SETSCI Conference Proceedings,
4 (1): 612-614, 2019

ISAS 2019
3rd International Symposium on Innovative Approaches in Scientific Studies
Section: Engineering and Natural Sciences
April 19 – 21, 2019, Ankara, Turkey

MPAA Rating Prediction Based on Deep Learning

Caner BALIM[1*], Ugur GUREL[2+]

[1]Afyon Kocatepe University
[2]Eskisehir Osmangazi University
[*]Corresponding author: cbalim@aku.edu.tr
[+]Speaker: cbalim@aku.edu.tr
Presentation/Paper Type: Oral / Full Paper

Abstract – The Motion Picture Association of America(MPAA) is used in the United States to rate a film's compatibility via its content. These systems evaluate ratings according to the scenes in the movies. In this work, a deep learning based classification technique is proposed to differentiate between MPAA ratings (G, PG, PG-13, R, NC-17) via subtitles. Movie subtitles which generally locate bottom of the screen to show character dialogs in movies or series. The syntax of each subtitle files are learned through Word2Vector. We then constructed a binary classifier based on the preceding representation dataset. We studied the performance of different classifiers with stratified ten-fold cross-validation. The model has been validated with experiments on English subtitle dataset. According to the experiment results, our proposed method was achieved more than 59\% accuracy rate.

Keywords – Film rating, Word2vec, Deep Learning, Machine Learning, Classification

I. INTRODUCTION

Motion picture rating systems used to help humans understand movie contents. These systems is designed to classify films with regard to compatibility for audiences in terms of issues such as sex, violence, substance abuse, profanity, impudence or other types of mature content. Also, these systems help parents to choose convenient films for their children.

The MPAA is one of the movie picture rating system that are used in United States [1]. It is a certification and rating agency operating since 1922. MPAA logo is at the bottom of the casts at the end of American cinema movies. MPAA has five categories:

- General Audiences(G): There are no uncomfortable scenes for children. Anyone can watch these movies.
- Parental Guidance Suggested(PG): There is no sex or drug scene in this category. Violence and horror scenes can be minimally found.
- Parents Strongly Cautioned(PG-13): Violence, drugs, sex categories may be implicitly found.
- Restricted(R): 17-year-old child can only watch these movies with their parents. Drugs, sex etc. scenes may be found.
- NC-17: This movies are clearly adult. Children are not admitted in these movies.

In order to prevent the film from getting a lower MPAA rating, the film makers are cutting some scenes in order to get more box office revenues. When the films were examined, it was seen that the films with the most G and PG labels were found. Because of this issue, especially the number of NC-17 labeled films is quite low.

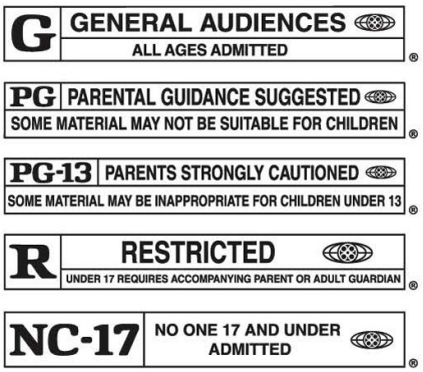

Fig. 1 MPAA Film Ratings System

There is a lot of research on the films such as movie genre classification, movie revenue prediction etc. in literature. In the last a few year, deep learning studies have gained intensity with the advantage of GPU technologies. As a result of our research, although there are deep learning based studies in the literature, we have not found any studies developed by our techniques on MPAA rating detection. But there are studies in which the MPAA rating is used as feature.

Kim et al. (2015) proposed a method for box office prediction [2]. They used features such as MPAA ratings and budget, as well as different representations of star power in actresses and actors.

Ozkan et al. (2018) tried to guess the box office revenue through the movie posters [3]. In this study, Inception-V3 model was used for feature extraction.

Sivaraman and Somappa proposed a method on movie trailer classification [4]. They classified trailers from scenes such as action, drama, horror etc. As a result of the study, it was seen that LSTM networks were successful in video classification.

Jain [5] proposed a method for classifying movie trailers into 5 movie genres, namely action, comedy, horror, drama and music. He use a neural network with 21 inputs (7 low-level visual features and 14 audio features) for classification.

