Film Synopsis Analysis using Robust Clustering and Robust Regression

Juliana Henao[1]

Advisor:
Henry Laniado [2]
Mauricio Toro [3]

Research practice 2
Final Report
Mathematical Engineering
Department of Mathematical Sciences
School of Sciences
Universidad EAFIT

NOVEMBER 2021

[1]Mathematical Engineering student. Universidad EAFIT jhenaoa4@eafit.edu.co (CvLAC: https://scienti.minciencias.gov.co/cvlac/visualizador/generarCurriculoCv.do?cod_rh=0001839296)
[2]Mathematical Science Department. School of Sciences. Universidad EAFIT. hlaniado@eafit.edu.co
[3]Informatics and Systems Department. School of Engineering. Universidad EAFIT. mtorobe@eafit.edu.co

Abstract

Film Industry is one of the most important industries in the world, making a big impact in our society. But film making requires many resources. Producers are interested in finding the future commercial performance of a film project. With the help of Natural Language Processing and robust clustering and regression methods; the rating given by the audience can be predicted and patterns can be found within the films that have the most success in the commercial market. This research uses some of these techniques, on the data set formed by the processing of the synopsis of a film.

Keywords: NLP, Films synopses, Robust Clustering, Robust Regression, BERT.

1 Introduction

Cinematography is a way of art that everyone has enjoyed in their lives, the seventh art has a direct impact in society, culture and economics. That is one reason why the film industry is one of the most important in the world and worth billions of dollars. But it is well known that the cost of the production of a film can be very high, and the budget directly affects the audience rating; this is why the accuracy when predicting financial performance is important when making an investment.

Usually, producers and investors make the investment decision based on the plot or the synopses proposed by the director or the writer. But when commercial objectives are involved, personal opinions can overlook important details. Here, data science takes place as a very good alternative.

Data science is a very useful tool to find implicit patterns that are intuitively perceived but difficult to grasp on their own. So, Machine Learning, Deep Learning and Artificial Intelligence have reached the enough development to be applied to creative industries like music, painting, literature, and of course cinematography, which seemed impossible a few decades ago.

One application in the motion picture industry is predicting which films people will want to see. Researchers from the 20th Century Fox film studio say that understanding the detailed audience composition is important for movie studios that invest in stories with uncertain commercial outcomes (Hsieh *et al.*, 2018). This film positioning can be based on many different criteria, based on genre, on directors value, synopses and others. Besides, the analysis of audience success based on genre can be done in many different ways, by analysing different components of a movie or any kind of motion picture. For instance, analysing movie synopses, video data, scripts and movie trailers.

For these reasons, there is a great opportunity for research in this field. In this project the approach that is going to be taken is the synopses analysis using Natural Language Processing (NLP), Robust Clustering Methods and Robust Regression Methods. Since much of a film's most relevant information is found in its synopsis, it becomes interesting not only to identify clusters within the data but also to explore the relationship between the plot and the film's rating. This is very useful when a producer wants to predict the rating a movie is going to receive, without having to produce the entire film, this is one motivation of taking this approach.

Having this in mind, this paper includes four sections. First, a research on the literature, about similar works on this matter that have contributed to the field. Then, the solution method is

presented, describing the techniques and algorithms that were developed during the research. The two following sections of the paper are the results, showing the performance of the algorithms, and conclusions, summarizing the principal findings and proposing future research.

2 State of the art

Natural Language Processing has been studied and applied in many fields, including sentiment analysis- where an NLP model can predict the kind of sentiment that a piece of text it operates upon expresses, virtual chat-bots- which are robots that interact with humans via text, having the ability to understand and provide logical responses to text messages sent by humans, speech recognition - technology commonly used in speech-to-text converters, now commonplace in mobile phones and video streaming websites, and several others. The most important techniques in NLP are (Zappy AI, 2021):

- Recurrent Neural Networks (RNN's) are variants of regular feedforward fully-connected neural networks having memory in their models.
- Attention Mechanism that allowed for more parallelism than RNN's during training a model (Vaswani *et al.*, 2017).
- The BERT model, which stands for 'Bidirectional Encoder Representations from Transformers, published by Google Research, uses the Transformer architecture with multi-headed attention. (Devlin & Chang, 2018).

