# TMUNLP System at the NTCIR-14 STC3 Task

Yan-Chun Hsing[1], Chien-Hung Chen[1], and Yung-Chun Chang[1,2*]

[1]Graduate Institute of Data Science, Taipei Medical University, Taipei, Taiwan
[2]Pervasive AI Research Labs, Ministry of Science and Technology, Taiwan
{m946106006, m946106007, changyc}@tmu.edu.tw

**Abstract.** This paper presents our approach to the Chinese emotional conversation generation (CECG) subtask in the short text conversation (STC) task at NTCIR-14. The official training data contains 600,000 pairs of post and response from Weibo, from which we remove noisy data and train our model. The proposed methods to generate responses mainly include retrieval-based and generation-based approaches. We construct a sequence-to-sequence-based model, which is commonly used in generation-based methods, to generate responses that contain emotions of our choosing. Besides, we also propose a refined distributed emotion vector (RDEV) representation model, which is an emotion detection method based on valence and arousal, to improve the responses so that they would contain appropriate content as well as adequate emotion. RDEV combines convolutional and recurrent neural networks, and performs remarkably well on the dataset of emotion analysis in Chinese weibo texts task in NLPCC 2014. Our final evaluation results achieve an average score of 0.32 from three annotators. The performance of our system is very promising, although not the best among the competing teams in this challenge.

**Keywords:** valence-arousal, emotion detection, seq2seq
**Task Name**: Short Text Conversation Task (STC-3)
**Subtasks:** Chinese Emotional Conversation Generation (CECG)

## 1 Introduction

With the popularity of mobile devices and social media platforms (Twitter, Weibo, etc.,) social interactions between modern human has changed drastically, and the amount of conversation data on the Internet has increased greatly. Communicating through short text conversations (STC) has become an important way of emotional sharing [1][2]. In view of the advancement of artificial intelligence (AI) and natural language processing (NLP), chatbots can effectively talk with people. In order to facilitate the talking functionality of the robots with users, it is necessary to understand the cognitive behaviors of humans. Expressing and understanding emotions is one of the

[*] Corresponding authors. Fax: +886-2-6638-2736 ext. 1184 (Y.C. Chang).
E-mail address: changyc@tmu.edu.tw (Y.C. Chang).

most important characteristics of humans. Therefore, natural language understanding techniques need to be realized first, so that the chat robot can obtain a high Emotional Quotient (EQ), which represents the degree of Emotional Intelligence [3].

There are two major methods to generate a response to a piece of text, such as a short post on an online social media, including the retrieval-based method (RBM) and generation-based method (GBM) [4]. RBM needs a database of pairs of post and response in advance, and then builds the index for the pairs using information retrieval (IR) techniques. After receiving a post, RBM exploits some IR algorithms such as BM25 [5] to search for the most similar sentence in the database and retrieve the corresponding response. Therefore, the responses in the index are reusable, and we are capable of controlling the content of the response. However, RBM cannot deal with conversations outside of the fields in the training data efficiently [6][7].

A common approach for GBM is building a sequence-to-sequence (Seq2Seq) model, which employs neural networks [8]. Seq2Seq is based on recurrent neural networks (RNN) and it needs abundant conversation data to learn how to respond. At its core, Seq2Seq is composed of an encoder and a decoder [9]. Encoder works by reading words in a post one by one and transform them into a “context vector” of a fixed length. The decoder then treats the context vector as the input and generate a sequence of responses. An RNN model can only output responses of the same length as a post, but a Seq2Seq model can generate responses of various lengths through bucketing and padding methods. Note that RNN is not the only possible option for building the encoder and decoder, since we can alternatively use a convolutional neural network (CNN) or recursive neural network (RecNN) within the same framework [10]. Also, GBM can generate responses with more resilience even if the input is from a different domain, however, we cannot control the content of the output of the model [11].

