Romance or Thriller? Movie Genre Prediction from Audio, Visual, and Text Features!

Alex González

1 Introduction

1.1 Overview

In this document is described the implementation of a Neural Network able to **predict a movie genre**. Most importantly, several properties of the data set were analysed through three different research questions or hypotheses. For this purpose, it was utilised a data subset derived from external resources, containing metadata, visual and audio features. These resources are presented as part of the related work revision in Section 1.2.

The hypotheses of this research are explained in Section 2. The remainder of the document is structured as follows: Section 3 describes the model developed, including the characteristics of the data set utilized and the neural network implemented. In Section 4 is presented the analysis carried out with the data set to generate relevant knowledge from the neural network model. Section 5 contains a discussion of the findings found in the previous section. Finally, in Section 6 the conclusions of the research are presented.

1.2 Related work and context

The data set provided for this research was extracted from the following resources:

- "MMTF-14K: A Multifaceted Movie Trailer Feature Dataset for Recommendation and Retrieval" [1]
- "MovieLens Datasets: History and Context" [2]

These studies are part of a broader investigation in this field in the last years. As context, the recommendation systems are a key element in the battle for user attention that content providers fight nowadays. In other words, more successful recommendations mean more time spent in front of a screen, more user's retention, and finally, more revenue for the content providers.

An example of this is the recommendation system that runs on YouTube, perhaps one of the most sophisticated systems to suggest videos to the users. This system is one the most important competitive advantages of this company. In "Deep Neural Networks for YouTube Recommendations" [3] it is described the system at a high level.

The improvement of the recommendations systems is a business priority in this industry. For example, in 2017 Netflix changed its whole rating system from a five-point scale (five-star) to a "thumbs up/down" rating [4]. The reason, it was not clear enough for the users that Netflix was making customized recommendations, therefore the users were not very likely to do many evaluations.

Those examples are taken directly from the content industry. However, extensive researches have been conducted in this field. The two papers already cited are instances of the progress in this topic. Additionally, it is necessary to mention two prominent investigations. First, in "Studying aesthetics in photographic images using a computational approach" [5], it is described a significant correlation between visual properties with a computational approach. This technique is the same used to process videos in later studies. Finally, Deldjoo [6], the same author of [1], makes a valuable contribution in "Exploring the Semantic Gap for Movie Recommendations" exploring how the content features of movies can be extracted and utilised to improve recommendations.

2 Hypothesis

The following are the hypotheses analysed in this research:

- Relevance of types of feature; for example, ¿are the visual features more relevant than the audio features? Or, according to the terminology used by Deldjoo in [6] ¿how does "the wisdom of the crowd" (e.g., tags) perform against the *Mise-en-Scene* features (video, audio)?
- Differences in the predictability of the movie genres. As hypothesis, some movie genres are more feasible to predict than others. These genres are presumed to

have specific characteristics that are common to all movies in the category.

- Feature relevance: some features might be not relevant at all. For example, the title of the movies are a source of diversity and creativity, it is difficult to expect useful trends in the movie titles that might be relevant for genre prediction.

3 Model

3.1 The data set

The originally provided data set consists in 132 fields distributed as follow:

- 5 metadata features: fields as title, year of release and a list of tags generated by humans.
- 107 visual features: floating value extracted from the movie trailer.
- 20 audio features: floating value extracted from the movie trailer.

Regarding the metadata features, the most important element is the "tag" field. This feature was processed to create 201 new fields, with one column for each tag concept. For the purposes of this report, these tags are referenced as **Tag features**.

Considering seven fields that were discarded, finally the data set contains 326 features to develop a machine learning model. 270 features were preselected according to its Mutual Information value. At the same time, the training data set contains 5.240 instances.

3.2 The neural network

Respect to the model, it was implemented a neural network in Keras[1] with 2 hidden layers. This model was selected due to its advantage for automatic feature learning, as it is suitable to analyse a great quantity of features (270). Apart from this, the neural networks are known to have very good performance.

Regarding the model results, an accuracy of 0.40 was obtained in the training phase and up to 0.42 in the validation phase. For the purposes of this report, this version will be referenced as the '**best model**'.

3.3 Baseline and benchmark

As baseline it was implemented a Zero-R method, predicting the most common class in the training data (Romance). This baseline achieves an accuracy of 0.17, which is considerably less than the accuracy achieved by the best model.

A latent benchmark of this effort corresponds to the results of the Kaggle competition which contextualize this research. In general, the levels of accuracy exhibited in that competition are similar to the achieved in this research. In fact, the mode for this metric is close to 0.4. However, this comparison corresponds to a preliminary evaluation.