In this study, we put forward a new movie rating detection technique via movie subtitles. Word2Vec is used for feature extraction. It realizes higher accuracy of about 98.69\%. Regular Expression (Re) and natural language processing techniques was used to text pre-processing step. At last, to verify the model, English subtitles dataset was used.

The rest of the paper is organized as follows. Materials and method are explained in detail (Dataset, pre-processing, feature extraction with Word2Vec, Classification) at section 2. In Section 3, experimental setting and performance result are illustrated. And finally, Section 4 concludes the paper.

II. Materials and Method

A. Materials

According the our research, we didn't find any bench-marked public movie MPAA ratings and subtitle dataset available. For this reason we decided to collect a new dataset of subtitles. Our dataset consists of English subtitles extracted from different web sites [6]-[7], distributed among five categories, namely G, PG, PG-13, R and NC-17. See Table 1 for more information about the dataset. The reason why the NC-17 category is less is that the filmmakers cut some scenes and evaluate them as unrated .

Table 1. Movie rating dataset

Label	Counts
G	1859
PG	1357
PG-13	823
R	692
NC-17	43
Total	**4810**

B. Method

The following sections aim to explain proposed method on which the problem of this work is based. Section B.1 is focused text processing for word2vec performance improvement. The goal of Section B.2 and B.3 is then to give an introduction word2vec technique and understanding of the classification step.

Figure 2 shows the workflow of our method.

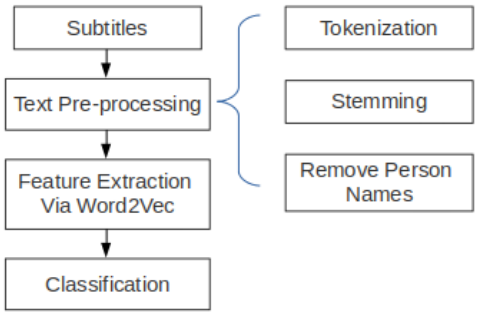

Fig. 2 Proposed Algorithm

B.1 Text Pre-processing

Movie subtitles can includes in a variety of forms from a list of individual words, different symbols and person names. Because of this, transforming subtitles into something an algorithm can digest it a complicated process. The sample subtitle file is showed in Figure 3.

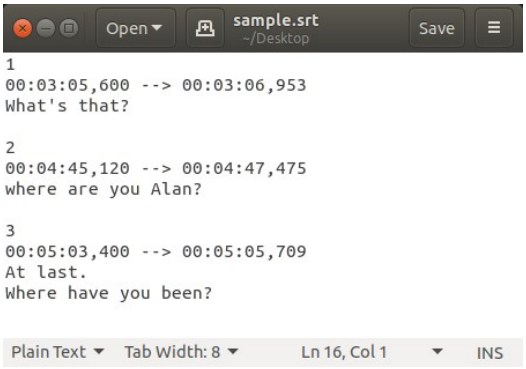

Fig. 3 Sample subtitle file

In this study, we have used Word2Vec algorithm to transform texts to vector. Before feature extraction, We was applied several pre-processing steps for performance improvement in our model:

- Removes all non symbol characters.
- Lowercase all words.
- Tokenization of words is done to separate words using white space as the delimiter.
- Stemming which reducing words to their word stems [8].
- Removes all person names [9].

B.2 Feature Extraction via Word2Vec

With the advancing Graphical processing unit (GPU) technologies deep learning applications which lasted for weeks, started to result in shorter periods. One of the best aspects of deep learning is that it performs very successful determinations in the feature extraction stage for machine learning. In this study, the Word2Vec model introduced by Google in 2013 was used for feature extraction step [10].

Word2Vec is an algorithm that generates vectorized representations of words, based on their linguistic context. The general operation logic is similar to the artificial neural network. First, random weights are determined. Afterwards, forward-propagation and loss are calculated and back-

propagation algorithm is applied and weights are updated. This process is repeated up to the number of epoch. By default Word2Vec has the epoch count of 5. Increasing the number of Epoch by 15-20 may improve the model. Because Word2Vec contains a lot of matrix multiplication and has a large network structure, the process of calculating gradients is very complicated, difficult and slow.

There are two types of sub methods in Word2Vec: CBOW (Continuous Bag of Words) and Skip-Gram. Two methods are generally similar. Figure 4 shows CBOW and Skip-Gram model workflow.

