GU J Sci, Part C, 11(1): 252-262 (2023)

Gazi University

Journal of Science

PART C: DESIGN AND TECHNOLOGY

http://dergipark.gov.tr/gujsc

A Smart Movie Suitability Rating System Based on Subtitle

Murat IŞIK[1*]

[1]*Kirsehir Ahi Evran University, Faculty of Engineering and Architecture, Department of Computer Engineering, 40100, Merkez/KIRŞEHİR*

Article Info

Research article
Received:21/07/2022
Revision:0711/2022
Accepted:21/11/2022

Keywords

Machine Learning
Subtitles
Movie Ratings
Parental Guidelines
NLP
Natural Language Processing

Abstract

With the enormous growth rate in the number of movies coming into our lives, it can be very challenging to decide whether a movie is suitable for a family or not. Almost every developed country has a Movie Rating System that determines movies' suitability age. However, these current movie rating systems require watching the entire movie with a professional. In this paper, a model has been developed which can determine the rating level of the movie by only using its subtitle with no professional help. To convert the text data to numbers, TF-IDF (Term Frequency Inverse Document Frequency) vectorizer, WIDF (Weighted Inverse Document Frequency) vectorizer, and GWS (Glasgow Weighting Scheme) have been used. RF (Random Forest), SVM (Support Vector Machine), KNN (K-Nearest Neighbour) and MNB (Multinomial Naive Bayes) have been utilized to find the best combination that achieves the highest results. The accuracy of the results has been achieved of 85%. The result of the proposed classification approach is promising, and the model can be used by the movie rating committee for pre-evaluation.

Cautionary Note: There may be some inappropriate words in the paper that one may find offensive; however, this cannot be avoided because of the nature of the work.

1. INTRODUCTION

Movies produced in the world are growing at an exponential rate so the number of movies coming into our lives is increasing day by day. It can be very difficult to monitor if a movie is suitable for the family or not. That's why there is a system named "Movie Ratings" in almost every country. The Movie Ratings also called Parental Guidelines or Rating Certificate or Parental Certificate, provide information about the content and age-appropriateness of the movies. Before the movies are released, they are evaluated based on their content. Although many countries have their own movie rating processes and approaches, the drawback of these approaches is that they require a professional involvement which causes and consumes great time and effort. In this study, our motivation is to reduce the consumed time and effort by developing a model using the dialogues of characters in the movies.

In a movie, a story is proceeded by dialogues of characters also known as subtitles. These subtitles can contain detailed descriptions of all the information related to the movie [1], [2]. Hence, we can consider the subtitles as they constitute large amounts of conversational, interesting resources for dialogue modelling [3] and information about the movie. This means the subtitles of the movies can be used to analyse the content of the movies. By doing so, it's possible to identify various types of patterns. The use of these patterns enables extracting much information associated with the movies such as finding anomalies, correlations, similarities and so on. Therefore, many researchers have developed such techniques that extract meaningful information from movies for genre classification, browsing, searching, and indexing [1].

A model has been developed to classify a movie automatically into its suitability rating degree for the parental guide without any professional help by using its subtitle. For the suitability rating degree, The Motion Picture Association of America (MPAA) system has been used. The proposed model utilizes machine learning algorithms (ML) and natural language processing (NLP).

[*]Corresponding author, e-mail: muratisik@ahievran.edu.tr

DOI:10.29109/gujsc.1146352

2. RELATED WORKS

Since determining the rating certificate of a movie just by using its subtitles is a novel idea, there are some gaps in the literature. However, there are other works to determine violent scenes of a movie using subtitles or image processing techniques on the video itself. Additionally, there have been some other papers for searching, extracting, and analysing various types of patterns using ML algorithms, Deep Learning techniques and NLP on movie subtitles.

Shafaei et al. [28] developed a model to predict MPAA ratings based on movie dialogues. They have used tagged dialogues which show the level of Violence, Profanity, Nudity, Frightening and Alcohol as the training data. With a recurrent neural network-based architecture, they have achieved an F1-score of 81%. Khan et al. [29] proposed a violence detection scheme to eliminate violent scenes from movies to stop underaged people from watching them. They segmented the entire movie into shots and classify these shots into violence and non-violence classes using a lightweight DL model. Shafaei et al. [30] developed a multi-modal DL pipeline addressing the movie trailer age suitability rating problem. They attempted to combine video, audio, and speech information. Shafaei et al. [33] proposed RNN-based architecture to classify a movie into its corresponding suitability age class of MPAA by using movie scripts. They have achieved a weighted F1-score of 78%.

Vajjala and Meurers [4] have created a model to analyse linguistic features of subtitles to measure the “Readability” degree for spoken language from the perspective of understanding. Their main idea was to classify a TV show for ages as a matter of understanding. They classified the shows into three classes and achieved an accuracy of 95.9%. The paper proves that subtitles can be used to classify a movie as a matter of understanding. Boguszewski et al. [5] applied different approaches to detect offensive and hateful speech. They have achieved a F1-Score of 77%. The paper shows that the subtitles can be used to specify if a movie has hateful dialogue.

