Food-Related Sentiment Analysis for Cantonese

Natalia Klyueva[1], Yunfei Long[2], Chu-Ren Huang[1], Qin Lu[2],
[1]Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University
{natalia.klyueva,churen.huang}@polyu.edu.hk
[2]Department of Computing, The Hong Kong Polytechnic University
{csylong,csluqin}@comp.polyu.edu.hk

Abstract

In this paper, we describe a work in progress on collecting data from the domain of food reviews in Hong Kong for sentiment analysis and the initial experiments conducted on the data. Another goal of this study is to use the described data to create a sentiment lexicon for Cantonese, which will serve as a resource for opinion mining related machine learning experiments.

1 Introduction

Sentiment analysis is one of the most popular topics in Natural Language Processing, as the opinions of people present a very valuable asset for companies to develop and improve their products based on customer expectations and feedback. Luckily for this research area, product review comments and feedback are likely to be posted on the internet, free and in big quantity. User generated content is especially large in the domain of customer reviews, and that will be the source of our data - reviews in the food domain.

The approaches in automatic text classification according to their polarity traditionally involved supervised machine learning methods. Various linguistic characteristics (like positive or negative words, part-of-speech patterns that can characterize polarity) of the text that potentially contained opinion bits served as features for a classifier.

Nowadays, the most widely used technique to encode words, sentences and even texts – word embeddings – Latest development in deep learning has found its way into the sentiment analysis in which words are represented by a dense vector learnt from embedding models such as word2vec or Doc2Vec.

One of the "linguistic" knowledge that can potentially improve the accuracy of sentiment analysis are sentiment lexicons or more complex emotion lexicons. They are integrated into the machine learning pipeline, e.g. (S.K.Rastogi et al., 2014). Sentiment lexicons are important resources for sentiment analysis. These lexicons consist of predetermined list of words assigned to sentiment labels or values, which is baseline for many machine learning based methods(Liu and Zhang, 2012; Tabak and Evrim, 2016; Long et al., 2017). Depending on sentiment models, there are two mainstream labeling schemas. The first schema is representing affective meanings of words by discrete sentiment labels, such as as positive, negative, etc (Ekman et al., 1983). The second schema is to represent affective meanings by means of more comprehensive multi-dimensional representation models, like the valence-arousal dominance model (VAD) (Russell, 1980) and the evaluation-potency-activity model (EPA) (Heise, 1965).

Sentiment lexicons are heavily investigated in English (Gilbert, 2014; Li et al., 2017) and Chinese (Wang and Ku, 2016). However, building sentiment lexicon for low-resources languages or dialects is not an easy task. That is why we decided to collect and investigate the text in Cantonese. Cantonese is a variety of Chinese spoken in the city of Guangzhou (historically known as Canton) and its surrounding area in southeastern China, including HK SAR and Macao SAR [1]. Compared to the studies on Man-

[1]https://en.wikipedia.org/wiki/Cantonese

32nd Pacific Asia Conference on Language, Information and Computation
The 25th Joint Workshop on Linguistics and Language Processing
Hong Kong, 1-3 December 2018

darin, studies on Cantonese are still in the infant stage and there is no available Cantonese sentiment lexicon useful for Cantonese sentiment analysis. In this paper, we present the initial attempts to build the lexicon with the help of students annotation.

This paper is structured as follows. First, in Section 2 we describe how we collected and processed the data from the web. In Section 3 we show the data pre-processing and implementation of a baseline classification models. In Section 4 we derive a sentiment lexicon that can be integrated into the machine learning pipeline to improve the baseline results. So far this experiment is yet to be completed. We conclude and show the plans for future in Section 5.

2 Data collection

The data were collected from the web site with reviews of food and restaurants in Hong Kong `www.openrice.com` using python scrapper (scrapy), with the permission of the portal owner to perform scientific (non-commercial) experiments with the data. The special restriction was made on user-sensitive data, like nickname and amount of money spent.

According the the customer reviews, the restaurants are assigned with a rank and the search can be done both according to the restaurant IDs and food name. The reviews are written mostly in Cantonese Chinese, but also in English and Mandarin Chinese. They feature exactly the information needed to perform sentiment analysis: text reviews and the scores that reflect customer opinion (sentiment). Figure 1 demonstrates the example of a review.

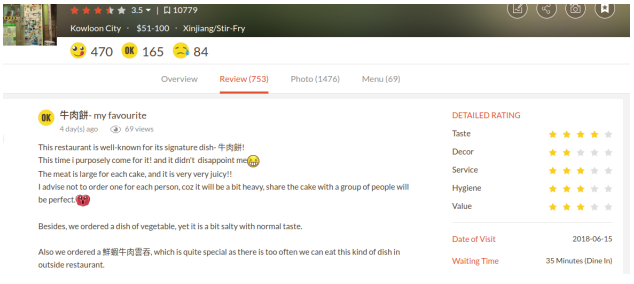

Figure 1: Example of a customer's review on openrice

During the crawling, the information from the web was extracted and stored into *json* format. The format of the crawled data is the following:

```
{"res_name": "清 真 牛 肉 Islam Food", "res_id": "1441",
"taste": 5, "environment": 3,
"service": 3, "sanitation": 3,
"worthy": 3, "title": "不能 的爆汁牛肉!!!",
"comment": "九 城. 不只是泰 菜出名. 有不少地道特色的餐 都非常有名. 人人都著朝....", "pictures": ...}
```

The statistics of the crawled data is presented in Table1

comments	360,701
restaurants	4,924
customers	about 60,000

Table 1: Crawled data: basic statistics

The raw data contained comments in 3 languages: Cantonese, Mandarin and English. Code-switching (e.g. mix of English and Cantonese) was observed in a number of cases. For our experiments we selected the comments that were mostly in Cantonese and could involve minor chunks/words in English.

