Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 189 (2021) 242–249

Procedia Computer Science

www.elsevier.com/locate/procedia

5th International Conference on AI in Computational Linguistics

Combining Text and Images for Film Age Appropriateness Classification

Le An Ha[a,*], Emad Mohamed[a]

[a]*Research Institute of Information and Language Processing, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1LY, United Kingdom*

Abstract

We combine textual information from a corpus of film scripts and the images of important scenes from IMDB that correspond to these films to create a bimodal dataset (the dataset and scripts can be obtained from https://tinyurl.com/se9tlmr) for film age appropriateness classification with the objective of improving the prediction of age appropriateness for parents and children. We use state-of-the art Deep Learning image feature extraction, including DENSENet, ResNet, Inception, and NASNet. We have tested several Machine learning algorithms and have found xgboost to yield the best results. Previously reported classification accuracy, using only textual features, were 79.1% and 65.3% for American MPAA and British BBFC classification respectively. Using images alone, we achieve 64.8% and 56.7% classification accuracy. The most consistent combination of textual features and images' features achieves 81.1% and 66.8%, both statistically significant improvements over the use of text only.

© 2021 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of the 5th International Conference on AI in Computational Linguistics.

Keywords: age appropriateness classification; deep learning; bi-modal classification; image classification; text classification;

1. Introduction

The question "Is this film appropriate for my children of X years of age?" frequently arises in parents' minds. Up till now, age-appropriateness of films has been recommended by censorship bodies, in the form of age rating certificates. In the United States and the United Kingdom, these age rating certificates are issued mainly by two organizations: the Motion Picture Association of America (MPAA) in the United States of America and the British Board of Film Classification (BBFC) in the United Kingdom. The two "censorship" bodies base their ratings on the film content and provide descriptions for each certificate. Different ratings for the US and UK and their interpretations can be found in Table 1. The BBFC define their classification as "the process of giving age ratings and content advice to films and other audiovisual content to help children and families choose what's right for them and avoid what's not.". The

* Corresponding author. Tel.: +44 1902 518705
E-mail address: L.A.Ha@wlv.ac.uk

1877-0509 © 2021 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of the 5th International Conference on AI in Computational Linguistics.
10.1016/j.procs.2021.05.087

classification is, in principle[1], based on the content of the films. As a result, we hypothesise that it is possible to use automatic methods to perform the classification. This, in turn, would, among other things, improve the consistency and productivity of the classification process. An automatic classifier would also provide insights into the differences in the perception of appropriateness in different countries or decades (e.g. if a machine classifier trained on data from one decade performs differently on data from different decades, we could infer that there are some differences in human perceptions across different decades, as similar texts and images, as determined by machine classifiers, are now looked at differently by human classifiers). The contribution of factors such as the country of the censor board, the time the film was produced, and the quantified content of violence or explicit material could also form the basis of various studies in Digital Humanities and Computational Social Science. While not the main focus of this research, such aspects could be very important to the understanding of the making, reception, and perception of films in different times and cultures.

Previous research indicates that using the textual content of the films alone, it is possible to build classifiers that could perform the classification fairly accurately for various aspects of the film [13, 9, 8]. Mohamed and Ha [10] compiled a dataset of film scripts and their age-appropriateness ratings, developed various classification models and reported fairly good accuracies (79.1% accuracy for American MPAA and 65.3% accuracy for British BBFC) using TFIDF values of character based ngrams as features. In this paper, we try to see whether using image features extracted using state-of-the-art image feature extraction could improve the classification performance further. From a human perspective, we know that vision adds more information, and should thus improve the classification accuracy, a fact also supported by machine vision research [11, 15, 12]. Our research focuses on whether the use of current state-of-the-art image feature extraction could improve the automatic classification models. If it could, we then can have further evidence that these image feature extraction methods can capture abstract concepts such as age-appropriateness. We add images to Mohamed and Ha's dataset by using the Internet Movie Database (IMDB) to extract images associated with each film. We then use state-of-the-art image feature extractors to extract vectors representing the images, combine these vectors with textual vectors, and investigate the impact of these image feature vectors on the accuracy of the classifiers. The contributions of this paper are: (1) a bi-modal datatset combining images and texts for 17000 films, (2) a new application domain for Deep Learning to the Humanities in the field of Film Studies showing that DL can perform what has so far been a human-only activity, and (3) The introduction of deep learning methods to the Digital Humanities, which has so far been limited.