As with NLP, with supervised and unsupervised learning, there has been much research in the art field. This includes projects such as categorizing music by similar audio features done by Dua (2020) and researching clustering approaches to select a set of exemplar images to present in the context of a concept done by Jing *et al.* (2010) and published by Google Research. But also these techniques have been applied in the cinematographic field, for instance, the 20th Century Fox film studio is using artificial intelligence to predict what films people will want to see (Hsieh *et al.*, 2018). Other research has analyzed this pattern using movie synopses (Campo *et al.*, 2018a) and movie trailers as in the article proposed by Campo *et al.* (2018b) and Lee *et al.* (2018).

In addition, by deepening the field of this project, ScriptBook (2020) provides artificially intelligent script analysis, AI driven content validation automated story generation, they have been building AI that analyzes and comprehends screenplays for decision support. In addition, AI Life has a script analysis project that extracts relevant information from scripts to describe it, compare it, summarize it, rate it and recommend adequate or similar alternatives (AI Life, 2020).

3 Solution method

In this research practice different algorithms which use Natural Language Processing were implemented, Robust Clustering Methods and Robust regression methods.

In this regard, the project has been divided into four phases as follows.

3.1 Data Collection

To conduct this research it was necessary to use the data files, i.e. the synopses to be analyzed and classified, provided by IMDb (2021), which are public access.

There was no specific data set containing the necessary labels for this project (title, rating, genres and synopsis). Which led to use web scrapping techniques to grasp this information from IMDb (2021).

The data was extracted from the Most Popular Movies and TV Shows tagged by different keywords in IMDb (2021).

3.2 Natural Language Processing with RoBERTa

The NLP uses a concept called word embedding. Word embeddings are representations of words as vectors, learned by exploiting vast amounts of text. Each word is mapped to one vector and the vector values are learned in a way that resembles an artificial neural network (Özgür Genç, 2019).

In order to get these vectors called tokens, the pretrained optimized method RoBERTa was used, an improved version of the self-supervised NLP technique BERT mentioned earlier.

RoBERTa builds on BERT's language masking strategy, in which the system learns to predict intentionally hidden sections of text within otherwise unannotated language examples (Liu *et al.*, 2019). RoBERTa was selected for this work because, as Liu *et al.* (2019) say, it modifies key hyperparameters in BERT, including removing BERT's next-sentence pretraining objective, and training with much larger mini-batches and learning rates. As it was shown in the paper of Liu *et al.* (2019), "this allows RoBERTa to improve on the masked language modeling objective compared with BERT and leads to better downstream task performance".

Hence, the implemented algorithm is described as follows:

1. Initialize the model using the library Transformers/ Hugging Face, and the pretrained model RoBERTa.
2. Calculate the tokens using the encoder that the model provides.
3. For each sentence of the text, take the average of the tokens. This is because the memory is not enough to compute all.
4. Detach the output from PyTorch.
5. Save the data in a CSV file.

3.3 Robust Clustering

Once the vectors representing the synopses were obtained, a robust clustering algorithm—described below—was implemented.

After obtaining the clusters, given the high number of columns, a Principal Component Analysis

(implemented in MATLAB) was performed. This dimensionality reduction allowed for graphical visualization of the clusters.

Trimmed k-means

This method is based on the work presented by García-Escudero *et al.* (2010). Trimmed K-means, is a method based on the same principles as the well-known K-means, with the difference that in this one, when calculating the centers, outliers that can affect the mean are trimmed. This method looks for the centers that solve the following minimization problem

$$\arg \min_{Y} \quad \min_{m_1,\ldots,m_k} \sum_{x_i \in Y} \min_{j=1,\ldots,k} ||x_i - m_j||^2$$

Where Y are subsets of size $[n(1-\alpha)]$ of the data sets. And the distance is calculated with the Euclidean Norm.

Taking this into account the algorithm to implement this method is the following:

1. Initial k centers are drawn at random.
2. Concentration steps.
   (a) A partition $\{H_1, ..., H_k\}$ is made, where H_j contains the observations closer to center j than to the other centers.
   (b) The set H is created, formed by the $[n(1-\alpha)]$ observations closest to the respective center.
   (c) The centers are updated such that each center is the sample mean of the observations in H_j.
3. Step 1 and step 2 are repeated several times in order to minimize the objective function.

This algorithm was implemented on MATLAB.

3.4 Robust Multiple Regression

In order to full fill the objective of predict the rating of a film having only the plot, the vectors encoding the synopses were used as explanatory independent variables, with the rating serving as the response variable.