Hesham et al. [6] developed a model to create a trailer from a video using its subtitles in 2018. The model was tested with movies, and they achieved an accuracy rate of 89% in classifying the movies into their corresponding genres but when it comes to creating an automatic trailer, they only achieved an accuracy of 47% comparing the original movie trailer. Their research aims to create a Trailer using subtitles without any professional interfere. The study shows that trailer frames can be detected just by using subtitles. Li et al. [9] propose a framework for classifying videos. Their aim was to automatically annotate and organize the videos using their named entity. They used several ML algorithms and achieved an accuracy of 43.60% - 46.58%. Katsiouli et al. [2] proposed an innovative method for the unsupervised classification of video content by applying NLP techniques to their subtitles. They utilized the TextRank algorithm, W3D technique and WordNet domains. They achieved various accuracy scores based on movie kind/genre. These studies aim to classify the videos/movies into their corresponding genre by using subtitles or named entities.

Bougiatiotis and Giannakopoulos [7] developed a model to determine the similarity between the movies in 2016. In their work, they utilize NLP and a topic modelling algorithm, namely Latent Dirichlet Allocation (LDA). Their model calculates the existence of a similar correlation between movies. Scaiano et al. [8] attempted to segment movies into scenes using subtitles. Cosine and the WordNet similarity measure have been calculated to specify the segments based on TextTiling which is for text segmentation. Their research aims to calculate a similarity score between movies. These studies show that the subtitles contain enough information about the movie to infer similarities. But they didn’t work on Parental Guidelines.

Jenkins et. al. [10] studied violence in movies in 2005. They aimed to determine whether MPAA rating system distinguishes among the 3 primary rating categories (PG, PG-13, and R) with respect to violence. They worked on a sample of 100 films, a total of almost 2143 bodily violent actions from 1994. They concluded that the frequency of violence in films cannot be predicted by the rating system. The paper shows that determining only violent actions is not enough to predict the age rating. Martinez et. al. [34] develop a model to identify violence from the language used in movie scripts. Their approach was based on a broad range of features designed to capture lexical, semantic, sentiment and abusive language characteristics. SVM and RNN-based classification models were employed in their study, and they achieved a macro average F1 score of 60.4%.

Subtitles are textual versions of the dialogue in movies and provide condensed information about the contents of the movie [11]. Therefore, there are many works which use movie subtitles. Conventional strategies mostly focus on video classification [12], video segmentation [13], [14], parallel corpora [15], subtitles alignment [16], emotional classification [17] and so on. The novelty of this research is that the rating certificate of a movie for the parental guide can be simply defined just by using its subtitle. The comparison of the proposed model with conventional studies has been presented in Table 3.

Table 1. *Comparison of previous studies*

Study	***Aim***	***Difference***
Shafaei et al. [28]	*Predicting MPAA rating based on movie used tagged dialogues.* *(They achieved 81% weighted F1-score)*	*The model needs tagged data and only works with respect to some specific severity tag as Violence, Profanity, Nudity, Frightening and Alcohol.*
Khan et al. [29]	*Eliminating just violent scenes from movies for underaged people using a lightweight deep learning model on the segmented movie.* *(They achieved an accuracy of 96.3%)*	*They only evaluate violent scenes for children's suitability but there are maybe other scenes like drugging, nudity and so on.*
Shafaei et al. [30]	*Addressing the movie trailer age suitability rating problem using a multi-modal deep learning model on movie trailers.* *(They achieved 86.06% weighted F1-score)*	*They didn't use subtitles and their model only tested with trailers.*
Shafaei et al. [33]	*To predict the suitability of the movie content for children and young adults based on scripts.* *(They achieved weighted F1-score of 78%)*	*The scripts are the story written with planned dialogues associated with actions. The scripts may change during to film the movie. However, subtitles are the last form of dialogue.*
Martinez et al. [34]	*To predict if a movie is violent or not using scripts of movies.* *(They obtained a macro-averaged F1-score of 60.4%)*	*They didn't use raw data in their study and the classification result is not considerably high.*
Vajjala and Meurers [4]	*Determining a readability score for spoken language based on subtitles.* *(They achieved a classification accuracy of 95.9%)*	*The study is for evaluating subtitles in the name of just determining the reading capacity. They didn't interest in any suitable age or Rating Certificate.*
Boguszewski et al. [5]	*Detecting offensive and hateful speech.* *(They obtained a macro-averaged F1-score of 77%)*	*They didn't work on determining the age of suitability but only detecting offensive and hateful speech*
Hesham et al. [6]	*Classifying the movies into their corresponding genres and creating a trailer using subtitles.* *(They achieved an accuracy 89% for classification the movies into genres and an accuracy 47% for automatic trailer.)*	*The subtitles only used for making genre classification and creating trailers. They didn't work on the Parental Guidelines.*
Bougiatiotis and Giannakopoulos [7]	*Developing a model to determine the similarity between the movies using subtitles.*	*The subtitles were only used to calculate similarities between movies not to determine Parental Guidelines.*
Jenkins et. al. [10]	*Determining whether MPAA rating system distinguishes among the 3 primary rating categories (PG, PG-13, and R) with respect to violence.*	*They just work to determine in the name of violent scenes, not for the suitability age.*