The rest of this paper is organised as follows: section two introduces the data and the methods used in the research, section three outlines the results and provides analysis, examples, and the confusion matrix, followed by the conclusion and plans and suggestions for future work.

Table 1. Certificate classifications and average number of photos per film for each certificate rating

Country	Certificate	Meaning	#Photos
USA	G	General Audience	55.38
	PG	Parental Guidance	58.56
	PG-13	Parents Strongly Cautioned	72.49
	R	Restricted: Under 17 requires accompanying parent or adult guardian	48.98
	NC-17	Adults only: No One 17 and Under Admitted.	38.12
UK	U	Suitable for all	43.95
	PG	Parental Guidance	53.22
	12A,12	Cinema and video release suitable for 12 years and over	82.70
	15	Suitable only for 15 years and over	44.97
	18	Suitable only for adults	47.11
	R18	Adult works for licensed premises only	2

[1] We hope that the classifications are not influenced by any other factor, such as writers, principle actors, actresses, producers, and distributors. Having an automatic classifier would also allow us to check whether these factors influence the classifications.

2. Data and Methods

Mohamed and Ha's dataset was created using an INNER JOIN of two resources: film scripts and film certificates. *Film scripts* were obtained from the website www.springfieldspringfield.com, which unfortunately does not exist any more. The files, available in html, were converted into text and were run through a basic cleaning pipeline that involved transforming the utterances into proper sentences using the Spacy package [6]. Mohamed and Ha also removed non-dialogue elements from the scripts like scene descriptions and actor actions, a practise that we follow for two reasons: (1) these are not consistent across the film scripts as many films do not have them, and (2) because these are external to the film content proper. These scripts were combined with *IMDB Certificates*, which indicate, for each film, the age for which the film is appropriate. These certificates may vary by country and cut. For example, the film "The Hobbit: The Battle of the Five Armies" has been rated both PG-13 and R in the United States based on which cut is intended. The main certificate used on IMDB for both the UK and the USA is used, which in the case of this Hobbit film is 12A for the UK and PG-13 for the USA. We then collected *IMDB Images*, accompanying and characterizing main scenes of the film by downloading the images in the photo gallery of each film. The number of images accompanying each film description on IMDB is limited; and we understand that this limitation will affect the accuracy of prediction. We nonetheless hypothesise that, even with the limited number of images, the combination of text and images will lead to better performance in classification since text alone could be ambiguous. The combination of these two modalities would then contribute to the disambiguation of otherwise difficult to interpret textual content, and will thus lead to better classification accuracy.

The IMDB certificates are used as labels, in what is known as *distant annotation*. The BBFC website explains that they use two raters for each film and when there is a dispute, a third, more experienced, rater steps in. This is very similar to human linguistic annotation. We do not know the inter-rater agreement, and thus are unable to determine the ceiling of human performance. Similar to [10], we use the following upper bounds and baselines:

The Upper Bound. The IMDB hosts certificates from 70 countries around the world. The upper bound takes as predictor variables all these certificates and as a target the country in question. If we want to predict the UK certificate, we use all the other certificates as features. This method achieves accuracies of 84.7% and 80% for the US and the UK (OtherCts in Table 5). Both experiments were performed using XGBoost, our best classifier for this task. *The baseline* for this paper is 55.0% and 41.8% for the USA and UK respectively, representing the majority classes ("R" for the US and "15" for the UK).

2.1. Dataset Description and Statistics

The dataset comprises 17018 titles. Transcripts of these titles contain a total of 181 million words. USA certificates are available for 8923 titles and British certificates are available for 10920 titles. 7068 titles have both countries' certificates. The mapping between the UK and the USA ratings is not one to one. A classifier that uses the UK ratings to predict the USA ratings would only have an accuracy of 80.6% (SingCt in Table 5). For each title in the dataset, we download images that belong to the title gallery excluding those images that are not part of the film itself, for example, those whose captions include descriptions such as "X at an event to promote the film Y" from IMDB. A total of 429050 photos have been collected, with an average of 46.94 photos per title. The average numbers of photos per title for each certificate rating can be found in Table 1. We use the same train (70%), test (20%), and dev (10%) subsets.

Table 2. Image feature extraction models, their sizes, and their performances in the ImageNet challenge.