Then, A multiple regression model was employed to model the linear relationship between the p tokens that encoded the synopses ($X = (X_1, \ldots, X_p)^\top$) where X_i represents the vector of each synopsis and the rating the film got in IMDb (Y).

The multiple regression model can be formally represented as

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \tag{1}$$

Three robust and non-parametric regression methods were implemented to estimate the parameters (β_0, β) in the Equation 1.

The estimation of these parameters can be calculated by solving the following optimization problem

$$\left(\widehat{\beta}_0, \widehat{\beta}^\top\right) = \operatorname{argmin}_{\left(\beta_0, \beta^\top\right) \in \mathbb{R}^{(p+1)}} \sum_{i=1}^{n} \left(Y_i - \beta_0 - x_i^\top \beta\right)^2$$

The solution of the above equation minimizes the sum squares of residuals, this solution can be written as

$$\widehat{\beta}_0 = \hat{\mu}_Y - \hat{\mu}_X^\top \widehat{\beta}_{\mathrm{LS}}, \quad \widehat{\beta} = \widehat{\Sigma}_{XX}^{-1} \widehat{\Sigma}_{XY}$$

where $\widehat{\Sigma}_{XX}$ and $\widehat{\Sigma}_{XY}$ are covariance matrix, and $\widehat{\mu}_Y$ and $\widehat{\mu}_X$ are empirical means of Y and X, respectively (Laniado *et al.*, 2020).

The first two regressions to review are variations of this multiple robust regression, this variations consist on the way of calculating the covariance matrix. This methods are explained bellow.

Pearson Regression

This regression use the Pearson´s rank correlation coefficient to calculate the covariance. This coefficient (ρ) is represented as:

$$\rho_{X,Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{\operatorname{Cov}(X,Y)}{\sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}}$$

For samples can be represented as

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

where,

- n is the sample size.
- x_i, y_i are individual sample points indexed with i.
- x denotes the sample mean defined by $x = \frac{1}{n} \sum_{i=1}^{n} x_i$.

Spearman Regression

This regression use the Spearman´s rank correlation coefficient to calculate the covariance. This coefficient (ρ) is represented as:

$$\rho = 1 - \frac{6 \sum D^2}{N\left(N^2 - 1\right)}$$

where D is the difference between the corresponding order statistics of x - y. N is the number of pairs of data

Nadaraya-Watson Regression

This is a nonparametric regression that represents the conditional expectation of a variable Y relative to a variable X in the following way

$$\mathrm{E}(Y \mid X) = m(X)$$

where m is an unknown function and it does not has to be lineal.

This function can use different kernels proposed in the literature this way

$$\widehat{m}_h(x) = \frac{\sum_{i=1}^{n} K_h\left(x - x_i\right) y_i}{\sum_{i=1}^{n} K_h\left(x - x_i\right)}$$

where K_h is a kernel with a bandwidth h.

$$h = \sigma\Big(\frac{4n}{3}\Big)^{1/5}$$

The kernel used in this project is the Epanechnikov, written as

$$K(u) = \frac{3}{4}\left(1 - u^2\right) \quad \text{for} \quad |u| \leq 1$$

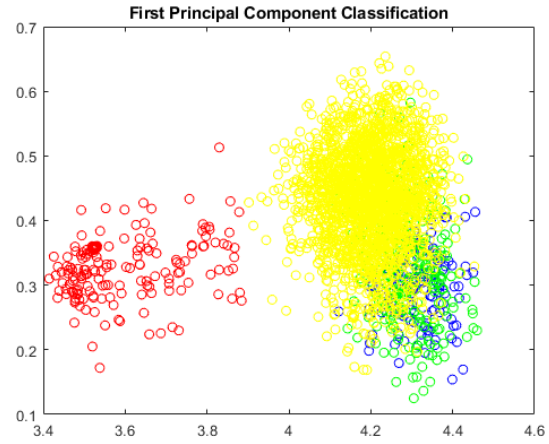
4 Results

In this section is described the performed experiments and the results obtained.

