IMDB Movie Review Text Mining

Yingjun Guan

May 2018

1. Abstract

The project is to expand our research on text mining from movie information with the objective of helping all the movie production corporations and studios. The goal of the project is to provide a model procedure for digging the movie information and predicting the average rating for the movie according to the genre and the reviews of the movies. The data are extracted from the official website of IMDB. The project is based on the latest research on text mining, machine learning and sentiment analysis.

1. Introduction

The world of movies contains enormous information and is worth digging for prediction and recommendation. IMDB database [6] is one good example for extracting the relevant features, including both the item-based information (only related to the movies), for example, the title, the poster, the trailer, the genre, the cast, the director, etc., and the user-related information, such as name, gender, age, occupation, review, number of watching (the same movie), etc.

Our project plans to benefit both the movie production corporations and the audience. For movie production corporations, our goals include better predicting the movie ratings given the item-based information and the user-related information. Data mining and supervised machine learning techniques are applied for labelling and classification to the nominal features. Text mining and sentiment analysis techniques are used for the review context in the database. For each movie, some reviews are analyzed for sentiment analysis – both polarized results and sentiment intensity results are offered in the project for better analysis. The project also benefits the audience to see the distribution of the movie’s overall ratings. At the same time, the sentiment analysis can provide the key words of audience’s review, which offers them a direct and obvious introduction of the movie.

In this project, we select two main features apart from the *rating* of movies which we are interested in – the *genre* of movie and the movie *reviews* from various audience. As we mentioned before, the genre of the movie stands for the item-based features and the reviews of the movie represents the user-based features.

*The genres*

The genre of the movie is the first feature to be analyzed. As Sokappadu (2013) indicated, the movies in some certain genres have universally lower evaluations than in other genres. To dismiss the influence from the genres, our project selects three genres to analyze: the drama, the musical, and the horror. It can be instinctively noticed that the first two genres are close to each other while the horror movie is relatively far from the cluster of the first two. Our hypothesis is that the rating of movies in the same or similar genres may share the similarity in their comments and ratings, simultaneously, those movies which are in different genres may have different statistical performance in their reviews and ratings.

The data types of genres are categorical. Given that there are movies with more than one genre, all different genres will be treated as different features, for example, in this research each movie has features of “whether drama or not”, “whether musical or not”, and “whether horror or not”, separately. The analyses in this part will include the technologies of the data classification, data regression, statistical analysis, and so on.

*The reviews*

The movie reviews from the audience are the second feature to be analyzed. As the reviews contain the sentiments from the audience, sentiment analyses are of the core analyzing techniques. The potential features to be added for better analysis include but not limited to the users’ sentiments and the length of the reviews. The text data always occupy long time to preprocess and analyze, as our review data do. In this process, some stemming and stopping techniques can be utilized for better analyzing and word cloud, as a cool visualization and analyzing technique, may be one of the output for this part.

*The rating of the movies*

To make our research more convincing, our database includes movie with different average rating – our dataset includes those buzz beater movies, like “godfather”, “Shawshank’s redemption”, etc. While at the same time, similar number of normal movies and some notorious ones will also be included. For each movie in the database, we will also try to balance the influence of both positive and negative reviews, for better predicting the ratings.

The rating of the movies in this research is of numerical type (within the range of 0-10), we might keep it binary by splitting from 5.0, namely, label the movie “good” with an average rating higher than 5.0 and “bad” with an average rating lower than 5.0. In other cases, we can also take them numeric, the decision is to be made during the analysis. Since the data are already labelled from the database, cross-validation may be applied to justify our prediction results.

The main objectives include:

1. Movie review sentiment analysis: processing the text mining and sentiment analysis/ topic modeling from the review data.
2. Movie rating prediction: help predicting the new movies’ (test data) average rating given only genre and the reviews from the audience.
3. Data

The origin of all the dataset is IMDB [7], who is a public and free online movie database and provides a database of more than 4.7 million titles, 8.3 million personalities, as well as 3.5 million user reviews. With genres selection, only horror, musical and drama are of our concerns, whose reviews are calculated in the following table.

