Abstract

Movies have a lot of economic benefits, they also help to relieve stress, and are a great source of entertainment, not excluding their educational benefits. Although the effects of movies are beneficial, it is rather ingenuous to rule out certain negative consequences such as violence, adolescent sexual exposure, foul language, and alcoholism. Certain features, when implemented, can help to mitigate this. In this research, Movie scripts were collected and cleaned to classify movies into genres and age-appropriate ratings. Exploring Machine Learning algorithms such as XGBoost, K-Nearest Neighbors, and Support Vector Machines in comparison to word2vec, TFIDF bag-of-words, and Gensim's Doc2vec. For genre rating, a Support vector (SVM) Doc2vec model achieved an f1 score of (0.49) and the age-appropriate rating achieved an f1 score of (0.84) over SVM TFIDF bag-of-words

Introduction

It is very difficult to ignore the exhilaration and cinematic experience that motion pictures bring to our society; Its economic impact, and cultural cohesion, not to ostracise its educational benefits. The relevance is that these temporal and spatial dimensions of life play a role in influencing how people interact with one another, (Vagionis & Loumioti, 2011), such as what they share and the type of community they form or pull down. Furthermore, Ed Vaizey former minister for Culture, Communications, and creative industries in the UK reiterated this when he stated clearly that "British movies contributed to the economy as well as promoting cultural life, identity and diversity on the international stage" (Department for Digital, Culture, Media & Sport and Ed Vaizey. (2010).,). In 2017, the film industry generated a trade surplus for the UK of over £1.45 billion up from £926 million in 2016, it generated a turnover of £14.8 million and contributed £6 million to the GDP (British Film Industry, 2019). In addition, BFI statistics for 2019 show the film industry alone generated a 30% uplift for the UK economy (BFI, 2022). As advantageous as it seems, recent statistics on children's screen time show an alarming increase in the amount of time they spend watching movies or videos on electronic devices (Chen & Adler, 2019). While much of the information in these films and videos appear to be harmless, there is also harmful and inappropriate stuff that may have adverse effects on their behaviour. Watching some programs, for example, may encourage irresponsible sexual conduct and alcohol use in teenagers (Parkes et al., 2013; Hartley et al., 2014) or instill dread and fear in children. (Sultana et al., 2021). Some commercial promotions in movies endorse smoking and drinking by actors and portray it as a desirable behaviour amongst teens (Dalton et al., 2003). It is believed that movie images such as cartoons can also impair children's mental development (Habib & Soliman, 2015), being able to identify a movie's genres and encouraging age-appropriate content can drastically reduce these undesirable effects

In 1968, the Motion Pictures Association of America (MPAA) established a system of movie ratings for parents to determine the appropriateness of the content of movies for children and teenagers (Premiere Theatres, 2022). They examine each movie in terms of theme, language, nudity, sex, drug use, and violence. However, this seems like a very time-consuming and biased process, because there exists a difference in the cognitive development of children and teenagers (Judith Rosenthal, Henry Massie & Kenneth Wulff, 1980). This also explains why numerous movies on the internet have no ratings or incorrect ratings (Ihla, 2018; Banks,). To overcome this challenge, it is primary to investigate predicting the MPAA rating at any point, of which text classification and Natural Language Processing are effective (Luo, 2021). Furthermore, As a result, producers can utilise the expected rating to tailor screenplays to the target demographic.

A genre is very beneficial, it categorizes movies based on similarities in its elements and it is a movie's sales pitch, this explains why films that are difficult to categorize into a genre are often less successful (Park & Berger, 2010). After the covid-19 lockdown, the number of movie tickets sold in the U.S and Canada increased by more than 120% between 2020 and 2021, when it surpassed 498 million dollars, with Adventure being the highest-grossing movie genre. Titles with adventure and action elements are also on top on a global scale (*U.S. & canada: Film genres by total box office revenue 2022.*). Succinctly, A suitable model should be developed to help users identify movies by their genres to improve their experience, this will create modifications to aid analysts' decisions on movie production as well as understanding patterns for future decision making

The following Sub-sections can be found below.
Literature Review, Data Collection, Data Cleaning and Pre-processing, Data Transformation, Algorithms, Precision, Recall, F1 Metrics, and Classification Reports, Discussion, Conclusion, and Recommendation.

Literature Review.

The process of using text classification to predict movie genres and rating systems isn't novel research. Several researchers have tried to classify movie text using different machine learning algorithms. Here are some examples of related projects.

Blackstock & Spitz (2008) concentrated on classifying movie scripts into genres solely based on the script. They investigated the Naive Bayes Classifier as well as the Maximum Entropy Markov Model Classifier (MEMM). The inconsistency of the script format, data size, and the fact that movies had several genres posed the most difficult challenges. The MEMM used logistic regression to learn weights for the selected features. Although the MEMM outperformed a generative model despite not being conditioned on past decisions, its F1 scores (0.5196) and (0.4805) Naïve Bayes demonstrate that the features were generalizable across algorithms. Another great example was, Ho, (2011) who desired to speed up movie classification from synopsis' by developing a model comparable to an experienced human. He experimented with bag-of-words representation, utilizing the one vs all SVM technique, K-Nearest Neighbour (KNN), Parametric mixture model, and error backpropagation neural network, assuming that movies of the same genre should have comparable plots. Support Vector Machine (SVM) fared the best across all models, with a 55% accuracy, only after balancing the classes in the dataset. Additionally, Hoang, (2018) investigates ways of classifying films based on plot summaries. The strategies investigated were Naïve Bayes with Bag of Word features and pre-trained word2vec embeddings of plot summaries on an XG-Boost classifier and a GRU network. He found that word2vec and machine learning classifier XG-Boost were inadequate representations but the GRU network produced outstanding results, with a Jacquard index of 50%, an F-score of 0.56, and an 80.5% hit rate. Shafaei et al., (2020) sought to automate the prediction of movie suitability for children using the MPAA rating system to set a strong baseline for future unpredicted movies to help save cost. Using an LSTM model architecture with the attention that jointly models the genre and emotions in the script to predict an MPAA rating, they targeted movies based on conversational data, emotional dynamics between characters, movie genre, and similarity between multiple movies, the classification model achieved an 81% weighted F1 score. They determined that bad words could affect the MPAA rating, but that a threshold was insufficient to forecast rating with fair accuracy because word context varies. Additionally, Arikatla & Chinnapottu, (2021) whose objective was to develop a model that predicts a movie's title using fragments of random text that are comparable in meaning to the text from the screenplay corpus, utilizing classification techniques; Random Forest, Logistic Regression, Naïve Bayes, and Support Vector Machine. In their prediction, Naïve Bayes and Support Vector Machine achieved the maximum accuracy of 59.63%. it was said that due to the minute variances in the probability of the text, the random forest did not perform so well in movie text classification.