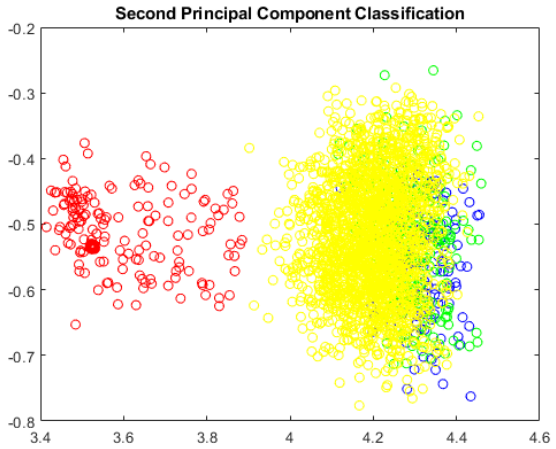
4.1 Robust Clustering

Trimmed k-means

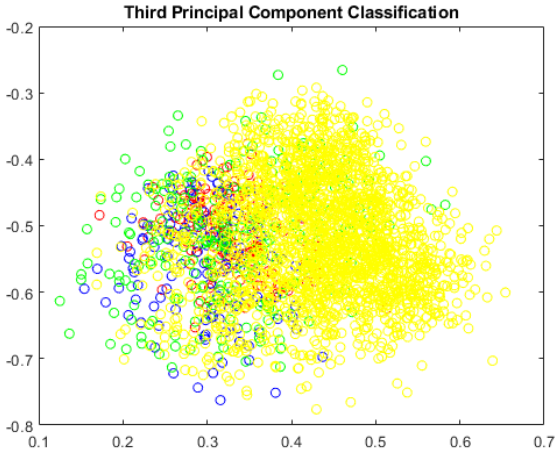
For this method the α used to trim the data was $\alpha = 0.1$.



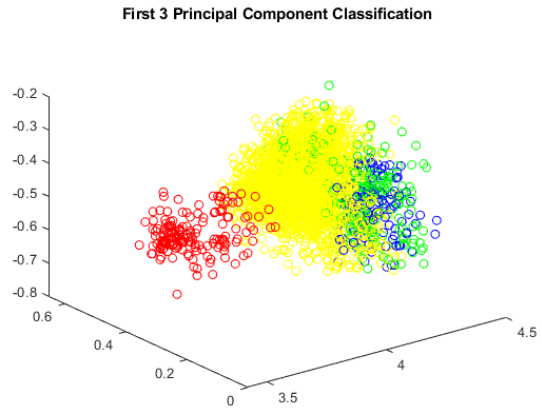
(a) First vs. Second principal component



(b) First vs. Third principal component



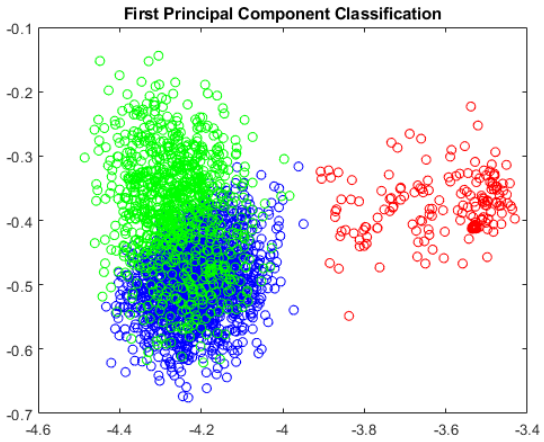
(c) Second vs. Third principal component



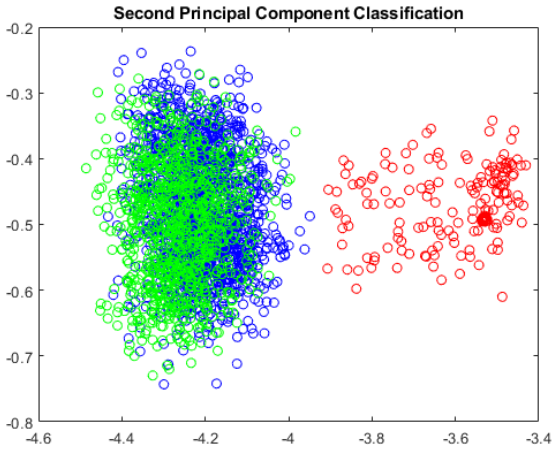
(d) First 3 principal component

Figure 1: Trimmed k-means clustering (4 groups), PCA view

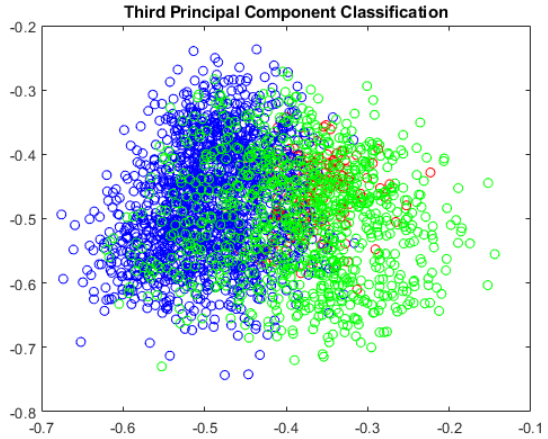
In Figure 1, it is evident that the red group can be well distinguished from the others. But the yellow, green and blue ones can not. The yellow one is better established, but the other two seems to be part of the same group. For this reason, the decision was made to reduce the number of groups to 3, as it is shown in Figure 2.



