A Comparative Study of Hollywood Movie Successfulness Prediction Model

Riefvan Achmad Masrury[1], Muhammad Apriandito Arya Saputra[2], Andry Alamsyah[3], Made Ayunda Sukma Primantari[4]
School of Economics and Business, Telkom University, Indonesia
riefvanvanos@telkomuniversity.ac.id[1], m.apriandito@gmail.com[2], andrya@telkomuniversity.ac.id[3], madeayunda97@gmail.com[4]

***Abstract*— The movie industry is a highly competitive industry with a lot of new movies are queued to be released each year. Movie making is subject to potential profits or loss in the magnitude of billion of dollars making this industry very risky. Predicting the successfulness of movie based on its financial performance prior to the release date is valuable in order to reduce number of uncertainties faced by decision makers such as producers, distributors, and exhibitors. Using the concept of machine learning, we suggest a classification model to predict the successfulness of a Hollywood Movie using Artificial Neural Network, Naïve Bayes and Support Vector Machine. The objective of this research is to compare classification algorithms performance for predicting the successfulness of Hollywood movies before they are being released. Artificial Neural Network produces the best model in terms of performance in predicting a movie successfulness. Reaching 80% of accuracy and having above 80% of F-measure, precision, and recall suggests that Artificial Neural Network is a good model to assist producers, distributors and exhibitors assess risks.**

Keywords—Hollywood Movie, Prediction, Neural Networks, Naïve Bayes, Support Vector Machine

I. INTRODUCTION

The movie industry is a highly competitive industry with a lot of new movies are queued to be released each year. According to President and CEO of the Motion Picture Association of America (MPAA), Jack Valenti, nobody knows how a movie will perform in the marketplace until it's released in the theater and come between the audience and the screen [1]. The movie production is divided into 3 stages. There are pre-production, production and post-production stages. Pre-production starts when the first time producers, screenwriters, and studio executives discuss about the potential concepts and actors for their movie. This phase also includes the process of greenlighting financing [2]. Production is where the actual process of shooting, including acting, cinematography, costume design, directing, lighting and design activities. The important part of the movie making process begins after filming is complete in the post-production phase. There are processes of editing, finishing and releasing the movie to audience, including the distribution and exhibition process [3]. The movie is distributed through sales, trailers and publicity then it is exhibited through theaters (cinemas), DVD/VCR/Blue-ray and TV [4].

Hollywood is the biggest movie industry reaching up to $11 billion in the last 10 years. However, the last reported that in the 2017, the United States box is decreased from the highest record in 2016. This movie industry can give profit or loss up to billion dollars [5]. The uncertainty of whether a movie will be successful or not does makes this industry very risky. But several researches show that by studying the features of the movie through a prediction, we can find out whether a movie is going to be successful or not [6]. The definition of a movie successfulness is relative, some movies are considered as successful according to the financial performance of the gross that the movies make, while some may not support the business, but they are called successful for a popularity and high ratings. In this study, we define the movie successfulness based on the gross that movies make for the box office. The successful movies stay in the theaters for a month, while those that fail can disappear under two weeks [2]. Therefore, releasing an unsuccessful movie can lead to high risk of failure of the producers and distributors.

It is worth noting that predicting the revenue before the theatrical release must leverage data that are available only before the movie is released. The accurate estimation of the movie box office revenue, mainly before the movie is released in the theater is a more challenging problem for the movie industry [7]. However, predicting a movie successfulness based on its financial performance prior to the release date is valuable in order to reduce number of uncertainties faced by decision makers such as producers, distributors, and exhibitors. Producers can secure investment capital and distributors can make a better decision regard to what and when the movie will be released in the theater or exhibited by exhibitors. Predictions can also help the distributor's decision when changing the time period if a movie turns out to be successful or not. By planning releases in a more systematic way, the prediction is considered as important as the producer, director and actor who make a succesful movie.

Machine learning is a method associated with large-scale automatic data analysis through supervised learning or unsupervised learning [8]. Using the concept of machine learning, computer can extract patterns in data and use it to perform a prediction task, one of the well-known models in machine learning is classification. Classification is a supervised learning model which learns from labeled data. There are many types of classification algorithms such as Decision Tree, K-Nearet Neighbors, Naïve Bayes, Support Vector Machine, Artificial Neural Network and etc. Accuracy of a classification model can be maximized using features extraction and engineering [7].