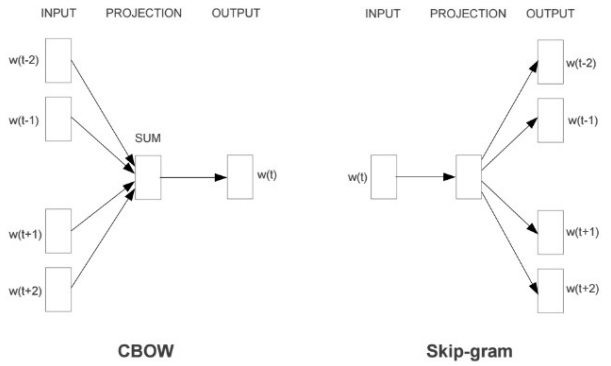

Fig 4. CBOW and Skip-Gram (https://rohanvarma.me/Word2Vec/)

After pre-prosessing step, we trained the Word2Vec model on the cleaned messages. Then the words in each message are converted into a 1x300-dimensional vector by means of Word2Vec. Therefore we was extracted 300 features to predict MPAA ratings.

B.3 Classification

After the feature extraction step, we are used for input of several machine learning classifiers such as Support Vector Machines(SVM), Decision Tree(DT) and Naive Bayes(NB), along with the label of MPAA ratings.

Also we use cross-validation which is a statistical method, used in applied machine learning to compare and select a model for a given predictive modeling problem. In this study, the performance of each classifier was assessed with a stratified 5-fold cross-validation method. Stratified k-fold cross-validation was selected to make accurate evaluations on unbalanced datasets.

III. PERFORMANCE EVALUATION

We run the experiments on Ubuntu 18.04 operation system with Inter(R) Core(TM) i7 CPU of 16 GB. We use Python programming language. Python Gensim Library was used for Word2Vec Feature extraction [11].

To evaluate the performance of our created classification and make it comparable to current approaches, we use Recall (Sensitivity), Precision, F-measure and Accuracy (ACC) to measure the capability of classifiers.

Confusion Matrix after classification and cross validation step is shown in Table 2.

Table 2. Confusion Matrix

	G	NC-17	PG	PG-13	R
G	235	34	70	12	21
NC-17	1	6	0	0	1
PG	36	27	55	14	1
PG-13	12	52	24	42	34
R	41	95	9	39	87

Figure 3 show results of different classifiers.

Table 3. Evaluation Results with different classifiers

	SVM	DT	NB
Accuracy	0.59	0.39	0.41
Precision	0.47	0.29	0.39
Recall	0.42	0.29	0.38
F1-Score	0.43	0.29	0.34

As a result of evaluations, SVM get best result on subtitle datasets.

IV. CONCLUSION

In this work, we explored the issues on the current movie classification techniques, and suggested a new MPAA rating detection method based on deep learning. After text pre-processing step, we utilized WordVector technique for converted them into high-dimension vectors. Also, three different machine learning algorithm were applied to classify movie MPAA ratings. The evaluation result shows that SVM algorithm predicted 59\%.

In future, we will collect more data. In addition, we want to develop a model which have movie subtitles and trailer visual feature.

REFERENCES

[1] https://www.mpaa.org/
[2] T. Kim, J. Hong, and P. Kang. (2015). Box office forecasting using machine learning algorithms based on SNS data.
[3] K. Ozkan, O. N. Atak, and S. Işik. "Using movie posters for prediction of box-office revenue with deep learning approach." 2018 26th Signal Processing and Communications Applications Conference (SIU). IEEE, 2018.
[4] K. S. Sivaraman and G. Somappa (2016). MovieScope: Movie trailer classification using Deep Neural Networks.
[5] S. K. Jain, Movies Genres Classifier using Neural Network, in: Proceedings of the International Symposium on Computer and Information Sciences, 2009, pp. 610–615.
[6] https://www.imdb.com/
[7] *https://www.opensubtitles.org*
[8] M. F. Porter, (1980). An algorithm for suffix stripping. *Program*, *14*(3), 130-137.
[9] J. R. Finkel, T. Grenager, and C. Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pp. 363-370.
[10] T. Mikolov,, Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111-3119).
[11] R. Rehurek and P. Sojka 2010. Software Framework for Topic Modelling with Large Corpora. Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, pp. 45-50