A Surrogate-based Generic Classifier for Chinese TV Series Reviews

Abstract

With the emerging of various online video platforms like Youtube, Youku and LeTV, online TV series' reviews become more and more important both for viewers and producers. Customers rely heavily on these reviews before selecting TV series, while producers use them to improve the quality. As a result, automatically classifying reviews according to different requirements evolves as a popular research topic and is essential in our daily life. In this paper, we focused on reviews of hot TV series in China and successfully trained generic classifiers based on eight predefined categories. The experimental results showed promising performance and effectiveness of its generalization to different TV series.

Introduction

With Web 2.0's development, more and more commercial websites, such as Amazon, Youtube and Youku, encourage users to post product reviews on their platforms BIBREF0 , BIBREF1 . These reviews are helpful for both readers and product manufacturers. For example, for TV or movie producers, online reviews indicates the aspects that viewers like and/or dislike. This information facilitates producers' production process. When producing future films TV series, they can tailer their shows to better accommodate consumers' tastes. For manufacturers, reviews may reveal customers' preference and feedback on product functions, which help manufacturers to improve their products in future development. On the other hand, consumers can evaluate the quality of product or TV series based on online reviews, which help them make final decisions of whether to buy or watch it. However, there are thousands of reviews emerging every day. Given the limited time and attention consumers have, it is impossible for them to allocate equal amount of attention to all the reviews. Moreover, some readers may be only interested in certain aspects of a product or TV series. It's been a waste of time to look at other irrelevant

ones. As a result, automatic classification of reviews is essential for the review platforms to provide a better perception of the review contents to the users.

Most of the existing review studies focus on product reviews in English. While in this paper, we focus on reviews of hot Chinese movies or TV series, which owns some unique characteristics. First, Table TABREF1 shows Chinese movies' development BIBREF2 in recent years. The growth of box office and viewers is dramatically high in these years, which provides substantial reviewer basis for the movie/TV series review data. Moreover, the State Administration of Radio Film and Television also announced that the size of the movie market in China is at the 2nd place right after the North America market. In BIBREF2 , it also has been predicted that the movie market in China may eventually become the largest movie market in the world within the next 5-10 years. Therefore, it is of great interest to researchers, practitioners and investors to understand the movie market in China.

Besides flourishing of movie/TV series, there are differences of aspect focuses between product and TV series reviews. When a reviewer writes a movie/TV series review, he or she not only care about the TV elements like actor/actress, visual effect, dialogues and music, but also related teams consisted of director, screenwriter, producer, etc. However, with product reviews, few reviewers care about the corresponding backstage teams. What they do care and will comment about are only product related issues like drawbacks of the product functions, or which aspect of the merchandise they like or dislike. Moreover, most of recent researchers' work has been focused on English texts due to its simpler grammatical structure and less vocabulary, as compared with Chinese. Therefore, Chinese movie reviews not only provide more content based information, but also raise more technical challenges. With bloom of Chinese movies, automatic classification of Chinese movie reviews is really essential and meaningful.

In this paper, we proposed several strategies to make our classifiers generalizable to agnostic TV series.

First, TV series roles' and actors/actresses' names are substituted by generic tags like role_i and player_j, where i and j defines their importance in this movie. On top of such kind of words, feature tokens are further manipulated by feature selection techniques like DRC or INLINEFORM0 , in order to make it more generic. We also experimented with different feature sizes with multiple classifiers in order to alleviate overfitting with high dimension features.

The remainder of this paper is organized as follows. Section 2 describes some related work. Section 3 states our problem and details our proposed procedure of approaching the problem. In Section 4, experimental results are provided and discussed. Finally, the conclusions are presented in Section 5.

Related Work

Since we are doing supervised learning task with text input, it is related with work of useful techniques like feature selections and supervised classifiers. Besides, there are only public movie review datasets in English right now, which is different from our language requirement. In the following of this section, we will first introduce some existing feature selection techniques and supervised classifiers we applied in our approach. Then we will present some relevant datasets that are normally used in movie review domain.

