Movie Genre Prediction from plot summaries by comparing various classification algorithms

Aziz Rupawala
Pillai's HOC college of engineering and technology
University of Mumbai
Mumbai, India
azizrupawala786@gmail.com

Dhruv Pujara
Pillai's HOC college of engineering and technology
University of Mumbai
Mumbai, India
dhruvpujara06@gmail.com

Mustakim Shikalgar
Pillai's HOC college of engineering and technology
University of Mumbai
Mumbai, India
shikalgar.mustakim@gmail.com

***Abstract*—Any crowd, before viewing a film peruses the motion picture plot outlines. Movie plots portray the setting of a film, yet additionally the class. In this task, we try to perform grouping of motion picture classification by utilizing the plot outlines. We perform characterization on an extensive range of classifications. We would do as such by utilizing unique grouping algorithms like SGD, Multinomial Naive Bayes, Random Forest, and Logistic Regression to locate the best result possible. With this, a great deal of time and working hours can be spared.**

***Index Terms*—Movie Genre classification, Multinomial Naive Bayes, Logistic regression, Random Forest, Stochastic Gradient Descent**

I. Introduction

Film plot rundowns reflect the class of the motion pictures such as romance, drama, comedy, etc., in a way that individuals can effortlessly seize the category of the film. Classifying movies has been a great trouble for watch-lists creators as they had to go through the entire movie plot and determine the genre manually.

In the publication, there exist several works that carry out film genre categorization, which makes use of an expansion of assets like audio [1], video, and literature from posters [2] and summaries [3]. It very well may be construed whether a plot rundown shows the genre of the film to which it has a place. Thus, this strategy can be useful during the training of film plots.

These genre classification algorithms, after predictions, select the best-suited results and plot the graph of genres with which the movie is associated.

Though the work revolves around genre classification, it is mainly associated with the selection of best-suited algorithms and comparisons between their classification capabilities. Every algorithm provides different value to the work and towards the output as well.

II. Methodology

A. Corpus construction and Preprocessing

The corpus "movies_metadata.csv" for the model is acquired from the website "https://www.kaggle.com." The corpus consists of around 45,000 distinct items, containing movie plots and genres associated with the respective movies along with 17 redundant features. Out of the 18 genres that were present in the corpus, we have used 12 genres for training the model. The critical reason for the elimination of excessive genres is the insufficient amount of data points for making proper predictions. The following table depicts the statistics of the corpus used.

TABLE I
Conveyance of the value counts for every genre

Genre	Count
Drama	11966
Comedy	8820
Action	4489
Documentary	3415
Horror	2619
Crime	1685
Thriller	1665
Adventure	1514
Romance	1191
Animation	1124
Fantasy	704
Total	**40393**

After data collection, we have removed the redundant rows which are not required for the training of our model. Furthermore, we have converted the plot texts into lower cases; also, we have removed all the null values from the corpus. In addition to this, with the help of NLTK[1], we have discarded all the abbreviations and stop words that are irrelevant to the training of the model. Thus, we have obtained the cleaned corpus consisting of relevant features, which are movie name, plot/summary, and genre, which is required for the training of our model.

[1]https://www.nltk.org/

B. Model

The model consists of various distribution algorithms viz.,

1) **Multinomial Naive Bayes:** Multinomial Naive Bayes is a group of algorithms dependent on applying Bayes theorem with a strong(naive) assumption, that each

component is free of the others, to predict the class of a given example. It furnishes the classification with the most elevated likelihood determined with Bayes calculations as the yield. Naive Bayes algorithms have been effectively applied to numerous areas, especially Natural Language Processing(NLP).

2) **Logistic Regression:** Logistic Regression (LR) [5] is the advancement of linear regression techniques, which is used when the yield(target) is categorical. This algorithm is broadly utilized in different data mining and AI issues where Logistic Regression depicts the target variable with at least one predictor factor.
3) **Random Forest:** According to Leo Breiman, random forests [6] are a mix of tree combinations to such a degree that each tree depends upon the estimations of an irregular vector tested autonomously and with a similar appropriation for all trees in the forest. A Random Forest is a classifier comprising of an assortment of tree-structured classifiers h(x,Θk), k=1, ... where the Θk are autonomous indistinguishably circulated random vectors, and each tree makes a single choice for the most well-known class at input x.
4) **Stochastic Gradient Descent:** Stochastic Gradient Descent(SGD) [7] is a variation of the Gradient Descent(GD) algorithm in which the updating of coefficients occurs after each iteration rather than at the end of each batch. This helps in much faster learning since a very less number of passes through the dataset are required to reach a sufficiently right amount of set of coefficients.

