Valence-Arousal Ratings Prediction with Co-occurrence Word-embedding

Yan Gao, Jing Xu, Xu Yang, Bin Xu
SAS R&D (Beijing) Co Ltd, SAS Institute Inc.

Introduction

Sentiment analysis draws increasing attention of researchers in wide-ranging fields. Compared with the commonly-used categorical approach representing affective states as a few discrete classes, the dimensional approach represents emotions as continuous numerical values in multiple dimensions, such as valence-arousal (VA) space. It can thus provide more fine-grained sentiment analysis. However, affective lexicons with VA ratings are very rare. This limitation makes the dimensional approach hard to use in reality. This study proposes a VA ratings prediction method combining word2vec and KNN.

Data

We used two data sets to build word2vec. The first one is the Google Crawl archive for February 2016 from Common Crawl, from which we extracted a size of 13.5G traditional Chinese texts. The other one is Sinica Corpus 4.0 that we have purchased.

The third data set used in the study is the NRC Emotion Lexicon which contains 1835 positive (tagged 'P') words and 2317 negative (tagged 'N') words.

We also used the train data set cvaw1_utf8 and the test data set DSAW_Test_NewInput provided by the competition.

Methods

First, we used SAS tokenizer to segment the Google Crawl archive above mentioned, and then combined them with Sinica Corpus to build a word embedding model with GloVe. We extracted vectors of affective words from NRC and the shared task data, namely Affective_Word_Vector.

Second, we divided the train data set cvaw1_utf8 into two parts, 80% for train and 20% for validation.

Third, we trained the KNN models and the Neural Network models respectively. In the KNN models of Valence, to solve the sentiment-ignorance problem of the co-occurrence word embedding model, we adjusted the neighbors' valence ratings based on the sentiment polarity of the target word. If the polarity was positive and the valence rating of its neighbor was less than 5, 10 minus valence rating was assigned to its neighbor in our formula. The similar adjustment was applied to the valence prediction of the target words with negative polarity.

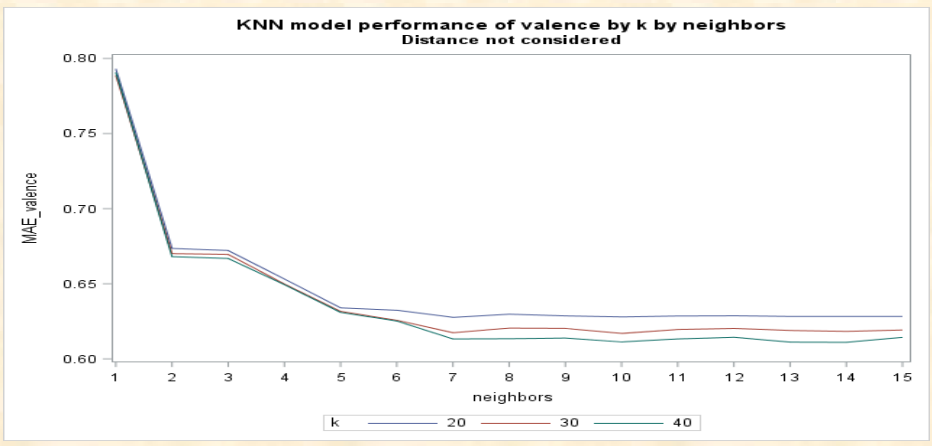
Last, we evaluated all KNN models with validation data and chose the best KNN models.

Results

We used two measures in the model evaluation. First, to measure the accuracy we used MAE (Mean Absolute Error) as in the challenge. Second, we evaluated the model's generalization ability. We used two data sets, i.e., one was the validation data and the other was NRC sentiment words.

We compared KNN models and Neural Network models with validation data and the results showed that KNN models outperformed Neural Network models.

Meanwhile, we predicted VA ratings of NRC lexicons and we thought VA ratings prediction was reasonable based on the scatter plot. So we used the KNN models in the challenge task.