4 Analysis

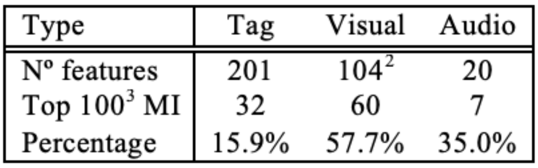
One of the disadvantages of neural networks is its lack of interpretability. In other words, it is difficult to understand the specific contribution of each feature to the model. For this reason, an analysis of the relevance of the features/type of features will be carried out through:

- Mutual Information score
- Impact in the model performance dropping the features/type of features

The limitations of these approaches are discussed in Section 5.

4.1 Feature type relevance

In Table 1 is presented for each type, the percentage of features that are in the top 100 attributes preselected by its Mutual Information (MI) score. Most of the characteristics correspond to those of visual type:



Type	Tag	Visual	Audio
Nº features	201	104[2]	20
Top 100[3] MI	32	60	7
Percentage	15.9%	57.7%	35.0%

Table 1 Perc. of selected features in top 100 by type

[1] Keras is an open-source neural-network library written in Python.

The accuracy obtained implementing the model without each feature type is presented in Table 2:

Model	Accuracy
Without Tag	0.22
Without Visual	0.39
Without Audio	0.40
Without Visual-Audio	0.39
Best model	0.40

Table 2 Impact of feature type in performance

These results are quite surprisingly. It seems that the visual and audio features only are contributing to improve the model accuracy in one percentual point. Alternatively, it might be a sub utilization of these features in the model design.

4.2 Predictability by genre

The Table 3 shows the metrics for predicted genres in the validation set.

Genres	Prec.	Recall	F1-score
Action	0.00	0.00	0.00
Adventure	0.00	0.00	0.00
Animation	0.00	0.00	0.00
Children	0.20	0.33	0.25
Comedy	0.47	0.47	0.47
Crime	0.50	0.20	0.29
Document.	0.64	0.50	0.56
Drama	0.32	0.67	0.44
Fantasy	0.80	0.44	0.57
Film_Noir	1.00	0.25	0.40
Horror	0.38	0.38	0.38
Musical	0.17	0.10	0.12
Mystery	0.80	0.22	0.35
Romance	0.36	0.35	0.36
Sci_Fi	0.73	0.69	0.71
Thriller	0.32	0.43	0.37
War	0.56	0.48	0.51
Western	0.00	0.00	0.00

Table 3 Predictability by genre

As it is possible to observe, some genres are more predictable than others. At the same time, it is expected that, in general, the model can make better predictions for genres with a higher number of instances in the training data set. To explore this, the Figure 1 plots the f1-score obtained for each genre and the number of instances.

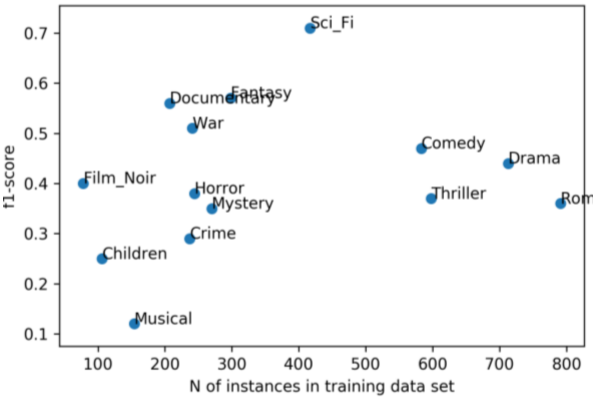


Figure 1 Nº of instances in train data set by f1-score

In general, with more data the model performs better predicting the genre. Even though, it is observed a group of genres with a poor prediction performance given their number of instances in the training data set: Comedy, Drama, Romance and Thriller. To double click on these findings, Table 4 exhibits a subsection of the confusion matrix. It is possible to explain most of the errors in this confusion matrix section through two findings:

- The Romance movies tend to be predicted as Comedy or Drama (**bolded**).
- The low performance for Thriller genre is explained because it is a common false positive of other categories (underlined).

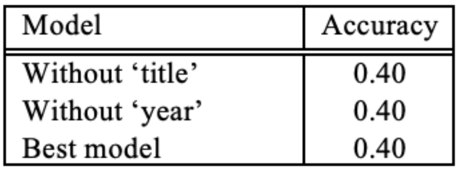
An interpretation for these findings is offered in Section 5.2.