These are developments in the field of movie success prediction, movie recommendation systems, and movie genre prediction. This thesis, this research aims at developing a model for the

prediction of content rating for a movie as well as its genre type from the scripts. Notwithstanding, the above research papers, it has been identified that Support Vector Machine (SVM), Naïve Bayes, XGBoost, KNN, and Logistic Regression are classifiers that can be used for training on movie scripts.

Data Collection.

The MPAA rating of screenplays and genres may be challenging as there's a total reliance on a machine with no human input. There were several salient decisions made for this project to improve its accuracy. This project intends to build the dataset from scratch so the scripts have to be collected.

The movie scripts and genres were extracted from the IMSDB website using request and beautiful soup python libraries. Additionally, the age rating was extracted against the title from IMDB. Since I was primarily interested in the age rating from the IMDB website, the design decision was to scrape only the content rating even though there exist multiple attributes per movie on the website. I implored the try and except functionality in python to refrain from error. This was due to irregularity on both websites which later generated missing entries in our dataset.

Initially, after extraction, we had a total of 1213 movies and 1077 scripts, as some movies were mentioned on the website, but their scripts and age-rating were not provided. I set out to remove these empty scripts and ages, further reducing it to a total of 1046 rows and 4 columns.

In summary, data includes the title, age rating, genre, and script. The movies were mostly in English, with script length ranging from 1 to 25080.

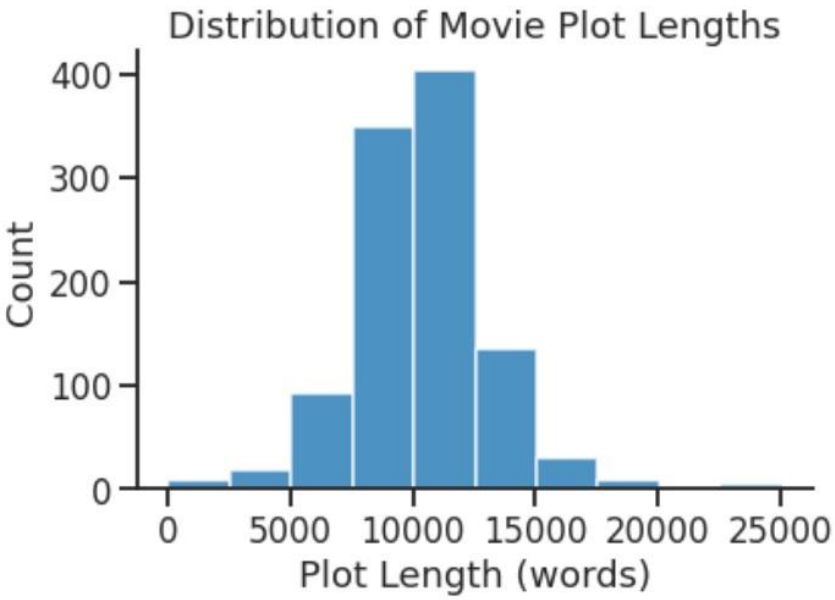


Figure 1 Representation of Plot Lengths

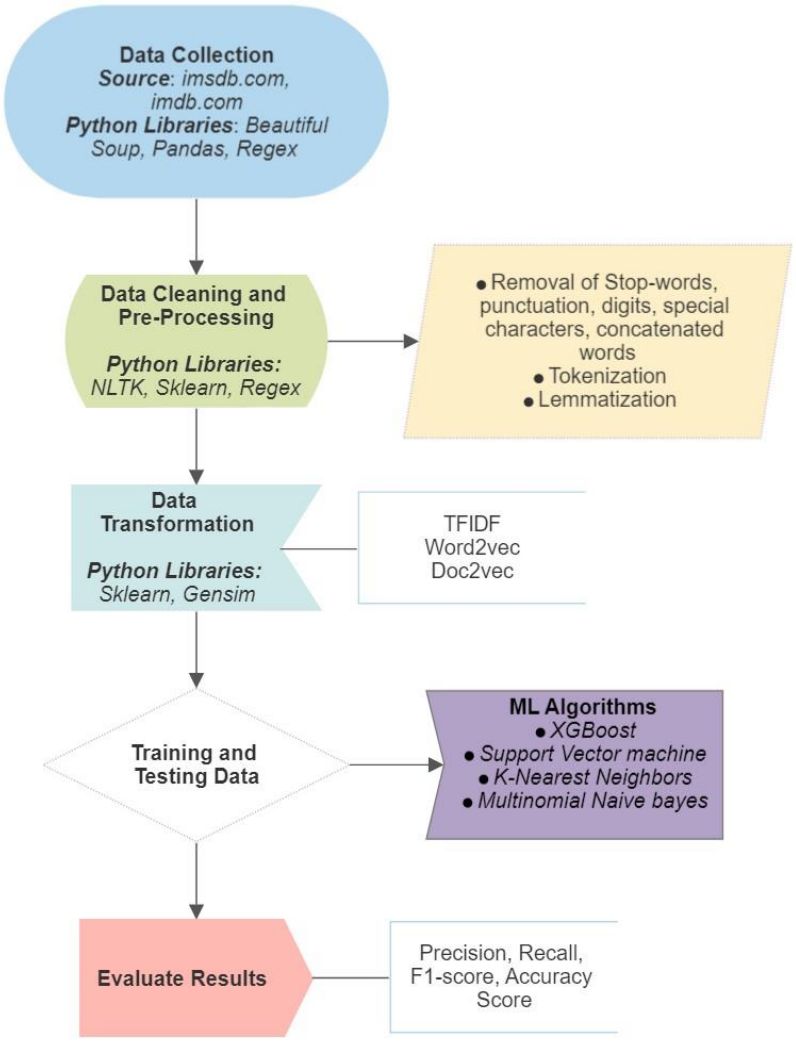


Figure 2 Flow chart for the Automation of Movie Script.

Data Cleaning and Pre-processing.

To train a model successfully, the script data must be pre-processed because it contains a lot of undesired data such as stop words, punctuations, and contracted words such as (can't, won't, didn't) that are difficult for computers to interpret. NLP methods using NLTK, and regular expression libraries were employed to do this.