Table 1. number of reviews in IMDB for the genre of drama, musical and horror.

|  |  |
| --- | --- |
| Genre | Number of reviews |
| Drama | 1,219,402 |
| Musical | 8,864 |
| Horror | 99,662 |

Each review in the target genre that interests us can be retrieved through API [8] for downloading and filtering. For each review, the structure includes title of the movie, user name, review content and the individual review rating, as shown in the example figure below. For future work, the review content will be analyzed, and sentiment analysis will be conducted. The individual review rating can act as a label to exam the sentiment analysis.



Figure raw data example of movie review

To better fulfil the final prediction of overall rating of movies. For each movie, general information including the overall rating can be retrieved as shown in figure below, where the feature of genres and of overall ratings will be applied for mining. The sentiment results will be imported to the table of our interests before analyzing.

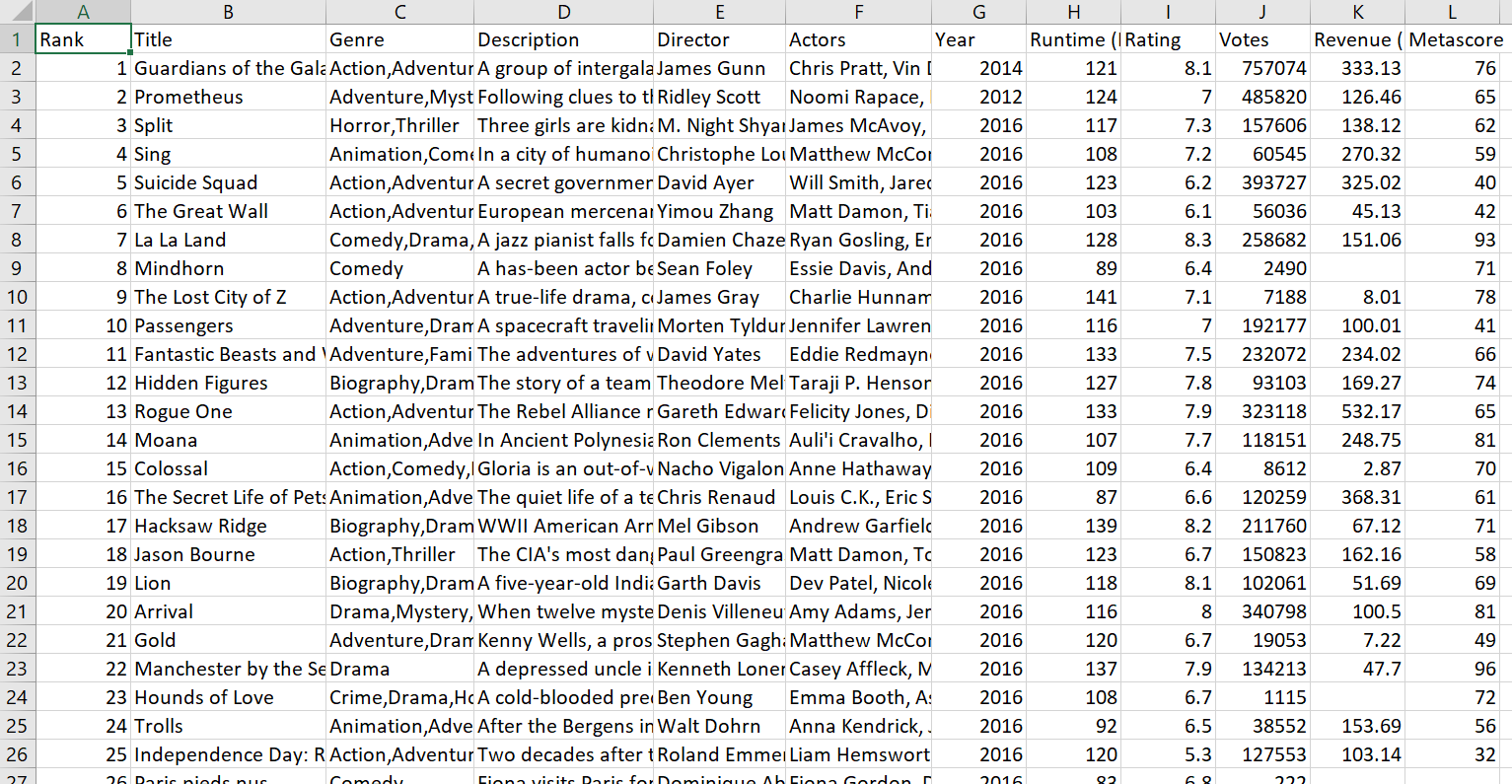


Figure sample of dataset containing other movie information

Noticing the rating here are numeric, they will be transformed into binary (polarity) before analysis, ie, each movie will be labelled as either “good movie” (rating over 5 out of 10) or “bad movie” (rating below 5 out of 10) as a feature given the overall rating. To improve the quality of results, movies with scores to both ends (closer to 10 and 1) will be selected according to the sorting and filtering for the use of classification and prediction.

