Effects of film characteristics on consumer behavior; A text classification approach

Farhad Khalilzadeh
Department of Computer Science, Hacettepe University, Ankara, Turkey
farhad.khalilzadeh@gmail.com

Abstract. Film industry has lots of consumers worldwide. Studying these consumers behavior is one of the interesting subjects in business management and other disciplines like computer sciences. This paper analyzes the effects of film characteristics on audiences by text classification based algorithms. These algorithms are designed to classify the documents into classes automatically. The main data base is extracted from IMDB, which is a popular online film information database. By extracting the film features as texts and giving them as an input to classification algorithms such as SVM algorithm, we show the relationship between these features and consumers behavior which is shown by audience's ratings. Features such as director, stars and genre are most focusing features that are given to the algorithms. By focusing on Oscar winner films we extracted some other detailed features and find their trace on audience's behavior. We also compare the features effect with each other and find the most important groups of them.

Keywords: consumer behavior, film industry, text classification

1 INTRODUCTION

All over the world every year there are lots of film productions and the investing budget in this industry is growing rapidly reaching multi billion dollars. In this industry the consumers are the audiences which are very important part of this industry because of the importance of gaining their satisfaction. Studying film consumer's behavior is one of the interesting subjects in business management and other disciplines like computer sciences. Understanding the relationships between film viewer's satisfaction and the film characteristics is very important to predict a film's success because of the billion dollar budgets.
Designing a system which can understand the relationship between consumers and film characteristics could help the producers to gain maximum profit of the market. This paper analyzes the effects of film characteristics on audiences by text classification based algorithms. The main data base is extracted from IMDB, which is a popular online film information database. By extracting the film features as texts and giving them as an input to classification algorithms such as SVM algorithm, we show the relationship between these features and consumers behavior which is shown by audience's ratings.

This paper is arranged as follow: section 2 describes the basics of consumer behavior analysis and text classification process. Section 3 describes details of proposed method. Section 4 provides experiment results. Section 5 covers conclusions.

2 BASICS OF TERMS

2.1 Consumer behavior

Consumer behavior is one of the important concepts in marketing discipline. One of the first usage of this concept can be seen at 1950's when the marketers used Sigmund Freud's ideas in their business. Consumer behavior is one of the most challenging concepts which effects people and their purchases.

Wilkie and Salmon's definition: Consumer behavior is people's physical, emotional and mental activities during selection, purchase, use and discard of goods and services in order to satisfy their needs and desires.

In another definition consumer behavior is a collection of activities which are done to achieve, use and discard the goods and services. These activities consist of decisions made before and after the achievement, consumption and wasting processes.

The marketing discipline especially in the field of consumer behavior, has lots of opportunities learning from other disciplines such as sociology, psychology, anthropology, economics and etc. . So in consumer behavior studies we need to consider the society, psychology and economic conditions of audiences of movies.

2.2 Film viewer as a consumer

The effects of movies and films on society became one of the important subjects of discussion and the relationship between them have been studied carefully. The film viewers or audiences are considered as consumers. This effect is bio-directional meaning that the films represent a society's cultural situation. This representation is came to existence by the society and also consumed by itself. The society or film viewer also has an undeniable effect on film production industry.

2.3 Film Characteristics

At nowadays film industry a film does have lots of special technical characteristics from a movie expert's professional point of view. On the other hand regular audiences consider some simple and obvious characteristics of films to care. These characteristics can be addressed as: actors and actresses – directors – Genre – title – writer –duration – lightening and so on. The audiences rate a film considering these characteristics of it.

2.4 Text Classification

Text classification is an automatic system that its input is text documents and the output is the predefined classes or topics of the input documents. It can provide conceptual views of document collections and has important applications in the real world.

In the text classification each document can be in multiple classes, exactly one classes, or no class at all. By means of machine learning, the system should take the input samples and learn some classifier, this classifier performs the class assignments to the documents automatically. This is a supervised learning problem.

2.5 Text classification key methods

In text classification there are lots of classes of learning techniques that have been used. These include at the very least probabilistic methods, regression methods, decision tree and decision rule learners, neural networks, batch and incremental learners of linear classifiers, example-based methods, support vector machines, genetic algorithms, hidden Markov models, and classifier committees (which include boosting methods).

The procedure of text classification task consists of several steps.

A) Finding suitable corpus
B) Extracting features
C) Feature Transformation
D) Making a Classifier Model
E) Training and Testing the Model
F) Evaluating

3 THE PROPOSED METHOD

This paper considers the film characteristics as some text strings and tries to predict film viewer ratings based on these characteristics. It also looks for the most effecting characteristic among others by using greedy backward

algorithm. The characteristic's list is considered as features of the system and the ratings are considered as the output of it. It uses SVM algorithm to predict the ratings based on features.

3.1 Database

The IMDB (Internet Movie Database) is an online and most famous database of films that includes almost everything about the films. The best thing of it is ratings of movies. Each film viewer can give its rating to the film and IMDB have its own special formula to calculate the overall rating for a film based on these ratings. This paper uses these ratings and also some other information about films.
The total of 400 films are selected for training set and also another 200 films for test set. The film information that are used as film characteristics are as: Film Title, Film Actors, Film Director, Film Genre. The film ratings are divided to 4 classes as below:

Poor Class = 0 to 2.5
So-So Class= 2.6 to 5
Good Class = 5.1 to 7.5
Excellent Class = 7.5 to 10
The films are labeled as one of these classes based on its rating.