Most of the emotional semantic identification datasets in NLP are from SemEval or social media, for instance, Sina Weibo, Plurk, or Facebook. The wide variety of their content exceeds the scope of movie reviews or hotel reviews. Therefore, a more delicate detection is needed to achieve effective emotional recognition results. For example, semantic relation classification based on dictionaries [12], including Chinese dictionaries such as the E-HowNet[†] or NTU Sentiment dictionary (NTUSD[‡])[13]. Moreover, feature extraction methods, including PMI, LLR, or Chi-square, are often employed to extract keywords for specific emotions. Emojis that often appear on social networking sites are also considered as an important indicator to measure emotions [14] ◦

Figure 1 illustrates the task definition of the CECG subtask of STC3 at NTCIR-14, in which we only consider one round of conversation instead of multiple rounds, and each round is formed by two short texts. More specifically, the first text is regarded as the input (referred to as the “post”) from a user and the latter an output given by the computer (referred to as the “response”). For each input, participants are required to produce five different outputs, each containing a different emotion. They must express the correct emotion in addition to maintaining conversational fluency. In this competition, our team separates the task into two stages. We first generate Chinese emotional

[†] http://ehownet.iis.sinica.edu.tw/index.php

[‡] http://academiasinicanlplab.github.io/#download

conversation using a Seq2Seq model, and then refine it through emotion detection methods. We describe in detail the methods and processes of our conversational system in the following section.

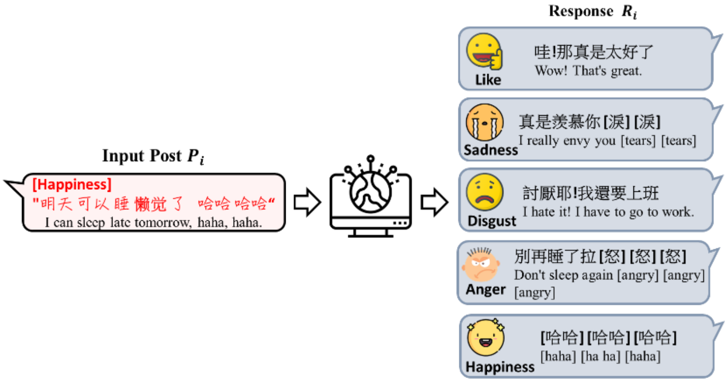

Fig. 1. The task definition of Chinese emotional conversation generation

## 2 Methodology

The CECG task requires participants to generate responses containing five different emotions to the post, including “like,” “sadness,” “disgust,” “anger,” and “happiness.” It also emphasizes that the content of the response has to not only contain a specific emotion but also conform to contextual semantics and fluency. Figure 2 shows the architecture of the proposed method. We first obtain several responses through encoding the post, post emotions, and response emotions into vectors, and then send them as input to the generation model. However, even after filtering, there are many unsuitable sentences in the responses due to the noise in the training materials. In light of this, we create heuristic rules for filtering out noise and obtaining qualified training instances. Next, the Seq2Seq approach is adopted for training a generation model. In order to improve the quality of the generated responses for emotional conversation, we propose a valence and arousal-based emotion detection model for response refinement. The following sections provide detailed description of major components in our system.

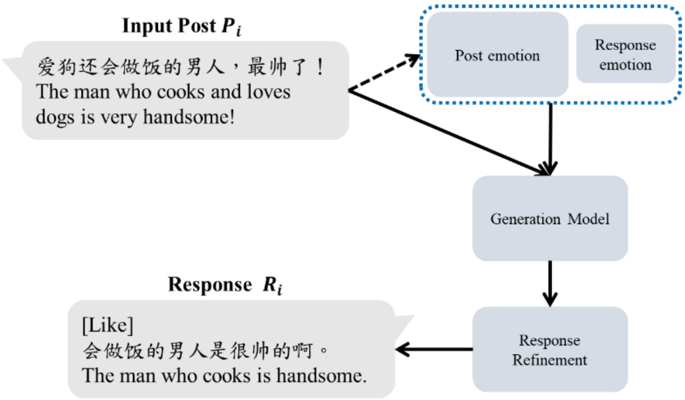

Fig. 2. The system structure of the proposed method.

## 2.1 Learning the Generation Model for Chinese Emotional Conversation

The training set of the CECG competition contains 600,000 pairs of sentences. In order to reduce training time and increase efficiency, we devise a refinement process to the original data. More specifically, we conduct the following steps to filter out data with inadequate quality.