1. Data Preprocessing

In this project, multiple steps are covered in the data preprocessing. The purpose of the preprocessing is to extract useful word features for further analysis from raw data shown in figure 1. In order to get word-based features, text analyzing procedures are utilized and the following steps are realized one by one.

* String extraction

All the movie reviews save in the format of html files. The information includes the title, the author, the year, the review, the reviewer information (some leaving blank), and the potential personal score for the movie. To extract useful information, regular expression is being used. (All the scripts will be provided later both in the appendix and in my personal GitHub repository.)

Instead of the review sentences, we even split the sentences into different word vectors for future usage. Therefore, after extraction, three columns have been provided, the title (serving as movie identifier), the raw review, and the word features from the raw review. A sample has been shown in the figure below.

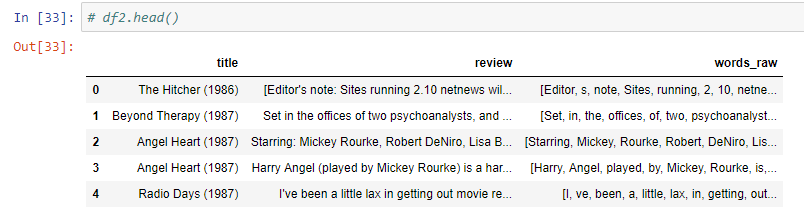


Figure example of string extraction of movie reviews

* Digits removal

There are lots of digits and numbers used in the movie reviews. However, most of them are irrelevant to the sentiment analysis. In this case, all the numbers (ie: 1, 2, 1.3, 523, etc.) and the words indicating numbers (one, two, ten, hundred, etc.) are removed from the database. Again, in this step, all the functions are written in python, and the scripts will be provided together with the report.

* Case normalization

To effectively reduce the dimension of the term-document matrix, all the words have been transformed into lower case. Thus, all the capitals have been changed. However, we have to pay attention to the potential damage or unnecessary drawback for some meaning words: take US and China as an example, they could be transformed into lower case, however, the lower-case words are previously created and having different meanings. Although there is some influence on this preprocessing, we still apply the technique.

* Stemming

Stemming is the method to further combine words in the same word family (from the same word root). Noticing that there are different stemming rules in use now. This project selects the Porter Method to stem in this paper. Again, we’d like to mention the drawback of stemming here that i) some of the words look alike but are not in the same family indeed, eg: universal, university; ii) words after stemming are not natural words which cause problems in recognition, eg: beautifully 🡪 beauty. Although there are some drawbacks on the stemming, we consider its influence is ignorable because what we care is only the sentiment and stemming seldom changes the sentiment of the words.

* Stopping

Stopping indicates the removal of the most commonly used words, which either has no meaning (eg: the, of, an, etc.) or makes very small contribution to differentiate different documents because it appears almost anywhere. Usually stopping is one of important steps in text mining. However, in sentiment analysis, stop words are kept because removing words like “not, never” may cause dramatic (if not opposite) change in the sentiment of the paragraph.

* Threshold adjustment

Similar to the idea discussed in stopping, we set the threshold to further eliminate the words appeared too much (more than 95 percent of all documents) or too little (less than 5 percent). In either case, words contribute little. And this threshold is processed in Oracle database.

