Movie Genre Classification using Metadata, Audio and Visual Features

1 Introduction

The classification of movie genres has been a widely studied case of Machine Learning (ML) in the past few years, given the growing market for movie recommendation systems. In particular, authors have been using different ML models and feature types taken from metadata, audio and visual data in order to classify movie genres. The purpose of this paper is to evaluate the prevalence of metadata and audio-visual feature types on the performance of the proposed ML models.

1.1 Relevant Literature

A novel approach has been applied in movie genre classification by using Convolutional Neural Networks by Simoes et. al. (2016), where the authors extracted a dataset from 3,500 movie trailers and then utilized a CNN architecture to perform the classification task.

A more advanced approach has been considered by Ertugrul and Karagoz (2018), where the authors worked with a Bi-directional Long Short-Term Memory network to classify movie genres. The dataset was based on movie plot summaries, from which they extracted sentences to be used as features.

1.2 Dataset

The dataset (Deldjoo et al., 2018; Harper and Konstan, 2015) was represented by 5,240 training instances and 299 validation instances, both including features and labels. Each movie is labelled with a single movie genre out of 18 different labels (*Action*, *Adventure*, *Animation*, *Children*, *Comedy*, *Crime*, *Documentary*, *Drama*, *Fantasy*, *Film_noir*, *Horror*, *Musical*, *Mystery*, *Romance*, *Sci_fi*, *Thriller*, *War*, *Western*). Each movie contains a unique ID and three different type of features:

- Metadata: including *title*, *year* of release and human annotated *tags*.
- Visual: including 107 pre-computed visual features extracted from movie trailers.
- Audio: including 20 pre-computed audio features extracted from movie trailers.

1.3 Evaluation metrics

We provide a set of evaluation metrics in order to assess the classification task for this paper:

- Accuracy: the proportion of instances that have been predicted correctly.
- Learning curve: represents the training and validation accuracy for different levels of training instances.
- Macro average precision: the proportion of instances that have been predicted correctly over the total positive values, averaged among all classes.
- Macro average recall: the proportion of instances that have been predicted correctly over the actual positive values, averaged among all classes.
- Macro average F1 score: the harmonic mean of precision and recall, averaged among all classes.

2 Hypothesis

We provide two hypothesis for this paper:

- Will our proposed models outperform the baseline model?
- Will there be a significant difference in performance of ML models when using metadata-features-only compared to using audio-visual features and a combination of both?

3 Methodology

This section provides an overview of the steps considered to obtain results for the classification task.

3.1 ML Algorithms

We decided to include the following ML algorithms in order to compare their performance:

- Zero-R (baseline model).
- Logistic Regression (LR) (linear).
- Support Vector Machine (SVM) (linear).
- XGBoost (XGB) (non-linear).

The above algorithms were programmed in Python language using the Jupyter platform. The selection of parameters was performed based on expert knowledge of the models and the given dataset. In particular, LR parameters were set to multinomial for solving the multiclass classification task and solver to newton to get smoother results. Moreover, SVM kernel function was set to linear and one-vs-rest-classifier approach was used to perform a multiclass classification. Furthermore, XGB learning rate was set to a small number to prevent overfitting and the maximum depth of the was set to a small number to save memory.

3.2 Data Pre-processing

We performed data cleaning tasks in order to run our models correctly, for which we removed 3 training instances and corrected the year fields for 5 instances. We then used 5,237 instances for our experiments.

3.3 Feature Engineering

We extracted features from the metadata by generating a vectorized text representation of the *tag* and *title* features. The following steps indicate how this approach was undertaken to create a corpus of words:

- Remove numbers and punctuations.
- Change data from uppercase to lowercase.
- Split titles and tags into different words.
- Filter stop-words.
- Apply stemming process to reduce the extracted words.

After the above steps, we applied tf-idf technique to measure the relative importance of each word in the corpus for each instance. Moreover, all the new metadata features were standardized.

The year feature was transformed into string type and then one-hot-encoded in order to be processed by the models correctly.

Audio-visual features were standardized for the models to perform correctly and efficiently.

For the features that were not seen in the validation set but seen in the training set, we created new features filled with zero values. Finally, the countdown of features reached to 2,305, including all new metadata and audio-visual features.

