Analysis on Li Qingzhao's Poems: Discussion on the Limitations of Applying Sentiment Analysis on Chinese Literatures of the Song Dynasty

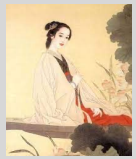
Alexandra Hoi Ting LAW
The University of Hong Kong

Abstract

Natural language processing (NLP), concerning the interactions between computers and human languages, has been widely used in processing and analyzing large amounts of data in human languages. NLP not only enables computers to understand the contents of human languages in different media, for example documents and audio, its subfields have been universally employed to handle lexical tasks. Sentiment analysis, which has been used to quantify and study affective states and subjective information, is an example. Although a rising number of packages and algorithms for sentiment analysis have been introduced to further enhance the accuracy and usefulness of the analysis, they are not that developed for analyzing Chinese characters compared to that of English. SnowNLP, one of the few packages for sentiment analysis in Chinese, despite being acclaimed as having satisfying accuracy in the analysis, its efficiency in analyzing ancient Chinese characters is still questionable. This paper aims to evaluate the efficiency and accuracy of sentiment analysis on classical Chinese literature of the Song Dynasty by using SnowNLP to analyze all poems (Song Ci) of Li Qingzhao, a famous female poet in the Song Dynasty. In this paper, an experimental analysis of all poems of Li by SnowNLP are discussed, followed by the evaluation based on the analysis results. Finally, suggestions on ameliorating the outcomes of sentiment analysis are provided.

Keywords: Natural Language Processing, Sentiment Analysis, Chinese sentiment analysis, SnowNLP model, Naive Bayes Algorithm

Backgrounds



Li Qingzhao is a famous poet in the Song Dynasty. Her styles of writing varied with her marriage status and experiences of leaving her hometown after the collapse of Song. The sentiments of her existing 50 poems have been popular research targets in various research areas, including Chinese Literature, History and Asian Studies.

Song Ci, uses a form of poetic meters including variable line-length, fixed-rhythm and fixed-tone, is a type of lyrics poetry in classical Chinese literature. Ci was used to express the emotions of the poets through the depiction of scenes and historical figures.

Introduction

What is Sentiment Analysis/ opinion mining ?

- studies people's sentiments and feelings in statements
- are used in marketing, psychology, and education
- are applauded with its usefulness and accuracy
- can be applied in different languages

Significances of conducting this research/ Why this research needs to be conducted?

- the effectiveness of sentiment analysis in analyzing Chinese is under discussion
  - complexity of written Chinese
  - undeveloped databases for Chinese characters
- growing use of packages and algorithms for sentiment analysis in Chinese characters, e.g. SnowNLP
  - effectiveness disputable
- insufficient usage of SnowNLP in analyzing classical Chinese

What is SnowNLP?

- Python library enabling translation between traditional and simplified Chinese characters, text categorization sentiment analysis and pinyin-characters conversions
- This paper uses the SnowNLP tool to perform sentiment analysis on all 50 poems written by Li Qingzhao

Focus of this research

- investigates how does SnowNLP perform in conducting sentiment analysis on Classic Chinese
- discusses the causes of SnowNLP having deviations in sentiment analysis on Classic Chinese
- introduces methods of how SnowNLP improve its accuracy in analyzing Classic Chinese

Methods

Assumptions of the research

- literature reviews on every single poems in the dataset analyze all the sentiments of the poem accurately
  - the literature reviews serve as the only indicator to evaluate the accuracy of sentiment analysis
- the contents and wordings of the poems in the dataset are as same as what Li wrote in the Song Dynasty
  - the research aims at evaluating the accuracy of SnowNLP on Chinese literature in the Song Dynasty.

Methods of maintaining the credibility of research

- Rerun and logical check of algorithms used are conducted to make sure no syntax error can be found
- Contents of the data are crosschecked with at least three literacy sources

Research Process

The results of sentiment analysis on 47 poems of Li are analyzed as either 'Happy' and 'Sad'. Literature reviews are conducted on all poems to understand the sentiments of the poems. The results from literature preview and SnowNLP will be compared to evaluate the accuracy of the package.