- **The same word can only be repeated for at most 3 times:** The post and response from Weibo are similar to oral data, which often contain repeated words or emojis so as to reinforce the emotion, such as the underlined words in the following sentences: “对不起 我 笑 了 哈 哈哈 哈哈 哈哈 哈哈 哈哈 哈哈 哈哈 orz”, “好像 有点 扭 到 脚 了 … … [ 泪 ] [ 泪 ] [ 泪 ] ”. Despite the function of these repeated words in increasing the degree of the emotion, we can only gain limited information as the number of repetitions increases. Thus, we prune words that are repeated for more than three times to at most three times.
- **Remove stopwords in the post:** To facilitate learning of only the most significant information by the model, we remove stopwords in the post. In contrast, we keep the stopwords in the responses in order to ensure the sentences generated from our model is fluent enough. This process lowers the computing resources effectively and maintains output quality at the same time.
- **Retain shorter sentence:** After the above two processes, the average and maximum length of the posts are 17 and 108 words, respectively; the average and maximum length of responses are 11 and 496 words, respectively. Considering the assumption that this data has characteristics of oral communications, we only retain posts with less than 20 words and responses with less than 25 words. Note that posts are shorter than responses in average due to the stopword removal process.

After the data cleaning process, the training set contains 377,385 pairs of sentences. We select 50,000 pairs from the processed data as input, and transform it into character vectors. Figure 3 illustrates our generation framework based on the seq2seq model, which consists of an Encoder and Decoder. In our model, the Encoder, which employs a bi-directional LSTM, converts the input character vectors into a context vector, and the Decoder, which is a uni-directional LSTM, transforms the context vector back to output words.

Recall that the task objective is to generate five responses with a specific emotion. Therefore, in order to output emotional response, we transform the emotions in the post through one-hot encoding and response into a vector and concatenate it with the context vector in our model. Note that we merge both post and response emotions into the model instead of only using that of the response, because the post emotion would influence the words used in the response with a specific emotion. In other words, depending on the emotion of the post, we will use different terms with different emotional intensities to create a response.   Because we only utilize a part of the training data, the possibility of encountering unseen words in the test set needs to be considered. In order to reduce the impact of unseen words, we sequentially drop one word in a post in the test set to create a new input before sending it into the model. In this way, some of the responses may be identical to one another, however, we have a better chance of finding higher quality responses due to the dropping out of unknown words that have no useful information. Subsequently, we develop an emotion recognition model to identify the best one from this set of responses. This recognition model is introduced in the following section.

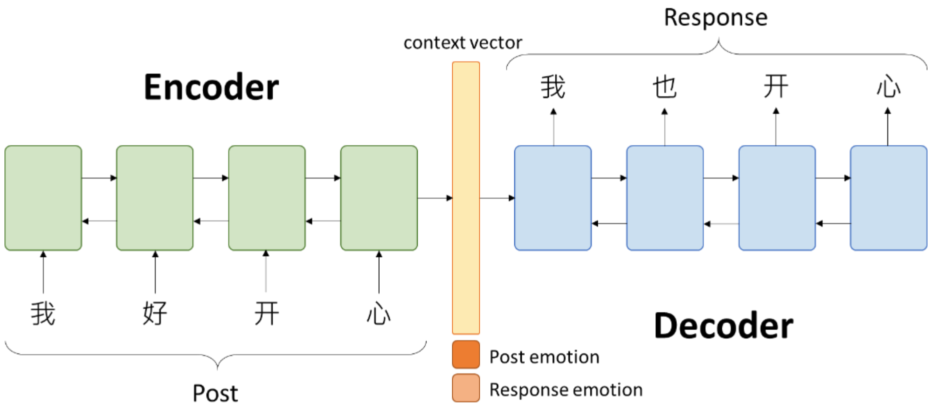

Fig. 3. Generation Model for Chinese Emotional Conversation

### 2.2 Response Refinement using an Emotion Recognition Approach

Emotion recognition refers to the application of semantic analysis and other NLP or text mining technologies to identify the opinions or emotions expressed by a piece of text. Considering that the most crucial element of the CECG task is the understanding and expression of emotions, we therefore propose a response refinement process that selects the response with the most appropriate emotion, as shown in Figure 4. This process is based on the “dimensional sentiment” model that consists of valence (V) and arousal (A) values, which relates to the polarity and the intensity of the sentiment, respectively. The dimensional approach represents affective states as continuous numerical values from 1 to 9 on each dimension, thus allowing for more fine-grained sentiment analysis [15]. Therefore, we utilize this model to evaluate the emotions in responses in a two-dimensional fashion.