In order to test the prediction on available movie data, an experiment is carried out in this paper. By constructing a classification model, a successfulness of Hollywood Movie is predicted using popular algorithms in classification, namely the Artificial Neural Network, Naïve Bayes and Support Vector Machine. We collect IMDb data having attributes of genre, director, actors, release date, budget, gross and MPAA rating. The objective of this research is to compare classification algorithms performance for predicting the successfulness of Hollywood movies before the theatrical release.

II. LITERATURE REVIEW

Earlier study suggests that machine learning is widely used in predicting successfulness. Artificial Neural Network, Naïve Bayes and Support Vector Machine are the most popular models to classify movie successfulness. The previous studies were applied to the different data, such as

978-1-5386-8052-0/19/$31.00 ©2019 IEEE

ShowBiz Data, Opus Data and also IMDb. In this study, we only use movie data available before the theatrical release such a genre, budget, actor's star power, director's star power, month and MPAA rating from IMDb which are different from the rest of the previous study.

A. Naïve Bayes

Naïve Bayes is a well-known model for classification in terms of supervised learning [9]. It is an algorithm for classification models that based on the Bayes theorem, this algorithm captivates attention for its simplicity and performance. Naïve Bayes algorithm uses probability methods and statistics that predict opportunities based on previous events. The Naive Bayes algorithm assumes each feature is independent to others.

In terms of predicting the successfulness of a movie, the Naïve Bayes algorithm calculates the probability of the features inherent in a movie, then looks at the prediction class which has the maximum posterior probability. However, in predicting movie successfulness, the application of Naïve Bayes is still limited. Flora et al [10] in their study, show that Naïve Bayes implemented to classify movies according to 7 categories of its gross. The latter shows that Naïve Bayes has a better accuracy compared to other models in the study. They suggest that those models work best in complementing a human's analysis and intuition, preventing unconscious or conscious biases.

B. Support Vector Machine

Support Vector Machine is a supervised learning method for classification and regression. In terms of classification, Support Vector Machine minimizes the error and maximizes the margin [11]. Support Vector Machine learns from training set that contains a set of examples where each example is represented by point consisted in the model. Support Vector Machine maps the input by maximizing the difference of examples from different classes. So new examples are set to a particular class according to which side of the class they're closer to [12]. The concept of Support Vector Machine can be simply explained as an effort to find the best hyperplane that functions as a separator of two classes in input space.

In terms of movie prediction, the Support Vector Machine approach is commonly used for two classes' problem, such one provided by Rhee and Zulkernine, in [13] to predict the movie successfulness. The movies are classified into Hit and Flop classes. They use the movies released in 2011 to 2015 with 14 attributes after the movie released and they run it through Neural Network and Support Vector Machine. It shows the good accuracy for Support Vector Machine and Neural Network.

C. Artificial Neural Network

Artificial Neural Network is an information processing technique or approach inspired by biological nervous system, especially in human brain cells processing information. The key element of this technique is the structure of information processing systems that are unique and varied for each application. The Neural Network consists of many information processing elements (neurons) that are interconnected and work together to solve a problem, one of it is classification.

The neural network starts with a random set of weights that are adjusted by constructing the parameters such as number or hidden layers, number of neurons, learning rate, and etc. to map the input and the output. The model is able to be constructed in a way that it can adjust its weights by calculating new samples as the new data become available [14]. ANN has several benefits such as easily-adjust learning, self-organized, real-time operation [15].

The Neural Network is widely used in a movie successfulness prediction. Sharda and Delen [14] in their study, predict the movie successfulness based on its gross according to the attributes available before the theatrical release. The prediction problem is converted into the classification where the movies are classified into 9 classes according to the box office gross. They reach the success rate of a 36.9% using a multilayer perceptron network which has the highest result compared to the other models. Similar to the work by Sharda and Delen [14], Galvao and Henriques [16] propose a Multi-Layer (MLP) neural network to classify movies into 9 classes of its gross. They reach the better accuracy compared to Sharda and Delen [14] using movie attributes after the theatrical release. Zhang et al [17] also proposed neural network to predict the movie successfulness based on its gross classes. Zhang et al maximize the accuracy of the model by selecting the weights for the proposed model using statistical methods. Although they reach the better accuracy compared to Galvao and Henriques [16], they classify the movies into 6 classes using a rather small dataset. Those works show that neural network results the best accuracy among other models [17].

