Movie Success and Rating Prediction Using Data Mining Algorithms

S. Pirunthavi[1*], R.P. Vithusia[1,] K. Abishankar[1,] E.M.U.W.J.B. Ekanayake[2] and M. Yanusha[1]

[1*] *Uva Wellassa University of Sri Lanka*
[2] *Sabaragamuwa University of Sri Lanka*

This project developed the models to predict the success and the ratings of a new movie before its release. Since the success of a movie is highly influenced by the actor, actress, director, music director, and production company, those historical data were extracted from the Internet Movie Database (IMDb). The Box Office Mojo stores information about the cost of production of a movie and the total income of the movie. This information is helpful to determine whether the movie is successful or not in terms of revenue. A threshold was defined on revenue based on heuristics to categorize the movie into success or failure. Teasers' and trailers' comments were extracted from YouTube as those are very helpful to rate a movie. The keywords were extracted from the user reviews using a Natural Language Processing (NLP) technique and those reviews were categorized into positive or negative based on the sentimental analysis. A Random Forest Algorithm was trained using the features extracted from IMDb to predict the success of a movie. Further, the Naïve Bayers model was trained using the user reviews extracted from YouTube to predict the rating of a movie. The models were tested on real datasets and the accuracy of those were evaluated respectively. Finally, two conclusions have been met that the rating of a new movie cannot be predicted in advance through the YouTube trailers' and teasers' comments and the success of a new movie can be predicted in advance by using the data or features collected from online. The performances of the models are decent enough compared to the existing models in the literature. The Success Prediction model can be used as an early assessment tool of movies since it has gained 70% overall accuracy and hence, useful for the people in the movie industry and the audience of the movies. YouTube allows us to extract a limited number of user comments and hence, this factor could be negatively affected by the accuracy of the movie rating prediction.

Keywords: Rating prediction, Data mining algorithms