3.2 Support Vector Machines

The support vector machine (SVM) algorithm tries to find a maximum margin between different classes. It just uses the support vectors (the closest samples to the hyperplane). The data will be transformed into a higher dimension in which the classes are separable. In SVM, if the data is not separable we use a kernel function to map the data in a new space where they are separable. The main reason of developing a kernel function is to map the data into a (generally) higher dimensional space in order to detect structure in data more easily.

3.3 Greedy backward elimination

To find the degree of importance of film characteristics, the correlation between features and the target output is tested. The greedy backward elimination algorithm is to get best feature subset among all features. This algorithm gets all the features of films and in a loop, at each iteration it removes the worst feature in the feature set. By this method the least and the most important characteristics of films can be detected.

3.4 Evaluation

Evaluations applied the following relations:

TP (True Positive), FP (False Positive), FN (False Negative), TN (True Negative)

$$\text{Precision} = \frac{TP}{TP + FP} \qquad (1)$$

Precision shows the fraction of rated films that are relevant (are truly rated).

4 RESULTS

The algorithm is tested by 2 kind of features. The first one used the primary features explaind in database section (Actors, director, genre and writer). In second one all the Oscar winner films are replaced by non-Oscar winner films and a new feature (The film Budget) is added to primary features. The results shows that the SVM algorithm based on primary features reaches 51% of precision and the algorithm based on added new feature reaches 51% of precision. The correlation between the features and the output also shows that the importance of features that effects the film viewer's ratings are as below from highest to lowest respectively:

1-Director 2- Writer 3- Actors 4-Gener

Table 1: the results of algorithm

	SVM (by primary features)	SVM (By Oscar features)
Precision	49.2%	51%

5 CONCLUSION

The experiments show that expanding the input film features to the algorithm has an improving effect on rating prediction. Also the effects of film characteristics on film viewer ratings (consumer behavior) are obviously high in a way that just considering one more film characteristic to the model improves the prediction rate. Results also show that film viewers mostly pay attention to the director of films. The writer, actors and genre of films are on next steps.

To improve the prediction model and also understanding the effects of other film characteristics, a much bigger database should be considered and more film features should be tested.

Acknowledgments

This research was supported by BİDEB branch of Turkey TÜBİTAK organization. I thank my professor İlyas Çiçekly from Hacettepe University who provided insight and expertise that greatly assisted the research.

References

Darin Im, Minh Thao, Dang Nguyen, "Predicting Movie Success in the U.S. market," Dept.Elect.Eng, Stanford Univ., California, December, 2011

Lashkari, M.; Nokhchiyan, A. (2009) "Increasing Marketing performance with identifying consumer behavior", Journal of the Iranian Chamber, No. 38,14-17

Nabi Zadeh, Mahmoud (1994), "Models of consumer behavior," Social Science letter,267-284

Iman Khan, N (2008), "Consumer Behavior in Digital Marketing", Journal of Management, Issue 11, 81-88

Salar, J. (2006) " Mix Relationship of marketing and consumer behavior", Tadbir, No. 176,59-64

Lashkari, M.; Nokhchiyan, A. (2009) "Increasing Marketing performance with identifying consumer behavior", Journal of the Iranian Chamber, No. 38,14-17

Iman Khan, N (2008), "Consumer Behavior in Digital Marketing", Journal of Management, Issue 11, 81-88

Pei-Yi Hao, Jung-Hsien Chiang, Yi-Kun Tu, Hierarchically SVM classification based on support vector clustering method and its application to document categorization, Expert Systems with Applications, Volume 33, Issue 3, October 2007, Pages 627-635, ISSN 0957-4174, 10.1016/j.eswa.2006.06.009.

Fabrizio Sebastiani, Text Categorization, Dipartimento di Matematica Pura e Applicata Universit`a di Padova 35131 Padova, Italy

www.IMDB.com

Cristianini, Nello; and Shawe-Taylor, John; An Introduction to Support Vector Machines and other kernel-based learning methods, Cambridge University Press, 2000. ISBN 0-521-78019-5

Sagar V. Mehta, Rose Marie Philip, Aju Talappillil Scaria, "Predicting Movie Rating based on Text Reviews," Dept.Elect.Eng, Stanford Univ., California, December, 2011

Deniz Demir, Olga Kapralova, Hongze Lai, "Predicting IMDB Movie Ratings Using Google Trends," Dept.Elect.Eng,Stanford Univ., California, December, 2012

Suhaas Prasad, "Using Social Networks to improve Movie Ratings predictions," Dept.Elect.Eng, Stanford Univ., California, 2010

Thorsten Joachims, “Text Categorization with Support Vector Machines: Learning with Many Relevant Features”, 1998

Farhad Khalilzadeh was born in Iran in 1985. He received the Associated Degree in Software Engineering from the University of Applied Scientific and Technology, Tabriz, Iran, in 2006, and the B.S. degree in the same field at UCNA, Tabriz,Iran in 2008. In 2011 he started combined Master and Ph.D. degrees in Artificial Intelligence at Hacettepe University, Ankara, Turkey where he is currently a Ph.D. candidate.

His main areas of research interests are machine learning, NLP, pattern recognition and signal processing.