3. METHOD

During the application phase of the study, a computer with Intel (R) Core (TM) i7-10750H CPU (2.6 GHz) processor feature, 32.00 GB of installed memory (RAM), and 64-bit Operating System, and 8GB Nvidia GPU system have been used. Also, each implementation in the study has been performed in python language with NLTK (Natural Language Toolkit) and spacy library. Figure 1 illustrates the proposed model architecture. The very first step is to collect the data and the second step is the pre-processing of the data which is one of the major steps. The last one is the training of the model. Each step will be detailed below.

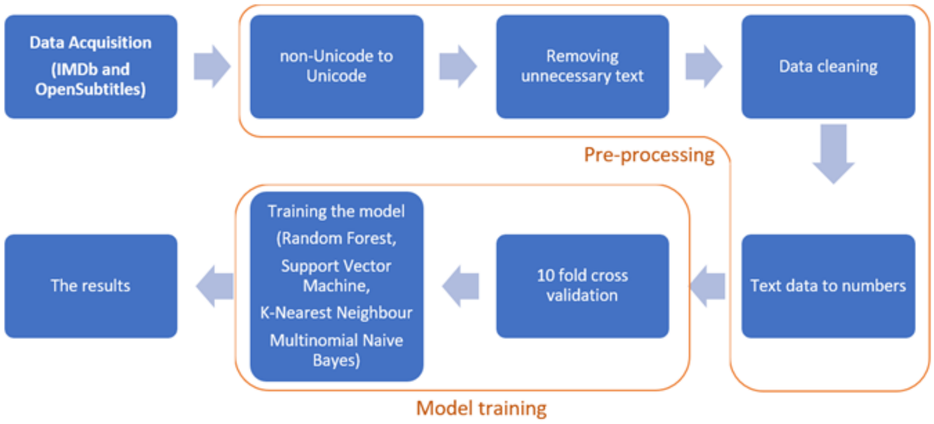

Data Acquisition (IMDb and OpenSubtitles)
non-Unicode to Unicode
Removing unnecessary text
Data cleaning
Pre-processing
Text data to numbers
10 fold cross validation
Training the model (Random Forest, Support Vector Machine, K-Nearest Neighbour Multinomial Naive Bayes)
The results
Model training

Figure 1. *Model architecture diagram*

3.1. Data Acquisition

Sometimes the determined rating may change over time as the age suitability rules are updated or if a movie is popular and it has high quality. Therefore, a couple of certain movie selection criteria have been used. IMDb movie quality measure also known as IMDb score was used to choose the movies. IMDb is one of the most popular platforms [18] that contains a great deal of information [19] about movies including the Parental Guidelines for many countries. IMDb has also reliable valid measures of movie quality [20] based on users' views. 279 movies have been used in compliance with the following criteria:

- To be 7.0 or more quality scores with at least 1000 users' views. The reason is that sometimes age suitability certificates may change after the movie is released in particular if the movie is popular and it has high quality.
- To be released between 2000 and 2022. The reason is that sometimes the determined rating may change over time as the age suitability rules are updated.

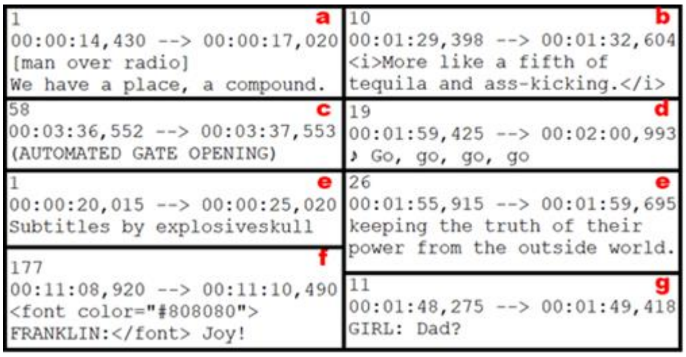
OpenSubtitles has been used to get the subtitles of the movies. OpenSubtitles provides a large collection of user-contributed subtitles in various languages for movies and TV programs [21]. Over 6 million subtitles in 65 different languages are available on the Web Page [22]. English subtitles of the chosen movies are downloaded from OpenSubtitles. It is noteworthy that sometimes a movie may have multiple subtitles for the same language, especially if the movie is relatively popular. In such cases, the best quality one has been preferred according to the subtitle quality rating of OpenSubtitles which is based on user reviews. USA rating class, UK rating class, names of the screenplay writer and IMDb score information were added to the subtitle files after they have been downloaded. The dataset has also been shared on the WEB for other research purposes [34].