3.4 Feature Selection

We selected a filtering based approach for feature selection, namely Analysis of Variance (ANOVA), in order to make a ranking of the features that result to be more discriminating of the classes. After obtaining the ranking, we conducted a training experiment on the training set for each model, including the whole dataset, for different feature sizes (keeping the ranking), using cross-validation technique to obtain stable results (see Table 1).

	Accuracy			
Feat. Nº	**0-R**	**LR**	**SVM**	**XGB**
50	0.15	0.32	0.26	0.33
100	0.15	0.36	0.30	0.36
150	0.15	0.38	0.32	0.36
200	**0.15**	**0.39**	0.33	**0.38**
250	0.15	0.38	0.33	0.38
300	0.15	0.38	0.33	0.38
400	0.15	0.37	0.33	0.38
500	0.15	0.37	**0.34**	0.38
750	0.15	0.35	0.33	0.38
1000	0.15	0.34	0.32	0.38
1500	0.15	0.33	0.32	0.38
2000	0.15	0.32	0.31	0.38

Table 1 – Accuracy values for increasing number of features ranked with ANOVA, after 10-fold Cross Validation.

In most cases, the higher result was obtained with the first 200 features, with the exception of SVM. However, the difference between the highest value for accuracy with SVM and the one for 200 features was considered negligible for analysis purposes. Thus, we considered the 200 highest ranked features for our following experiments when using all features. For metadata-features-only, we removed the audio-visual features from the set and ranked the best remaining features using the same approach, and keeping the top 200. In the case of using audio-visual features only, we considered all the 127 features, because reducing the features in such context only led to lower accuracy scores.

3.5 Analysis – Training Phase

In this section we performed an analysis of our selected models to have a perspective on how they learn in the training phase according to consecutive increases in the amount of training data, that is, how the

accuracy evolves.

In the case of the LR and SVM models (see Figure 1 and Figure 2), it was observed that validation accuracy converged in a similar manner when including all features compared to metadata-features-only, achieving an approximate value of 0.40 and 0.32 for validation accuracies, respectively. In contrast, when including audio-visual features only, it was observed that the validation accuracy was lower, meaning that the error rate increased, which means that the use of audio-visual features tended to a higher bias for both models. Finally, it was observed that the use of metadata-features-only slightly increased the gap between training and validation scores in comparison to the other sets of features, leading to a higher variance in both LR and SVM models.

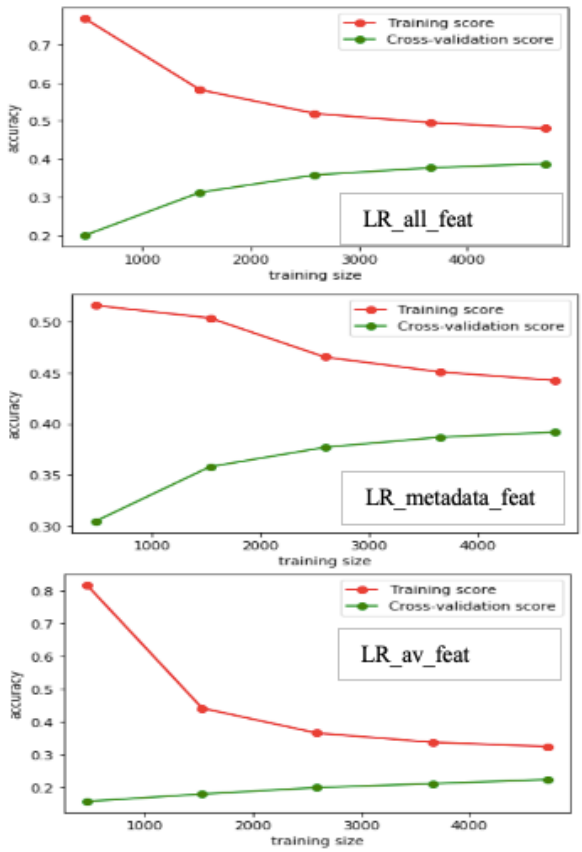

Figure 1 – Learning curves for LR model when including all, metadata and audio-visual features respectively.

