Sentiment Analysis on Movie Scripts and Reviews

Utilizing Sentiment Scores in Rating Prediction

Paschalis Frangidis, Konstantinos Georgiou, and Stefanos Papadopoulos

School of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece
{frangidis,georgiouka,stefanospi}@csd.auth.gr

Abstract. In recent years, many models for predicting movie ratings have been proposed, focusing on utilizing movie reviews combined with sentiment analysis tools. In this study, we offer a different approach based on the emotionally analyzed concatenation of movie script and their respective reviews. The rationale behind this model is that if the emotional experience described by the reviewer corresponds with or diverges from the emotions expressed in the movie script, then this correlation will be reflected in the particular rating of the movie. We collected a dataset consisting of 747 movie scripts and 78.000 reviews and recreated many conventional approaches for movie rating prediction, including Vector Semantics and Sentiment Analysis techniques ran with a variety of Machine Learning algorithms, in order to more accurately evaluate the performance of our model and the validity of our hypothesis. The results indicate that our proposed combination of features achieves a notable performance, similar to conventional approaches.

Keywords: Natural Language Processing · Sentiment analysis · Machine learning · Prediction

1 Introduction

Movie scripts are an interesting source of text, due to the diverse display of sentiments expressed in them. In principle, they are a storytelling device where the screenwriter is trying to convey something meaningful. The script's "emotional charge" is usually a tool to achieve the aforementioned goal or a byproduct of the process- and indeed a very important and powerful one. It is probably the most immediate point of resonance and communication with the audience. Movies usually contain scenes where emotions alter dynamically, between happiness and sadness, calmness and anger in order to aid the narrative progression, while some works are characterized by an overarching emotional 'weight', such as sadness in a tragedy-based film. In order to achieve this goal, the script needs to be written in a manner that captures the appropriate sentiments and allows the actors to portray it in their performances. It is very common nowadays, for the audience and critics alike to express their enjoyment of a movie or lack thereof, in the form of online reviews. In sites like "Metacritic" or "Rotten Tomatoes" reviews, analyses and criticisms are collected to create an average estimation of how popular and well-received

Published by Springer Nature Switzerland AG 2020
I. Maglogiannis et al. (Eds.): AIAI 2020, IFIP AICT 583, pp. 430–438, 2020.
https://doi.org/10.1007/978-3-030-49161-1_36

a movie is. This information has already been exploited in many applications including movie recommendations and genre classification systems. However, most reviewers do not simply state their like or dislike of a movie and instead tend to focus on the different feelings that the movie evoked in them. It would be an interesting inquiry to analyze the intended emotion of a movie's script, compare it to the received emotional response of the reviewer and study whether the unison or dissonance between the two are correlated or not. The primary goal of this study is to accumulate movie scripts and their respective reviews in order to examine the validity of the aforementioned statement and examine whether the relationship between the intended emotional weight of the movie and the received emotion of the reviewer can help in accurately predicting movie ratings. Our proposed model was tested in conjunction with more conventional approaches by running multiple experiments with different combinations of features like Vector Semantics, typical Sentiment Analysis and using different Machine Learning algorithms.

The remaining chapters are structured as follows: in Sect. 2, we present some related work and pinpoint our contribution to the topic. In Sect. 3, we analyze the dataset and present our key features, and in Sect. 4 the methodology applied is explained. In Sect. 5, the basic results of our study are presented and discussed, while in Sect. 6 we provide our main conclusions, accompanied by some interesting future work suggestions.

2 Background and Related Work

Sentiment analysis, as a field of study, has met a significant increase in its applicability in various sectors, as it can easily become cross disciplinary and facilitate different processes [8, 15, 16]. Its simplicity and straightforward approach have established it as a respected factor of text analysis. Indicative sectors where sentiment analysis is implemented are business systems, marketing campaigns and recommender systems [3].

