Predicting IMDb Rating of Movies by Machine Learning Techniques

Conference Paper · July 2019

DOI: 10.1109/ICCCNT45670.2019.8944604

3 authors:

Warda Ruheen Bristi
Daffodil International University
4 PUBLICATIONS **9** CITATIONS
SEE PROFILE

Zakia Zaman
Daffodil International University
4 PUBLICATIONS **7** CITATIONS
SEE PROFILE

Nishat Sultana
Daffodil International University
4 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Predicting IMDb Rating of Movies by Machine Learning Techniques

Warda Ruheen Bristi
Department of CSE
Daffodil International University
Dhaka, Bangladesh
wardaruheen39@gmail.com

Zakia Zaman
Department of CSE
Daffodil International University
Dhaka, Bangladesh
zakia.cse@diu.edu.bd

Nishat Sultana
Department of CSE
Daffodil International University
Dhaka, Bangladesh
nishatsultana241@gmail.com

***Abstract*—Film Industry is not only a industry or a centre of entertainment, rather it is now a centre of global business. All over the world is now excited about a movie's box office success, popularity etc. A huge data is available online about these movies success or popularity. We have used hollywood movie list from Wikipedia and their rating from IMDb movie rating website to create our data set. Then machine learning classification algorithms are applied of the data set. Lastly an efficient model is developed to predict a movie's IMDb rating. The model gives good classification measures with the data set.**

***Index Terms*—Movie Rating, Machine Learning, Prediction, Box Office, IMDb**

I. Introduction

Now a days movies are not the only source of recreation, rather it is one of the major sources of global commerce and marketing. Movies create a new craze among people specially young people. Not only movie directors and box office officials are concerned with the success of movies but general people also. People used to talk about these in social medias. Therefore analysis of social media data about movies is recently popular among the data analysts. Other than this there remains some other scopes like analyzing a director's previous success histories or a actor's previous popularity etc. Again the analysis may be different on different countries. Naturally peoples from all the regions of the world do not react in the similar way. Movies are now available on internet. There are platforms like IMDb (Internet Movie Database) [1], Rotten Tomatoes [2], Metacritics [3] etc. where people can share their reviews about movies. Day by day these platforms are becoming popular since people are getting honest reviews there. So, huge data is available online about reviews and ratings of movies. In this paper this data about movie rating is analyzed to predict rating of movies.

There are a good number of studies to predict the movie success rate as people are too much excited about films. Very few studies [1] [2] include the features (director, screen play, actor, actress, genre etc.) of a movie to predict the success rate. So, in this paper, the focus is to predict success rate based on movie's own features. For example, Director, Screenplay, Actors/Actress, Country and Genre. Again, IMDb is a popular platform which is growing day by day. Generally people shows much interest if the IMDb rating of a movie if high. Therefore, the ultimate motive is to find out an model that can efficiently predict IMDb rating of a movie.

II. Related Work

Several works are done to predict movie success rate before a movie is released. Many researchers have used various machine learning techniques to predict the success rate. A data about movie attributes rather than social media data is used to analyze and found that logistic regression gives 84% accuracy level [3]. Pramod, Abhisht and Geetha shows that Fuzzy logic gives high accuracy for categorizing predictions [4]. Machine learning approaches are applied on synthetic dataset to build efficient structure for prediction using IMDB is used [3]. Various machine learning algorithms are being used to predict movie success rate [5]. Depending on various movie attributes mathematical models are being implemented to determine movie success [6].

Another study shows the market shares of domestic and international movies in Russia, where the international movies are ahead from the year of 2002 to 2014 [1]. This paper also distinguished three factors behind success of a movie. They are, budget, brands like popular actor/actress, directors and viewer's review. Lastly the model concludes that the sanction and budget has comparatively high effect on success of movies.

An interesting work about analyzing moviegoers taste along with behavior is done recently. Here at first individuals taste is analyzed to make a model then aggregate to predict box office result [7].

Indian film industry is vast and it's impact and influence is huge. So data mining techniques are used to predict whether a Bollywood movie will be a blockbuster one or not [8].

A good number of studies are done based on social media data analysis [9] [10]. The data may be a review or comments or reactions for an event etc.

