Korean Movie Sentiment Prediction

Sooeun Oh
Master of Science Analytics
Georgetown University, Washington, D.C.
so472@georgetown.edu

Abstract

Movie rating prediction is one of the most interesting topics in machine learning field, because we have continuously increasing amount of data available online. At the same time, movie rating plays an important role as we can learn whether others recommend the movie or not. It is easy to find ourselves that we highly tend to rely on others' opinions when we search movie to watch. In this study, we explore various machine learning (ML) tools and natural language processing (NLP) methods to perform a prediction analysis of movie ratings. Prior to the prediction process, this paper presents data collection and step of data pre-processing with the introduction of Python libraries and packages available. The results show the experiment with various deep learning models and their accuracies.

1 Introduction

In recent decades, there have been many sources available online that provide us with the information of different items and services. There may be bunch of information such as price, review, rating, and related item. Review and rating of previous users are the most significant factors for us to determine the willingness to purchase the item. This system can be well applied to movie industry. We often read the review and rating of the movie in order to decide to watch the movie. In other words, movie is the product that we highly rely on the opinion of others via review and rating system. Therefore, sentiment analysis of movie review and opinion as well as the prediction analysis of movie rating has been popular since they can be used to classify movies into different categories and to improve the recommendation system. Natural language processing (NLP) tool are well-known to conduct the analysis. While there have been many successful attempts at sentiment analysis and prediction of English text-based content, few attempts have been made to classify text in Korea.

The research aims to develop a model using machine learning techniques to predict the rating of movie review in Korean language. I obtain the data via web-scraping and pre-processed the data including cleaning, creating a new feature, and tokenization. Then, I train a classifier based on logistic regression to predict movie ratings into three classes: negative, neural, and positive. I further implement additional classifiers such as Naïve Bayes, Ridge Regression, and stochastic gradient descent to compare the accuracy and select the best classifier that achieves the highest accuracy.

2 Method

2.1 Data Collection

This research uses 9,220 movie reviews taken from Naver, a South Korean online search engine. Naver is often called as 'the Google of South Korea' as it handled 74.7% of all web

searches in South Korea as of September 2017(Wikipedia). Since Naver does not provide a zip file of movie review data for users, the data is collected via URL by using Beautiful Soup, which is a Python library for parsing HTML documents. Beautiful Soup creates a parse tree for parsed pages that can be used to extract data from HTML, which is useful for web scraping.

After collecting 9,220 movie reviews, which is the maximum number of movie reviews to crawl allowed by Naver, the following data columns of each movie review are stored:

- User ID
- Movie Review
- Rating
- Title of the movie

The collected data is stored as a json file.

2.2 Pre-Processing

Since the research aims to predict the sentiment of movie based on the rating, the dataset needs to be pre-processed so that I can train a machine learning model efficiently. Therefore, I only keep movie review and rating in the dataset. As the rating is comprised of 0 to 10 scale, I create a new feature 'class' to classify the rating into three different categories of ratings: negative (0-4), neutral (4-8), and positive (8-10). I analyze the dataset and note the distribution of the dataset is highly skewed toward positive class of rating. Figure 1 shows the richness of the positive review but lack of negative and neutral review.

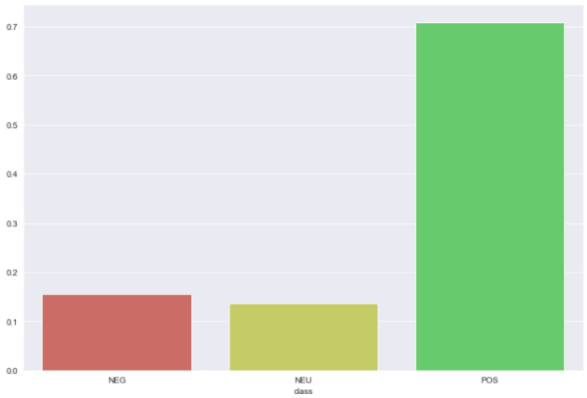

Figure 1. Distribution of Class

Stop Words. Movie review may contain stop words like 'the', 'is', or 'are', as I have not filtered out any stop words from the text. While NLTK(Natural Language Tool Kit) package provides a list of stop words in English text to be filtered out, Korean does not have a pre-made set of stop words[1]. As there is no universal list of stop words in natural language processing field, I create and define the stop words that meet the following conditions.