In recent years, sentiment analysis in conjunction with NLP and ML techniques have been used in an array of different applications concerning movie scripts and their reviews. They have been used in order to identify patterns in movie structures [2] and learn to predict the following emotional state based on the previous [6], showing respectively that 'successful' movies follow specific narrative progressions and have a certain "flow" and consistency in the way that emotional states unfold. Additionally, it has been observed that existing binary ("positive/negative") sentiment analysis techniques have a 'positivity bias', favoring the learning of positive emotions over negative ones, hence "underestimating" the latter's existence. This can create a significant discrepancy in accuracy of around 10 to 30%. It is shown that taking into consideration meta-features (capital letters, punctuation and parts of speech) helps mitigate this problem [4].

Other approaches have tried to classify reviews based on semantically similar words [3] in order to detect communities of reviewers via clustering. Some interesting alternatives propose the identification of the driving aspects of a movie, them being mainly associated with the screenplay, the acting and the plot or some more particular characteristics (e.g. music, effects) [7], and how they are reflected on the reviews in sentiments. Results indicate that the acting and the plot are usually the most important factors that influence a review. However, variations in these studies which examine the same aspects in different movie genres may alter the driving factors in respective reviews. Typically, the methodology implemented to perform sentiment analysis, directly involves classic text mining solutions such as N-Gram extraction or Part-of-Speech tagging in combination with Naïve Bayes (NB) or Recurrent Neural Networks (RNN) [9]. However, different techniques, such as the use of a Gini Index or Support Vector Machine (SVM) approaches [10, 11] have increased the accuracy of results. The Skip Gram and Continuous Bag of Words (CBOW) models [3] have also shown promising results.

A recent research venture [1] has tried a mixture of techniques to extract the emotion out of reviews. Using a combination of emotion lexicon and word embeddings in order to extract reviews' sentiment, they acquire a satisfactory level of prediction to the reviewers' binary score of the movie. Moreover, it provides solutions for both English and Greek datasets. This particular approach is going to be extended in the current research. Regarding the classification of movie scripts based on NLP, there have been attempts in using subtitle files of different movies to ultimately predict their genre [8]. The premise is that via sentiment analysis of the subtitles' context, sentiment can be extracted and thus, the genre of the movie can be identified. Initial findings show that these techniques yield better results when applied to action, romance or horror movies. However, further analysis could strengthen the prediction margin.

The arising importance of sentiment analysis in movie and review evaluation has also prompted well known and respected online competition networks like Kaggle [11] to organize initiatives in order to suggest innovative solutions to the problem. It is evident, though, that in order to produce a well-documented solution, features have to be carefully selected, taking into consideration the variance of terms, the handling of negation and the treatment of opinion words [10].

Our paper contributes to the current scientific framework by exploiting data both from movie scripts and their corresponding reviews and by applying two specialized lexicons in order to conduct sentiment and emotion analysis. We believe our research to be of notable importance as it can indicate a new approach to the problem which treats the review text and the movie script as coexisting entities and merges them for optimal results.

3 Data Collection and Preprocessing

In order to gather data for movie scripts and movie reviews, we used simple scrapers in Python to crawl specific web pages. For the movie scripts, we used the IMSDB website, which contains more than 1100 movie scripts and drafts. Our reviews were gathered from the Rotten Tomatoes website, a well-known source for reviewing films, also by using a web scraper. The scripts gathered were in.txt format and the reviews were saved in a CSV file. Before continuing to the preprocessing phase, we noticed a significant imbalance in our review dataset which contained 55.886 fresh/positive movies and 22.193 rotten/negative. This imbalance would create inaccurate predictions in our model and in order to correct it, we performed undersampling, keeping all of the "rotten" and 25.000 of the "fresh" reviews. "Fresh/Rotten" values were transformed into 0 and 1 and were used as the model's target variable.