SnowNLP library

- based on the naive Bayes algorithm
  - a model of conditional probability, whose process of classifying two sentiments, c_1 and c_2, which have the characteristics of $w_1, \dots, w_n$, is

$$P(c_1|w_1 \dots w_n) = \frac{P(w_1, \dots, w_n|c_1)}{P(w_1,\dots,w_n)} \times P(c_1)$$

where $P(w_1, \dots, w_n) = P(w_1, \dots, w_n|c_1) \times P(c_1) + P(w_1, \dots, w_n|c_2) \times P(c_2)$

The above formula calculates the probability of each poem achieving positive sentiment according to the SnowNLP library. The SnowNLP library, whose value is between 0 to 1, generates the tendency result which is a probability of having positive sentiments. If the value is greater than or equal to 0.6, the contents are likely 'Happy', else it is a 'Sad' statement.

Results

Left: The Sentiments of the 47 Li Qingzhao's Poems analyzed by Literature Review
Right: The Sentiments of the 47 Li Qingzhao's Poems in traditional Chinese analyzed by SnowNLP

Left: The Changes of Sentiments of Li Qingzhao's Poems analyzed by Literature Review
Right:The Changes of Sentiments of Li Qingzhao's Poems in Traditional Chinese Analyzed by SnowNLP

Highlight

"Melody of Becoming Drunk in the Shade of Blossoms" (醉花陰) – One of the most famous poem of Li Qingzhao

- Expresses the loneliness of the writer when her husband left her for work
- 0.988 probability for "Happy" sentiment in SnowNLP

Results Reflection

Causes of the deviations in analysis

- poems written in the Song Dynasty heavily relies on the depicture of the environment to illustrate the emotions of the writers - implicit expression of emotions
- literature metaphors are used in the Song poems
- the trained dictionaries of the SnowNLP library is not well-developed
  - dictionaries mainly consist of everyday Chinese, but not the vocabulary used in classical Chinese literature

Suggestions of Improving the Accuracy of Conducting Sentiment Analysis with SnowNLP on Classics Chinese

- Train databases/ dictionary before conducting sentiment analysis with SnowNLP
  - existing databases of SnowNLP is not developed in terms of numbers of vocabulary and language setting
  - users can build or prepare a database or a dictionary based on the contents they are analyzing
- Make large amounts of contents shorter
  - to spot out which lines are the most influential to affect the results of the sentiment analysis

Conclusions

- SnowNLP performs less satisfactory in the analysis of Chinese Literature than that of everyday Chinese
- Potential causes include the uses of literature metaphors, implicit language expressions, and unmature dictionaries built for analyzing classics Chinese in SnowNLP
- Applying specific dictionaries for analysis and the break-down of long contents help enhance analysis

Future Research Suggestions

- investigate other causes of having insufficient outcomes of SnowNLP when conducting sentiment analysis on classics Chinese with other literature sources
- explore other ways of enhancing the quality of conducting sentiment analysis of SnowNLP
- research on building more powerful database/ dictionaries of SnowNLP
- investigate the efficiency of SnowNLP on other styles of literatures, e.g. online comments

References

Liu, B., & Zhang, L. (2012). A survey of opinion mining and sentiment analysis. Mining Text Data, 415–463. https://doi.org/10.1007/978-1-4614-3223-4_13

J. Wu, K. Lu, S. Su and S. Wang, "Chinese Micro-Blog Sentiment Analysis Based on Multiple Sentiment Dictionaries and Semantic Rule Sets," in IEEE Access, vol. 7, pp. 183924-183939, 2019, doi: 10.1109/ACCESS.2019.2960655

Chen, C., Chen, J., & Shi, C. (2018). Research on credit evaluation model of online store based on SnowNLP. E3S Web of Conferences, 53, 03039. https://doi.org/10.1051/e3sconf/20185303039