Feature selection

Feature selection, or variable selection is a very common strategy applied in machine learning domain, which tries to select a subset of relevant features from the whole set. There are mainly three purposes behind this. Smaller feature set or features with lower dimension can help researchers to understand or interpret the model they designed more easily. With fewer features, we can also improve the generalization of our model through preventing overfitting, and reduce the whole training time.

Document Relevance Correlation(DRC), proposed by W. Fan et al 2005 BIBREF3 , is a useful feature selection technique. The authors apply this approach to profile generation in digital library service and news-monitoring. They compared DRC with other well-known methods like Robertson's Selection Value BIBREF4 , and machine learning based ones like information gain BIBREF5 . Promising experimental results were shown to demonstrate the effectiveness of DRC as a feature selection in text field.

Another popular feature selection method is called INLINEFORM0 BIBREF6 , which is a variant of INLINEFORM1 test in statistics that tries to test the independence between two events. While in feature selection domain, the two events can be interpreted as the occurrence of feature variable and a particular class. Then we can rank the feature terms with respect to the INLINEFORM2 value. It has been proved to be very useful in text domain, especially with bag of words feature model which only cares about the appearance of each term.

Supervised Classifier

What we need is to classify each review into several generic categories that might be attractive to the readers, so classifier selection is also quite important in our problem. Supervised learning takes labeled training pairs and tries to learn an inferred function, which can be used to predict new samples. In this paper, our selection is based on two kinds of learning, i.e., discriminative and generative learning algorithms. And we choose three typical algorithms to compare. Bayes BIBREF7 , which is the representative of generative learning, will output the class with the highest probability that is generated through the bayes' rule. While for the discriminative classifiers like logistic regression BIBREF8 or Support Vector Machine BIBREF9 , final decisions are based on the classifier's output score, which is compared with some threshold to distinguish between different classes.

TV series Review Dataset

Dataset is another important factor influencing the performance of our classifiers. Most of the public available movie review data is in English, like the IMDB dataset collected by Pang/Lee 2004 BIBREF10 . Although it covers all kinds of movies in IMDB website, it only has labels related with the sentiment. Its initial goal was for sentiment analysis. Another intact movie review dataset is SNAP BIBREF11 , which consists of reviews from Amazon but only bearing rating scores. However, what we need is the content or aspect tags that are being discussed in each review. In addition, our review text is in Chinese. Therefore, it is necessary for us to build the review dataset by ourselves and label them into generic categories, which is one of as one of the contributions of this paper.

Chinese TV series Review Classification

Let INLINEFORM0 be a set of Chinese movie reviews with no categorical information. The ultimate task of movie review classification is to label them into different predefined categories as INLINEFORM1 . Starting from scratch, we need to collect such review set INLINEFORM2 from an online review website and then manually label them into generic categories INLINEFORM3 . Based on the collected dataset, we can apply natural language processing techniques to get raw text features and further learn the classifiers. In the following subsections, we will go through and elaborate all the subtasks shown in Figure FIGREF5 .

Building Dataset

What we are interested in are the reviews of the hottest or currently broadcasted TV series, so we select one of the most influential movie and TV series sharing websites in China, Douban. For every movie or TV series, you can find a corresponding section in it. For the sake of popularity, we choose “The Journey of Flower”, “Nirvana in Fire” and “Good Time” as parts of our movie review dataset, which are the hottest TV series from summer to fall 2015. Reviews of each episode have been collected for the sake of dataset

comprehensiveness.

Then we built the crawler written in python with the help of scrapy. Scrapy will create multiple threads to crawl information we need simultaneously, which saves us lots of time. For each episode, it collected both the short description of this episode and all the reviews under this post. The statistics of our TV series review dataset is shown in Table TABREF7 .

