Sentiment Polarity Analysis of Chinese Movie Reviews

Shilpa Balan
Department of Information Systems
California State University, Los Angeles

Jia Yangyang
Department of Information Systems
California State University, Los Angeles

Kathryn Joy Mak
Department of Management
California State University, Los Angeles

Linray Song
Department of Management
California State University, Los Angeles

Keywords: polarity, sentiment, movie reviews, recall, precision

Extended Abstract

Customer's opinions have always been an important piece of information in decision making process. The sentiment analysis of the web content has been of significant interest to researchers in the last decade [1]. It is important to understand the sentiments of users toward various products and services because it enables improved contextual advertisement, recommendation systems and analysis of the market trends [2].

For this study, we attempt to predict the sentiment polarity of Chinese movie reviews from the web. The reviews are obtained from the IMDB movie website that contains ratings by users for the popular Chinese movies [3]. The movies ranged between years 1971 and 2015. For our study, we selected the top 100 Chinese movies by user ratings.

First, we attempted to examine the movie reviews in Chinese language itself. We then used Google Translate to translate these movie reviews to English. However, due to the limitations of Google Translate, we were unable to look at the reviews in Chinese language. For example, a Chinese movie titled '*Ying xiong ben se*' is translated by Google Translate as '*Hero'*. However, the correct translation for this movie title in English is '*Heroic Nature*'. Hence, due to this limitation, we analyzed the Chinese movie reviews that were already available in English.

To evaluate our sentiment analysis framework, we use a sentiment analysis software, *SentiStrength* that is free for academic research [4]. This software is written in Java. SentiStrength developed by Thelwall et al. [5] is one of the powerful sentiment analysis tools for English in the literature.

SentiStrength estimates the strength of positive and negative sentiments in short texts. SentiStrength reports binary (positive/negative) and trinary (positive/ negative/ neutral) classification results.

For our data set of the top 100 Chinese movie reviews, we found that SentiStrength detected 80 positive reviews and 20 negative reviews using the binary classifier methodology. Using the trinary classifier methodology, 59 movie reviews were found positive, 25 were found negative and 16 movie reviews were found neutral.

Examples of positive movie reviews we found from our data using SentiStrength are: 1) '*I just saw this film today. I was totally captivated*', 2) '*If you love Kung Fu films you will definitely enjoy this one.'*

Examples of negative movie reviews we extracted from our data are: 1) '*Well I was a little bit disappointed after I saw this movie*', 2) '*Unfortunately the rest of the film doesn't hold up to this quality*'.

Further, we evaluated the performance in terms of accuracy. The accuracy is computed by the ratio of the number of movie reviews whose polarity is correctly predicted to the total number of reviews [6]. For example, using the binary classifier methodology of the SentiStrength software, the number of True Positives are 77, True Negatives are 8, False Positives are 3 and False Negatives are 12. Using the trinary methodology, we see that the number of True Positives are 42, True Negatives are 19, False Positives are 17 and False Negatives are 16. The recall of the binary and trinary classifier methodologies of the SentiStrength software for our movie review data set is found to be comparable at 86% and 87% respectively. However, the precision of the binary methodology (96%) is higher than that of the trinary methodology of the SentiStrength software for our data set. When comparing the recall and precision of the two methodologies, we eliminated any reviews detected as neutral by SentiStrength.

In this study, we have proposed a framework for sentiment analysis of Chinese movie reviews. For our analysis, we used the corpus of words trained by SentiStrength. For future, we plan to use our customized repository of words along with SentiStrength.

References

[1] Bo Pang, Lillian Lee. Opinion mining and sentiment analysis. Found. Trends Inf. Retr. 2, 1–135 (2008).

[2] Avni Vural. Sentiment-Focused Web Crawling. Thesis. Retrieved from http://etd.lib.metu.edu.tr/upload/12616409/index.pdf (2013).

[3] IMDB Movie Review Data set. Retrieved from https://www.imdb.com/list/ls064849128/ (2019).

[4] SentiStrength, Retrieved from http://sentistrength.wlv.ac.uk (2019).

[5] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai. Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology 61(12):2544–2558 (2010).

[6] Vural A.G., Cambazoglu B.B., Senkul P., Tokgoz Z.O. A Framework for Sentiment Analysis in Turkish: Application to Polarity Detection of Movie Reviews in Turkish. In: Gelenbe E., Lent R. (eds) Computer and Information Sciences III. Springer, London, (2013).