C. Evaluation Method

We have used multiple classifiers, but we intend to use only the best classifier, which is suitable for predicting a particular genre. For comparing the classifier results for each genre, we are using AUC(Area Under the ROC Curve) to evaluate the results given by the classifiers for each genre.

1) **Receiver Operating Curve(ROC):** An ROC curve [8] is a graph that projects 2 parameters, the True Positive Rate(TPR) and False Positive Rate(FPR).

   TPR can be described as follows :

$$TPR = \frac{TruePositive}{TruePositive + FalseNegative}$$

   FPR can be described as follows :

$$FPR = \frac{FalsePositive}{FalsePositive + TrueNegative}$$

2) **Area Under the ROC Curve**: AUC calculates the area beneath the whole ROC Curve. It gives a vast proportion of execution overall conceivable classification limits. A method of explicating AUC is that it tells how much the model is capable of distinguishing among different classes. Higher AUC score represents the precision in predicting true as true and false as false.

III. Experiments and Results

After the pre-processing step, the data set is divided into subsets of data items according to each genre, which is a total of 13 subsets of the pre-processed corpus. Followed by this, the first subcorpus is then split into two separate parts, one for training purpose and the other for testing purpose. The training data is then passed to all the 4 classification models for training. After the completion of the training of the 4 individual models, the AUC for each classification algorithm is calculated using the testing data.

Next, we compare the AUC scores which we've obtained after testing. Then the algorithm with the highest score along with the corresponding genre for which the result is obtained is stored in a result dictionary. In the dictionary, the key is the genre, and its value is the resultant algorithm which we have obtained after comparing the algorithms. Similarly, the whole process is repeated for all the remaining 12 genres to obtain the corresponding best-suited algorithm with the highest AUC score. These results are also then stored in the result dictionary.

The following flow diagram explains the process of obtaining the result dictionary.

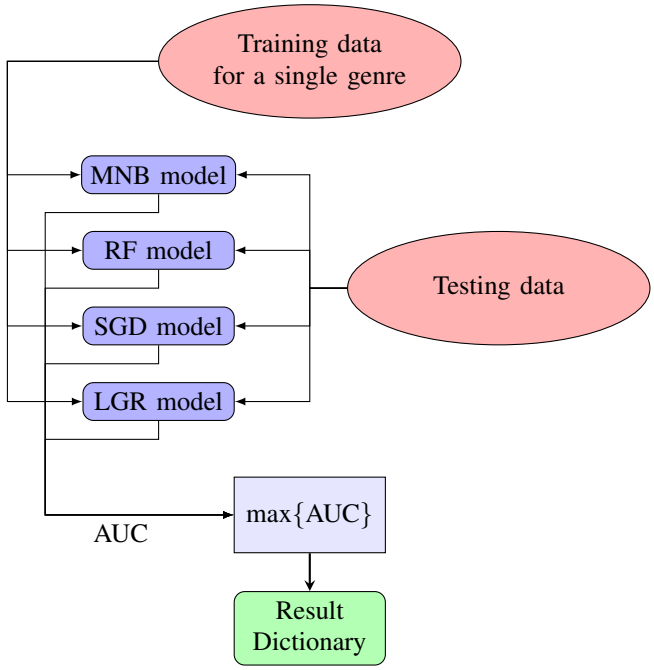

Fig. 1. Flowchart for the first phase

Concerning the resultant dictionary we obtained, instead of creating a multi-label classifier, which can tell if the input plot contains elements of Action, Horror, etc., we build 13 different classifiers. Each of these classifiers tells if that plot contains elements of that genre or not. For obtaining the genre of a new input plot, the plot is provided to these 13 classifiers, which individually calculate the probabilities. Then the maximum of these probabilities is selected, and the

genre with the highest probability is the best-suited output according to the model.

The following flow diagram depicts the above process of obtaining the best-suited output genre.