D. Evaluation Measurement

For a simple binary classification problem, the performance of the model was measured using Confusion Matrix using numbers of True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN) for accuracy measurement. Confusion matrix is a summary of prediction results on a classification problem. Basically, confusion matrix contains information that compares the model's prediction with the real values (Table 1). Definitions of prediction outcomes are presented below:

1) True Positive (TP): the number of correct prediction towards positive class.

2) False Positive (FP): the number of false prediction towards positive classBudget: budget is amount of resources to make a movie.

3) True Negative (TN): the number of correct prediction towards negative class.

4) False Negative (FN): the number of false prediction towards negative class.

TABLE I. PARAMETER MEASUREMENT

		Predicted	
		Successful	**Not Successful**
Actual	**Successful**	True Positive	False Negative
	Not Successful	False Positive	True Negative

Based on the value of TP, FP, TN and FN, we can obtain the value of accuracy, precision and recall. Those parameters successfully generate the performance of algorithm in a classification model. Table II shows the parameters used in this study to measure the evaluation of each model.

TABLE II. PARAMETER MEASUREMENT

Parameter	Description
Precision	The ratio of true prediction of a single class to total prediction towards the class
Recall	The ratio of true prediction of a single class to total population of the class
F-measure	The weighted average or mean of precision and recall
Accuracy	The ratio of true prediction outcomes to total population

The values of all parameters were calculated using TP, FP, TN and FN numbers in the confusion matrix. The mathematical equations of the parameters are expressed as below:

$$\text{Precision} = TP/(TP + FP) \quad (1)$$

$$\text{Recall} = TP/(TP + FN) \quad (2)$$

$$\text{F-measure} = 2 \times \frac{(Precision \times Recall)}{(Recall + Precision)} \quad (3)$$

$$\text{Accuracy} = (TP + TN)/(TP + FP + TN + FN) \quad (4)$$

The model is validated using a cross validation. A cross-validation is used as it tends to give a less bias accuracy estimation compared to other methods. In this work, 10-folds cross-validation is done with a stratified sampling method to rearrange the dataset so that each fold makes a good representation of the whole. From the total 10-folds of cross validation, the subset of training data includes 9-folds and the subset of testing data includes 1-fold.

III. METHODOLOGY

We dispart the study into 4 stages as shown in Fig. 1.

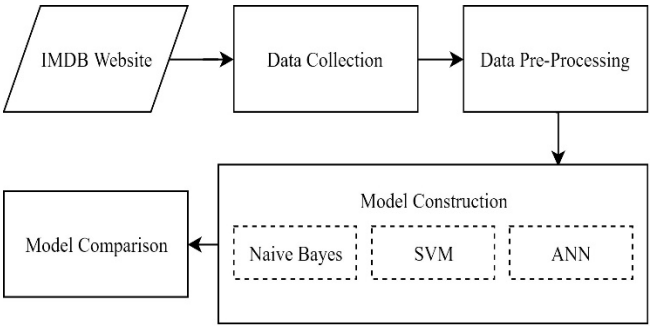

Fig. 1 Research Stages

A. Data Collection

The movie data are retrieved from IMDb using web-scraper application software. We only selected the top grossed 150 English movies released in the United States each year ranged from 2008 to 2017. Our dataset contains 1500 raw movie data. The movie data included information as follows:

1) MPAA Rating: Movie Picture Association of America (MPAA) rating is a movie category based on its rate of violence, obscene language and sexual content. There are 4 types of MPAA rating listed.

2) Genres: genre is used to categorize the content of the movie. The movie can have more than one content category outright. There are 18 types of genre listed.

3) Budget: budget is the amount of resources to make a movie.