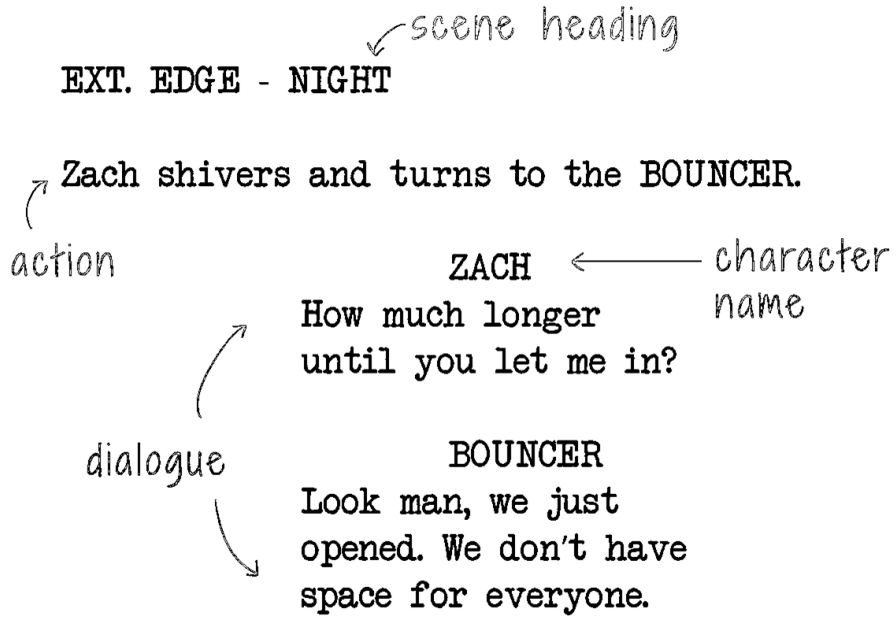


Figure 3 Script Sample

A script is divided into parts, scene lines, action lines, dialogue lines, extension lines, and transition lines. The conversations in this project were in lower case, and every other portion was in upper case, with names as well as seen above. To limit the recurrence of names, a function was created that captures all uppercase characters and returns those words that appeared in less than 10 counts. Any word occurring above 10 counts was treated as a stop word. All stop words were further eliminated from the dataset using NLTK library. The Regex library, was used to eliminate punctuations, digits, and special characters and concatenated words based on patterns noticed in the dataset. Upon removal, data returned were in lower case. After cleaning, the data was then tokenized, this is a process of splitting them into single tokens in a dataset (Webster & Kit, January 1, 1992). After that, the tokenized data is lemmatized. Lemmatization is the process of reducing words to their simplest form, accurately recognising the components of speech and their meaning in a phrase (Khyani & B S, 2021).

Data Transformation.

1. **Term Frequency Inverse Document Frequency (TFIDF Bag-of-words):** is a numerical measure that ranks the importance of a text in a document based on its frequency. Scripts were transformed using scikit's learns TFIDF vectorizer, to decrease dimensionality in the vector representation, a min_df, and max_df score was set to ignore terms that appear several times and ones that appear less frequently. After transforming the vectors, each script included a vector representation (1 x 37673).

2. **Word2vec (W2V):** This is an embedding technique in natural language processing that use a shallow neural network model to learn word connections from a huge corpus of text. The w2v plot matrices and vectors made were created using the google news W2V model. It is trained on more than a 100-billion-word news corpus. When training the model, the assumption was that the semantic quality of each word might indicate the direction of one or more genres/ages. It contained words after training for each plot as a 300-dimensional vector.

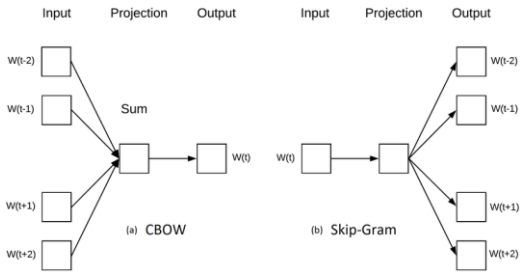


Figure 4 Word2vec representation

3. **Doc2vec (D2V):** This is an unsupervised approach that generates vectors for phrases, paragraphs, and documents in a manner like the word2vec technique. The vectors produced by doc2vec can be used to compare the similarity of sentences, paragraphs, or documents. doc2vec is a python package by Gensim. Each script was treated as though it were a single word. A vector size of 20 fits our genres for the doc2vec model.

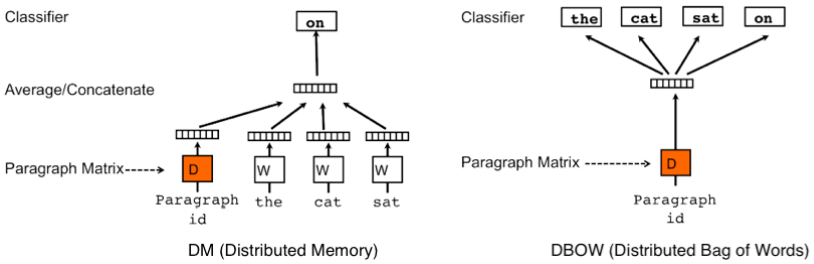


Figure 5 Doc2vec representation

To be compatible with the classifiers, the MinMaxScaler was used to scale the vectors (W2V and D2V) to a list from an array of arrays.

Algorithms

1. **XGboost** is a gradient boosting technique. It is an ensemble of decision trees added one at a time to fit and correct prediction errors made from prior models (Aziz & Dimililer, 2021). Our vectors (TFIDF, W2V) were applied to the model. The disadvantage of XGboost is that it is very prone to overfit the training data.

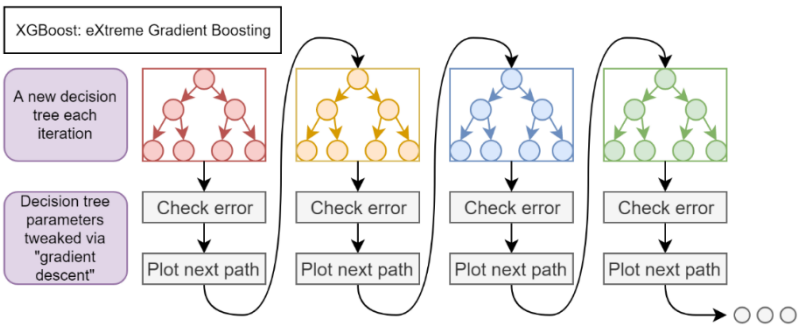


Figure 6 XGBoost

2. **K-Nearest Neighbors (KNN)** is used to classify unlabelled observations by assigning them to the class of the most similar labeled data. It works off the assumption that similar points can be found near one another. In the scripts, because of the ambiguity, KNN is used for classification because of its predictability and interpretability.