Almost every country has its own movie rating system, but a movie may not have any rating certificate if it is not released in the country. In this study, the USA certificate rating system has been used to train and test the proposed model, since almost every movie has a rating certificate in this country. MPAA assigns age-based ratings for every film that is released in the United States [10]. Table 2 shows the Ratings Certificate System and the number of films ever made so far into their corresponding ratings based on IMDb. The table also lists the number of movies in compliance with the selection criteria and the distribution of the dataset. As it can be seen, the dataset is unbalanced due to the selection criteria. Besides, most of the movies are in R Class, then PG-13 class and PG Class. G Class are generally animations and usually, their scores are less than 7.0. There are not many NC-17 Class movies and almost none of them has a higher score of 7.0. Therefore, the number of G and NC-17 Classes are not enough to train the model. So, only "R", "PG-13" and "PG" Classes which are also known as primary rating categories [10] have been used.

Table 2. *USA Movie Ratings Certificate [23]*

Rating Certificate	*Suitable Ages (MPAA)*	*Movie Count*	*Movies match the criteria*	*Distribution of the Dataset*
G	*Nothing that would offend parents for viewing by children.*	*1538*	*30*	*0*
PG	*Parents are urged to give "parental guidance." May contain some materials parents might not like for their young children.*	*5130*	*242*	*51*
PG-13	*Parents are urged to be cautious. Some material may be inappropriate for pre-teenagers.*	*5401*	*557*	*103*
R	*Contains some adult material. Parents are urged to learn more about the film before taking their young children with them.*	*17318*	*980*	*125*
NC-17	*Adult. Children are not admitted.*	*77*	*7*	*0*
	TOTAL	***29464***	***1816***	***279***

3.2. Pre-Processing



```
1
00:00:14,430 --> 00:00:17,020
[man over radio]
We have a place, a compound.
```

```
10
00:01:29,398 --> 00:01:32,604
<i>More like a fifth of
tequila and ass-kicking.</i>
```

```
58
00:03:36,552 --> 00:03:37,553
(AUTOMATED GATE OPENING)
```

```
19
00:01:59,425 --> 00:02:00,993
♪ Go, go, go, go
```

```
1
00:00:20,015 --> 00:00:25,020
Subtitles by explosiveskull
```

```
26
00:01:55,915 --> 00:01:59,695
keeping the truth of their
power from the outside world.
```

```
177
00:11:08,920 --> 00:11:10,490
<font color="#808080">
FRANKLIN:</font> Joy!
```

```
11
00:01:48,275 --> 00:01:49,418
GIRL: Dad?
```

Figure 2. *Different types of subtitles*

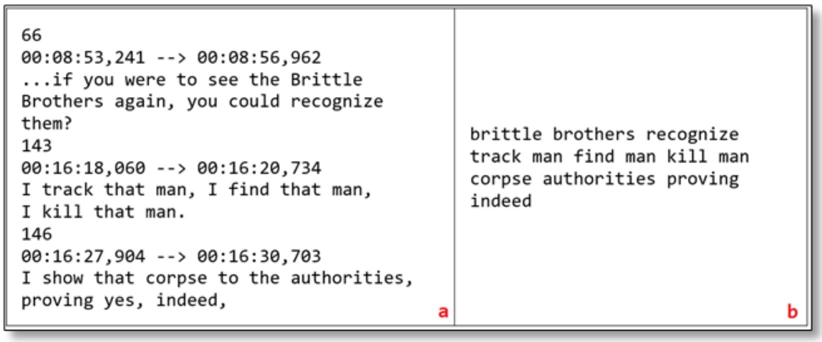
There are a variety of subtitle formats like SRT, SUB, SSA and so on. Each of them has its own layout. SRT format which is the most regular and reachable one has been used for the study. Figure 2 shows some different content types of SRT format. SRT files consist of three parts. The first one is a number that shows the sequence of subtitles. The second one is the interval duration for subtitles to be appeared on the screen and the last one is the subtitle itself.

Before training the model, the subtitle files must be undergone several pre-processing steps. The first step is to fix encoding problems. The subtitles have specific formats and encodings [24] based on the language. Especially, older files rely on language-specific encodings instead of Unicode. Since we are only working in the English language, all the files must be Unicode. Dealing with non-Unicode files is a difficult and error-prone process, also opening non-Unicode files and reading non-Unicode characters are also other tough problems to deal with. In this step, all non-Unicode characters are converted to Unicode characters.

The second step is to remove unnecessary texts from Subtitle files. Because some contents in the subtitles give us no useful information about the movie rating certificate, such as time duration, subtitle sequence number, subtitled song lyrics (see Figure 2d) and attribution to the maker (see Figure 2e) and so on.

The third step is text cleaning using the advantage of regular expressions. Data cleaning is a process that includes removing unwanted characters from text data. The aim of this process is to prepare the raw text for analysing in machine learning algorithms. If we subdivide this cleaning process into six steps: (1) removing sound expressions which written between brackets and parentheses (see Figure 2c and Figure 2a) (2) removing the tag of the current speaker's name which is written in the uppercase form (see Figure 2g) (3) removing HTML tags which written between '<' and '>' (see Figure 2f and Figure 2b) (4) removing non-alpha and non-ASCII characters (5) removing punctuation marks and emoticon character encodings. (6) removing numbers.