III. Problem Domain

This paper focuses on finding a model to predict a movie's IMDb rating using the movie's information like studio, director, screenplay, actor, actress, genre and country as input

[1] https://wwww.imdb.com
[2] https://www.rottentomatoes.com
[3] https://www.metacritic.com

features. In data set, title of the movie and year of the movie release is also stated. The model is built using machine learning techniques. It is trained and tested by the data set prepared from Wikipedia and IMDb movie rating website. Hence, the model built here is learned and tested by supervised learning technique. The data set created here contains the class attribute “rating” which can be flop, below average, average and hit, is measured by the rating from IMDb website. List of Hollywood movies released in the year of 2018 are fetched from Wikipedia. Then the ratings of each of those movies are collected one by one from IMDb website. Therefore the data set is a real data set and contains data from two different sources. Then a number of popular machine learning classification algorithms are applied on the data set. But as the data set prepared is imbalanced, so techniques are used to make it a balanced one before applying classification algorithms. To get satisfying accuracy measures, resampling without replacement is done with the data set. Lastly, the accuracy measures of different algorithms are compared. The measures include accuracy [11], kappa statistics [12], root mean squared error [13]. Therefore, the model can predict movie’s rating with high accuracy.

IV. Proposed Methodology

The working method for this work involves few steps. The methodology is shown in figure 1. The steps are described below.

- Data Extraction
- Data Preprocessing
- Applying Machine Learning Techniques
- Comparing the results of different algorithms

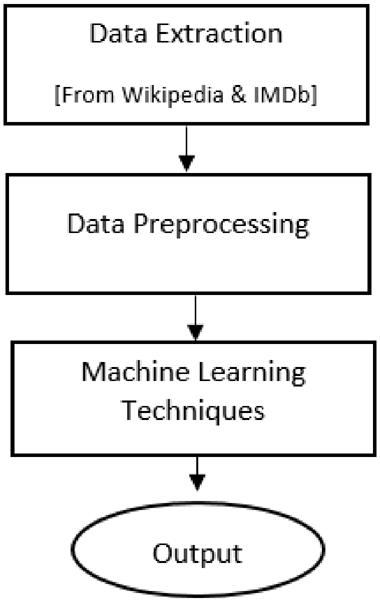

Fig. 1. Working Methodology

An algorithm to develop the efficient model is illustrated in algorithm 1. This algorithm shows every details of our working procedure.

Algorithm 1 Algorithm for developing the model

```
Prepare data set
Check Minority
If needed apply SMOTE algorithm until the minority class becomes equal to the size of it’s closest class
Classification
$Accuracy \longleftarrow 0$
while True do
    Resample Data
    Call (Classifier)
    if % of correctly classified Instance >Previous Accuracy Measure then
        $Accuracy \longleftarrow \% \ of \ correctly \ classified \ Instance$
    else
        Break
    end if
end while=0
```

After completing the preprocessing step, classifiers are used in a repeated style. Each time data is resampled without replacement and classifiers are used. Each time accuracy is recorded. This repeating process is ended when accuracy doesn’t increase anymore. Thus, the highest accuracy found from each classifier is recorded. In the algorithm, a while loop is used which repeats this process and the loop breaks if accuracy doesn’t increase at a step. Therefore the highest accuracy is recorded lastly.

A. Data Extraction

Data is extracted from Wikipedia and IMDb movie rating website. We have merged data from two platforms for our data set. Note that, data about only Hollywood movies released on the year of 2018 is extracted from Wikipedia. About 250 Hollywood movies are released on the year of 2018[4]. The extracted data from wikipedia contains title of the movie, studio, cast and crew, genre, country, month and date of release, year. The cast and crew column of wikipedia data contains director, screenplay and cast list of each movie. Therefore, obviously it is a tough job to separate director, screenplay, actor and actress of each movie in separate columns.

In the preprocessed data, we replaced the cast and crew feature by director, screenplay, actor and actress features. Among the male actors in the list, the main male actor is selected and the main female character is selected. This selection is done from the IMDb website. Each movie information from IMDb is fetched and main male or female character is set from there. Therefore, only one main male and one main female character of each movie is considered in our data set. Date and month of movie release is not considered in our work, so this feature is removed from data set.

