IMDB Movie Review Mining and Sentiment Analysis

Yingjun Guan

Doctoral Student, School of Information Sciences, University of Illinois at Urbana-Champaign

ILLINOIS
School of Information Sciences

Motivation

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.

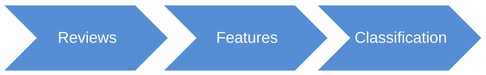
Our project plans to benefit both the movie production corporations and the audience. For movie production corporations, our goals include better predicting the movie ratings given the item-based information and the user-related information. Data mining and supervised machine learning techniques are 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.

Keyword: Sentiment analysis; opinion mining; NLP; text mining; data mining; IMDB movie review analysis.

Target

Target 1: polarity classification from reviews

- Extract features from movie reviews to build models for classifying polarity overall ratings.
  [text classification problem]



- Starts from raw review strings. (Data Cleaning)
- Extract words as features (improve feature selection)
- Generate the binary polarity labels for ratings.

Target 2: supervised learning for ratings

- Extract binary sentiment index from movie reviews, together with genres, to predict the overall rating.
  [data mining problem]

Introduction

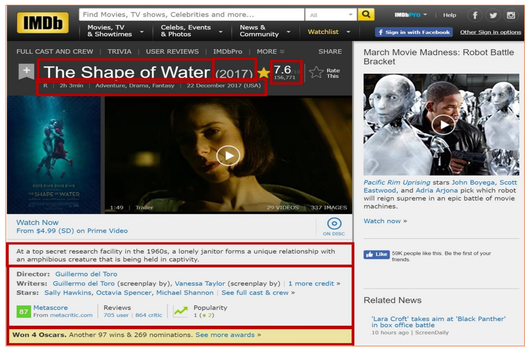
The world of movie is mysterious and interesting. As we notice, it is hard to precisely predict the movie performance because there are too many factors affecting the rating of movies, some of which are item-based information (related to the movies) and some are user-based attributes.

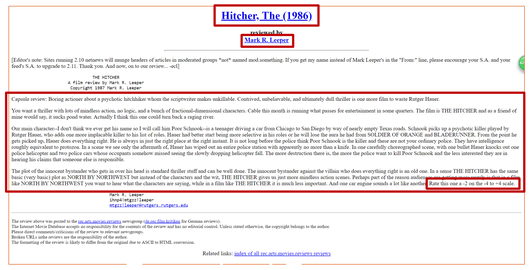
User-based	Item-based
• Gender • Age • Nationality • Occupation • Review • Times of watching • …..	• Title • Poster • Trailer • Genre • Cast • Director • ……

In target one, we focus only on the user-based information (review as an example): the obstacles include how to extract both the information from any single specific audience and the general opinion from all. In target two, we try to combine the influence from both user-based information (reviews) and the item-based information (genres) to predict the overall ratings.

Data

IMDB [1] 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. Each review in the target genre that interests us can be retrieved through API [2] for downloading and filtering.





Results

- **Feature Engineering**

To better select features, three groups of features have been prepared for training the model of movie reviews. The first group is based on basic pre-processing, the second on tf-idf algorithm, the third applying the common sentiment lexicon (SentiWordNet [3]).

Basic processing	TF-IDF	Sentiment lexicon
• Digits. • Cases. • Stemming • ~~Stopping~~ • Threshold.	• Using tf-idf method to extract top features (100 vs 500)	• General sentimental analysis. • Corresponding 100 and 500.

- **Classification models**

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.

- **Review Sentiment Classification Results.**

Accuracy	DT	NB	SVM	GLM	Averaged
Tfidf – 100	59%	59%	61%	63%	60%
Tfidf – 500	63%	71%	73%	77%	73%
Senti – 100	51%	57%	53%	51%	53%
Senti - 500	51%	60%	59%	70%	60%

➢ GLM > SVM > NB > DT
➢ Number of features influencing the performance of the model.
➢ Corpus based features (tf-idf) perform better than general features (sentimental lexicon).
➢ Although the lexicon features have been through the same preprocessing (stemming, figure, lower case, etc.) There's still a large amount of **helpless** information for the given database.

- **Overall Rating prediction**

The are 54.5% results for the prediction is not ideal: best performance of all four models

	HORROR	DRAMA	MUSICAL	SENTI_SCORE	CLASS
1	0	1	0	0.8544	pos
2	0	1	0	0.9854	pos
3	0	1	0	0.9918	pos
4	0	1	0	0.9774	pos
5	1	0	0	-0.9575	pos
6	0	1	0	0.9892	pos
7	0	0	0	-0.8462	pos
8	0	1	0	0.9945	pos

Models	Correct Predictions %
CLAS_DT_11_5	54.2303
CLAS_NB_11_5	52.7990
CLAS_GLM_11_5	54.5802
CLAS_SVM_11_5	46.3104

Conclusions and future work

1. The text mining section works on the sentiment analysis and the hands-on practice on how to better extract the features for the project.
2. 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.
3. In feature selection, apart from the tf-idf method which has been applied, other selection methods can be utilized, for example: information gain.
4. The data mining section in the project focuses on predicting the overall rating for each movie from its genre and the review sentiments. The results are convincing yet can be improved.
5. 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.
6. More information can be found in my GitHub [4] and through my personal webpage [5].

References

[1] https://github.com/richardasaurus/imdb-pie
[2] Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction. The Mathematical Intelligencer, 27(2), 83-85.
[3] http://sentiwordnet.isti.cnr.it/
[4] https://github.com/yingjun2/
[5] https://ischool.illinois.edu/people/phd-students/yingjun-guan

ILLINOIS