- Stop words
  - I. The length of words is ≥ 2
  - II. List of Stop words: 으로 이 에 가 을 는 은 다 이다 게 돼 있지 없지 에서 입니다 „ 였고 않는 하고 같아요 였다 때문 하는 명의 왔습니다 으로서 이렇게 그렇게 저렇게 그런 그래도 된다 번은 에선 가면 뭔가 하시는데 싶게 이라 줍니다 봤는데 이제 이었어요 었어요 크히히 나니 있는 라고 이었습니다 이런게 하며 아니라며 웬만하면 있네요 였습니다 로다 이라는 할만큼 이라 해서 하고 사 고 하라 하기 주는 보니 하게 까지 해서인지 였네요 에게 에 게 . ~ ~~ ~~~ 하며 하고 했네 ...? ..? ....? 라니 더니 그냥 없다 있다 하면 엇 급 진심 따름 이나 그리고 도 나 너 우리 ! !! !!! ? 함 임 거 겠죠 더 였어요 의 ~ 이런 저런 는데 됐 없는 수 한당 들

Tokenization. KoNLPy is a Python library for natural language processing of the Korean language that offers a variety of tools including POS tagging, morphological analysis, or tokenizer as we have such functions in a leading platform, NLTK, in English. Inside the KoNLPy, Okt class, an open source Korean tokenizer, is available to parse phrases of the movie reviews to morphemes.

Vectorization. With the help of TfidfVectorizer of scikit-learn library, I convert a collection of movie review texts into a matrix of TF-IDF features. As a part of parameters, I constrain to build a vocabulary that only consider the top 300 features ordered by frequency across the corpus.

[1] Stop words are highly frequent words such as 'the', 'is, or 'are' that do not have specific semantics.

2.3 Classifier

There are many different data mining, machine learning, and natural language processing tools available. For this study, I select four different classifiers to implement the prediction as shown in the Table 1.

Classifier
Logistic Regression
Ridge Regression
Naïve Bayes
Stochastic Gradient Descent

Table 1. List of Classifier

3 Results

To evaluate the performance of the methods, I use the accuracy on the test data set with different classifiers. The performance of each classifier can be evaluated by the fraction of correctly predicted movie rating from the test dataset. The overall accuracy of all classifiers implemented is presented in Figure 2. Though logistic regression achieves the best accuracy over the rest of the classifiers, the model obtains around 75% accuracy across the classifiers. As I note the skewed distribution of the dataset via Figure 1, I also want to see how accuracy changes in different classes: negative, neutral, and positive. As expected, the accuracy of positive rating prediction is always higher than that of neutral and negative rating prediction regardless of the classifiers as shown in Figure 3. Thus, the study supports an idea that positive rating has enough training set to evaluate accuracy while negative and neutral rating do have not enough data to give us a descent rate of accuracy.