Model	Short description	Size	Depth	Top-1acc	Top-5acc
NASNetMobile	Architecture search	23M	382	77.4	91.9
Dense169	Densely connected	57M	169	76.2	93.2
InceptionV3	Efficient implementation of Resnet	92M	159	77.9	93.7
ResNet152V2	Residual NN	232M	152	78.0	94.2
NASNetLarge	Architecture search	343M	527	82.5	96.0

2.2. Methods

a) Texts: Mohamed and Ha tried a variety of classification methods from both traditional machine learning and Artificial Neural Networks. They concluded that the best setting is to use character ngrams tf-idf as features, and XGBoost as classifier, achieving an accuracy of 79.1% when predicting USA certificates, and 65.3% when predicting British certificates. We have replicated their experiments and we have reached the same results using textual features. The next section will combine these textual features with image features and will also explore the use of images alone in film age appropriateness classification.

b) Images only: Recent advances in machine vision have produced models that almost surpass human performance in image object recognition tasks, specifically the ImageNet challenges. Information needed to distinguish between all the 1,000 classes in ImageNet is also often useful to distinguish between new kinds of objects. Such information can be harvested from the outputs of penultimate layers of models originally trained to distinguish between all the classes in ImageNet. We use these outputs as our image feature extractors. Specifically, we use NASNetMobile [18], Dense169 [7], InceptionV3 [14], ResNet152V3 [5], and NASNetLarge [18]. These models represent the state-of-the-art in image object recognition (Table 2). Keras implementations of these models[2] are used. For each of the images, we produce a feature vector; we then pool feature vectors of all the film's images first, and use the resulting vectors as input for certificate classifications. We try mean, median, and max pooling, and find mean pooling to be the best. We also try dimension reduction methods such as PCA as pooling methods; and find that they also are inferior to mean pooling. We also produce an ImageConcat vector, which is the concatenation (stacking the vectors horizontally) of the pooled vectors produced by individual feature extraction models.

Film age classification using images may not be as easy as it sounds. The reason for this is that in an R-rated movie, most of the images may be innocent, the equivalent of PG-rated, but only some may contain violence or explicit references. This poses even a bigger a challenge to our experiments since we use only the set of images provided by IMDB, which, for various reasons, may not contain the most violent or explicit images in the film. It is thus useful to check the accuracy of using only the images as per category using a balanced dataset. To build our balanced dataset, we first choose titles of which we have at least 40 images. We then build a balanced training set of 450 titles of each rating. We then choose 40 random images for each title to form the training set. Similarly, from the titles that have at least 40 images in the test set, we choose a set of 150 random titles for each rating. For each of these titles, we pick 40 random images. They form our test set. We perform this experiment only for the USA certificates. We only use three ratings in this experiment: PG, PG-13, and R. The two other categories have not been used due to the small number of films in the categories, which make it impossible to balance them. In experiment ImagePool, we pool feature vectors of all the film's images first then classify the pooled vectors, while in ImageIndividual, we classify individual images then count how many times images belonging to a film have been classified as belonging to a specific rating, and then take the rating with the most count as the predicted rating for each film.

c) Text and Images combined: For each title, the character based ngram TFIDF vector and the image vector are concatenated into a single vector, and fed into XGBoost. Other classification algorithms such as Random Forests and Logistic Regression have also been experimented with, but the results are inferior to those of XGBoost. While TFIDF is not usually thought of as comparable to word embeddings, Mohamed and Ha's experiments show that in this specific case, word embeddings (from BERT and ELMO) were not as good as this traditional method. In the experiments we ran, word embeddings did not produce good results. The use of XGBoost was also beneficial in other ways. Since the TFIDF vector is very large, corresponding to the vocabulary size of X words, neural network implementations in Keras and PyTorch did not scale well, unlike XGBoost and similar algorithms that can deal with a large number of textual features.

d) Evaluation metrics: For the balanced image experiment, we use the standard precision, recall, f-measure, and overall accuracy. For other experiments, we use the standard metrics of accuracy and the Area Under the Curve of the Receiver Operating Characteristic (AUC), which incorporates the trade-off between precision and recall. Two settings for evaluation of accuracy are used: strict accuracy (Acc in Table 5) is the normal accuracy, and relaxed accuracy (RelaxAcc), in which a prediction of a certificate that is either the same as, or only one age rating higher or lower than, the true certificate, is considered correct. While the relaxed accuracy is in common use in Machine Learning, it

[2] https://keras.io/applications/

Table 3. Results of the balanced experiment, ImagePooled

Experiment	P	R	F	Train size	Test size
PG	0.69	0.65	0.67	450	150
PG-13	0.52	0.56	0.54	450	150
R	0.60	0.58	0.59	450	150
accuracy			0.60	1350	450

Table 4. Results of the balanced experiment, ImageIndividual

Experiment	P	R	F	Train size	Test size
PG	0.75	0.59	0.66	450	150
PG-13	0.49	0.55	0.52	450	150
R	0.53	0.58	0.55	450	150
accuracy			0.57	1350	450

is especially important in the context of film ratings due to the differences among countries. This relaxed evaluation thus mirrors the state of the data set.

