Prediction of movies rating
Machine Learning for Natural Language Processing 2020

Camille JEHLE
3A ENSAE voie PPE
camille.jehle@ensae.fr

Margaux THOREZ
3A ENSAE voie DSSA
margaux.thorez@ensae.fr

Abstract

We predict films rating using, in addition to the usual variables such as the runtime, variables created with Sentiment Analysis and Topic Model. These variables help us to have a better prediction but the increase is unfortunately not significant.

1 Problem Framing

Our project is to predict films rating using the overview variable and written reviews found on Amazon. We use the IMDB database and Amazon reviews. The data and the notebook of the project are available on GitHub. [1]

We want to predict the average note given by users on the IDMB website (for films with more than 100 votes). To improve our predictions, we use Amazon reviews of the movies [2]

2 Experiments Protocol

We use a Sentiment Analysis model on the Amazon review data to get the polarity of each review. We use two different approaches : one using Word2Vec as words embedding technique and a second one using the pretrained classifier of BERT (Devlin and Toutanova, 2018). We were finally able to create a new variable representing for each film the mean polarity of its reviews.

We also use a Topic Model (Blei, 2012) with 7 topics (a number that gives a good coherence score and a word distribution that looks sensible) on the overview of each film to create new variables.

The last step is the prediction of the ratings for each film. We split our data in train and test datasets. Different models are trained two times, with and without the NLP-created variables.

3 Results

We compare the models using the Mean Absolute Error (MAE) and the Mean Squared Error (MSE), for each model, and using or not the variables created using the NLP tools.

Model	MSE	MSE NLP
Linear Regression	0.334	0.334
XGBoost	0.260	0.246

Table 1: Comparison of the 2 models

The best model is XGBoost using the variables created using NLP new variables is the one that gives the best prediction. However the difference between models using or not our new variables is not significant.

4 Discussion/Conclusion

We were able to implement a sentiment classifier with an average accuracy of 92%. However, the use of NLP in our project don't allow us to improve significantly our predictions.

Several problems could explain why the predictive power of our NLP-created variables is low. The sample of Amazon reviews might not be representative of the entire reviews : we need more data. Another problem might be we've trained our classifier with IDMB review but Amazon review contains not just comment about the movie but also sometimes information on the time delivery : this might noise our predictions.

We believe our variables remains however interesting : it can be useful for example to predict TV consumption of television viewers if we know the programs they use to see but not the genres of these programs.

[1] https://github.com/Margauxxxxxxx/NLP_Project/blob/master/NLP_Project.ipynb

[2] We suppose we have a random sample of the total overview, so we can get a feeling of Amazon's users on the movies

References

David M. Blei. 2012. Probabilistic topic models. *doi:10.1145/2133806.2133826*.

Chang M.-W. Lee K. Devlin, J. and K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.