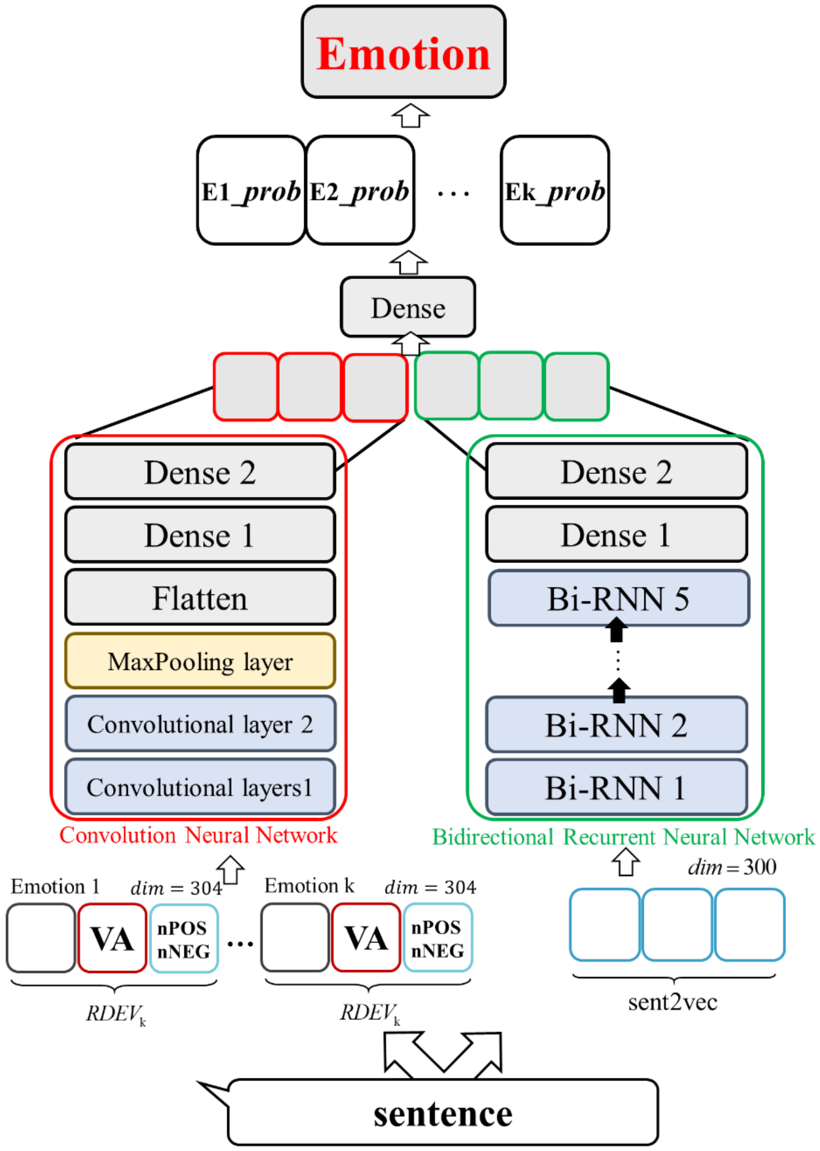

Fig. 4. Emotion detection model structure

Specifically, we first train a model for VA prediction using the corpus from the Dimensional Sentiment Analysis for Chinese Words (DSA_W) shared task held by the

2016 International Conference on Asian Language Processing (IALP). Next, we transform text sequences to vectors using the proposed refined distributed emotion vector (RDEV) representation approach, which integrates word embedding with the predicted valence-arousal scores (denoted as *V* and *A* in Figure 4) as well as the number of positive and negative words (denoted as *nPOS* and *nNEG* in Figure 4.) Therefore, the input is a 304-dimension vector of RDEV for a response. We set the number of 1D filters in the first convolutional layer as 200 with a size of 5. These filters perform convolutions on the input vector and generate feature maps. Next, the generated feature maps are refined through 100 filters with the same region size; a one-max pooling layer subsequently performs sampling over each map to identify the largest value from each feature map. Two dense layers with 100 and 64 dimensions are used in the end for feature refinement. Meanwhile, a five-layer bidirectional recurrent neural network (Bi-RNN) is adopted for learning context features. Finally, we concatenate features extracted from CNN and RNN as the input to the last layer. The final layer projects the combined feature vector once again and applies the softmax function to calculate the probability of this response belonging to one of the emotion categories. Based on the above architecture, we generate one response for each required emotion, and return the one that has the highest score as the response for a specific emotion.