The scraper utilized for obtaining the movie scripts has a quite simple and easy to understand structure. It uses the Beautiful Soup package and redirects to the IMSDB website, locating all the script titles, categorized in alphabetical order. The scraper's first operation is to store all the <a> tags, corresponding to movie titles, in a list. Afterwards, it iterates the list of titles and redirects to the corresponding URLs, where it loads scripts in.txt files and creates a directory to store them. The scraper keeps only data present in <body> tags and opts to ignore tags like <head> or <footer>.

To obtain the corresponding movie reviews, we modified a Beautiful Soup based scraper which collected review texts and stored them to a shared CSV file. Our data were limited due to the reason that Rotten Tomatoes only contains an excerpt of the original review. During the collection phase, we filtered reviews which contained no text, as we deemed them unsuitable for our model.

In the preprocessing phase, we applied sentence splitting and word tokenization. Punctuation marks and stop-words were removed but only after experimenting and keeping track of the results with their inclusion. We then created a punctuation-removal list which consisted of all the punctuation marks we aimed to remove. At first, only exclamation marks were excluded from the punctuation list and the words "not" and "but" from the stopword-removal list in order to study their effects on the VADER sentiment analysis, as discussed in the next section. Their effect on the predictive accuracy was considered negligible, hence all stop words and punctuation marks were eventually removed. Finally, we applied part-of-speech tagging and lemmatization because we noticed that the lexicon-based sentiment analysis tools we selected lacked many derivatives and would lose out on a significant segment of our dataset.

4 Methodology

Our proposed model is based on calculating the emotional weight of a movie's script and combining it with the emotion expressed by the reviewer. For Sentiment and Emotion Analysis, we considered many different tools but ended up selecting VADER and NRC. VADER (or Valence Aware Dictionary and sEntiment Reasoner) is a binary Sentiment Analysis tool using a dictionary approach, containing 7.518 uni-grams including punctuation, slang words, initialisms, acronyms and emoticons. VADER receives a sentence as input and returns 4 values, negative, neutral, positive and compound which is the 'normalized weighted composite score'. Each output is ranging from $-1{,}5$ to $+1{,}5$, from 'Very Negative' to 'Very Positive', but we normalized them into a range of $\{0$ to $1\}$. VADER is widely used [12, 13] and preferred as a sentiment analysis tool because of its advanced heuristics. It includes 5 built-in and pre-trained heuristics taking into consideration punctuation marks (especially exclamation points), capitalization, degree modifiers (boosters and dampeners) Shift Polarity (with words like "but") and Negation Handling using tri-grams.

NRC, on the other hand, is an emotion analysis tool, categorizing a sentence into eight emotions (anger, fear, anticipation, trust, surprise, sadness, joy, disgust) and two sentiments (positive, negative). Regarding the selected emotions in the NRC lexicon, there is a consensus among Psychology Researchers about Joy, Sadness, Anger, Disgust, Surprise and Fear being categorized as 'Basic Human emotions' showing their importance and universal nature [14]. In addition to these six emotions, the NRC contains words about Trust and Anticipation and although they may not be considered as 'basic emotions' by everyone, we believe them to be important especially for analyzing texts of movie scripts and their reviews. This lexicon contains more than 14000 words, with each one being scored on every emotion. We had to calculate the average of these ten emotion scores for every review, after having summed the scores for every word. We hypothesize that NRC is better suited for the model we are trying to build due to the complexity of movie scripts but both sentiment analysis tools were tested empirically.