3 Baseline Model for Sentiment Analysis

The standard approach to sentiment analysis in the domain of reviews generally includes the basic task - training the model to predict the sentiment score of a review. The first step includes text data encoding as vectors powered by Doc2Vec/Word2Vec models. [2] The output is a sentiment score/mark - which can be either binary (positive/negative) or on a larger scale including neutral.

In the initial stage of this experiment, we have selected only one type of label - the overall rating of the restaurant as displayed by the emoji in the interface.

We believe it to be the most representative evaluation tag of the review, the one that really reflects the *sentiment/opinion* of a customer.

In sentiment analysis, the classification can be performed either on a binary scale (positive vs. negative) or on a multi-class scale that is suggested by

[2] `https://github.com/RaRe-Technologies/gensim/blob/develop/docs/notebooks/doc2vec-lee.ipynb`

the data: rating from 1 to 5 stars. In our experiments we will compare both types. [3]

3.1 Data Preprocessing

In order to feed the data to the model, we needed to convert it to the format required by the framework. First, we filtered out the data points that do not contain reviews or are in a language other than Cantonese and created two sets: one for multi-class classification, another for binary classification, the statistics is presented in Table 2.

	multi-class	binary
whole data	360,000	n/a
cleaned data	26,549	4,450

Table 2: Summary of the corpus size, number of reviews for data with multi-class classification (review scores on slace 1-5), and for the data with binary (1 or 5 star score) classification

The second step was text tokenization. As the Doc2Vec input should present sentences segmented into words, we tokenized Chinese text using Jieba segmentor (`https://github.com/fxsjy/jieba`). The website did not specify any metrics to measure accuracy of the tool, so we just relied on it as the most popular tokenizer for Chinese.

3.1.1 Text Encoding: Doc2Vec

In our experiment – unlike in the original Doc2Vec tutorial – we used unlabeled documents-reviews (the reviews are not labeled). After preprocessing the reviews (cleaning, tokenization), we trained the model on the openrice data using Doc2Vec implementation. We choose the vector size of 100, so that each review is encoded into the array of this size. After training the model, we substituted the reviews in the corpus with the respective vector according to the model and thus the text data became numerical and could had served as input for standard classifiers.

The side-effect of the trained model is its capability to represent relations between words or sentences (Mikolov et al., 2013). We tested on the following cases just to demonstrate the power of word analogies as measured in the vector space:

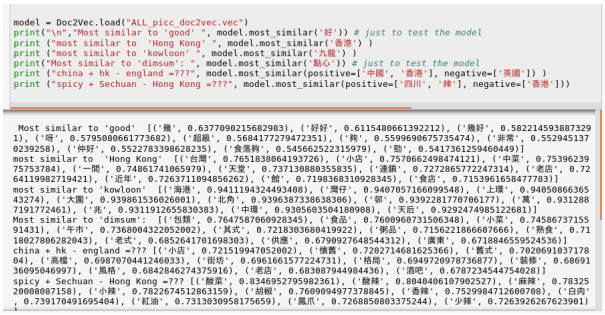

Figure 2: Lexical relations obtained from the model using the function *most_similar*

3.2 Preliminary Experiments

In our preliminary experiments, we trained three classifiers using *sklearn* python library:

- Logistic regression (logreg)
- Support Vector Machines (SVM)
- k-Nearest Neighbours

We conducted the following experiments; results are summarized in Table 3:

- **Multi-class classification**. For this, we used all the reviews that were ranked on the scale from 1 to 5 (in stars). The average accuracy for the classifiers was about 0.5.

  Such a low score can be attributed to the fact that the sentiment scale was quite large - from 1 to 5, and there is not much typical 'emotion' expression in the neutral (score 3) reviews.

- **binary classification** We took only the reviews with a either very negative (1) or very positive (5) score. For the same experiments setup the accuracy was much higher than for the baseline with all 5 sentiment classes.

Above all, we tested those setups for the whole data and for the half-size data to explore correlation between data size and accuracy of classifiers. It demonstrates that with bigger data the accuracy does not change at all for multi-scale classification

[3]The scripts with a small sample of data are available at: `https://github.com/polyu-llt/openrice_annotations`

	all-multi	all-bi	half-multi	half-bi
Logreg	0.493	0.854	0.477	0.773
SVM	0.464	0.854	0.469	0.775
Kneigh	0.408	0.823	0.408	0.77

Table 3: ML experiments: Logistic regression, SVM and k-neighbours; for all data, and half-size data

and only gets a little bit higher for a binary classifier.