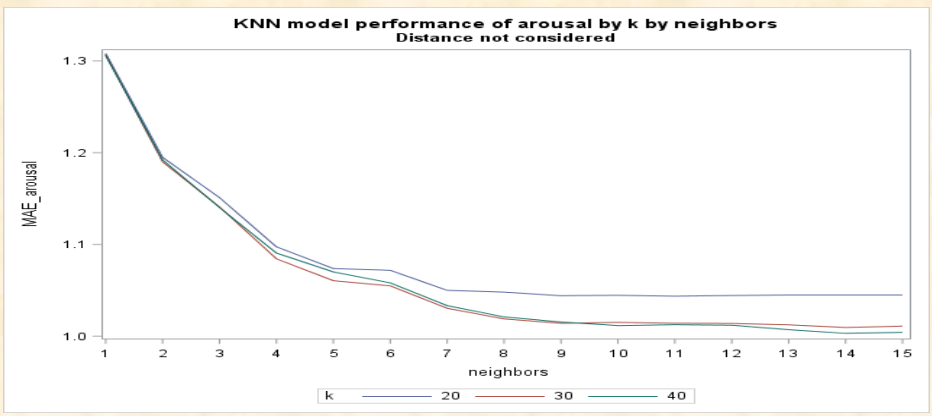


TABLE I. ACCURACY EVALUATION OF TRAIN DATA

Model	MAE of Valence	MAE of Arousal
Neural Network	0.365867	0.329759
KNN	0.4622995647	0.8197367743

TABLE II. ACCURACY EVALUATION OF VALIDATION DATA

Model	MAE of Valence	MAE of Arousal
Neural Network	2.0818307358	1.8582470789
KNN	0.6132667011	1.0332612214

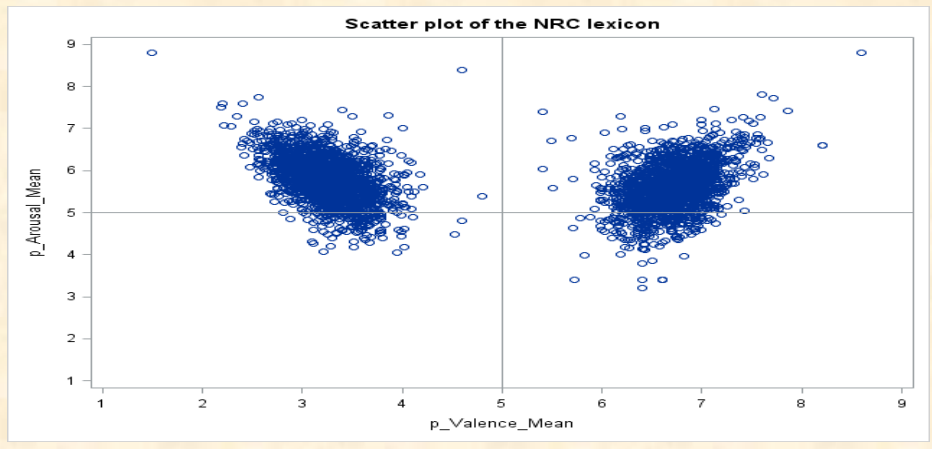
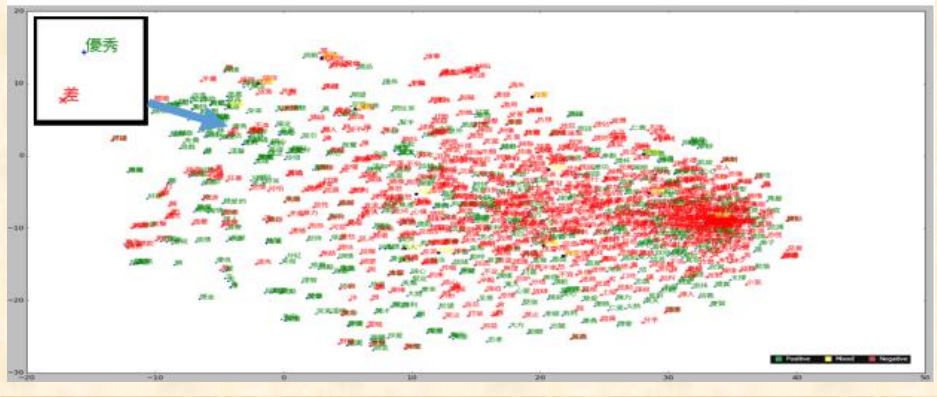


TABLE III. ACCURACY EVALUATION OF TEST DATA

	MAE of Valence	PCC of Valence	MAE of Arousal	PCC of Arousal
KNN model	0.805	0.845	1.334	0.304

T-SNE PLOT OF TRAIN LEXICON



Conclusion

This study presents a valence-arousal rating prediction model consisting of word2vec and KNN. Compared with Neural Network, KNN produces much better results regarding the generalization ability. With a group of constructed corpus with valence-arousal ratings, the experiment provides a feasibility evaluation for VA ratings prediction of new affective words. However, the experiment result shows that the MAE of Arousal and PCC of Arousal are far from satisfaction. There is still much room for improvement even for the performance of valence. Regarding the future work, we will explore how to apply the sentiment-specific word embedding to improve the VA ratings prediction ability.