Type	Comedy	Drama	Romance	Thriller
Comedy	25	3	6	2
Drama	7	27	3	4
Romance	**12**	**15**	16	6
Thriller	1	5	1	15

Table 4 Confusion matrix for specific features

4.3 Relevance of other features

In Table 5 is shown the accuracy of the model without considering specific features separately. As can be observed, the features 'title' and 'year' do not seem to be important at all for the model. Furthermore, for the field 'title' processed in the same way that the "tag" field, in order to generate additional features. This effort was unsuccessful due to the generated attributes was useless for improving the model.



Model	Accuracy
Without ‘title’	0.40
Without ‘year’	0.40
Best model	0.40

Table 5 Impact of features in performance

5 Discussion

5.1 Limitations of analysis

The limitations of the analysis carried out are:

- Mutual Information score: a known undesirable behaviour of this metric is that it might assign a high core to useless attributes like ‘id’. This kind of attributes has a great granularity and scores high for label prediction. However, it is not useful for the generalization of the model. Similarly, the field ‘year’ exhibits a suspicious high MI score but does not seem to have a concrete impact in the model.
- Impact in the model performance by dropping: the disadvantage of this approach is that it might underestimate the importance of the discarded attributes. If dropping an attribute ‘x’ does not have an impact in the model performance, perhaps it is because an attribute ‘y’ in the model has a high correlation with ‘x’. If 'y' was not in the model, the effect of dropping ‘x’ might be more significant and observable. However, it is difficult to consider all the possible correlations between features.

5.2 Hypotheses revision

Regarding the relevance of the type of features, the Tag fields seems to make the highest contribution to predict the movie genre. On the other hand, despite many visual features were selected for their MI score, these attributes had a limited impact for prediction. Similarly, the audio features exhibit even less merits to justify their presence in the model.

Related to the differences in the predictability of the movie genres, some genre types are more feasible to predict. A couple of explanations for the genres that resulted more unpredictable are the following:

- Many Romance movies are incorrectly predicted as Comedy. In fact, many movies are described as “romantic comedy”, which is subgenre by itself known as “romcom”[2]. Some iconic movies like “Pretty Woman” are classified in this category. The same explanation applies for “romantic drama”. For example, “Casablanca” is an iconic film classified with both genres.
- Many movies are incorrectly predicted as Thriller. In fact, the definition of this genres specifies that “Thriller films are typically hybridized with other genres”[3]. Given the above, it is reasonable that the prediction of this genre is difficult.

Finally, features as title and year were not helpful at all.

6 Conclusions

In this research, a series of analyses were carried out in order to improve and generate insights for movie genre prediction. A neural network model was developed for this task, utilizing a data set with tag, visual and audio features. The tag features were more suitable for this purpose. A different model design might take advantage of the visual and audio features. However, for the model implemented in this investigation, they don't make a substantial contribution.

On the other hand, some errors made by the model seems to be a consequence of the overlapping between genres. Perhaps a multilabel classification model deals more accurately with this issue. Alternatively, the data might be processed to obtain mutually exclusive movie genres.

Bibliography

[1] Y. Deldjoo, M. G. Constantin, B. Ionescu, M. Schedl and P. Cremonesi, "MMTF-14K: A Multifaceted Movie Trailer Feature Dataset for Recommendation and Retrieval," in *MMSys’18: 9th ACM Multimedia Systems Conference*, Amsterdam, Netherland, 2018.

[2] F. M. Harper and J. A. Konstan, "The MovieLens Datasets: History and Context," *ACM Transactions on Interactive Intelligent Systems (TiiS),* p. Article 19., 2015.

[3] P. Covington, J. Adams and E. Sargin,

[2] https://en.wikipedia.org/wiki/Romantic_comedy

[3] https://en.wikipedia.org/wiki/Thriller_film

"Deep Neural Networks for YouTube Recommendations," in *10th ACM Conference on Recommender Systems*, Boston, USA, 2016.

[4] N. Mcalone, "Netflix has replaced its 5-star rating system with 'thumbs up, thumbs down' -- here's why," Business Insider, 5 Apr 2017. [Online]. Available: https://www.businessinsider.com.au/why-netflix-replaced-its-5-star-rating-system-2017-4?r=US&IR=T. [Accessed 21 May 2020].

[5] R. Datta, D. Joshi, J. Li and J. Z. Wang, "Studying aesthetics in photographic images using a computational approach," in *European Conference on Computer Vision*, 2006.

[6] M. Elahi, Y. Deldjoo, F. B. Moghaddam, L. Cella, S. Cereda and P. Cremonesi, "Exploring the Semantic Gap for Movie Recommendations," in *Proceedings of the Eleventh ACM Conference on Recommender Systems*, 2017.