The last step in pre-processing part is to remove some words depending on their occurrence frequency to prevent overfitting and improve classification accuracy. Thus, some word has been ignored if the word count is less than 16 and more than 70% of the entire subtitle. A crucial point is that if a word is ignored or allowed mistakenly, some important terms may loose and thus may reduce the classification quality. Figure 3 shows an example of pre-processed data.



```
66
00:08:53,241 --> 00:08:56,962
...if you were to see the Brittle
Brothers again, you could recognize
them?
143
00:16:18,060 --> 00:16:20,734
I track that man, I find that man,
I kill that man.
146
00:16:27,904 --> 00:16:30,703
I show that corpse to the authorities,
proving yes, indeed,
```

```
brittle brothers recognize
track man find man kill man
corpse authorities proving
indeed
```

Figure 2. *A part of subtitle before (a) and after (b) pre-processing procedure*

Machine learning algorithms only work on numbers, not raw text data. In the literature, there are many different models for representing text documents with numbers. The selection of the model may significantly affect the accuracy rate. Among these models, TF-IDF vectorizer is one of the most famous algorithms used in text mining research [25] [26]. TFIDF vectorizer, WIDF vectorizer and Glasgow weighting scheme (GSW) have been used for classification. The dataset is limited due to the movie selection criteria therefore 10-cross validation has been used during the training to prevent overfitting.

3.2.1. Term Frequency Inverse Document Frequency (TF-IDF)

TF-IDF is a very common algorithm to transform the raw text into a meaningful representation of numbers which can be used for ML algorithms for prediction. The core idea of TF-IDF is not only to measure the word frequency but to measure the weightage of the word and this way calculate the importance of the word in the overall context. As it can be understood from Equation 1, TF-IDF concerns term occurrence across a collection of texts but a drawback of TF-IDF is that all the texts that contain a certain term are treated equally. So, TF-IDF does not distinguish between one occurrence of a term in a text and many [31]. In this paper, we call this problem as "Proportional Importance".

$$TFIDF(d,t) = TF(d,t).\log\left(\frac{N}{df(t)}\right) \qquad 1$$

Where N is the total number of files in the corpus. Then, df(t) is the number of documents containing the word t. TF(d,t) is obtained by dividing the number of occurrences of the word t in file d to sum of the occurrences of all words in the file d.

3.2.2. Weighted Inverse Document Frequency (WIDF)

WIDF's main purpose is to take into account the occurrence counts of a term. As it can be seen from Equation 2, WIDF overcomes the Proportional Importance problem by weighting terms that sum up to one over the collection of texts [31]. In this way, the greater number of appearance of words in the document will provide greater value relevance. Our purpose is to see if the occurrence count has any effect on this study.

$$WIDF(d,t) = \frac{TF(d,t)}{\sum_{i \in N} TF(i,t)} \qquad 2$$

Where d is a document collection, t is word or term, i is a related document. Then, TF (d, t) is the appearance of a word (t) in a document divided by TF (i, t), which is the total number of words (t) in the related document (i).

3.2.3. Glasgow Weighting Scheme (GSW)

As it can be seen from the Equation 3, GSW is similar to TF-IDF vectorizer. However, the difference lies in the frequency part of the equations, especially in the normalization [32].

$$GSW(d,t) = \frac{\log(TF(d,t)+1)}{\log(length\ of\ d)} \cdot \left(\log\left(\frac{N}{df(t)}\right)+1\right) \qquad 3$$

Where N is the total number of files in the corpus. Then, df(t) is the number of documents containing the word t. TF(d,t) is obtained by dividing the number of occurrences of the word t in file d to sum of the occurrences of all words in the file d.

Since the stop words like ‘the’, ‘you’, ‘can’ and so on will not have any good effect on the model, the standard spacy library for English has been used to remove the stop words with all vectorizers. As a classifier, RF, SVM, KNN and MNB have been chosen. They are well-known algorithms in the literature, and they have been widely used [27]. The hyperparameters used with ML algorithms will be detailed. Accuracy, F1-Score, precision, and recall have been used to compare and analyse the results.

3.3. Model Training

SVM is an algorithm that is based on statistical learning theory by finding a hyperplane in N-dimensional space that distinctly classifies the data points. With the SVM model, sigmoid type, 30-degree polynomial kernels have been used for training. RF is an ensemble of many decision trees. With the RF, 1000 trees have been used for training and the entropy function has been used as a quality measure. The core idea of KNN is that similar points are near to each other. So simply we can say that this algorithm performs classification based on the distance measure of the samples in the data points. With the KNN, 15 neighbours and Euclidean as a distance metric have been used. For closer neighbours to have a greater influence, the distance function was used as a weight function. MNB is a probability-based classification algorithm based on Bayesian theorem. With the MNB, smoothing parameter has been set to 0.9, 1.0 and 1.1 separately.