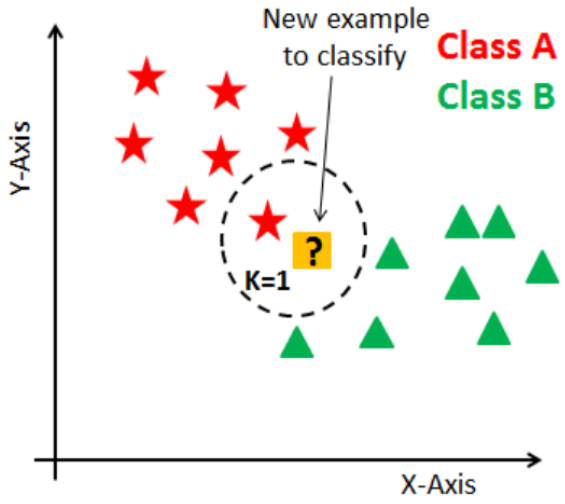


Figure 7 K-Nearest-Neighbors

3. **Support Vector Machine (SVM)** is a type of algorithm that separates classes using a decision boundary. This boundary is set in such a way that it differentiates features on a plane, there are simply the coordinates of individual observations on the hyperplane (Wright et al., 2013). In our analysis, the expectation was that SVM will successfully be able to analyse the data by maximising margins between the vector points (TFIDF, W2V, D2V), and the right boundary will predict the classes (age/genres)

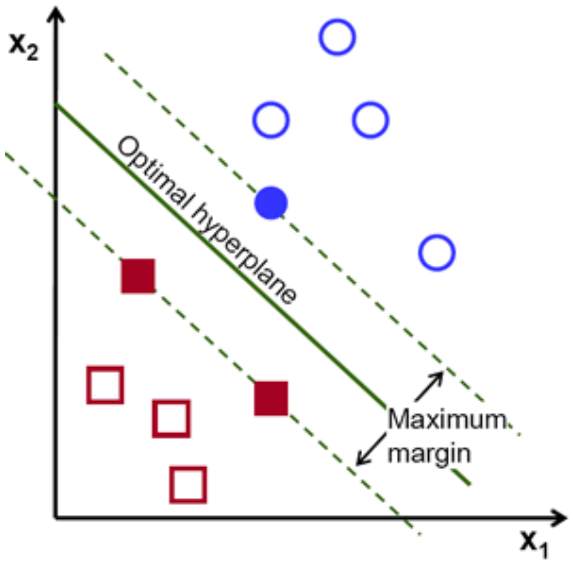


Figure 8 Support Vector Machine

4. **Naïve Bayes** is a classifier that assumes all predictors are independent that is the presence of one feature doesn't affect the presence of another. The multinomial or multimodal naïve bayes is a version of naïve bayes designed to handle word counts as its underlying method of calculating probability.

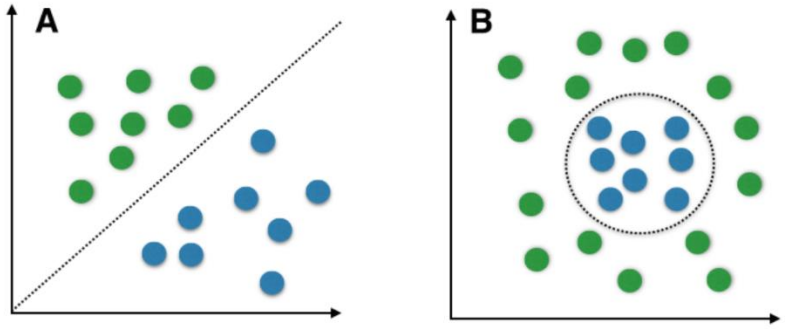


Figure 9 Naive Bayes

The models used for each predicting feature were Naive-Bayes with a cross-validated smoothing parameter (alpha), support vector machines (SVM) with cross-validated penalty parameter of the error term (C), KNN with an effective metric 'minkowski' and Xgboost with a learning rate of 0.02. Each of these models were implemented using scikit-learn's out-of-the-box model classes.

Function-based Modeling Pipeline

An "evaluate model" function was created, this divides the data into an 80% training set and a 20% test set. A random seed is established to make sure that our models are learning and making predictions on the same training and test sets. A model object, predictor set, response set, cross-validation argument, and model-specific cross-validation tuning parameters are the function's inputs.

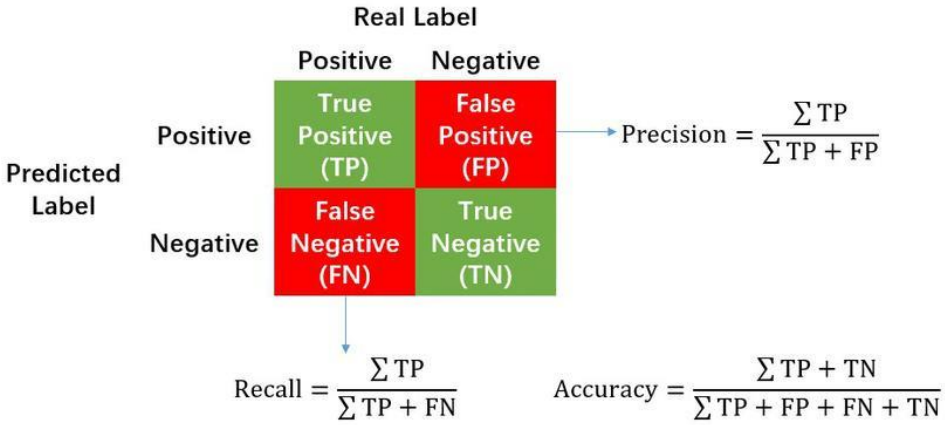
	Model	Parameters
Naïve Bayes	Multinomial Naïve Bayes()	‘alpha’: [0.01, 0.1,1.0]
KNN	KNeighborsClassifier (metric = ‘minkowski’)	‘n_neighbors’ : [4,5,7,9]
SVM	SVC (class_weight = ‘balanced’, kernel = ‘linear’)	‘C’: [0.01, 0.1, 1.0]
XGBoost	XGBClassifier (learning_rate = 0.02, n_estimators = 100)	‘max_depth’: [2,4,6,8], ‘gamma’ : [0.5,1,1.5,2]

In addition to a complete classification report for each genre in the dataset, the model produces a dictionary that includes the average precision and recall for each class. Modifications to the w2v and d2v arrays had to be made before been fed to the modelling function. Their values had to be normalised between 0 and 1 and transformed from an array of arrays to an array of lists in order to be compliant with the response variable's format. For scaling, MinMaxScaler from scikit-learn was implored. This outputs a dictionary of models and predictors after the data is processed, iterate over them, and save a dictionary of outcomes for additional analysis. (Nick Morgan, 2021).