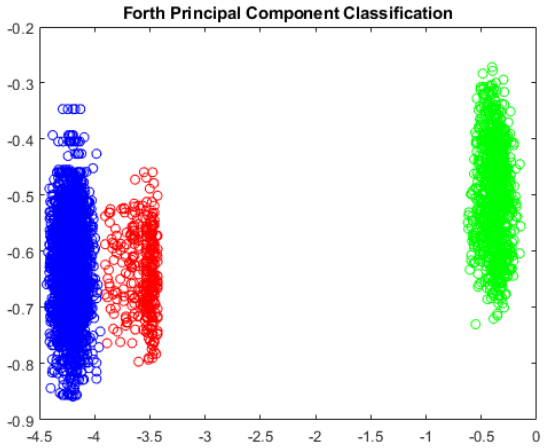
(a) First vs. Second principal component



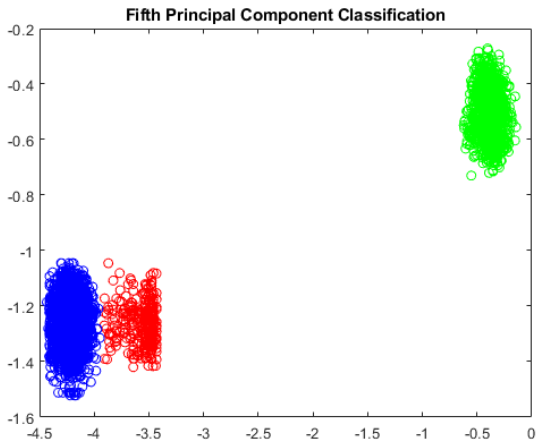
(b) First vs. Third principal component



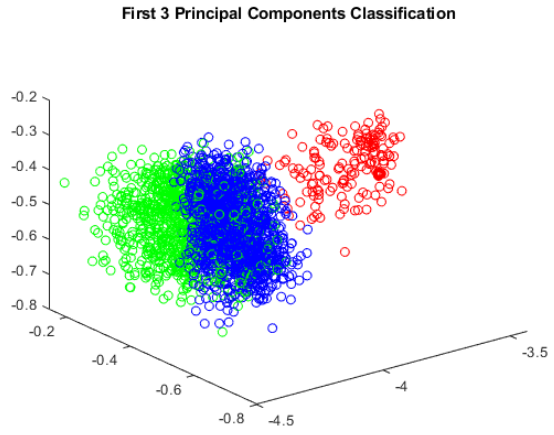
(c) Second vs. Third principal component



