DS Project Comments

The motivation of applying machine learning on the application is clear. Problem definition is clear – multi-class classification. Related work was carefully presented and discussed. Multiple datasets have been collected. Evaluation metrics are clearly given. The timeline looks good. Suggestions:

- From the title, I learned that the final solution would be an ensemble method. However, Random Forest (RF) or AdaBoost themselves are ensemble methods. The motivation of the technical contribution is not clear. Are you going to propose a novel ensemble method, or just implementing either RF or AdaBoost and comparing with other non-ensemble methods? If you are going to propose a novel method, I would suggest you to add “novel” into the title, and present how and why your approach will work.
- Multiple datasets are collected. Descriptions are given. Are you going to develop models separately on each dataset, which means, the datasets will be used separately to validate the same conclusion? Or, are you going to integrate the datasets to create a very large dataset and unify the label schema (the set of genres) for performing and evaluating machine learning solutions on the very large dataset?
- It is good to see a few feature examples have been identified to train the models. There is a risk that none of the features is correlated with the label or most of the features are redundant (highly correlated with each other). Perform data analysis to identify data quality issues when the datasets are ready.