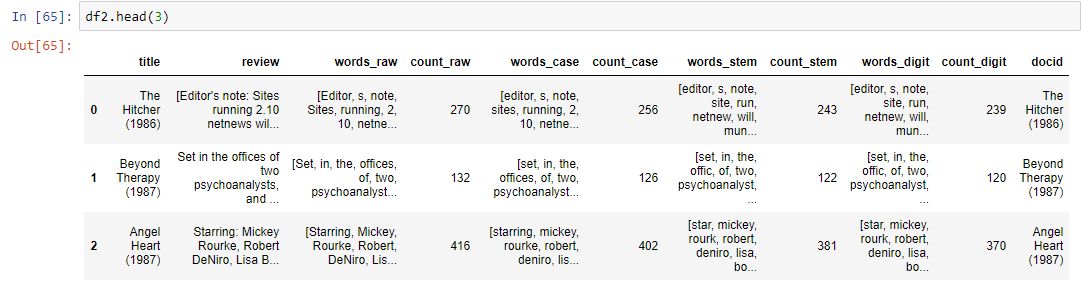


Figure sample of movie review preprocessing

The figure above indicates the results of preprocessing, apart from the threshold adjustment, all the displayed. And we can also notice that the number of words in each review keeps decreasing as we go through the steps of preprocessing.

1. Model and Results

With the pre-processed dataset, we have the data matrix containing both the document information and the term information. Importing the data into Oracle SQL developer to fetch the help from database, we transform the data to three columns: documented, termed, and term. From the example table shown above, we can easily retrieve the term and the document information as we want, which we will use different models introduced as below.

Table 2. Oracle database document-term table

|  |  |  |  |
| --- | --- | --- | --- |
|  | docid | termid | Term |
| 0 | 0 | 0 | Editor |
| 1 | 0 | 1 | S |
| 2 | 0 | 2 | Note |
| 3 | 0 | 3 | site |
| … | … | … | … |

* Feature selecting Models

1. TF-IDF

There are different feature selecting methods and, in this project, tf-idf model has been selected. TF-IDF is an excellent feature selection method for text mining, tf-idf is short for term frequency and inverse document frequency. Here, term frequency is the occurrence of each term (how many times it appears in the whole corpus), document frequency is the occurrence of documents which conceives the term (how many documents contain the certain term), and idf is the logarithm function of division from total number of corpus and the document frequency. The score of tf-idf is the product of term frequency and the inverse document frequency. All the formulas have been shown as below and been programmed in Oracle SQL Developer (the scripts have been provided together with other files.)

For each term, the score of tfidf takes only the highest score in case of the repetition in later feature selection.

1. Mature Sentiment Lexicon

At the same time, a mature feature list has been prepared from SentiWordNet [10], a resource of feature lexicon specially designed for sentiment analysis in the field of opinion mining. In the list, there are two parts, one is the word list for positive effects sorted the influential value; while the other is similar but for the negative effects.

In this project, the lexicon has been selected as features to predict the sentiment of the movie reviews. This is a general lexicon used in the world of sentiment analysis. Therefore, we select it as the benchmark to evaluate our work.

* Sentiment analysis Results

With two models introduced above, finally we prepare four groups of features for our prediction. Tfidf-100, tfidf-500, senti-100, senti-500. They separately from tfidf or sentiment lexicon models, with a different number of top features for selection (100 and 500).

And then we apply the features to test our overall rating of the movies (the labels are binary – good or bad). In the prediction, we applied 5-fold cross validation, and used four different algorithms (Decision Tree, Naïve Bayes, Support Vector Machine, Generalized Linear Model) to utilize the prediction. The final results are shown as below.

Table 3. sentiment analysis results

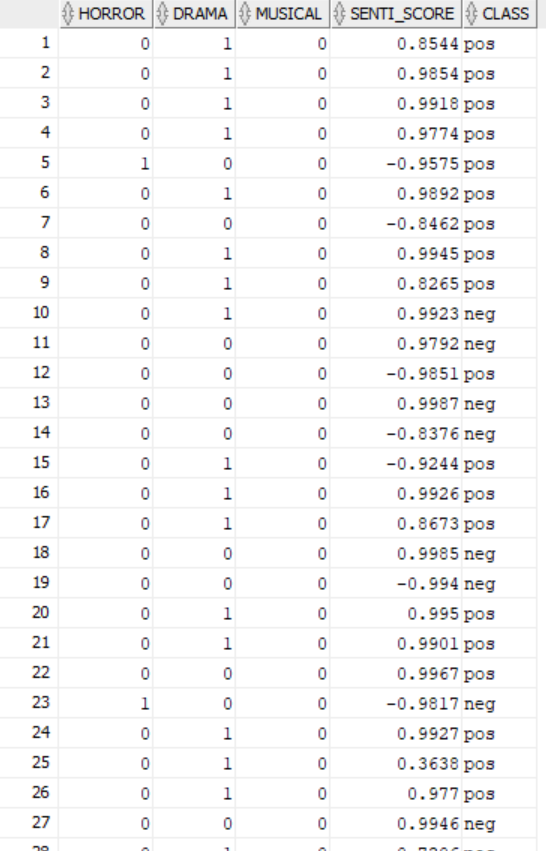
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | DT  Accuracy | NB  Accuracy | SVM Accuracy | GLM Accuracy | Averaged Accuracy |
| Tfidf – 100 | 59% | 59% | 61% | 63% | 60% |
| Tfidf – 500 | 63% | 71% | 73% | 77% | 73% |
| Senti – 100 | 51% | 57% | 53% | 51% | 53% |
| Senti – 500 | 49% | 59% | 53% | 55% | 54% |