(d) First vs. Fourth principal component



(e) First vs. Fifth principal component



(f) First 3 principal component

Figure 2: Trimmed k-means clustering (3 groups), PCA view

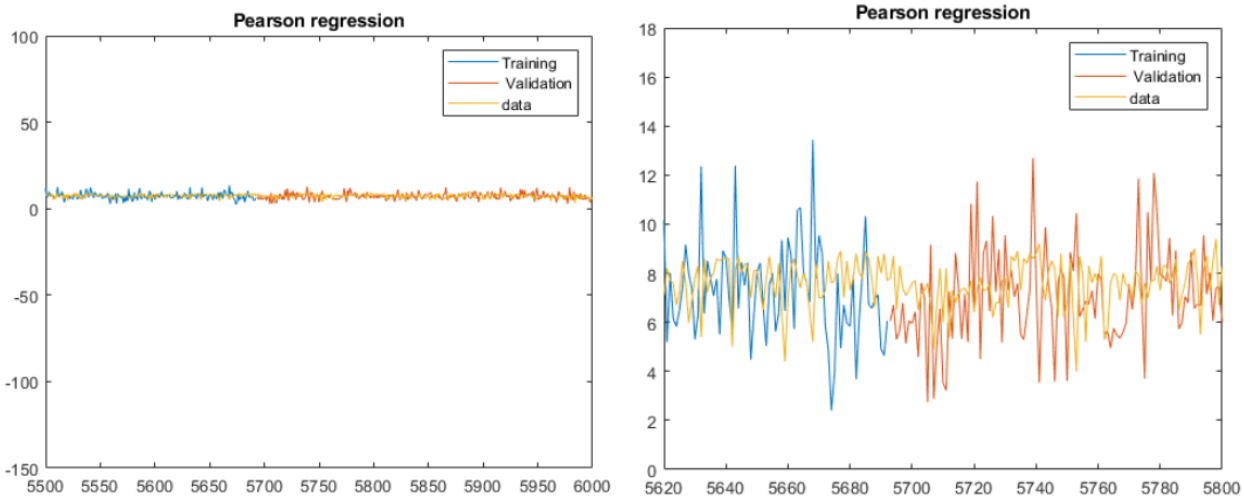
In Figure 2, the hypothesis suggested above is confirmed, it can be seen that there are clearly 3 groups in this data set. Since it is know that this data contains the codification of the synopses of many films, it can be said that there are three main topics that these synopses address.

4.2 Robust Regression

For the regressions, the first 80% of the data was taken, i.e. the ratings, to train the model and calculate the betas of the regression. The last 20% was used to validate the model with the betas calculated before.

Pearson Regression

In Figure 3, it can be seen the real ratings of the films and the estimation using the Pearson regression. The estimation is not extremely accurate, nevertheless, it follows the same trend of the real data.



(a) Data from 5500 to 6000

(b) Close up

Figure 3: Pearson regression

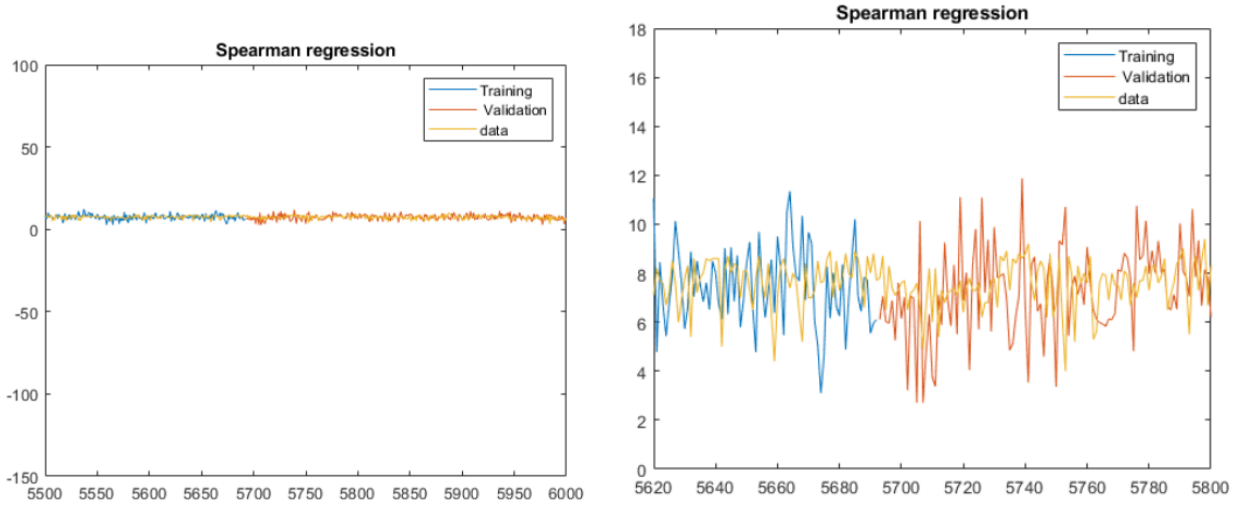
In order to see the correlation between the predicted data and the real rating, a scatter plot was made. In Figure 4, it is shown that there is no correlation between this variables, this leads to conclude that this regression is not the most appropriate for this data set.

Figure 4: Pearson regression

The correlation of the validation data and the predicted on is $\rho = -0.0089$. And the mean absolute percentage error is MAPE = 0.2446. This supports the hypothesis made above.