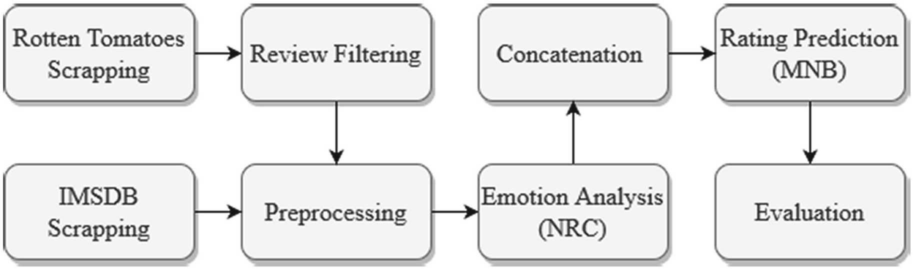

Fig. 1. Applied workflow

Our model requires the concatenation of the emotionally analyzed of both the movie scripts and their reviews using the NRC lexicon (Fig. 1). In order to evaluate our model more reliably, we reproduced many different conventional methods for movie rating predictions. As features we used Vector Semantics like CountVectorizer and TF-IDF in combination with NRC and VADER. Indicatively, the experimental combinations that were tested on the movie reviews alone were:

1. TF-IDF.
2. CountVectorizer.
3. VADER.
4. VADER combined with NRC.
5. VADER (including stop words 'not and but') combined with NRC.
6. CountVectorizer (CV) combined with VADER.
7. CV combined with NRC.
8. CV combined with VADER and NRC.

The dataset was split into train and test data, by 75% and 25% respectively. Finally, regarding the selection of machine learning algorithms we experimented with the following algorithms: Multinomial Naive Bayes (MNB), Logistic Regression (LR), SVM and Multilayer Perceptron (MLP). All algorithms were able to return satisfactory results but MNB was selected because it consistently resulted in higher predictive accuracy (Table 1) and was significantly faster computationally. We selected to test the four machine learning models against two feature combinations, CV as a representative of Vector Semantic approaches and CV in combination with NRC to add information about the emotions into the equation.

Table 1. F1-scores of the four compared ML models.

	MNB	LR	MLP	SVM
CV	**0.807**	0.791	0.789	0.565
CV with NRC	**0.768**	0.747	0.702	0.422

5 Results

The evaluation of the aforementioned experiments was conducted with the use of Accuracy, Precision, Recall and F1 as the selected metrics. The bar plot below (Fig. 2), presents an indicative selection of experiments, from the total that were performed, which we considered to be the most important. Based on the visualized results, it seems that the CountVectorizer method produces slightly better results than TF-IDF method, possibly because of better document representation. The VADER lexicon showed the worst performance when used on its own. Additionally, it appears that its heuristics, despite increasing the intensity of the sentiment, did not play a crucial role in improving the predictive accuracy. VADER's low performance was expected, as we initially hypothesized that binary sentiment analysis (positive/negative) was too simplistic to capture the complexity of a movie's script. We believe that the problem lies in the interpretive ambiguity of binary sentiment approaches. A negative sentiment score for a review can have two very different meanings. It may show legitimate dislike for the movie or it may express the reviewer's experience of sadness or horror which can technically be considered as "negative" emotions but are expected in an effective and well-made drama or horror movie respectively. Furthermore, confirming our hypothesis, the NRC lexicon greatly outperforms VADER by 10%, proving the importance of taking into consideration a multitude of emotions.

An unexpected finding is that the combination of Vector Semantics approaches with sentiment lexicons yields significantly different results based on which lexicon is applied. The VADER lexicon slightly enhances the performance of the CountVectorizer - if only by 0.1% - while the NRC lexicon reduces the performance. This can possibly be attributed to the relative increase in complexity of the model when combined with the multiple emotions of the NRC lexicon, which led to constrained results. Finally, our proposed model, which took into consideration both the review and the script emotions without any other NLP technique, managed to reach impressive precision percentages, only 0.8% lower than the best performing model but didn't quite reach the same level of accuracy. While not being the best performing model, it indicates the potential and validity of our initial hypothesis which dictated that the relationship between the expressed emotion of the movie and the received emotion by the reviewer can be a potent predictor for movie's rating.

Generally, Vector Semantics approaches (TF-IDF and CountVectorizer) performed much better than simply using sentiment and emotional lexicons. They seem better at identifying the importance of each word inside the text which helps the machine learning model to comprehend its structure and any possible hidden patterns. Sentiment analysis tools are already being used in a variety of cases, but they are still in a transitional stage.