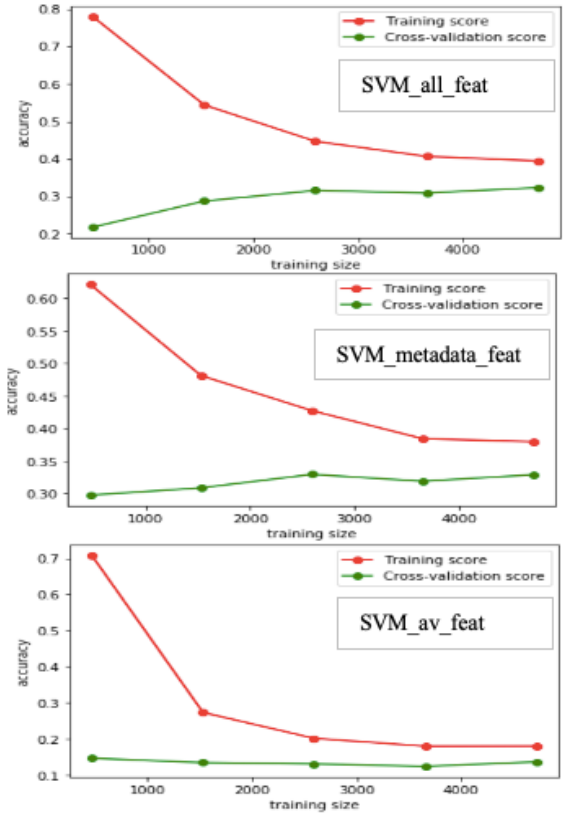

Figure 2 – Learning curves for SVM model when including all, metadata and audio-visual features respectively.

In the case of the XGB model (see Figure 3), it was observed that overall accuracy converged similarly when including all features compared to metadata-features-only, achieving an approximate value of 0.40 for validation scores in both cases. However, using all features tended to a larger gap compared to metadata-features-only, thus leading to a higher variance, but keeping a similar error level, thus having a similar bias level. In terms of using audio-visual features only, it was observed that validation accuracy was significantly lower and the gap increased, thus leading to a higher variance level and a higher bias compared to using all features and metadata-features-only.

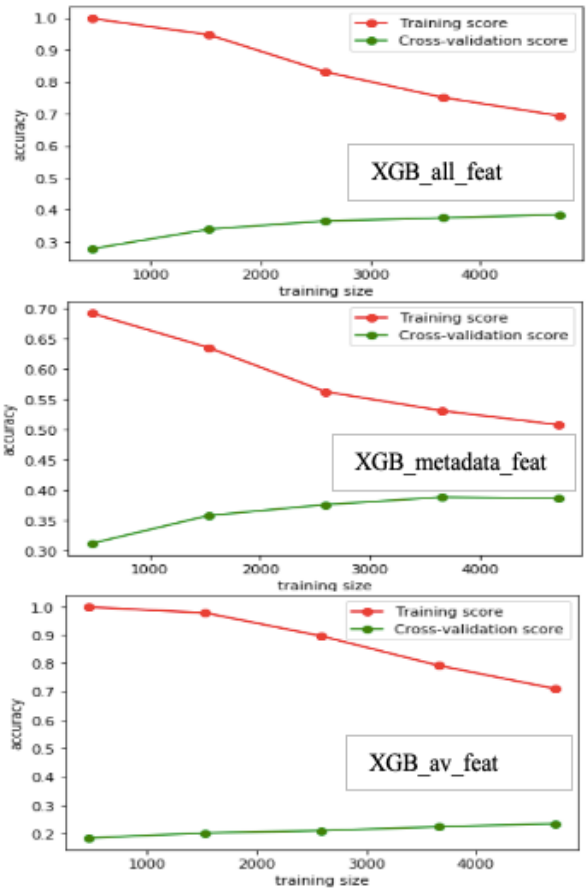

Figure 3 – Learning curves for XGB model when including all, metadata and audio-visual features respectively.

Among all the models, both LR and XGB with metadata-features-only showed a higher performance in training phase, although both having a similar and considerable overall error rate. Furthermore, when comparing LR and XGB in terms of gaps, it was observed that XGB presented slightly higher gaps when using all features and metadata-features-only, and a considerably higher gap when using audio-visual features only.

3.6 Analysis – Validation Phase

In this section we performed an analysis of the validation dataset for each model. For all that, we used the overall accuracy and the macro averaged values of precision, recall and F1 score (see Table 2).