From the table, we can observe that: a) tfidf model performs better than the benchmark of sentiment lexicon as shown in the table. (Both in top 100 and 500); b) top 500 performs better than top 100 groups in both tfidf and sentiment lexicon, indicating the positive influence when increasing the number of features; c) among all four algorithms, GLM>SVM>NB>DT.

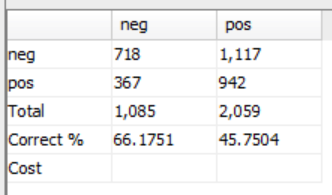
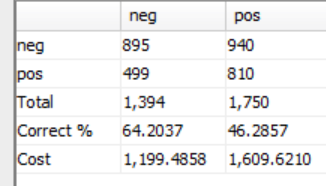
* Data Mining Results

In this section, we use the genre and the movie review sentiment score to predict the final overall rating for the movies. Sample dataset have been shown as below, where we focus on three kinds of genres, horror, drama and musical. The sentiment score is the intensity score. And the “Class” is the label of overall rating, “pos” indicating positive rating (good movie), “neg” indicating negative overall rating (bad movie).

Table 4. sample dataset for data mining



Again, we used decision tree model and naïve Bayes models to predict the label. And this time, the results are shown as below, we can see that the accuracy of negative is around 65% while positive is around 45%.



Decision Tree model Naïve Bayes model

1. Conclusion and future work

The text mining section works on the sentiment analysis and the hands-on practice on how to better extract the features for the project. The project successfully applies the words (after pre-processing) to features for sentiment analysis; import the common sentiment lexicon to analysis for study and made a comparison with different feature selection strategies. In this section, the number of features selected can be adjusted more. So far, top 100 and top 500 features have been selected for analysis and comparison. However, with a large database as used in this project, 500 may not be enough. Moreover, in feature selection, apart from the tf-idf method which has been applied in this passage, other selection methods can be utilized, for example: information gain.

The data mining section in the project focuses on predicting the overall rating for each movie from its genre and the review sentiments. The project applies Naïve Bayes and the Decision Tree as the models. The results are convincing yet can be improved. For future work, we can not only use sentiment parameters as binary variables (positive and negative), but also as a numeric data, which takes into consideration of the intensity of the sentiments. More supervised learning algorithms rather than Naïve Bayes and the Decision Tree can be applied to improve the results.

References

[1] Wiebe, J. (2000). Learning subjective adjectives from corpora. Aaai/iaai, 20(0), 0.

[2] Thet, T. T., Na, J. C., & Khoo, C. S. (2010). Aspect-based sentiment analysis of movie reviews on discussion boards. Journal of information science, 36(6), 823-848.

[3] Zhuang, L., Jing, F., & Zhu, X. Y. (2006, November). Movie review mining and summarization. In Proceedings of the 15th ACM international conference on Information and knowledge management (pp. 43-50). ACM.

[4] Sokappadu, B. N. (2013, Jul 22) Why are horror movies mostly rated less than 7 in the IMDb? Retrieved from https://www.quora.com/Why-are-horror-movies-mostly-rated-less-than-7-in-the-IMDb#

[5] The Internet Movie Database. IMDB.com, Inc, 2009. Web. 29 Oct. 2009. <http://www.imdb.com>

[6] IMDB.com. “Press Room.” IMDB, www.imdb.com/pressroom/stats/.

[7] <https://github.com/richardasaurus/imdb-pie>

[8] https://github.com/yingjun2/

[9] Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction. The Mathematical Intelligencer, 27(2), 83-85.

[10] http://sentiwordnet.isti.cnr.it/