4.RESULTS

Table 3 shows the Classification Performance. Since 10-cross validation on the test/train dataset has been used, all the results are the macro average of 10-cross samples. An accuracy of 85% has been achieved with RF classifier using GSW as a feature extractor. A similar accuracy of 83% has also been obtained with SVM algorithm with TF-IDF. Although the model is trained with various types of hyperparameters with KNN, the accuracy never gets better than 45% with GSW and WIDF vectorizers. 85% accuracy shows how close the predicted rating classes are to their true classes, while precision of 87% shows how close the predicted classes to each other.

Figure 4 visualizes and summarizes the performance of RF algorithm which has the highest accuracy with GSW as a future extractor. Since 10-Cross validation has been used during training process, the confusion matrix is concatenated matrix which is generated by all confusion matrices from 10 samples. 2% of R Class movies, 1,65% of PG-13 Class movies and 5,6% pf PG Class movies has been misclassified.

Table 3. *Classification Performance*

Vectorizer	***Classifier***	***Macro average Accuracy***	***Macro average Precision***	***Macro average Recall***	***Macro average F1-score***
TF-IDF	*SVM*	*0.82*	*0.86*	*0.83*	*0.83*
	RF	*0.83*	*0.88*	*0.83*	*0.84*
	KNN	*0.74*	*0.79*	*0.72*	*0.73*
	NB	*0.77*	*0.83*	*0.75*	*0.76*
WIDF	*SVM*	*0.76*	*0.80*	*0.76*	*0.77*
	RF	*0.84*	*0.87*	*0.85*	*0.85*
	KNN	*0.45*	*0.25*	*0.35*	*0.23*
	NB	*0.75*	*0.77*	*0.78*	*0.76*
GSW	*SVM*	*0.74*	*0.81*	*0.73*	*0.74*
	RF	*0.85*	*0.87*	*0.85*	*0.85*
	KNN	*0.45*	*0.21*	*0.35*	*0.24*
	NB	*0.70*	*0.69*	*0.74*	*0.69*

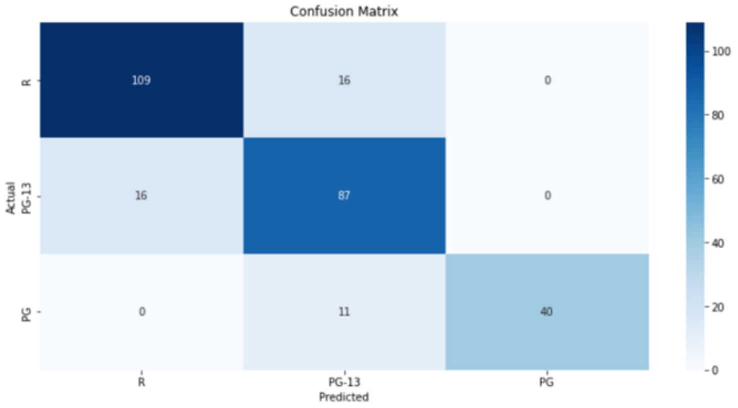


Figure 3. *Confusion matrix of RF algorithm with GSW*

5. DISCUSSION

In this study, four different classifiers namely, SVM, RF, KNN, MNB have been used with three different text feature extractors TF-IDF, WIDF, GSW. Each classifier has been tried with each feature extractor, hence there are 12 different results. Choosing the films in accordance with the film selection criteria has caused to limit the amount of data. The limited training data can cause undesired overfitting. A cross-validation procedure has been utilized to prevent this issue during the training process.

G Class are generally animations and there are not many NC-17 Class movies. Also, almost none of them is suitable for the movie selection criteria. So, the proposed model does not work on NC-17 Class and G Class movies. It only deals with three classes; R Class, PG 13 Class, and PG Class which are known as primary rating categories [10]. As a matter of fact, most of the movies are already in these rating classes.

The proposed model with RF algorithm and GSW has reached to the highest accuracy of 85%. Although some algorithms other than RF with GSW in this study appear to give similar high accuracy, their recall and F1-score are lower. The results validate that RF algorithm with GSW can be used on the subtitles to classify movies into their Rating Certificate of MPAA. In other words, the proposed model can accurately predict a movie to its corresponding suitability class with 85% probability.

The comparison of the proposed model with similar studies has been presented in Table 3. Since none of these studies in Table 3 has classified movies into their corresponding age certificate based on subtitles, it is therefore not possible to make quantitative comparisons with the proposed strategy. As has been discussed, determining the rating certificate of a movie through its subtitles is a novel strategy in this paper.