Basic Text Processing

Based on the collected reviews, we are ready to build a rough classifier. Before feeding the reviews into a classifier, we applied two common procedures: tokenization and stop words removal for all the reviews. We also applied a common text processing technique to make our reviews more generic. We replaced the roles' and actors/actresses' names in the reviews with some common tokens like role_i, actor_j, where i and j are determined by their importance in this TV series. Therefore, we have the following inference DISPLAYFORM0

where INLINEFORM0 is a function which map a role's or actor's index into its importance. However, in practice, it is not a trivial task to infer the importance of all actors and actresses. We rely on data onBaidu Encyclopedia, which is the Chinese version of Wikipedia. For each movie or TV series, Baidu Encyclopedia has all the required information, which includes the level of importance for each role and actor in the show. Actor/actress in a leading role will be listed at first, followed by the ones in a supporting role and other players. Thus we can build a crawler to collect such information, and replace the corresponding words in reviews with generic tags.

Afterwards, word sequence of each review can be manipulated with tokenization and stop words removal. Each sequence is broken up into a vector of unigram-based tokens using NLPIR BIBREF12 , which is a

very powerful tool supporting sentence segmentation in Chinese. Stop words are words that do not contribute to the meaning of the whole sentence and are usually filtered out before following data processing. Since our reviews are collected from online websites which may include lots of forum words, for this particular domain, we include common forum words in addition to the basic Chinese stop words. Shown below are some typical examples in English that are widely used in Chinese forums.
INLINEFORM0

These two processes will help us remove significant amount of noise in the data.

Topic Modelling and Labeling

With volumes of TV series review data, it's hard for us to define generic categories without looking at them one by one. Therefore, it's necessary to run some unsupervised models to get an overview of what's being talked in the whole corpus. Here we applied Latent Dirichlet Allocation BIBREF13 , BIBREF14 to discover the main topics related to the movies and actors. In a nutshell, the LDA model assumes that there exists a hidden structure consisting of the topics appearing in the whole text corpus. The LDA algorithm uses the co-occurrence of observed words to learn this hidden structure. Mathematically, the model calculates the posterior distribution of the unobserved variables. Given a set of training documents, LDA will return two main outputs. The first is the list of topics represented as a set of words, which presumably contribute to this topic in the form of their weights. The second output is a list of documents with a vector of weight values showing the probability of a document containing a specific topic.

Based on the results from LDA, we carefully defined eight generic categories of movie reviews which are most representative in the dataset as shown in Table TABREF11 .

The purpose of this research is to classify each review into one of the above 8 categories. In order to build reasonable classifiers, first we need to obtain a labeled dataset. Each of the TV series reviews was labeled by at least two individuals, and only those reviews with the same assigned label were selected in our training and testing data. This approach ensures that reviews with human biases are filtered out. As a result, we have 5000 for each TV series that matches the selection criteria.

Feature Selection

After the labelled cleaned data has been generated, we are now ready to process the dataset. One problem is that the vocabulary size of our corpus will be quite large. This could result in overfitting with the training data. As the dimension of the feature goes up, the complexity of our model will also increase. Then there will be quite an amount of difference between what we expect to learn and what we will learn from a particular dataset. One common way of dealing with the issue is to do feature selection. Here we applied DRC and INLINEFORM0 mentioned in related work. First let's define a contingency table for each word INLINEFORM1 like in Table TABREF13 . If INLINEFORM2 , it means the appearance of word INLINEFORM3 .