Precision, Recall, F1 Metrics, and Classification Reports.

A classification report is inclined to assess the accuracy of neural networks and classification algorithm predictions. It displays the recall, support, F1, and precision scores.

Precision and recall are great when measuring imbalanced data, where a class or multiple classes are in larger quantities than others, Precision is defined as the accuracy of positive predictions. obtained by TP/ (TP + FP) and Recall is the classifier's ability to detect all positive instances. That is, TP/(TP + FN) A model that has high precision but low recall returns fewer results but most of the predictions are accurately predicted when compared to the training labels. However, the F1 score is the weighted mean of recall and precision.



Discussion

Case 1: Age Prediction

For this project, the rating categories were limited to (PG, 12, 15, and 18). Initially, categories available after scraping were 18, PG, 12A, 12, U, 15, AA, PG-13, TV-MA, A, X, R, not-rated, tv-14, Adult, and Passed.

Age	Meaning	Replacement for this project
U	universal suitable for everyone	PG
PG	Parental Guidance is advised	PG
12A, 12	age 12 or older can view unsupervised	12
PG-13, 13	Aged 13 or over can view unsupervised	12
AA, TV-14 14	Aged 14 or over can view unsupervised	15
15	aged 15 or over can view unsupervised	15
TV-MA, 17	Age 17 and above can view unsupervised	18
A, X, R, 18, Adult, R -18	Mature adults only	18

Not rated and empty values were eliminated because the information couldn't be randomly inputted. (Motion picture association film rating system. 2022)

Age - Rating	PG	12	15	18
Count	123	213	449	261

This is the count of the movies after grouping them.

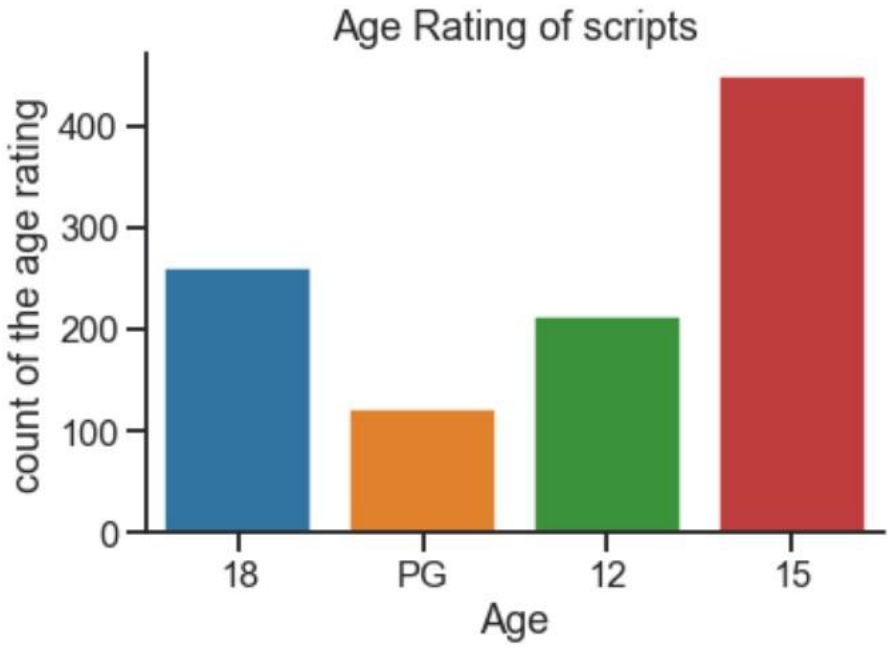


Figure 10 Age Rating

The movies aged 15 and over constituted 42% of the entire data, 24% of the movies were aged 18 and above, 20% were aged 12 and over and 12% were rated PG. The dataset was then normalised to return a balanced data using sklearn.utils resample method, results obtained are as shown below.

Results

	train_recall_score	test_recall_score	train_precision_score	test_precision_score	train_f1_score	test_f1_score
Naive-Bayes-movie_bow	0.959610	0.663889	0.982913	0.906449	0.971097	0.760644
Naive-Bayes-movie_w2v_mean	0.001393	0.005556	0.207173	0.266667	0.002766	0.010884
Naive-Bayes-movie_doc_vec	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
KNN-movie_bow	0.658078	0.472222	0.909607	0.714886	0.752227	0.554578
KNN-movie_w2v_mean	0.600975	0.427778	0.886458	0.687336	0.697062	0.503936
KNN-movie_doc_vec	0.635794	0.427778	0.902746	0.731919	0.738652	0.520210
SVC-movie_bow	1.000000	0.822222	0.968557	0.869728	0.983801	0.842649
SVC-movie_w2v_mean	0.607242	0.583333	0.349310	0.329246	0.437199	0.411617
SVC-movie_doc_vec	0.264624	0.241667	0.168109	0.171640	0.205216	0.200351
XGBoost-movie_bow	0.988162	0.708333	1.000000	0.903688	0.994020	0.779504
XGBoost-movie_w2v_mean	0.985376	0.622222	1.000000	0.883687	0.992630	0.671177
XGBoost-movie_doc_vec	0.798747	0.250000	1.000000	0.702279	0.887133	0.367410

The TFIDF bag of words performed best against all the other vector transformations, with Doc2vec performing least as seen in Naïve Bayes where no fitting and prediction occurred. All the classifiers appear to have overfitted to the training data. SVM generalized slightly better, The Analysis above explains that SVC performed the best, with test data TFIDF bag of words F1 - (0.8426). The Doc2vec worked best in the KNN model, based on the test f1 score as regards the other models. The Word2vec performed best with the XGBoost although there seem to be a serious case of 100% overfitting to the training data. Considering the variables in the research carried out by Shafaei et al., (2020) to predict MPAA rating, their research considered factors such as emotional dynamics between characters, movie genres, and similarities between movies. This could be variables that could be explored in the future to obtain better accuracies.

```
Results:  SVC-movie_bow
              precision    recall  f1-score   support

          PG       1.00      1.00      1.00       353
          12       0.98      1.00      0.99       363
          15       0.92      1.00      0.96       349
          18       0.97      1.00      0.98       371

   micro avg       0.97      1.00      0.98      1436
   macro avg       0.97      1.00      0.98      1436
weighted avg       0.97      1.00      0.98      1436
 samples avg       0.98      1.00      0.99      1436

              precision    recall  f1-score   support

          PG       0.99      0.91      0.95        96
          12       0.95      0.84      0.89        86
          15       0.76      0.66      0.71       100
          18       0.78      0.91      0.84        78

   micro avg       0.87      0.82      0.84       360
   macro avg       0.87      0.83      0.85       360
weighted avg       0.87      0.82      0.84       360
 samples avg       0.79      0.82      0.80       360
```

A classification report illustrating the best model achieved over TFIDF bag of words and SVM algorithm based on the F1 score. The test data appears to have very high precision over all classes with precision (PG) almost at a 100%, It is noticed that the ages (PG, 12, 18) that were resampled have a higher recall score compared to age (15) that wasn't resampled.