Spearman Regression

As in the Pearson regression, it can be seen the real data and the estimation have similar behaviour and tendency, without being extremely accurate. This result can be seen in Figure 5



(a) Data from 5500 to 6000

(b) Close up

Figure 5: Spearman regression

As with the Pearson regression a scatter plot was made in order to see the correlation between the predicted data and the real rating. In Figure 6, it can be seen that there is no correlation, this

leads to conclude that this regression is also not the most appropriate for this data set. This can be concluded also by seeing the correlation of the validation data and the predicted one, that is $\rho = 0.0211$. And the mean absolute percentage error is MAPE $= 0.2236$.

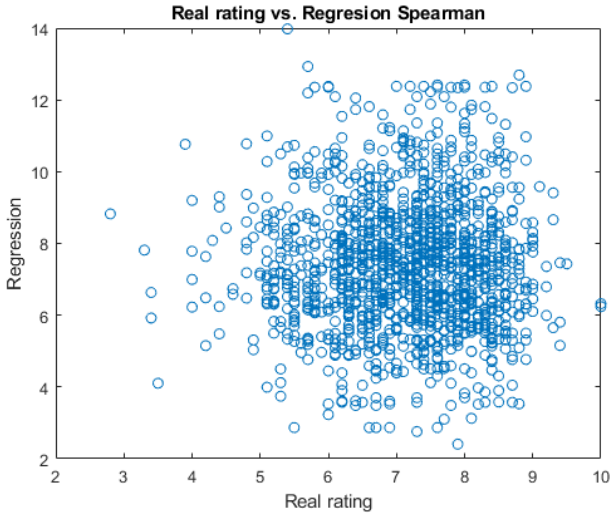
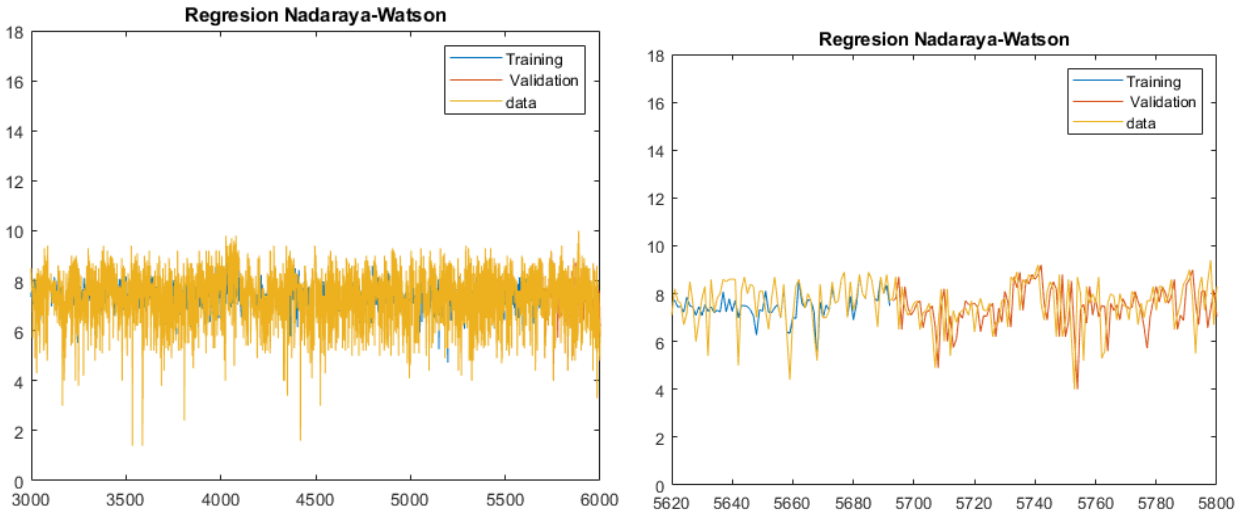


Figure 6: Spearman regression

Nadaraya-Watson Regression

Unlike the other two regressions, this Kernel regression gave much better results than the others. In Figure 7, the fitting of this regression is much more accurate.



(a) Data from 3000 to 600

(b) Close up

Figure 7: Nadaraya-Watson regression

To check the conclusions made by seeing the graph, a scatter plot between the real ratings and

the predicted ones is done in Figure 8. Here it seems that a linear relationship between these two variables exists.