NRC, especially, doesn't take at all into consideration the context of a word - it simply returns its pre-classified score - which we believe to be the most important reason for its lacking performance. Another factor that probably contributed to our model's restrained performance was the absence of the totality of each review. We only had a small excerpt from the whole review provided by Rotten Tomatoes. By analyzing the review in its totality it's possible that our model's predictive accuracy would have improved.

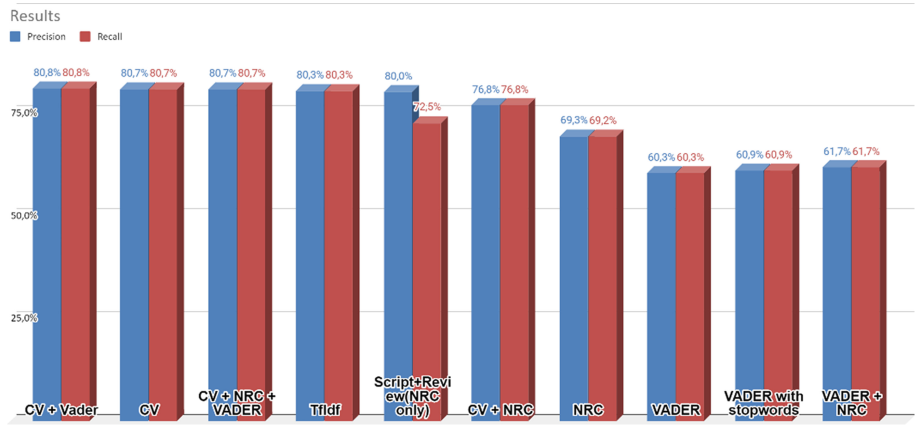

Fig. 2. Evaluation scores of the proposed model and the baseline methods. The proposed model is referred as "Script + Review (NRC only)".

6 Conclusions – Future Work

This paper explored the effect of incorporating emotion and sentiment analysis in predicting movie ratings and more specifically, the combinations of the movie script's intended emotion and the emotional response of the reviewer. Data were collected from relevant sources and contained a fair amount of scripts and reviews. We tried several machine learning models and experimented with the usage of sentiment and emotion analysis with two well-known lexicons, one intended for sentiment analysis and one for emotion analysis. We examined their performance on their own and in combination with Vector Semantics tools. The performance of our proposed model was actually quite impressive, almost reaching the best performing model in precision but we believe that the results were limited by the lack of negation handling and the fact that we used only excerpts of the reviews, provided by Rotten Tomatoes, and not the whole text. We believe that the accuracy of the model would improve if we were able to obtain the whole body of the reviews and thus have a larger corpus for analysis. However, the results may still end up being unsatisfying, due to the absence of negation handling and polarity shift which should also be tested. Of course, such enhancements would require different preprocessing of the data by keeping appropriate stop words or applying different representation methods.

Regarding suggestions for improving similar models in the future, we thought it would be useful to incorporate character and contextual analysis. Although sentiment can indeed be extracted from simple words, there should be additional weight to some character names. For example, the names of some famous villains can easily be associated with a specific sentiment, like fear, and thus should be considered non neutral words when conducting an analysis of sentiment. This can lead to better results in many cases and better characterization of the script's context as a whole.

Another possible expansion of the model would be the implementation of different document representation methods, such as Word2Vec embeddings or the use of N-Grams. It would be interesting to explore how different methods affect the quality of results and whether the combinations of lexicons and approaches would have similar outputs. Finally, we considered that a possible practical use of our model would be to expand its use not only in reviews but other forms of criticism like YouTube Comments or Tweets and apply the same methodology to give an early estimation of a movie's box office success or failure. It is obvious that such predictions will only be vague, as a movie's gross depends on a variety of factors (marketing, production, cast, merchandise etc.) but this model can certainly serve as a supplementary tool for further assurance of success.