Case 2 - Genre prediction

More films had numerous genres than just one genre, as represented below. A function was created to split the movie genres based on its occurrence. As many as seven genres were assigned to some films.

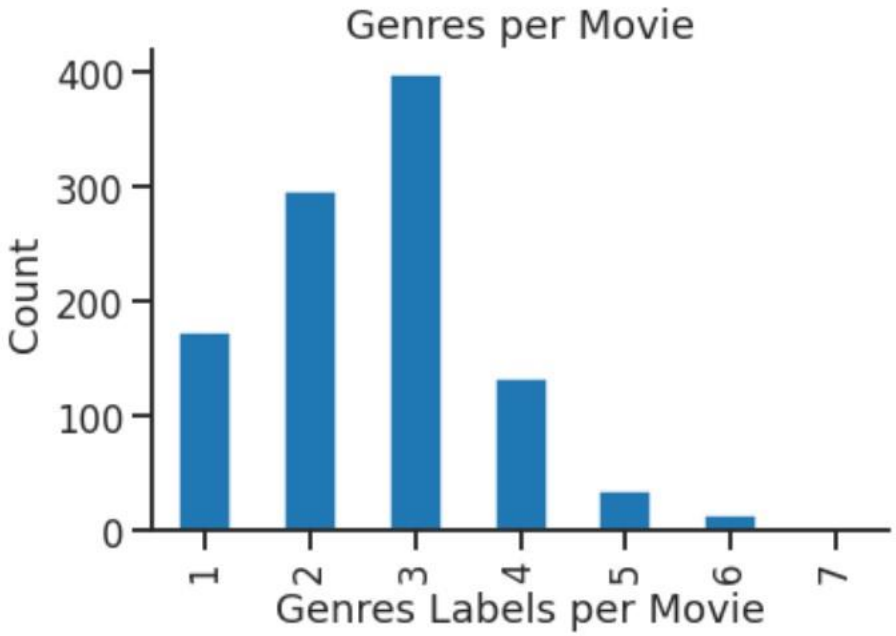


Figure 11 Genre Count per Movie

The movies after splitting, were of 20 categories as listed below, with drama being the most occurring genre type.

Genre	Frequency
Action	263
Adventure	152
Animation	38
Biography	3
Comedy	352
Crime	208
Drama	590
Family	36
Fantasy	92
Film-Noir	4
History	3
Horror	137
Music	28
Mystery	105
Romance	190
Sci-Fi	139
Sport	2
Thriller	356
War	28
Western	14

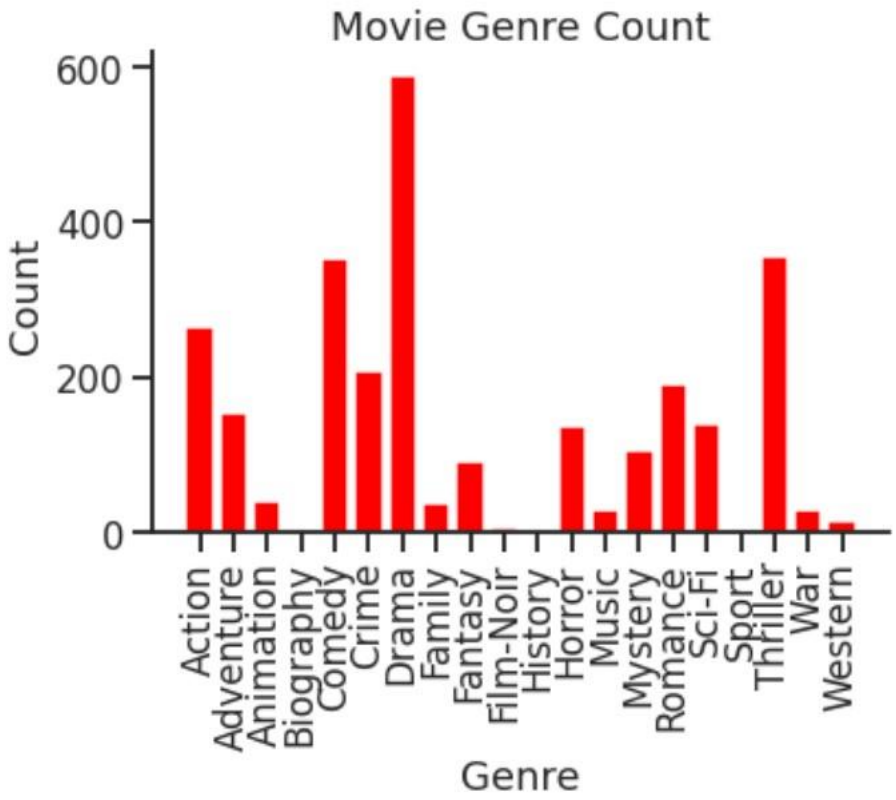


Figure 12 Movie Genre Count based on occurance

To easily explore our data, a transformation using scikit-learns Multilabel Binarizer (MLB) was applied to the updated genre list, splitting each row entry into single word values, then applying the binarizer. This was essential for allowing the model to employ one vs rest classification and precision-recall evaluation for each genre. Normally a different classifier will be built to predict each class but, One vs rest allows us to easily interpret how a model fits each genre in our dataset. I trained on a 50/50 split and 70/30 split but obtained the best results with an 80 training set and 20 testing set split. The MLB approach assigns genres to films based on feature values (TFIDF scores, W2V, D2V) where each feature is independent of another feature. Each genre is assumed to be independent of another. A movie classified as "Reservoir Dogs" has a genre of crime, action, and thriller. Due to XG-boost's ability to Overfit, as seen in the initial case for age prediction. It was purposely omitted in this analysis.