## 3 Results and Discussion

For this competition, we implement the proposed model using Keras[§], a Python deep-learning library. The maximum sentence lengths of post and response are set to 20 and 25, where longer documents are truncated, and shorter sentence are padded with zeros. The 300-dimension pre-trained word embeddings are used for vector generation. The training lasts for at most 20 epochs or when the accuracy of the validation set starts to drop. Moreover, we adopt the corpus of emotion analysis in Chinese weibo texts task in NLPCC 2014[**] to verify the effectiveness of our emotion recognition model for responses. The performances of our model in terms of Precision, Recall, and $F_1$-score are shown in Table 1. We observe that the $F_1$-score can reach up to 69% for the emotion "Like." However, our model cannot successfully identify the emotion "Anger," and only obtains an $F_1$-score of 16%.

Table 1. Performance of the Emotion Recognition model using the NLP&CC dataset.

| **Emotion** | **Precision** | **Recall** | **$F_1$-score** |
|---|---|---|---|
| Like | **74%** | **65%** | **69%** |
| Sadness | 54% | 53% | 54% |
| Disgust | 45% | 63% | 52% |

[§] https://keras.io/

[**] http://tcci.ccf.org.cn/conference/2014/

| | | | |
|---|---|---|---|
| Anger | 39% | 10% | 16% |
| Happiness | 56% | 76% | 64% |

Table 2 lists the performance of our model in this competition, in which “Label0” represents the number of responses that do not exhibit coherence and fluency, “Label1” represents the number of responses that are coherent and fluent but do not agree with the required emotion, and “Label2” stands for the number of responses containing all required criteria. Among the 1,000 responses that we submitted, there are 777 Lable0, 126 Label1, and 97 Label2. These results indicate that there is a large room for improvement on the coherence and fluency of our generated responses. We can increase the amount of training data or add syntactic features into the model to advance the performance in coherence and fluency. Moreover, as we examine the results of each individual emotion category more closely, we notice that the performances of “Like” and “Happiness” are better than “Sadness,” “Disgust,” and “Anger.” Therefore, we randomly select a pair of post and response generated by our model to observe the result. As shown in Table 3, the responses for “Sadness,” “Disgust,” and “Anger” are similar, which indicates that our model cannot discriminate among these emotions well. In addition, we make a similar observation in Table 1, that is, the performances for emotions “Sadness,” “Disgust,” and “Anger” are worse than others. From the above observations, we conclude that enhancing the quality of responses for these three emotion categories is key to an overall promotion of the performance of our model.

Table 2. Performance of the proposed model on the official submission to the CECG subtask.

| Emotion | Label0 | Label1 | Label2 | Total | Overall score | Average score |
|---|---|---|---|---|---|---|
| Like | **164** | 9 | 27 | 200 | 63 | 0.315 |
| Sadness | 163 | 26 | 11 | 200 | 48 | 0.24 |
| Disgust | 158 | **42** | 0 | 200 | 42 | 0.21 |
| Anger | 154 | **42** | 4 | 200 | 50 | 0.25 |
| Happiness | 138 | 7 | **55** | 200 | **117** | **0.585** |
| total | 777 | 126 | 97 | 1000 | 320 | 0.32 |

Table 3. Samples of a post and responses for different emotions generated by the proposed model.

| | | | |
|---|---|---|---|
| post | 怪 梦 一个 接 一个 , 睡眠 质量 差劲 ! [ 怒 ] [ 怒 ] [ 怒 ] [ 怒 ] [ 怒 ] | | |
| | Response1 | Response2 | Response3 |
| Like | 加油哦 | 你在哪里 | [爱你][爱你][爱你] |
| Sadness | 我也不知道[泪][泪] | 你是谁啊 | 你是谁的黑了吗?我还没睡 |

| | | | |
|---|---|---|---|
| | | | 呢。。 |
| Disgust | 怎么了? | 你是谁啊 | 你是干嘛的? |
| Anger | 怎么啦? | 你是谁啊 | 你要干嘛去了? |
| Happiness | [哈哈][哈哈][哈哈] | 回家[嘻嘻] | 回家里啦 |