Acknowledgements. This paper was directly based on an assignment we completed during our "Text Mining and Natural Language Processing" Course in the Data and Web Science MSc program of the School of Informatics of the Aristotle University of Thessaloniki. We would like to personally thank Dr. Grigorios Tsoumakas, our supervising professor, who provided direction, reviewed the project and encouraged us to submit it as a conference paper.

References

1. Giatsoglou, M., Vozalis, M., Diamantaras, K., Vakali, A., Sarigiannidis, G., Chatzisavvas, K.: Sentiment analysis leveraging emotions and word embeddings. Exp. Syst. Appl. **69**, 214–224 (2017)
2. Lee, S., Yu, H., Cheong, Y.: Analyzing movie scripts as unstructured text. In: Proceedings of IEEE Third International Conference on Big Data Computing Service and Applications 2017 (BigDataService), pp. 249–254. IEEE, San Fransisco (2017)
3. Chakraborty, K., Bhattacharyya, S., Bag, R., Hassanien, A.E.: Comparative sentiment analysis on a set of movie reviews using deep learning approach. In: Hassanien, A.E., Tolba, Mohamed F., Elhoseny, M., Mostafa, M. (eds.) AMLTA 2018. AISC, vol. 723, pp. 311–318. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-74690-6_31
4. Kim, J., Ha, Y., Kang, S., Lim, H., Cha, M.: Detecting multiclass emotions from labeled movie scripts. In: IEEE International Conference on Big Data and Smart Computing (BigComp), 2018, pp. 590–594. IEEE, Shanghai (2018)
5. Kim, D., Lee, S., Cheong, Y.: Predicting emotion in movie scripts using deep learning. In: IEEE International Conference on Big Data and Smart Computing (Bigcomp) 2018, pp. 530–532. IEEE, Shanghai (2018)
6. Sahu, T., Ahuja, S.: Sentiment analysis of movie reviews: a study on feature selection & classification algorithms. In: International Conference on Microelectronics, Computing and Communications (MicroCom) 2016, pp. 1–6. IEEE, Durgapur (2016)

7. Parkhe, V., Biswas, B.: Sentiment analysis of movie reviews: finding most important movie aspects using driving factors. Soft Comput. **20**, 3373–3379 (2015). https://doi.org/10.1007/s00500-015-1779-1
8. Mesnil, G., Mikolov, T., Ranzato, M., Bengio, Y.: Ensemble of Generative and Discriminative Techniques for Sentiment Analysis of Movie Reviews. CoRR (2014)
9. Sureja, N., Sherasiya, F.: Using sentimental analysis approach review on classification of movie script. Int. J. Eng. Dev. Res. **5**, 616–620 (2017)
10. Manek, A., Shenoy, P., Mohan, M., Venugopal, K.R.: Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier. World Wide Web **20**, 135–154 (2016)
11. Rotten Tomatoes Movie Database. https://www.kaggle.com/ayushkalla1/rotten-tomatoes-movie-database. Accessed 07 Jan 2020
12. Park, C., Seo, D.: Sentiment analysis of twitter corpus related to artificial intelligence assistants. In: 5th International Conference on Industrial Engineering and Applications (ICIEA). 2018, pp. 495–498. IEEE, Singapore (2018)
13. Newman, H., Joyner, D.: Sentiment analysis of student evaluations of teaching. In: Penstein Rosé, C., et al. (eds.) AIED 2018. LNCS (LNAI), vol. 10948, pp. 246–250. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-93846-2_45
14. Kowalska, M., Wróbel, M.: Basic Emotions. In: Zeigler-Hill, V., Shackelford, T. (eds.) Encyclopedia of Personality and Individual Differences. Springer, Cham (2017)
15. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Found. Trends® Inf. Retriev. **2**, 1–135 (2008)
16. Medhat, W., Hassan, A., Korashy, H.: Sentiment analysis algorithms and applications: a survey. Ain Shams Eng. J. **5**, 1093–1113 (2014)