Results

	train_recall_score	test_recall_score	train_precision_score	test_precision_score	test_f1_score
Naive-Bayes-movie_bow	0.949059	0.531194	0.936506	0.668107	0.591836
Naive-Bayes-movie_w2v_mean	0.213860	0.213904	0.330067	0.260032	0.234723
Naive-Bayes-movie_doc_vec	0.215695	0.213904	0.121264	0.122231	0.155566
KNN-movie_bow	0.501606	0.433155	0.782519	0.675291	0.527776
KNN-movie_w2v_mean	0.418541	0.310160	0.766685	0.545837	0.395555
KNN-movie_doc_vec	0.346489	0.272727	0.707799	0.379157	0.317254
SVC-movie_bow	0.991280	0.650624	0.912535	0.670511	0.660418
SVC-movie_w2v_mean	0.685177	0.629234	0.487589	0.479446	0.544221
SVC-movie_doc_vec	0.684718	0.677362	0.435649	0.418123	0.517069

From the figure above, Naïve Bayes has the highest F1 score but because most movies are composed of numerous genres, there is some correlation of genres in the data. However, naive bayes assumes perfect independence, causing overfitting. So also, the SVC movie bow shows that the hyperplane was biased toward the dominating features.

Best results

The results with the highest mean precision and highest mean recall.

	train_recall_score	test_recall_score	train_precision_score	test_precision_score	test_f1_score
KNN-movie_bow	0.501606	0.433155	0.782519	0.675291	0.527776
SVC-movie_doc_vec	0.684718	0.677362	0.435649	0.418123	0.517069

The Classification report of the KNN models training data above and the test data represented below

Results: KNN-movie_bow

	precision	recall	f1-score	support
Action	0.84	0.61	0.71	208
Adventure	0.70	0.25	0.37	122
Animation	1.00	0.25	0.40	32
Biography	0.00	0.00	0.00	3
Comedy	0.82	0.39	0.53	278
Crime	0.71	0.39	0.50	156
Drama	0.75	0.86	0.80	470
Family	1.00	0.07	0.14	27
Fantasy	0.90	0.12	0.22	73
Film-Noir	0.00	0.00	0.00	2
History	0.00	0.00	0.00	3
Horror	0.96	0.38	0.55	115
Music	0.00	0.00	0.00	25
Mystery	1.00	0.07	0.13	83
Romance	0.77	0.13	0.22	156
Sci-Fi	0.85	0.43	0.57	110
Sport	0.00	0.00	0.00	2
Thriller	0.68	0.79	0.73	280
War	1.00	0.05	0.09	22
Western	0.86	0.50	0.63	12
micro avg	0.76	0.50	0.60	2179
macro avg	0.64	0.26	0.33	2179
weighted avg	0.78	0.50	0.56	2179
samples avg	0.73	0.54	0.59	2179

	precision	recall	f1-score	support
Action	0.65	0.44	0.52	55
Adventure	1.00	0.27	0.42	30
Animation	0.50	0.17	0.25	6
Biography	0.00	0.00	0.00	0
Comedy	0.76	0.35	0.48	74
Crime	0.61	0.27	0.37	52
Drama	0.67	0.78	0.72	120
Family	0.00	0.00	0.00	9
Fantasy	0.67	0.11	0.18	19
Film-Noir	0.00	0.00	0.00	2
History	0.00	0.00	0.00	0
Horror	0.75	0.14	0.23	22
Music	0.00	0.00	0.00	3
Mystery	1.00	0.05	0.09	22
Romance	0.43	0.09	0.15	34
Sci-Fi	0.91	0.34	0.50	29
Sport	0.00	0.00	0.00	0
Thriller	0.63	0.74	0.68	76
War	0.00	0.00	0.00	6
Western	1.00	0.50	0.67	2
micro avg	0.67	0.43	0.53	561
macro avg	0.48	0.21	0.26	561
weighted avg	0.68	0.43	0.48	561
samples avg	0.66	0.48	0.51	561

There appears to be overfitting in this model, the weighted train scores for precision (0.78) and recall are (0.50) whereas the test scores for precision (0.68) and recall (0.43). The model has the best precision out of all the models, meaning many true positives and few false positives. Looking at the test report, the model classified Adventure, and mystery western

100% meaning every time it predicted either one of them, it was predicted correctly. Most of the models were biased toward the drama, although the precision score is relatively low, it still has a high recall score, meaning it was predicted correctly in a lot of cases.
This model was unable to successfully predict biography, family, film noir, music, sport, and war. This is made evident by the value obtained in the F1 score [0.0, 0.0, 0.0, 0.0, 0.0, 0.0]

SVM doc2vec

Results: SVC-movie_doc_vec

	precision	recall	f1-score	support
Action	0.45	0.76	0.57	208
Adventure	0.25	0.60	0.36	122
Animation	1.00	0.03	0.06	32
Biography	0.03	1.00	0.07	3
Comedy	0.54	0.68	0.61	278
Crime	0.27	0.76	0.39	156
Drama	0.70	0.64	0.67	470
Family	0.08	0.70	0.14	27
Fantasy	0.14	0.64	0.23	73
Film-Noir	0.22	1.00	0.36	2
History	0.14	1.00	0.24	3
Horror	0.26	0.80	0.39	115
Music	0.11	0.32	0.16	25
Mystery	0.10	1.00	0.18	83
Romance	0.32	0.64	0.43	156
Sci-Fi	0.25	0.75	0.38	110
Sport	0.10	1.00	0.17	2
Thriller	0.46	0.71	0.56	280
War	0.33	0.05	0.08	22
Western	0.10	1.00	0.18	12
micro avg	0.30	0.68	0.41	2179
macro avg	0.29	0.70	0.31	2179
weighted avg	0.44	0.68	0.49	2179
samples avg	0.31	0.70	0.41	2179

	precision	recall	f1-score	support
Action	0.40	0.67	0.50	55
Adventure	0.24	0.57	0.34	30
Animation	0.00	0.00	0.00	6
Biography	0.00	0.00	0.00	0
Comedy	0.52	0.74	0.61	74
Crime	0.40	0.75	0.52	52
Drama	0.66	0.61	0.63	120
Family	0.07	0.44	0.13	9
Fantasy	0.15	0.63	0.24	19
Film-Noir	0.12	0.50	0.20	2
History	0.00	0.00	0.00	0
Horror	0.21	0.68	0.32	22
Music	0.04	0.33	0.07	3
Mystery	0.10	1.00	0.19	22
Romance	0.28	0.71	0.40	34
Sci-Fi	0.24	0.79	0.37	29
Sport	0.00	0.00	0.00	0
Thriller	0.52	0.74	0.61	76
War	0.00	0.00	0.00	6
Western	0.05	0.50	0.08	2
micro avg	0.30	0.68	0.41	561
macro avg	0.20	0.48	0.26	561
weighted avg	0.42	0.68	0.49	561
samples avg	0.31	0.70	0.41	561

Unlike the KNN, this model predicts no precision accurately for the test data. When this model makes a prediction, the prediction will be more accurate, the low precision scores and high recall scores prove that there are fewer false negatives. The model was unable to predict Animation, Biography, History, Sport, and War. Comparing the weighted F1 score of both test models SVM (0.49) and KNN (0.48) this model outperforms the KNN. Although this model has a lower precision score, the greater recall score and F1 score prove it is less likely to overfit. Albeit KNN has higher precision, this is a better result therefore the best model for this project

Comparing Embedding Vectorizers.

	min_test_precision	max_test_precision	mean_test_precision	min_test_recall	max_test_recall	mean_test_recall
bow	0.668107	0.675291	0.671303	0.433155	0.650624	0.538324
w2v	0.260032	0.545837	0.428438	0.213904	0.629234	0.384433
doc_vec	0.122231	0.418123	0.306503	0.213904	0.677362	0.387998

Taking a closer look at the data, Bag of words has the best performance with regards to precision and recall whereas doc2vec performed next and word2vec performed least.