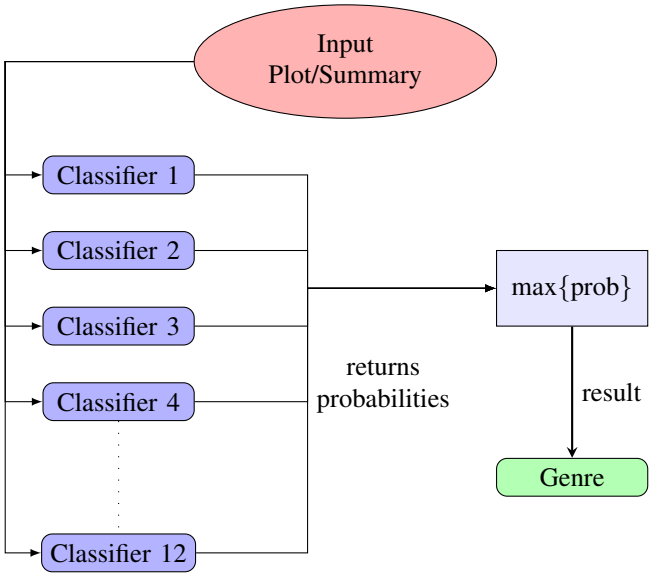

Fig. 2. Flowchart for the second phase

The table given below depicts the results obtained after training and testing the 4 models for each genre. The table depicts the AUC score, which is obtained for each of the genres for that particular algorithm. By looking at the table, it can be observed that Multinomial Naive Bayes model performs the best with the highest AUC score for each of the genres present in the corpus with an average AUC value of 0.70 whereas, Random Forest model is the worst performer with an average AUC score of 0.54. This information is then used to create individual classifiers for each genre, which then is used for predicting the genre for a plot summary provided as the input.

TABLE II
RESULTS

	SGDC	MNB	RF	LGR
Drama	0.68	0.68	0.59	0.67
Comedy	0.72	0.72	0.59	0.71
Action	0.74	0.76	0.55	0.73
Documentary	0.85	0.85	0.69	0.84
Horror	0.8	0.82	0.53	0.78
Crime	0.68	0.69	0.5	0.66
Thriller	0.61	0.63	0.53	0.59
Adventure	0.65	0.69	0.5	0.64
Romance	0.57	0.58	0.5	0.57
Animation	0.68	0.72	0.5	0.69
Fantasy	0.56	0.62	0.5	0.57
Science Fiction	0.67	0.73	0.5	0.69
Mystery	0.56	0.59	0.5	0.57
Average	0.674615	0.698462	0.536923	0.671428

IV. CONCLUSION

In this literature, we predict the movie genre from the plot summary of the movie by comparing 4 classification algorithms. The results show that the Multinomial Naive Bayes algorithm outperforms the other 3 classification algorithms and is better suited for the classification of the movie genre.

As future work, we are planning to increase the size of our corpus and the amount of movie genre tags that can be predicted from the current 13 possible classes.

REFERENCES

[1] Y.-F. Huang and S.-H. Wang, “Movie genre classification using SVM with audio and video features,” Active Media Technology, pp. 1–10, 2012.
[2] Z. Fu, B. Li, J. Li, and S. Wei, “Fast film genres classification combining poster and synopsis,” in International Conference on Intelligent Science and Big Data Engineering. Springer, 2015, pp. 72–81.
[3] Ali Mert Ertugrul and Pinar Karagoz, ”Movie Genre Classification from Plot Summaries Using Bidirectional LSTM,” 2018 IEEE 12th International Conference on Semantic Computing (ICSC).
[4] Shuo Xu, Yan Li, and Zheng wang, ”Bayesian Multinomial Naive Bayes classifier to Text Classification,” Institute of Haidian District, Beijing, People’s Republic of China.
[5] D. W. Hosmer and S. Lemeshow, “Applied Logistic Regression.” New York: John Wiley & Sons, Inc, 2000.
[6] Leo Breiman, ”RANDOM FORESTS,” Statistics Department University of California, Berkeley, CA 94720, September 1999.
[7] J. Kieffer and J. Wolfowitz, ”Stochastic estimation of the maximum of a regression function,” Cornell University.
[8] Metz, C. E. (1978). Basic principles of ROC analysis. Seminars in Nuclear Medicine, 8(4), 283–298.