4) Gross USA: gross is the revenue generated in the United States, usually in the period of 2 – 4 weeks

5) Actors and Director (Stars): actors and director are the people that have a big impact on movie successfulness. Many movies are known by its main actors or director. We have 3 main actors and a director in our dataset for each movie.

6) Released date: released date is the time when the movie is released in the theater.

B. Data Pre-Processing

Several steps are required that includes removal of missing values or noises which usually effects model performance. Pre-processing steps are listed below:

1) Missing value: incomplete data are removed from the dataset. There are 500 incomplete data removed so that leaves 1000 qualified ones to be processed further.

2) Engineered attributes: several new attributes are generated using original ones to extract new information [13].

a) Actors power score: we calculated how many times the main actor appears in our final dataset. The same calculation is done to the other 2 main actors. We then sum the number of the total appearance of all actors as actors power score for each movie.

b) Director power score: the director power score is calculated the same way as calculation of actor power score.

c) Month: months are extracted from release date attribut

d) Successfulness States: the classes of movie successfulness are created according to the amount of gross and budget of the movie. Successful class contains movie with gross that is more than the budget and unsuccessful class contains movie with gross that is less than the budget.

The final attributes used for the experiment can be seen in Table III as follows:

TABLE III. FINAL ATTRIBUTES

Attributes
MPAA rating
Genres
Budget
Classes
Actor power score
Director power score
Month of release

3) Normalization: the data are normalized using Min-Max normalization where X' represents the normalized values, X is the actual value of the attribute, X_{min} represents the minimum value of this attribute in the dataset and X_{max} represents the maximum value of this attribute in the dataset.

$$X' = (X - X_{min})/(X_{max} - X_{min}) \quad (4)$$

C. Model Construction

The classification models are constructed using Artificial Neural Network, Naïve Bayes and Support Vector Machine algorithm. In the model construction, the dataset is split into two subsets of data, training dataset and testing dataset. Ratio of training to testing data is set to 70:30. We selected a target "successful" as the target class for movie that have gross more than the budget and "unsuccessful" for movie that have gross less than the budget. We seek the best model to predict the successfulness of a movie.

D. Model Comparison

For assurance concerning a single model from several predictive model alternatives may give best prediction performance in predicting movie successfulness, three classification models are compared. The comparison of these models aims to discover the best model of classification to predict movie successfulness based on its gross. Table II. shows the parameters of the measurement used here.

IV. Result and Analysis

The experiment was conducted where the dataset is split in the ratio of 70:30 for the training and testing dataset. The 70% of the dataset is sufficient for a model learning in order to build the model. The latter 30% of the dataset is used in the interest of model validation. The training dataset loads the labelled data while the testing dataset loads the remaining unlabelled or unseen data. The number of samples used in this work is shown in Table IV.

TABLE IV. Data Partitioning

Data	Proportions	Total Samples
Training data	70%	667
Testing data	30%	286

The experiment resulted a confusion matrix as shown in Table V. It summarizes the score of the confusion matrix that generated by Artificial Neural Network, Naïve Bayes and Support Vector Machine. False Positive (FP) and False Negative (FN) shows the number of misclassified instances of these three models. Support Vector Machine is found to make the highest number of misclassifications.

TABLE V. Confusion Matrix Results

	ANN	Naïve Bayes	SVM
True Positive (TP)	90%	64%	52%
False Positive (FP)	10%	36%	48%
True Negative (TN)	58%	64%	55%
False Negative (FN)	42%	36%	45%

Support Vector Machine generated a 48% for False Positive (FP) and 45% for False Negative (FN) indicating that Support Vector Machine is not a strong model in classifying successful or unsuccessful movie classes. Subsequently, a cross validation is carried out on the testing data using the parameter in the Table II. in order to find out the best model of all. The Artificial Neural Network is discovered to generate the best model in terms of performance in predicting a movie successfulness with above 80% of accuracy. A complete result including precision, recall and f-measure is presented in Table VI.

TABLE VI. Cross Validation Result

Parameters	Cross Validation Results		
	ANN	*Naïve Bayes*	*SVM*
Accuracy	79.72%	64.34%	52.82%
Precision	81.69%	64.02%	52.33%
Recall	90.16%	65.70%	80.08%
F-measure	85.71%	64.85%	63.30%

Table VII. represents the summary of models used in this work. The results of accuracy mark out that Artificial Neural Network as the most suitable algorithm to predict the successfulness of the movie compared to the other 2 algorithms.