There has been previous work in combining texts and images for downstream tasks. Chen and Zhuge [2] combine text and image information to generate a multimodal summary comprising images and their captions. Rafkind et al. [11] combine text and image features to classify images in bioscience literature. Taniguchi et al. [15] and Sakaki et al. [12] combine text and image classifiers to identify the gender of Twitter users. They classify the images first, then pool the image classifications to classify the users, whereas we pool the image feature vectors first. We tested the former methods (classifications and then pooling), and found them to be inferior to pooling first (the accuracy for US certificates, image only, classification of individual images first: 59% compared to classification of pooled vectors: 62%). Generating captions from images has also gathered attention recently [17, 4, 3]. Ailem et al. [1] learn textual and visual representations jointly; this leads to competitive performance on tasks of assessing pairwise word similarity and image/caption retrieval.

3. Results & Analysis

Tables 3 and 4 present the results for two experiments using a balanced dataset of the categories PG, PG-13, and R. We can see that when the data is balanced, it is easier to classify PG then R then PG-13. This may be due to the fact that PG-13 is a confused category that has elements of both PG and R. In a PG film, one does not expect to see images of violence, gore, or sex, making it more consistent, whereas in R films innocent images may also be found, hence the easier classification of PG vs. R films. To give some examples of the classifications assigned by our image classifier versus the true category, figure 1 shows a number of images predicted as PG-13 and the true category of the film they come from. For example, the third image on the first row, which comes from the R-rated film "Courage Under Fire" (1996)[3] has been classified as PG-13. From a human perspective, the image does not show any violence or explicit material.

Tables 5 shows the results of our experiments, and figure 2 shows the confusion matricies. When image feature vectors are combined with text vectors, the performances of the classifiers are approaching or surpassing those using ratings from one country to predict those of another country (SingleCt in the table). Around 95% or more of the predictions are within one rating of the correct ones. Despite its incomplete nature, visual data, in the form of extracted feature vectors, do help improve the accuracy of the prediction of age rating certificates when combined with TFIDF. Only InceptionV3 shows statistically significant improvements in accuracy of predictions for both the USA

[3] https://www.imdb.com/title/tt0115956/

Fig. 1. Some examples of PG-13 classification by our balanced image classifier (text before the colons) versus the true category of the film they come from (text after the colons). From top to bottom, left to right: images from films ”Electric Dreams” (1984), ”Jumanji” (1995), ”Courage Under Fire” (1996), ”The Parent Trap” (1998), ”Not Another Teen Movie” (2001), and ”Stuck on You” (2003)

Fig. 2. Confusion matrix, best combinations of features for USA and UK ratings predictions. Rows are for true ratings, and columns are predicted ratings. We can see that most predictions are within one rating of the true target value. The fact that NC-17 and 18 rated films are downgraded to R and 15 may be a result of IMDB not having true image representatives of these two adult categories.

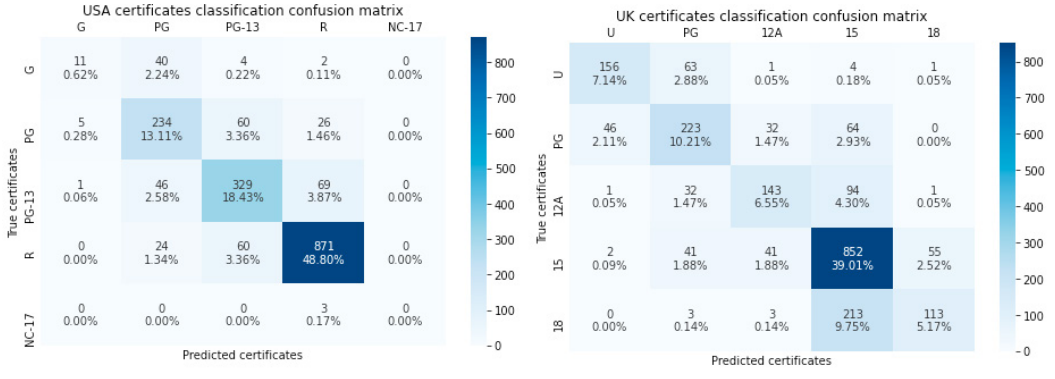

Table 5. Results from our experiments (*: statistically significant improvements ($p < 0.05$) over [10]'s best results, which were achieved using char-ngram TFIDF as features (TFIDF), determined using boostrapping statistical significance tests), classifiers are trained on the train set, and tested on the test set. Dev set is used for hyperparameter optimisation. Acc and RelaxAcc numbers are percentages