## 4 Conclusions

The CECG task combines the generation of conversation and the element of emotion to form a novel contest. It envisions a scenario where the reply from a machine is no longer emotion-less text, but rather with empathy. But integrating natural language understanding (NLU) technologies and teaching the machine human emotions is still a challenging and ongoing research topic. Building a model that can appropriately provide feedback with emotions is one of the goals that many researchers are trying to achieve. In this paper, we describe our framework that contains a sequence-to-sequence structure with an encoder and a decoder to produce responses to a piece of text, and an emotion detection model to distill the most suitable sentence as the final response. The final evaluation results of our approach from the competition show an average score of 0.32. Although our system is not the best among the participants, we believe it still possesses great potentials. In the future, we will continue to improve the quality of the generation model in order to build a truly natural conversational agent.

## 5 References

1. Yi-Ciao Gu, Yuen-Hsien Tseng, Wei-Lun Hsu, Wun-Syuan Wu, and Hsueh-Chih Chen, "Development and Classification of a Chinese Humor Corpus", In Proceedings of the 20th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing'19), 2019.
2. Yuen-Hsien Tseng, Lung-Hao Lee, Yu-Ta Chien, Chun-Yen Chang, and Tsung-Yen Li, "Multilingual Short Text Responses Clustering for Mobile Educational Activities: a Preliminary Exploration", Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications, in conjunction with the 56th Annual Meeting of the Association for Computational Linguistics (2018): 157-164.
3. Xiaofei Lu and Berlin Chen, Computational and Corpus Approaches to Chinese Language Learning (ISBN: 978-981-13-3570-9), Singapore: Springer, 2019.
4. Yu Wu, Wei Wu, Zhoujun Li, and Ming Zhou. "Learning Matching Models with Weak Supervision for Response Selection in Retrieval-based Chatbots." arXiv preprint arXiv:1805.02333 (2018).
5. Stephen Robertson and Hugo Zaragoza. "The probabilistic relevance framework: BM25 and beyond." Foundations and Trends® in Information Retrieval 3.4 (2009): 333-389.
6. Yiping Song, Rui Yan, Xiang Li, Dongyan Zhao, and Ming Zhang. "Two are better than one: An ensemble of retrieval-and generation-based dialog systems." arXiv preprint arXiv:1610.07149 (2016).

7. Yaoqin Zhang and Minlie Huang. "Overview of {NTCIR-14} Short Text Generation Subtask: Emotion Generation Challenge" Proceedings of the 14th {NTCIR} Conference (2019).
8. Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks." Advances in neural information processing systems (2014): 3104-3112.
9. Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. "Emotional chatting machine: Emotional conversation generation with internal and external memory." Thirty-Second AAAI Conference on Artificial Intelligence (2018): 730-738.
10. Soroush Vosoughi, Prashanth Vijayaraghavan, and Deb Roy. "Tweet2vec: Learning tweet embeddings using character-level cnn-lstm encoder-decoder." In Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval. ACM (2016): 1041-1044.
11. Minghui Qiu, Feng-Lin Li, Siyu Wang, Xing Gao, Yan Chen, Weipeng Zhao, Haiqing Chen, Jun Huang, and Wei Chu. "Alime chat: A sequence to sequence and rerank based chatbot engine." In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Vol. 2. 2017.
12. Changliang Li and Teng Ma. "Classification of Chinese word semantic relations. In: National CCF Conference on Natural Language Processing and Chinese Computing." National CCF Conference on Natural Language Processing and Chinese Computing (2017): 465-473.
13. Lun-Wei Ku, and Hsin-Hsi Chen. "Mining opinions from the Web: Beyond relevance retrieval." Journal of the American Society for Information Science and Technology 58.12 (2007): 1838-1850.
14. Xianda Zhou and William Yang Wang. "Mojitalk: Generating emotional responses at scale. " arXiv preprint arXiv:1711.04090 (2017).
15. Liang-Chih Yu1, Jin Wang, K. Robert Lai and Xue-jie Zhang. "Predicting valence-arousal ratings of words using a weighted graph method." Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). Vol. 2. 2015.