TABLE VII. MOdel Evaluations

Models	Status	Total Data	Accuracy	F-measure
ANN	Correctly classified	228	80%	86%
	Incorrectly classified	58		
Naïve Bayes	Correctly classified	184	64%	65%
	Incorrectly classified	102		
SVM	Correctly classified	151	53%	63%
	Incorrectly classified	135		

V. Conclusion

Predicting the successfulness of movie based on its financial performance is essential to assist the decisions by producers, distributors and exhibitors where they can plan to release and exhibit the movie in a more systematic way. The prediction model prior to release date is more valuable in order to reduce risks faced by movie decision makers. Three popular classification algorithms are used to construct the prediction models.

ANN produces the best model in terms of performance in predicting a movie successfulness. Reaching 80% of accuracy and having above 80% of F-measure, precision, and recall suggests that ANN is a good model to assist producers, distributors and exhibitors assess risks. Although Naïve Bayes does not perform better than ANN, it produces almost homogenous values across performance metrics. Naïve Bayes model is also considered less demanding in terms of computing power as it is the fastest model to build. SVM is not a suitable model mainly because of low accuracy and precision performance. Speed of processing is also not SVM's strong point to consider using it for prediction.

References

[1] Valenti, J. Motion Pictures and Their Impact On Society, 1978.

[2] Bergan, R. The Film Book: A Complete Guide to The World of Film. New York: DK Publishing, 2011.

[3] Kerrigan, F. Film Marketing. Oxford: Elsevier, 2010.

[4] Squire, J. E. The Movie Business Book. New York: Focal Press, 2017.

[5] MPAA, "THEME Report: A Comprehensive Analysis and Survey of THEME". 2017.

[6] Jangga, S., Ranjan, A. & Shanmugan, S. “IMDB Film Prediction with Cross-validation Technique,” International Journal for Scientific Research & Development (IJSART), 2016.

[7] M. T. Lash & K. Zhao, “Early Predictions of Movie Success: The Who, What, and When of Profitability.” *Journal of Management Information Systems,* 2016.

[8] M. D. N. Arusada, N. A. S. Putri and A. Alamsyah, "Traning Data Optimization Strategy for Multiclass Text Classification," *International Confrence on Information and Commnication (ICoICT),* vol. 5, 2017.

[9] A. Alamsyah & N. Salma, “A Comparative Study of Employee Churn Prediction Model,” *International Conference on Science and Technology (ICST),* 2018.

[10] B. Flora, T. Lampo, and L. Yang, “Predicting Movie Revenue from Pre-Release Data,” CS229, Stanford University, 2015.

[11] D. K. Srivastava & L. Bhambu, “Data Classification Using Support Vector Machine,” *Journal of Theoretical and Applied Information Technology,* 2009.

[12] V. Subramaniyaswamy, M. V. Vaibhav, R. V. Prasad & R. Logesh, “Predicting Movie Box Office Success using Multiple Regression and SVM,” *International Conference on Intelligent Sustainable Systems (ICISS),* 2017.

[13] T. G. Rhee & F. Zulkernine, “Predicting Movie Box Office Profitability,” *International Conference on Machine Learning and Applications (ICMLA),* 2016.

[14] R. Sharda & D. Delen, “Predicting Box-Office success of motion pictures with neural networks,” *Expert Systems with Applications,* pp 243–254, 2006.

[15] A. Alamsyah & M. F. Permana, “Artificial Neural Network for Predicting Indonesian Economic Growth using Macroeconomics Indicators,” *International Symposium on Advanced Intelligent Informatics (SAIN),* 2018.

[16] Galvao, M. & Henriques, R. “Forecasting Movie Box Office Profitability,” *Journal of Information Systems Engineering & Management,* 2018.

[17] L. Zhang, J. Luo & S. Yang, “Forecasting Box Office Revenue of Movies with BP Neural Network,” *Expert Systems with Applications,* pp 6580-6587, 2009.