	USA (MPAA)			UK (BBFC)		
Input features	Acc	RelaxAcc	AUC	Acc	RelaxAcc	AUC
NASNetMobile	61.3	88.0	0.896	55.5	86.7	0.880
Dense169	62.7	89.2	0.899	56.3	86.7	0.886
InceptionV3	62.0	88.3	0.896	55.0	86.1	0.883
ResNet152V2	62.0	88.1	0.897	56.5	87.1	0.886
NASNetLarge	63.8	89.0	0.902	55.8	86.8	0.886
ImageConcat	64.8	89.4	0.905	56.7	88.0	0.893
TFIDF	79.1	96.2	0.962	65.3	94.2	0.930
TFIDF+NASNetMobile	80.2	96.1	0.964	*67.4	94.8	0.939
TFIDF+Dense169	*81.0	96.8	0.965	66.7	94.9	0.939
TFIDF+InceptionV3	***81.1**	**97.0**	**0.966**	*66.8	**95.2**	0.938
TFIDF+ResNet152V2	80.2	96.6	0.965	***68.1**	94.5	**0.941**
TFIDF+NASNetLarge	79.4	96.1	0.966	*66.9	94.6	0.939
TFIDF+ImageConcat	80.4	96.6	0.966	*67.2	94.8	0.940
SingleCt	80.6	95.1	0.957	61.8	91.5	0.899
OtherCts	84.7	96.7	0.978	80.0	96.3	0.972

and the UK. Other image feature extraction models provided statistically significant improvements for either the USA (Dense169) or the UK (NASNetMobile, ResNet152V2, NASNetLarge, and ImageConcat). Using visual data alone, ImageConcat provides the best results for both countries. Given that the categories of certificates follow a certain order with respect to age appropriateness, we have experimented with regression models such as Random Forest regression and XGBoost regression, and found them not to be as good as the classification models (73.1% vs 81.1% for USA and 58.1% vs 68.1% for UK). We have tried a ordinal regression method [16], which does not assume the distances between two consecutive classes are a constant (as normal regression methods do), to take advantage of the fact that the age-appropriateness is progressive, i.e. films suitable for a 12-year-old should also be suitable for a 15-year-old. The results are slightly worse than what we reported here with regard to accuracy (79.2% vs 81.1% for USA, 67.8% vs 68.1% for UK), but slightly higher with regard to relaxed accuracy (97.4% vs 97.0% for USA and 97.0% vs 95.2% for UK).[4]

4. Conclusion and future work

We have conducted experiments with the target of predicting the age rating of films based on images, and the combinations of text and images. Our experiments included ones on a general corpus as well as limited experiments on a balanced subset geared towards examining the errors produced by the classifier. Our results indicate that the combination of images and texts is better than either images or text alone, reaching an accuracy comparable to that of using ratings from one country to predict the ratings in another country in spite of the fact that we use only a very limited subset of the images that can potentially be used for such a task.

Our future work will focus on two aspects: (1) investigating the use of the whole video and audio of the film in age rating classification. We believe that with such an amount of data, we can produce results that are on par with, if not more accurate than, those produced by censorship bodies, and, when we reach the point where we can quantify the distribution of these materials in the film, we will (2) conduct computational social science analysis of the distribution

[4] The computational cost of this experiment is not expensive. We used a typical desktop PC with an Intel I7 5820K 16GB RAM, and a GTX 1080 Ti GPU. Training classification models takes about 4 to 8 hours each, whereas converting images to feature vectors takes about 12 hours.

of sex and violence in films and its relationship to cultural and country-based differences, for which we will use not only the textual and audiovisual data, but also the reports provided by parents on film contents. The two future concerns are both related to our desire to conduct *responsible* Computational/Digital Humanities research.

References

[1] Ailem, M., Zhang, B., Bellet, A., Denis, P., Sha, F., 2018. A probabilistic model for joint learning of word embeddings from texts and images, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium. pp. 1478–1487. URL: https://www.aclweb.org/anthology/D18-1177, doi:10.18653/v1/D18-1177.