[4] Data is collected from Wikipedia on 11 April 2019

TABLE I
Summary of Data set and it's attributes

Attribute	Description	No. of distinct values in Hollywood movie data set
Title	Name of the movie	242
Studio	Production Company	87
Director	Director(s) of the movie	237
Screenplay	Screenwriter(s) of the movie	232
Actor	Main male character of the movie	222
Actress	Main female character of the movie	217
Genre	Genre of the movie	29
Country	Country from which the movie is released	18
Year	The year of movie release. [Movies released in the year of 2018 is considered only]	1
Rating	Class of movie based on it's IMDb rating	4

Secondly, IMDb rating of each selected movie is extracted from IMDb website [5]. IMDb rating is becoming popular day by day. People used to give and trust IMDb rating [14]. That's why this platform is preferred in this work. IMDb rates each movie out of 10. Therefore ratings are classified into four classes, flop, below average, average and hit. The ranges of ratings for each class is represented in Table 2. That is, if a movie's IMDb rating is 3, then it's class will be flop. If the rating is 7, the class will be Average. So, the class feature of each movie in the data set is either flop or below average or average or hit, based on it's IMDb rating.

Therefore, the data set contains extracted real data from both Wikipedia and IMDb website. A Brief summary of data set and it's attributes are illustrated in Table 1.

TABLE II
Class consideration

Range of Rating	Class
0-3.5	Flop
3.6-5.8	Below Average
5.9-7.4	Average
7.5-10	Hit

B. Data Preprocessing

Data preprocessing means to prepare the data for classification. Data is processed according to the requirements of classification. Here, for preprocessing the data, instances with missing attributes are removed. Finally we got data set of 242 movies. The features of the processed data set are Title, Studio, Director, Screenplay, Actor, Actress, Genre, Country, Year, Rating. Here, Rating is the class attribute. The details are described in table 1. The data set now produced is an imbalanced data set, as there are only 3 movies of flop class and 138 average movies. For this, we applied SMOTE (Synthetic Minority Over-sampling Technique) [15] algorithm on our data set. SMOTE was applied with 700 as percentage as the closest class of flop class was hit class with 24 number of movies. Therefore, by sampling the size of flop class became equal to the size of hit class.

C. Applying Machine Learning Techniques

We have used Weka 3.8.3 tool [16] and applied five machine learning algorithms [shown in table 3] to build the model.

Among the classifiers Bagging and Random forest is ensemble method. Random forest starts classifying with multi decision tree. J48 also classifies using decision tree, it is often referred as statistical classifier [17]. IBK is K-Nearest Neighbour algorithm and it is a non-parametric method. Lastly a probabilistic classifier Naive Bayes, which is based on Bayes' Theorem.

10 fold cross validation is used without replacement. We repeated the classification using a classifier till the accuracy increases. The highest accuracy is recorded.

TABLE III
List of Classifiers used

Name of the Classifier
Bagging
Random Forest
J48
IBK
Naive Bayes

1) Bagging: Bootstrap aggregating is an AI gathering meta-calculation intended to improve the soundness and precision of AI calculations utilized in measurable arrangement and relapse. It likewise decreases change and abstains from over fitting [18]. Bagging is a standout among the best computationally escalated methodology to enhance unsteady estimators or classifiers, helpful particularly for high dimensional informational collection issues [18].

2) Random Forest: Random Forest or discretionary decision woods are a gathering learning procedure for portrayal. The backslide and various endeavors that works by using the planning time to build enormous decision trees [19]. The classes use the yielding technique (game plan) or mean figure (backslide) of the individual trees [19]. Random Forests are a mixture of tree pointers to such a degree. That the tree depends upon the estimations of a subjective vector analyzed self-sufficiently.

3) J48: The C4.5 algorithm is used for constitution to select trees. Trees are called as J48 in weka which are experimented [20]. Channels have similar characteristics as classifiers. That are composed in a chain of importance: full name of weka is called as J48 [20]. Information mining incorporates the conscious examination of gigantic enlightening lists.