			Models			
			0-R	**LR**	**SVM**	**XGB**
	Accuracy	All_feat	0.17	0.39	0.35	**0.42**
		Met_feat	0.17	0.38	0.32	**0.40**
		A-V_feat	0.17	**0.30**	0.17	0.25
Macro Average	Precision	All_feat	0.01	0.33	0.29	**0.39**
		Met_feat	0.01	0.36	0.28	**0.37**
		A-V_feat	0.01	**0.20**	0.19	0.13
	Recall	All_feat	0.05	0.29	0.29	**0.30**
		Met_feat	0.05	0.25	**0.31**	0.26
		A-V_feat	0.05	**0.23**	0.20	0.14
	F1 score	All_feat	0.02	0.29	0.27	**0.30**
		Met_feat	0.02	0.25	0.25	**0.28**
		A-V_feat	0.02	**0.21**	0.16	0.13

Table 2 – Accuracy and macro averaged values of precision, recall and F1 score for all the models.

According to Table 2, it was observed that the proposed models performed better compared to the 0-R baseline in all of the evaluated metrics. In general terms, it was observed that the performance of the proposed models tended to be better when utilizing all features and metadata-features-only, compared to using audio-visual features only. Thus, the use of metadata features tended to show lower generalization error for our proposed models, meaning that vectorized *tags*, *titles* and *years* tended to be better predictors for movie genre classes compared to audio-visual features.

In terms of overall accuracy, the XGB classifier was expected to outperform the linear models as it is an ensemble learner, which iterates sequentially to adjust the weights of each weak-learner to achieve better results at the end of the experiments. Thus, XGB classifier tended to show to the lowest generalization error, the lower bias and the highest accuracy among the datasets that included metadata. However, it was observed that LR tended to perform better when including only audio-visual features, compared to XGB. SVM model tended to show lower results in terms of overall accuracy.

In particular, averaged precision tended to be higher for XGB classifier when including all features and metadata-features-only, which means that metadata predictors tended to result in lower false positives for XGB model. In contrast, when using audio-visual features only, linear models tended to result in lower

false positives compared to XGB classifier, thus leading to higher averaged precision values.

In terms of averaged recall, when using all features the result among LR, SVM and XGB tended to be similar, which means that the amount of false negatives for the three models tended to converge to a similar value. Furthermore, linear models tended to perform better in terms of averaged recall when introducing audio-visual features.

In terms of F1 score, XGB tended to show better results when using metadata-features-only and a combination with audio-visual features. This means that the higher precision from XGB using such features tended to have a higher impact in averaged F1 score, thus overall false positives tended to be lower than false negatives. As previous analysis, linear models tended to show a better performance when using only audio-visual features.

Finally, it was important to notice that the difference in performance when using all features compared to the performance when using a combination of both types of features was not significantly high. This was related to our feature selection method, because when filtering all features, ANOVA tended to select mostly metadata features.

4 Improvements

We consider three areas of improvements for increasing the performance. Firstly, we consider that grid search algorithm could be applied when tuning the parameters of all of our models. Secondly, other techniques for feature reduction could be applied, such as LDA or PCA, although it might lead to lose interpretability. Finally, we suggest that other models could be considered as well, such as neural networks and its more advanced and complex versions for movie genre classification.

5 Conclusions

Our first hypothesis was empirically proven to be true, because the performance metrics of the proposed models were consistently higher compared to the baseline model. Our second hypothesis was not proven to be true, because there was not a significant difference between using metadata-features-only and using metadata combined with audio-visual features. However, we were able to show that there was a significant difference in performance measures when including metadata features compared to using audio-visual features only.

References

Deldjoo, Y., Constantin, M., Ionescu, B., Schedl, M., & Cremonesi, P. (2018). MMTF-14K: A Multifaceted Movie Trailer Feature Dataset for Recommendation and Retrieval. *In: Proceedings of the 9th ACM multimedia systems conference,* 450-455.

Ertugrul, A., & Karagoz, P. (2018). Movie Genre Classification from Plot Summaries using Bidirectional LSTM. *12th IEEE International Conference on Semantic Computing,* 248-251.

Harper, F., & Konstan, J. (2015). The MovieLens datasets: History and context. *ACM Trans. Interact. Intell. Syst.,* 1-19.

Simoes, G., Wehrmann, J., Barros, R., & Ruiz, D. (2016). Movie Genre Classification with Convolutional Neural Networks. *International Joint Conference on Neural Networks,* 259-26