i.e.*, $\hat{\mathbf{y}}^a$) while the random selection samples the same number of scenes from original scripts. As a result, Figure 3 shows that MARS consistently outperforms random selection when varying the input length. Such performance gap indicates that our design choice of producing explanations is effective to focusing on informative parts for rating prediction. Interestingly, when the number of scenes decreases, MARS still achieve better performance compared to random selection. This suggests that MARS is more robust to the input length, as it enables to selectively take the most informative parts in texts.