[2] Chen, J., Zhuge, H., 2018. Abstractive text-image summarization using multi-modal attentional hierarchical RNN, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium. pp. 4046–4056. URL: https://www.aclweb.org/anthology/D18-1438, doi:10.18653/v1/D18-1438.

[3] Chen, X., Zitnick, C.L., 2015. Mind's eye: A recurrent visual representation for image caption generation, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2422–2431. doi:10.1109/CVPR.2015.7298856.

[4] Fang, H., Gupta, S., Iandola, F., Srivastava, R.K., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., Platt, J.C., Zitnick, C.L., Zweig, G., 2015. From captions to visual concepts and back, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1473–1482. doi:10.1109/CVPR.2015.7298754.

[5] He, K., Zhang, X., Ren, S., Sun, J., 2016. Identity mappings in deep residual networks, in: Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV, pp. 630–645. URL: https://doi.org/10.1007/978-3-319-46493-0_38, doi:10.1007/978-3-319-46493-0_38.

[6] Honnibal, M., Montani, I., 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

[7] Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, pp. 2261–2269. URL: https://doi.org/10.1109/CVPR.2017.243, doi:10.1109/CVPR.2017.243.

[8] Martinez, V., Somandepalli, K., Tehranian-Uhls, Y., Narayanan, S., 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), Association for Computational Linguistics, Online. pp. 4780–4790. URL: https://www.aclweb.org/anthology/2020.emnlp-main.387, doi:10.18653/v1/2020.emnlp-main.387.

[9] Martinez, V.R., Somandepalli, K., Singla, K., Ramakrishna, A., Uhls, Y.T., Narayanan, S., 2019. Violence rating prediction from movie scripts. Proceedings of the AAAI Conference on Artificial Intelligence 33, 671–678. URL: https://ojs.aaai.org/index.php/AAAI/article/view/3844, doi:10.1609/aaai.v33i01.3301671.

[10] Mohamed, E., Ha, L.A., 2020. A first dataset for film age appropriateness investigation, in: Proceedings of The 12th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France. pp. 1311–1317. URL: https://www.aclweb.org/anthology/2020.lrec-1.164.

[11] Rafkind, B., Lee, M., Chang, S.F., Yu, H., 2006. Exploring text and image features to classify images in bioscience literature, in: Proceedings of the HLT-NAACL BioNLP Workshop on Linking Natural Language and Biology, Association for Computational Linguistics, New York, New York. pp. 73–80. URL: https://www.aclweb.org/anthology/W06-3310.

[12] Sakaki, S., Miura, Y., Ma, X., Hattori, K., Ohkuma, T., 2014. Twitter user gender inference using combined analysis of text and image processing, in: Proceedings of the Third Workshop on Vision and Language, Dublin City University and the Association for Computational Linguistics, Dublin, Ireland. pp. 54–61. URL: https://www.aclweb.org/anthology/W14-5408, doi:10.3115/v1/W14-5408.

[13] Shafaei, M., Safi Samghabadi, N., Kar, S., Solorio, T., 2020. Age suitability rating: Predicting the MPAA rating based on movie dialogues, in: Proceedings of the 12th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France. pp. 1327–1335. URL: https://www.aclweb.org/anthology/2020.lrec-1.166.

[14] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2818–2826. doi:10.1109/CVPR.2016.308.

[15] Taniguchi, T., Sakaki, S., Shigenaka, R., Tsuboshita, Y., Ohkuma, T., 2015. A weighted combination of text and image classifiers for user gender inference, in: Proceedings of the Fourth Workshop on Vision and Language, Association for Computational Linguistics, Lisbon, Portugal. pp. 87–93. URL: https://www.aclweb.org/anthology/W15-2814, doi:10.18653/v1/W15-2814.

[16] Wooldridge, J.M., 2010. Econometric Analysis of Cross Section and Panel Data. volume 1 of *MIT Press Books*. The MIT Press. URL: https://ideas.repec.org/b/mtp/titles/0262232588.html.

[17] Xu, K., Ba, J.L., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R.S., Bengio, Y., 2015. Show, attend and tell: Neural image caption generation with visual attention, in: Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37, JMLR.org. pp. 2048–2057. URL: http://dl.acm.org/citation.cfm?id=3045118.3045336.

[18] Zoph, B., Vasudevan, V., Shlens, J., Le, Q.V., 2018. Learning transferable architectures for scalable image recognition, in: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, pp. 8697–8710. URL: http://openaccess.thecvf.com/content_cvpr_2018/html/Zoph_Learning_Transferable_Architectures_CVPR_2018_paper.html, doi:10.1109/CVPR.2018.00907.