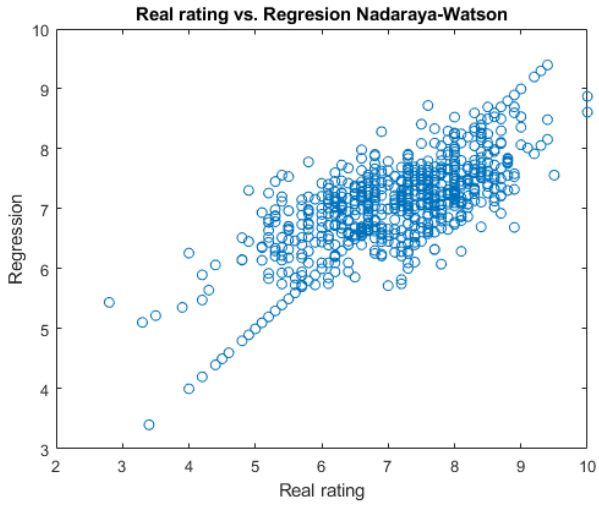


Figure 8: Nadaraya-Watson regression

The correlation of the validation data and the predicted is calculated as an indicator of the good fit of this model with the data. This coefficient is $\rho = 0.8534$. And the mean absolute percentage error is MAPE $= 0.0457$.

5 Conclusions and future research

One finding of this research was the fact that it was seem that there are 3 main topics that these synopses are about. These topics can define the genre of the film in matter. Also, for commercial purposes is valuable information knowing that the top films of the moment, basically, are about these 3 topics.

The robust regressions applied to coded synopses, can be a useful tool to predict the commercial outcome of a film. These methods gave good and not far from reality estimations of the ratings. Based on the results, it can be concluded that for the purposes of a film production company, the Nadaraya-Watson, is the most accurate to approximate the audience reception of the film before having to produce the entire film.

For a future research, it can be browsed an entire analysis of the whole audiovisual content of a film, and with the same objective, analyze the trailer of a movie before it is completely produced. Also, it might be interesting to explore another NLP techniques and to do a sentiment analysis of the different clusters found.

References

AI Life. 2020. *Script Analysis.*

Campo, Miguel, Espinozaa, JJ, Rieger, Julie, & Taliyan, Abhinav. 2018a. Collaborative metric learning recommendation system: Application to theatrical movie releases. *ArXiv*.

Campo, Miguel, Hsieh, Cheng-Kang, Nickens, Matt, Espinozan, JJ, Taliyan, Abhinav, Rieger, Julie, Ho, Jean, & Sherick, Bettina. 2018b. Competitive analysis system for theatrical movie releases based on movie trailer deep video representation. *ArXiv*.

Devlin, Jacob, & Chang, Ming-Wei. 2018. Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing.

Dua, Sejal. 2020. K-Means Clustering and PCA to categorize music by similar audio features.

García-Escudero, Luis Angel, Gordaliza, Alfonso, Matrán, Carlos, & Mayo-Iscar, Agustín. 2010. A review of robust clustering methods. *Advances in Data Analysis and Classification*, 89–109.

Hsieh, Cheng-Kang, Campo, Miguel, Taliyan, Abhinav, Nickens, Matt, Pandya, Mitkumar, & JJ, Espinoza. 2018. Convolutional Collaborative Filter Network for Video Based Recommendation Systems. *ArXiv*.

IMDb. 2021. *Sort by Popularity - Most Popular Movies and TV Shows tagged by keywords*.

Jing, Yushi, Covell, Michele, & Rowley, Henry A. 2010. Comparison of Clustering Approaches for Summarizing Large Populations of Images. *In: Proceedings ICME VCIDS*.

Laniado, Henry, Velasco, Henry, Toro, Mauricio, Leiva, Víctor, & Lio, Yuhlong. 2020. Robust Three-Step Regression Based on Comedian and Its Performance in Cell-Wise and Case-Wise Outliers. *MDPI, Mathemathics*.

Lee, Joonseok, Abu-El-Haija, Sami, Varadarajan, Balakrishnan, & Natsev, Apostol Paul. 2018. Collaborative deep metric learning for video understanding. 481–490.

Liu, Yinhan, Ott, Myle, Goyal, Naman, Du, Jingfei, Joshi, Mandar, Chen, Danqi, Levy, Omer, Lewis, Mike, Zettlemoyer, Luke, & Stoyanov, Veselin. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv*.

ScriptBook. 2020. *Deepstory*.

Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N., Kaiser, Lukasz, & Polosukhin, Illia. 2017. Attention is All You Need.

Zappy AI. 2021. *The Current State of the Art in Natural Language Processing (NLP)*.

Özgür Genç. 2019. The basics of NLP and real time sentiment analysis with open source tools.