In movies, words that appear too frequently or less may affect the accuracy of the classification results. For example, a specific person, place, or object name can be decisive for age suitability during the training. Thus, if there has been a word used less or more than a certain number through the entire subtitle, that particular word has been ignored in the proposed method. The reason is that there is a trade-off between the ignored word count and the accuracy of the results. In this paper, the optimum accuracy of the results has been achieved when the number of word counts is less than 16 and more than 70% of the entire word count. The accuracy in this case study has reduced down to 56% when the limits are not applied while the optimum accuracy has been achieved as 85%.

A drawback of the proposed model is that there could be some silent scenes where inappropriate content may appear which affect the rating directly but cannot be precisely evaluated with the model. Since the proposed model only works based on subtitles. To determine the suitability age of a movie in the conventional approach, every single part of a movie including silent scenes should be analysed in considering whether it contains any Sex, Nudity, Violence, Gore, Profanity, Alcohol, Drugs, Smoking, Frightening, Intense and so on. The proposed model has achieved an accuracy of 85% only with subtitles. The main reason for such high accuracy is because there is a typical conversation regarding any inconvenient scenes either before or after that particular scene including silent ones. By way of an example, if there is a drug usage on a silent scene, there would be a conversation either before or after the usage of the drug. In almost every inappropriate scene, this approach can be used. However, 16 R Class movies have been misclassified by the proposed model. The underlying reason is that the movies contain some sex and nudity without any related conversation. The mentioned drawback can be pruned in the next version of this paper by taking advantage of image processing and deep learning techniques on the frames of movies. The success of classification results from R to PG-13 and from PG-13 to PG gets lower gradually. This may be due to the distribution ratio in the dataset, which gets lower from R to PG-13 and from PG-13 to PG.

The golden method for rating a movie is watching the whole movie with a professional help and defining an age limit. The proposed model uses the weightage of words giving the importance of that particular word in the overall context. By utilizing those bags of words, the model has determined the rating of the movie with an accuracy of 85%. Therefore, the professional help is not always necessarily needed owing to the acceptable accuracy of the proposed strategy.

4.CONCLUSIONS

The movie rating system is the way for defining suitability age. Developed countries have their own strategies to determine suitability age. The main drawback of these strategies is the requirement of watching the entire movie with a professional. In this paper, a new model has been proposed to determine the Parental Guidelines of a movie without the need for a professional help. The result of the proposed classification approach is promising and can be used by the rating committee for pre-evaluation.

Future work may combine the proposed model with a new model which aim to silent scenes by taking advantage of image processing and deep learning techniques. In this way, it can be achieved to further increase the accuracy of the results.

REFERENCES

[1] Park SB, Kim HN, Kim H, Jo GS "Exploiting script-subtitles alignment to scene boundary dectection in movie". 2010 IEEE International Symposium on Multimedia, Taichung, Taiwan, 13-15 December 2010.

[2] Katsiouli P, Tsetsos V, Hadjiefthymiades S. "Semantic Video Classification Based on Subtitles and Domain Terminologies". KAMC 2007 Workshop on Knowledge Acquisition from Multimedia Content, Genoa, Italy, 5 December 2007.

[3] Lison P, Meena R. "Automatic turn segmentation for movie & tv subtitles". 2016 IEEE Spoken Language Technology Workshop (SLT), San Juan, Porto Riko, 13-16 December 2016.

[4] Vajjala S, Meurers D. "Exploring measures of 'readability' for spoken language: Analyzing linguistic features of subtitles to identify age-specific tv programs", 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR), Gothenburg, Sweden, 27 April 2014.

[5] von Boguszewski N, Moin S, Bhowmick A, Yimam SM, Biemann C. "How Hateful are Movies? A Study and Prediction on Movie Subtitles". arXiv preprint, 2108.10724(1), 2021.

[6] Hesham M, Hani B, Fouad N, Amer E. "Smart trailer: Automatic generation of movie trailer using only subtitles", IEEE 2018 First International Workshop on Deep and Representation Learning (IWDRL), Cairo, Egypt, 29-29 March 2018.

[7] Bougiatiotis K, Giannakopoulos T. "Content representation and similarity of movies based on topic extraction from subtitles", 9th Hellenic Conference on Artificial Intelligence, Thessaloniki, Greece, 18-20 May 2016.

[8] Scaiano M, Inkpen D, Laganiere R, Reinhartz A. "Automatic text segmentation for movie subtitles", 23rd Canadian Conference on Artificial Intelligence, Ottawa, Canada, 31 May - 2 June 2010.

[9] Li Y, Rizzo G, Redondo García JL, Troncy R, Wald M, Wills G. "Enriching media fragments with named entities for video classification", 22nd International Conference on World Wide Web (WWW13), Rio de Janeiro Brazil, 13 – 17 May 2013.

[10] Jenkins L, Webb T, Browne N, Afifi AA, Kraus J. "An evaluation of the motion picture association of america’s treatment of violence in pg-, pg-13–, and r-rated films", American Academy of Pediatrics, 115(5), 512-517, 2005.

[11] Park SB, Oh KJ, Kim HN, Jo GS. "Automatic subtitles localization through speaker identification in multimedia system". 2008 IEEE International Workshop on Semantic Computing and Applications, Incheon, South Korea, 10-11 July 2008.