Poor word2vec results could be attributed to its functionality and our corpus type. As mentioned above, because of the ambiguity in a script, the presence of dialogues from multiple people and separate emotions and context. There will be several unrelated sequences and patterns present in a movie, the word2vec sequence similarity approach will take all this information into account. This will explain why doc2vec performed better, because doc2vec trains data on a variable length, it does a holistic evaluation of similar data which in this instance is the entire script. Although TFIDF deals with data frequency, it doesn't put into account similarities between words.

Conclusion

Overall, outcomes obtained were satisfactory as the projects objectives were achieved, which were creating a reliable dataset from scratch and developing traditional multi-label classifiers. The dataset's flaw was the skewed data and the script being the only significant variable.

Recommendation

In the future, the polarity of the data could be put into consideration when making analyses, by tagging dialogues. This could help give an overview of the movies, another interesting extension could be the use of topic modeling to identify feature attributes in particular genres. This could tell us the level of impact certain words have effects on the dataset. Neural networks such as RNNs-Long Short Term Memory (LSTM and bi-directional LSTM) could be explored for their memory-like attributes, and Convolutional Neural Networks using pooling layers may be able to get more detailed information from the script data.

References.

U.S. & canada: Film genres by total box office revenue 2022. Available online: https://www.statista.com/statistics/188658/movie-genres-in-north-america-by-box-office-revenue-since-1995/ [Accessed Aug 21, 2022].

Motion picture association film rating system. (2022) .

Arikatla, G. & Chinnapottu, B. (2021) Movie prediction based on movie scriptsusing natural language processing and machine learning algorithms.

Aziz, R. H. H. & Dimililer, N. (2021) SentiXGboost: Enhanced sentiment analysis in social media posts with ensemble XGBoost classifier. *Journal of the Chinese Institute of Engineers,* 44 (6), 562-572.

Banks, D. Today's MPAA Ratings Hold Little Value for Parents. *Wired,* .

BFI (2019) The uk film economy.

BFI (2022) *BFI statistics for 2019 show film and high-end TV generates 30% uplift for UK economy.* Available online: /news/bfi-statistics-2019 [Accessed Aug 10, 2022].

Blackstock, A. & Spitz, M. (2008) Classifying movie scripts by genre with a MEMM using NLP-based features. *Citeseer,* .

Dalton, M. A., Sargent, J. D., Beach, M. L., Titus-Ernstoff, L., Gibson, J. J., Ahrens, M. B., Tickle, J. J. & Heatherton, T. F. (2003) Effect of viewing smoking in movies on adolescent smoking initiation: A cohort study. *The Lancet,* 362 (9380), 281-285.

Department for Digital, Culture, Media & Sport and Ed Vaizey. (2010). *The economic impact of UK film.* Available online: https://www.gov.uk/government/news/the-economic-impact-of-uk-film [Accessed Aug 10, 2022].

Habib, K. & Soliman, T. (2015) Cartoons' effect in changing children mental response and behavior. *Open Journal of Social Sciences,* 3 (09), 248.

Hartley, J. E., Wight, D. & Hunt, K. (2014) Presuming the influence of the media: Teenagers' constructions of gender identity through sexual/romantic relationships and alcohol consumption. *Sociology of Health & Illness,* 36 (5), 772-786.

Ho, K. (2011) Movies' genres classification by synopsis.

Hoang, Q. (2018) Predicting movie genres based on plot summaries. *arXiv Preprint arXiv:1801.04813,* .

Ihla, A. (2018) *Films that were rated totally wrong.* Available online: https://www.looper.com/112098/films-rated-totally-wrong/ [Accessed Aug 21, 2022].

Judith Rosenthal, Henry Massie & Kenneth Wulff (1980) A comparison of cognitive development in normal and psychotic children in the first two years of life from home movies. *Journal of Autism and Developmental Disorders,* 10 433-444.

Khyani, D. & B S, S. (2021) An interpretation of lemmatization and stemming in natural language processing. *Shanghai Ligong Daxue Xuebao/Journal of University of Shanghai for Science and Technology,* 22 350-357.

Luo, X. (2021) Efficient english text classification using selected machine learning techniques. *Alexandria Engineering Journal,* 60 (3), 3401-3409.

Nick Morgan (2021) *Python Movie Genre Predictor.* Available online: https://github.com/Nick-Morgan/Python-Movie-Genre-Predictor [Accessed May 10,2022].

Park, D. J. & Berger, B. (2010) Brand placement in movies: The effect of film genre on viewer recognition. *Journal of Promotion Management,* 16 428-444.

Parkes, A., Wight, D., Hunt, K., Henderson, M. & Sargent, J. (2013) Are sexual media exposure, parental restrictions on media use and co-viewing TV and DVDs with parents and friends associated with teenagers' early sexual behaviour? *Journal of Adolescence,* 36 (6), 1121-1133.

Premiere Theatres (2022) *The truth about ratings.* Available online: https://premierethea-tres.com/the-truth-about-ratings/ [Accessed May 6, 2022].

Shafaei, M., Samghabadi, N. S., Kar, S. & Solorio, T. (2020) Age suitability rating: Predicting the MPAA rating based on movie dialogues. *Proceedings of The 12th Language Resources and Evaluation Conference.*

Sultana, I., Ali, A. & Iftikhar, I. (2021) Effects of horror movies on psychological health of youth. *Global Mass Communication Review, VI,* .

Vagionis, N. & Loumioti, M. (2011) Movies as a tool of modern tourist marketing. *Tourismos,* 6 (2), 353-362.

Webster, J. & Kit, C. (January 1, 1992) Tokenization as the initial phase in NLP.

Wright, A., McCoy, A. B., Henkin, S., Kale, A. & Sittig, D. F. (2013) Use of a support vector machine for categorizing free-text notes: Assessment of accuracy across two institutions. *Journal of the American Medical Informatics Association : JAMIA,* 20 (5), 887-890.