Recall that in classical statistics, INLINEFORM0 is a method designed to measure the independence between two variables or events, which in our case is the word INLINEFORM1 and its relevance to the class INLINEFORM2 . Higher INLINEFORM3 value means higher correlations between them. Therefore, based on the definition of INLINEFORM4 in BIBREF6 and the above Table TABREF13 , we can represent the INLINEFORM5 value as below: DISPLAYFORM0

While for DRC method, it's based on Relevance Correlation Value, whose purpose is to measure the similarity between two distributions, i.e., binary distribution of word INLINEFORM0 's occurrence and documents' relevance to class INLINEFORM1 along all the training data. For a particular word

INLINEFORM2 , its occurrence distribution along all the data can be represented as below (assume we have INLINEFORM3 reviews): DISPLAYFORM0

And we also know each review INLINEFORM0 's relevance with respect to INLINEFORM1 using the manually tagged labels. DISPLAYFORM0

where 0 means irrelevant and 1 means relevant. Therefore, we can calculate the similarity between these two vectors as DISPLAYFORM0

where INLINEFORM0 is called the Relevance Correlation Value for word INLINEFORM1 . Because INLINEFORM2 is either 1 or 0, with the notation in the contingency table, RCV can be simplified as DISPLAYFORM0

Then on top of RCV, they incorporate the probability of the presence of word INLINEFORM0 if we are given that the document is relevant. In this way, our final formula for computing DRC becomes DISPLAYFORM0

Therefore, we can apply the above two methods to all the word terms in our dataset and choose words with higher INLINEFORM0 or DRC values to reduce the dimension of our input features.

Learning Classifiers

Finally, we are going to train classifiers on top of our reduced generic features. As mentioned above, there are two kinds of learning algorithms, i.e., discriminant and generative classifiers. Based on Bayes rule, the optimal classifier is represented as INLINEFORM0

where INLINEFORM0 is the prior information we know about class INLINEFORM1 .

So for generative approach like Bayes, it will try to estimate both INLINEFORM0 and INLINEFORM1 . During testing time, we can just apply the above Bayes rule to predict INLINEFORM2 . Why do we call it naive? Remember that we assume that each feature is conditionally independent with each other. So we have DISPLAYFORM0

where we made the assumption that there are INLINEFORM0 words being used in our input. If features are binary, for each word INLINEFORM1 we may simply estimate the probability by DISPLAYFORM0

in which, INLINEFORM0 is a smoothing parameter in case there is no training sample for INLINEFORM1 and INLINEFORM2 outputs the number of a set. With all these probabilities computed, we can make decisions by whether DISPLAYFORM0

On the other hand, discriminant learning algorithms will estimate INLINEFORM0 directly, or learn some “discriminant” function INLINEFORM1 . Then by comparing INLINEFORM2 with some threshold, we can make the final decision. Here we applied two common classifiers logistic regression and support vector machine to classify movie reviews. Logistic regression squeezes the input feature into some interval between 0 and 1 by the sigmoid function, which can be treated as the probability INLINEFORM3 . DISPLAYFORM0

The Maximum A Posteriori of logistic regression with Gaussian priors on parameter INLINEFORM0 is defined as below INLINEFORM1

which is a concave function with respect to INLINEFORM0 , so we can use gradient ascent below to optimize the objective function and get the optimal INLINEFORM1 . DISPLAYFORM0

where INLINEFORM0 is a positive hyper parameter called learning rate. Then we can just use equation (EQREF24) to distinguish between classes.

While for Support Vector Machine(SVM), its initial goal is to learn a hyperplane, which will maximize the margin between the two classes' boundary hyperplanes. Suppose the hyperplane we want to learn is INLINEFORM0

Then the soft-margin version of SVM is INLINEFORM0

where INLINEFORM0 is the slack variable representing the error w.r.t. datapoint INLINEFORM1 . If we represent the inequality constraints by hinge loss function DISPLAYFORM0

What we want to minimize becomes DISPLAYFORM0

which can be solved easily with a Quadratic Programming solver. With learned INLINEFORM0 and INLINEFORM1 , decision is made by determining whether DISPLAYFORM0

Based on these classifiers, we may also apply some kernel trick function on input feature to make originally linearly non-separable data to be separable on mapped space, which can further improve our classifier performance. What we've tried in our experiments are the polynomial and rbf kernels.