As we have mentioned in the introduction, one of the possible ways to improve precision of the prediction is introducing the sentiment lexicon – a list of words that will help to identify polarity of the comment. In the next section, we describe work in progress on the deriving the lexicon.

4 Sentiment Lexicon via Crowd-sourcing

In order to get the polarity (evaluative) words to form the lexicon, we asked the students from the linguistic department to annotate the evaluative words alongside with the polarity and the aspect – a word to which the candidate word/phrase is related.

4.1 Setup of the Experiment

60 students were asked to annotate an excel file containing 20 comments. To measure the inter-annotator agreement, comments were randomly assigned twice to two different students. The evaluative labels from crawled data were not shown to the students.

4.2 Annotation Guidelines

For each evaluative word found in a comment students were requested to attach:

- the aspect word (the object/service/meal) that is being evaluated
- the sentiment (evaluative) word
- the sentiment mark ('+' - positive; '-' - negative; '0' - neutral).

They had to put the above information into one of five columns–categories to which evaluation relates: 'taste words', 'overall words', 'environment words', 'service words', 'sanitation words' and 'worthy words'. Some evaluations could be 'general', not referring to any aspect, in this case the annotation had to be done without an aspect word:
Review: *Amazing experience*!!!
Annotation: overall column: Amazing experience+

The aspect words/phrases and sentiment words/phrases should be separated by a colon in case more than one is present in a review:
review: *Love the bread and the soup! The birthday soufflé dessert is surprisingly yummy! The staff though was so unfriendly and mean!*
Annotation:

- Taste column: bread:love+; soup:love+; soufflé dessert: surprisingly yummy+
- Service column : staff:so unfriendly-/mean-

Different examples of sentiment words should be separated by semicolon (';'). If for one aspect there were more evaluative words, they were separated by a pipe: ('|'):
Example: good and yummy beef, but was too salted
Annotation: Taste column: beef:good+|yummy+|too salted-

Example in Cantonese('taste' column):

(1) 羊架: 最欣+
roast-sheep-neck: most-appreciated
'roast-sheep-neck: most-appreciated'

4.3 Lexicon Cleanup

We collect 600 comments from 60 native speaker students. We take the intersection of lexicon between two different annotators. The intersected words are then fed into second round of human checking. Together we get 1,887 positive words (Positive_words.txt) and 858 negative words(Negative_words.txt).

The resource with raw (unprocessed) annotations as well as the cleaned lexicon can be found at [4].

The sentiment lexicon was intended to serve as additional data for machine learning experiments

[4] https://github.com/polyu-llt/openrice_annotations/tree/master/annotations

that we will describe in Future work section. However, it will have a value on its own, as data for linguistic analysis for a low-resource language.

5 Conclusion and Future work

In this paper, we presented our research project on food-related sentiment analysis for Cantonese language. We described how the data were collected from the web openrice, showed the preliminary machine learning experiments on the data. For binary classification the results displayed higher accuracy than in the case of multi-class classification.

As for future work, we plan to integrate the sentiment lexicon for Cantonese into the ML pipeline. We also illustrated the steps for creating the lexicon by means of crowdsourcing, as well as the guidelines presented to the students for the annotation of positive and negative words in the reviews.

6 Acknowledgements

This work has been supported by the postdoctoral fellowship grant of the Hong Kong Polytechnic University (PolyU RTVU), project code G-YW2P; and GRF grant (CERG PolyU 15211/14E, PolyU 152006/16E).

References

Paul Ekman, Robert W Levenson, and Wallace V Friesen. 1983. Autonomic nervous system activity distinguishes among emotions. *Science*, 221(4616):1208–1210.

CJ Hutto Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available at (20/04/16) http://comp. social. gatech. edu/papers/icwsm14. vader. hutto. pdf.*

David R Heise. 1965. Semantic differential profiles for 1,000 most frequent english words. *Psychological Monographs: General and Applied*, 79(8):1.

Minglei Li, Qin Lu, Yunfei Long, and Lin Gui. 2017. Inferring affective meanings of words from word embedding. *IEEE Transactions on Affective Computing*, (1):1–1.

Bing Liu and Lei Zhang. 2012. A survey of opinion mining and sentiment analysis. In *Mining text data*, pages 415–463. Springer.

Yunfei Long, Lu Qin, Rong Xiang, Minglei Li, and Chu-Ren Huang. 2017. A cognition based attention model for sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 462–471.

T Mikolov, W.-T Yih, and G Zweig. 2013. Linguistic regularities in continuous space word representations. pages 746–751, 01.

James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

S S.K.Rastogi, Rohit Singhal, and Anil Kumar. 2014. An improved sentiment classification using lexicon into svm. 95:37–42, 06.

Feride Savaroğlu Tabak and Vesile Evrim. 2016. Comparison of emotion lexicons. In *HONET-ICT, 2016*, pages 154–158. IEEE.

Shih-Ming Wang and Lun-Wei Ku. 2016. Antusd: A large chinese sentiment dictionary. In *LREC*.