[12] Agarwal R. "Video Classification into Academic and Entertainment using Subtitles", Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(11), 5633-5639, 2021.

[13] Lee AS, Oh H, Seo M. "ViSeRet: A simple yet effective approach to moment retrieval via fine-grained video segmentation", arXiv preprint, 2110.05146(2), 2021.

[14] Abdulhussain SH, Al-Haddad SAR, Saripan MI, Mahmmod BM, Hussien A. "Fast temporal video segmentation based on krawtchouk-tchebichef moments". Institute Electrical And Electronics Engineers, 8, 72347-72359, 2020.

[15] Lison P, Doğruöz AS. "Detecting machine-translated subtitles in large parallel corpora", 11th Workshop on Building and Using Comparable Corpora (BUCC 2018), Miyzaki, Japan, 8 May 2018.

[16] Saz O, Deena S, Doulaty M, Hasan M, Khaliq B, Milner R, Ng RWM, Olcoz J, Hain, T. "Lightly supervised alignment of subtitles on multi-genre broadcasts". Multimedia Tools and Applications, 77(23), 30533-30550, 2018.

[17] Topal K, Ozsoyoglu G. "Emotional classification and visualization of movies based on their IMDb reviews", Information Discovery and Delivery, 45(3), 149-158, 2017.

[18] Kumar HM, Harish BS, Darshan HK. "Sentiment Analysis on IMDb Movie Reviews Using Hybrid Feature Extraction Method", International Journal of Interactive Multimedia & Artificial Intelligence, 5(5), 109-114, 2019.

[19] Dhir R, Raj A. "Movie success prediction using machine learning algorithms and their comparison", 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC), jalandhar, india, 15-17 December 2018.

[20] Baugher D, Ramos C. "The Cross-Platform Consistency of Online User Movie Ratings", Atlantic Marketing Journal, 5(3), 9, 2017.

[21] Tiedemann J. "Finding alternative translations in a large corpus of movie subtitle", 10th International Conference on Language Resources and Evaluation (LREC'16), Portorož, Slovenia, 23-28 May 2016.

[22] OpenSubtitles.org, "Subtitles", https://www.opensubtitles.org, (31.03.2022).

[23] Motion Picture Association Inc, "The Voluntary Movie Rating System: How the Ratings Are Decided", https://www.motionpictures.org/film-ratings, (31.03.2022).

[24] Mangeot M, Giguet E. "Multilingual aligned corpora from movie subtitles", Information and Knowledge Processing Laboratory (LISTIC), 1, 6-14, 2005.

[25] Dadgar SMH, Araghi MS, Farahani MM. "A novel text mining approach based on TF-IDF and Support Vector Machine for news classification", 2016 IEEE International Conference on Engineering and Technology (ICETECH), tamil nadu india, 17-18 March 2016.

[26] Durahim AO, Setirek AC, Özel BB, Kebapci H. "Music emotion classification for Turkish songs using lyrics", Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi, 24(2), 292-301, 2018.

[27] Brigadoi I, Genre classification using syntactic features. MSc Thesis, Uppsala University, Uppsala, sweden, 2021.

[28] Shafaei M, Samghabadi NS, Kar S, Solorio T, "Age suitability rating: Predicting the MPAA rating based on movie dialogues", In Proceedings of The 12th Language Resources and Evaluation Conference, Marseille, France, 13-15 May 2020.

[29] Khan SU, Haq IU, Rho S, Baik SW, Lee MY, "Cover the violence: A novel Deep-Learning-Based approach towards violence-detection in movies", Applied Sciences, 9(22), 4963, 2019.

[30] Shafaei M, Smailis C, Kakadiaris I, Solorio T, "A Case Study of Deep Learning-Based Multi-Modal Methods for Labeling the Presence of Questionable Content in Movie Trailers", International Conference on Recent Advances in Natural Language Processing (RANLP 2021), Online, 1-3 September 2021.

[31] Tokunaga T, Makoto I, "Text categorization based on weighted inverse document frequency", In Special Interest Groups and Information Process Society of Japan (SIG-IPSJ), 1994.

[32] Sabbah T, Selamat A, Selamat MH, Al-Anzi FS, Viedma EH, Krejcar O, Fujita H, "Modified frequency-based term weighting schemes for text classification", Applied Soft Computing, 58, 193-206, 2017.

[33] Shafaei M, Samghabadi NS, Kar S, Solorio T, "Rating for parents: Predicting children suitability rating for movies based on language of the movies", arXiv preprint arXiv:1908.07819, 2019.

[34] Martinez VR, Somandepalli K, Singla K, Ramakrishna A, Uhls YT, Narayanan S, "Violence rating prediction from movie scripts", In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, No. 01, pp. 671-678), July, 2019.

[35] www.kaggle.com/dataset/e6440f4fb6d17b55e56ee8baffb55d9dc7931560b4b710608db33ab5c29296c7, E.T.: 16.07.2022