Experimental Results and Discussion

As our final goal is to learn a generic classifier, which is agnostic to TV series but can predict review's category reasonably, we did experiments following our procedures of building the classifier as discussed in section 1.

Category Determining by LDA

Before defining the categories of the movie reviews, we should first run some topic modeling method. Here we define categories with the help of LDA. With the number of topics being set as eight, we applied LDA on “The Journey of Flower”, which is the hottest TV series in 2015 summer. As we rely on LDA to guide our category definition, we didn't run it on other TV series. The results are shown in Figure FIGREF30 . Note that the input data here haven't been replaced with the generic tag like role_i or actor_j, as we want to know the specifics being talked by reviewers. Here we present it in the form of heat maps. For lines with brighter color, the corresponding topic is discussed more, compared with others on the same height for each review. As the original texts are in Chinese, the output of LDA are represented in Chinese as well.

We can see that most of the reviews are focused on discussing the roles and analyzing the plots in the movie, i.e., 6th and 7th topics in Figure FIGREF30 , while quite a few are just following the posts, like the 4th and 5th topic in the figure. Based on the findings, we generate the category definition shown in Table TABREF11 . Then 5000 out of each TV series reviews, with no label bias between readers, are selected to make up our final data set.

Feature Size Comparison

Based on INLINEFORM0 and DRC discussed in section 3.4, we can sort the importance of each word term. With different feature size, we can train the eight generic classifiers and get their performances on both training and testing set. Here we use SVM as the classifier to compare feature size's influence. Our results suggest that it performs best among the three. The results are shown in Figure FIGREF32 . The red squares represent the training accuracy, while the blue triangles are testing accuracies.

As shown in Figure FIGREF32 , it is easy for us to determine the feature size for each classifier. Also it's obvious that test accuracies of classifiers for plot, actor/actress, analysis, and thumb up or down, didn't increase much with adding more words. Therefore, the top 1000 words with respect to these classes are fixed as the final feature words. While for the rest of classifiers, they achieved top testing performances at the size of about 4000. Based on these findings, we use different feature sizes in our final classifiers.

Generalization of Classifiers

To prove the generalization of our classifiers, we use two of the TV series as training data and the rest as testing set. We compare them with classifiers trained without the replacement of generic tags like role_i or actor_j. So 3 sets of experiments are performed, and each are trained on top of Bayes, Logistic Regression and SVM. Average accuracies among them are reported as the performance measure for the sake of space limit. The results are shown in Table TABREF42 . “1”, “2” and “3” represent the TV series “The Journey of Flower”, “Nirvana in Fire” and “Good Time” respectively. In each cell, the left value represents accuracy of classifier without replacement of generic tags and winners are bolded.

From the above table, we can see with substitutions of generic tags in movie reviews, the top 5 classifiers have seen performance increase, which indicates the effectiveness of our method. However for the rest three classifiers, we didn't see an improvement and in some cases the performance seems decreased. This might be due to the fact that in the first five categories, roles' or actors' names are mentioned pretty frequently while the rest classes don't care much about these. But some specific names might be helpful in these categories' classification, so the performance has decreased in some degree.

Conclusion

In this paper, a surrogate-based approach is proposed to make TV series review classification more

generic among reviews from different TV series. Based on the topic modeling results, we define eight generic categories and manually label the collected TV series' reviews. Then with the help of Baidu Encyclopedia, TV series' specific information like roles' and actors' names are substituted by common tags within TV series domain. Our experimental results showed that such strategy combined with feature selection did improve the performance of classifications. Through this way, one may build classifiers on already collected TV series reviews, and then successfully classify those from new TV series. Our approach has broad implications on processing movie reviews as well. Since movie reviews and TV series reviews share many common characteristics, this approach can be easily applied to understand movie reviews and help movie producers to better process and classify consumers' movie review with higher accuracy.