4) IBK: A non-parametric estimation system that deals with k-nearest neighbours is called IBK. That is needed for portrayal and backslide. In the two cases, the data contains the k nearest planning points. That is of reference in the component space [21]. In k-NN request, the yield is a class interest. Pushed by applying Text Categorization to organizing Web

[5] Data is collected from IMDb on 12 April 2019

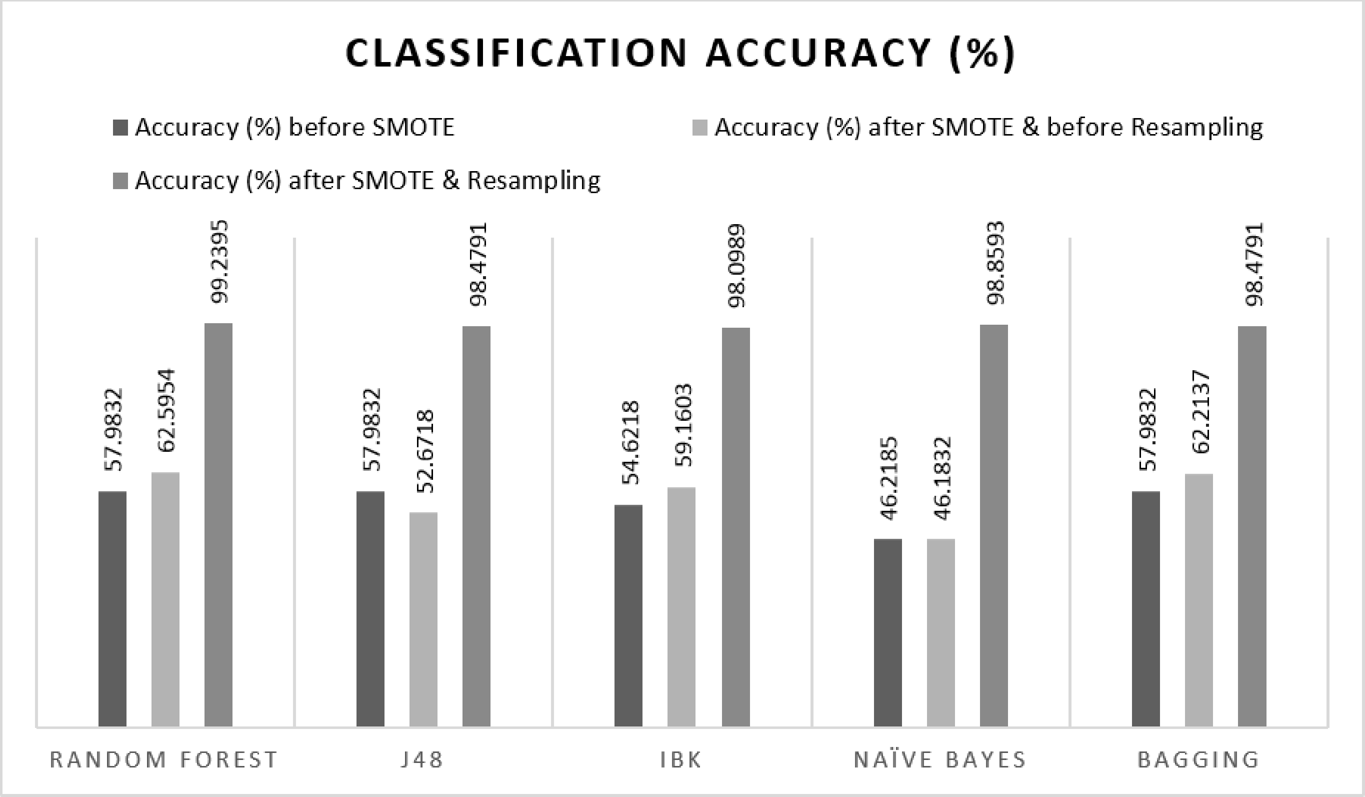

Fig. 2. Classification Accuracy of the classifiers

list things [21]. IBK selects the number nearest neighbours between 1.

5) Naive Bayes: A classifier is an AI model that is used to isolate different things. It is used to subject to explicit features. A Naive Bayes classifier is a anticipated AI model. This model is utilized for classification task. The core of the classifier depends on the Bayes hypothesis [22]. Implementing Bayes theorem, we can evaluate the prospect measure of A event, knowing the fact that B has taken place. Here, B is the proof and A would be the hypothesis [23]. The supposition is made based on the hypothesis. That is the predictors of the features are not dependent. That is existence of one fixed feature does not influence other. That's why naive is being called [24].