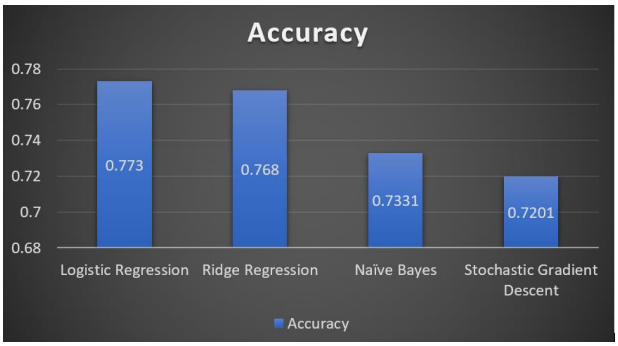

Figure 2. Overall Accuracy

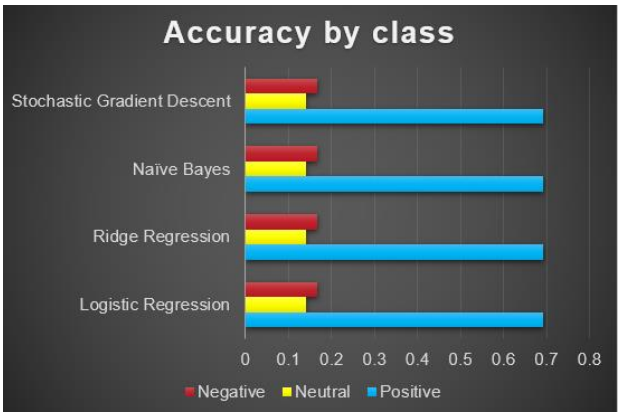

Figure 3. Accuracy by class

4 Discussion and Conclusion

From the obtained results, the research finds out that the performance of logistic regression achieves the best result even though positive reviews have a better prediction due to an enough training dataset regardless of the classifiers.

A key issue to this research is that the lack of dataset and the dataset is skewed. Due to the limitation of web-scrap by Naver, the dataset contains 9,220 movie reviews and ratings, which is the maximum number allowed. However, the performance of classifiers can be improved if I have more dataset available. While the dataset has enough positive ratings, negative and neutral ratings are comprised of only 30% of the dataset. Due to this nature, the accuracy of overall research is not descent. However, the positive accuracy for each classifier is always higher than negative and neutral accuracies. Therefore, the accuracy is highly dependent on the training set.

During the study, I observe that one review can have multiple opinions; even though the review has the corresponding rating, the review can still contain positive and negative aspects at the same time while the rating can be evaluated as neutral. It is interesting to see how to take such case into consideration and evaluation. Also, I learn that many reviews contain special characters that are difficult to evaluate the sentiment unless human carefully read and hand-assign the sentiment of the reviews. Special characters can deliver opinions but sometimes they do not. I put a lot of effort to keep my stop words to be unbiased by excluding most special characters.

Though the accuracy may be heavily relying on the training set, the research is meaningful as

one of a few researches in Korean natural language processing studies. Despite the limited size and variety of the dataset, the research is interesting and promising to develop a list of stop words in Korean and conduct prediction analysis of movie rating. As sentiment analysis and prediction in Korean language is a challenging subject, there are many topics to be studied and discussed in the future.

```
Review 김성오 배우분이 연기를 너무 잘해서 재밌게봤어요 모든 잘 어울리는 배우인거 같네요
Original Rating: POS
Predicted Rating: POS

Review 아름답고 신비한 영화였습니다
Original Rating: POS
Predicted Rating: POS

Review 언제나 다시봐도 명작 마지막 컷  주인공 가정부와  친구의 대화가 진한 여운을..
Original Rating: POS
Predicted Rating: POS

Review 영상미가 너무 아름다웟다
Original Rating: POS
Predicted Rating: POS

Review 편집을 심각하게 한건지 영화가 심각한건지 구분이 안된다.
Original Rating: NEU
Predicted Rating: POS
```

Figure 4. Example of Prediction

Acknowledgments

I thank Dr. Loehr for giving me this opportunity and his valuable comments and feedbacks on this research.

References

Battu Varshit and Batchu V. Vishal. 2018. Sentiment as a Prior for Movie Rating Prediction. India

Vasu Jain. 2013. Prediction of Movie Success using Sentiment Analysis of Tweets. Los Angeles, CA.

Hadi Pouransari and Saman Ghili. 2015. Deep Learning for Sentiment Analysis of Movie Reviews. Stanford University, Stanford, CA

Kub, A. 2018. Sentiment Analysis with Python (Part 1) – Towards Data Science. [online] Towards Data Science.

Wikipedia contributors. Wikipedia, The Free Encyclopedia.