V. Performance Metrics

To compute the performance of each classifiers we have used root mean squared error and kappa statistics along with classification accuracy. RMSE is applied to measure the difference of the actual value and the predicted value done by a classifier. RMSE value is always non-negative and value of 0 represents perfect fit.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{i=N}(\hat{y_i}-y_i)^2}{N}}$$

where, $\hat{y_i}$, is the value predicted by the classifier and y_i is the actual result. kappa statistics, [12] is a good measure for multi class and imbalanced data set. It calculates the inter-rater agreement for categorical items.

$$Kappa, k = 1 - \frac{1-y_o}{1-y_e}$$

where, y_o, is the relative observed agreement among raters and y_e is the hypothetical probability of chance agreement. Kappa statistics ranges from 0 to 1. Here 1 represent perfect agreement. The accuracy of the metrics are illustrated in details in figure 2 and 3.

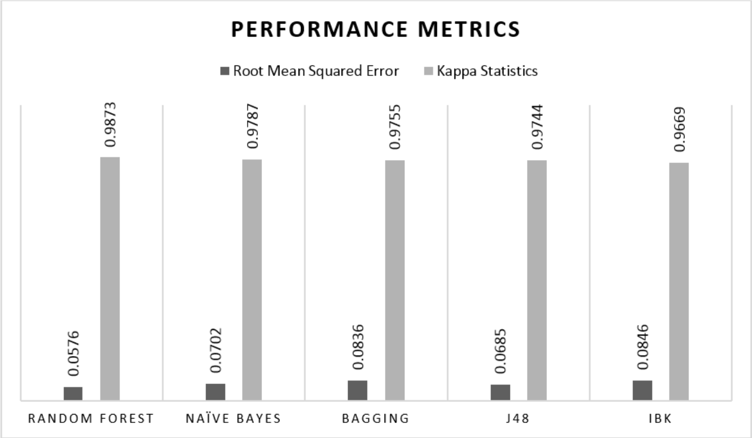

Fig. 3. Root Mean Squared Error and Kappa Statistics of the classifiers

VI. Result

This section describes about the accuracies of the classifiers at different stages of our work. We can divide our procedure in three stages. The first stage is to apply the classifiers on our data set before applying SMOTE technique. As our data set is imbalanced, so before applying SMOTE, it's accuracy is very low. The accuracies are illustrated in figure 2. Here, Random forest, J48 and bagging gives same classification accuracy and this is the highest among the five classifiers.

To make the data set a balanced one, SMOTE is applied afterwards. Thus, the accuracies increase. At this stage, again random forest gives highest accuracy. Like the previous stage, Naive Bayes gives the lowest accuracy.

Now at the final stage, satisfying outcome is acquired by applying classification algorithms by resampling the data set without replacement [25]. In algorithm 1, this resampling and determining the accuray is done in a loop. The loop continues when the accuracies increase in each turn. It breaks, when the accuracy starts to decrease. Thus, the highest accuracy obtained by each classifiers is listed and shown in figure 2. Among the five classifiers, random forest gives the highest classification accuracy. Though all the five classifiers give significantly good accuracy here, that is above 90%.

Figure 2 illustrates the kappa statistics and root mean squared error of the classifiers. Here, random forest again gives maximum kappa statistics and minimum root mean squared error. The lazy classifier IBK gives the minimum accuracy with highest root mean squared error.

Therefore, the performance metrics of the five classifiers state the conclusion that random forest classifier gives the best outcomes in terms of accuracy, kappa statistics and also root mean squared error for our data set.

VII. Conclusion

As the business market of film industries are becoming huge day by day, competition here is also growing complex. Therefore, predicting movie's rating is growing complex also. Our model is developed on a real world data set and it is collected from two platforms, Wikipedia and IMDb. The model can also be used to predict some other ratings like Rotten Tomato or Metacritic. Other than films, TV shows, music shows, etc. can be predicted by our model using the features of our model. [1] describes some features having more influence on movie success and some other features having less or no influence. According to [1], budget is having a small positive influence but cast or actor/actress doesn't have any influence on Russian film industry. Thus, our work can be done by a weighted feature classification to reflect these influences. Along with hollywood movie dataset, bollywood or other movie dataset can be used to make the model more efficient. The database can be enriched by including the movies released on recent years. That is, along with 2018, movies from 2017 or 2016 can be included.

References

[1] S. Gaenssle, O. Budzinski, and D. Astakhova, "Conquering the box office: Factors influencing success of international movies in russia," 2018.

[2] M. Saraee, S. White, J. Eccleston *et al.*, "A data mining approach to analysis and prediction of movie ratings," *Transactions of the Wessex Institute*, pp. 343–352, 2004.

[3] M. H. Latif and H. Afzal, "Prediction of movies popularity using machine learning techniques," *International Journal of Computer Science and Network Security (IJCSNS)*, vol. 16, no. 8, p. 127, 2016.

[4] S. Pramod, A. Joshi, and A. Mary, "Prediction of movie success for real world movie dataset," *Int. J. of Advance Res., Ideas and Innovations in Technol*, vol. 3, no. 3, 2017.

[5] R. Parimi and D. Caragea, "Pre-release box-office success prediction for motion pictures," in *International Workshop on Machine Learning and Data Mining in Pattern Recognition*. Springer, 2013, pp. 571–585.

[6] J. Ahmad, P. Duraisamy, A. Yousef, and B. Buckles, "Movie success prediction using data mining," in *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. IEEE, 2017, pp. 1–4.

[7] R. P. Ruhrländer, M. Boissier, and M. Uflacker, "Improving box office result predictions for movies using consumer-centric models," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2018, pp. 655–664.

[8] R. Khandelwal and H. Virwani, "Comparative analysis for prediction of success of bollywood movie," *Available at SSRN 3350907*, 2019.

[9] B. Çizmeci and Ş. G. Ögüdücü, "Predicting imdb ratings of pre-release movies with factorization machines using social media," in *2018 3rd International Conference on Computer Science and Engineering (UBMK)*. IEEE, 2018, pp. 173–178.

[10] S. Mundra, A. Dhingra, A. Kapur, and D. Joshi, "Prediction of a movies success using data mining techniques," in *Information and Communication Technology for Intelligent Systems*. Springer, 2019, pp. 219–227.

[11] R. Kohavi *et al.*, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Ijcai*, vol. 14, no. 2. Montreal, Canada, 1995, pp. 1137–1145.

[12] J. R. Landis and G. G. Koch, "An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers," *Biometrics*, pp. 363–374, 1977.

[13] N. Levinson, "The wiener (root mean square) error criterion in filter design and prediction," *Journal of Mathematics and Physics*, vol. 25, no. 1-4, pp. 261–278, 1946.

[14] S. Dooms, T. De Pessemier, and L. Martens, "Improving imdb movie recommendations with interactive settings and filters," in *8th ACM Conference on Recommender Systems (Poster-RecSys 2014)*, vol. 1247, 2014.

[15] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[16] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[17] Wikipedia contributors, "C4.5 algorithm — Wikipedia, the free encyclopedia," https://en.wikipedia.org/w/index.php?title=C4.5_algorithm&oldid=883549387, 2019, [Online; accessed 28-May-2019].

[18] P. Bühlmann, B. Yu *et al.*, "Analyzing bagging," *The Annals of Statistics*, vol. 30, no. 4, pp. 927–961, 2002.

[19] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[20] V. F. Rodriguez-Galiano, B. Ghimire, J. Rogan, M. Chica-Olmo, and J. P. Rigol-Sanchez, "An assessment of the effectiveness of a random forest classifier for land-cover classification," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 67, pp. 93–104, 2012.

[21] M. Radovanović and M. Ivanović, "Document representations for classification of short web-page descriptions," in *International Conference on Data Warehousing and Knowledge Discovery*. Springer, 2006, pp. 544–553.

[22] C. a. Ratanamahatana and D. Gunopulos, "Feature selection for the naive bayesian classifier using decision trees," *Applied artificial intelligence*, vol. 17, no. 5-6, pp. 475–487, 2003.

[23] A. McCallum, K. Nigam *et al.*, "A comparison of event models for naive bayes text classification," in *AAAI-98 workshop on learning for text categorization*, vol. 752, no. 1. Citeseer, 1998, pp. 41–48.

[24] H. Zhang, "The optimality of naive bayes," *AA*, vol. 1, no. 2, p. 3, 2004.

[25] S. Latif, "Customer annual income prediction using resampling approach," in *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*. IEEE, 2017, pp. 3865–3870.