Automatic emotional spoken language text corpus construction from written dialogs in fictions

Conference Paper · October 2017

DOI: 10.1109/ACII.2017.8273619

3 authors, including:

Ming Li
Duke Kunshan University
93 PUBLICATIONS **977** CITATIONS

Automatic Emotional Spoken Language Text Corpus Construction from Written Dialogs in Fictions

Jinkun Chen[*§], Cong Liu[†], Ming Li[*‡]
[*] *School of Electronics and Information Technology, Sun Yat-sen University*
[†] *School of Data and Computer Science, Sun Yat-sen University*
[§] *SYSU-CMU Shunde International Joint Research Institute, Foshan, China*
liming46@mail.sysu.edu.cn

***Abstract*—In this paper, we propose a novel method to automatically construct emotional spoken language text corpus from written dialogs, and release a large scale Chinese emotional text dataset with short conversations extracted from thousands of fictions using the proposed method. The emotional spoken language transcript resources in Chinese are relatively limited. However, constructing a large scale supervised corpus manually is neither efficient nor low-cost. This motivates us to try alternative efficient and effective approaches. First, we build a small scale emotion dictionary manually instead of a large scale corpus. Each word in dictionary has an emotion tag. Then, we use the emotional words to search emotional dialogs heuristically in fictions and classify them automatically. Second, we share our work to boost the performance of emotion recognition on spoken languages using the proposed new database. The labeled dialogs can be used for supervised learning while the unlabeled ones provide better word embeddings for the semantic level emotion recognition. We use the dialogs corpus as an auxiliary dataset in speech emotion recognition. We carry out experiments on automatic speech recognition (ASR) generated texts from the speech signals in Chinese Natural Emotional Audio-Visual Database (CHEAVD). It is an eight emotion states recognition task. We obtain a baseline average macro precision (MAP) of 37.08% and accuracy of 31.13% in terms of text-based method. With the labeled dialogs to pretrain neural networks and over-sampling the minority classes, we achieve an optimized MAP of 47.50% and the accuracy of 43.91%, which outperforms the baseline by 10.42% and 12.78% respectively.**

1. Introduction

Emotion recognition has been an important research topic in the past decades. Conventionally, emotion recognition systems on spoken language mainly rely on speech-based methods. In general, the speech-based methods extract the frame level or utterance level acoustic and prosodic features from audio signals and adopt various classification methods to predict the emotion labels [1]. However, the semantic level information is not well explored. Now, with the improvement of speech recognition performance, the recognized text transcript is more and more accurate. Therefore, emotion recognition on the automatic speech recognition (ASR) generated text becomes practical and has good potential to be fused with the speech-based methods to enhance the performance. Our work focus on the text-based methods in spoken language emotion recognition and try to make full use of the utterances in speech.

The text-based methods mainly deal with the written text corpora. The bag-of-words (BoW) model [2] is introduced to capture the word-level features of the linguistic material. The latent Dirichlet allocation (LDA) [3] is another common model used to estimate the distribution of words associated with hidden topics of documents [4].

Recently, methods based on word embeddings [5], [6], [7] and deep neural networks (DNN) have been proposed for text emotion recognition. With the word vectors as the features, different neural networks, such as, recurrent neural networks (RNN) [8], [9], long short-term memory (LSTM) [10], [11], [12], convolutional neural networks (CNN) [13], [14], CNN-LSTM [15] and Tree-LSTM [16], are introduced for text emotion recognition. Unlike the word vectors, paragraph vectors have captured the semantic, syntactic and the word order regularities of sentences [17]. The paragraph vectors can be directly fed to linear classifiers or neural networks for emotion recognition.

Some works of speech emotion recognition based on ASR generated text has been reported by Metze [18], but the results were not promising because of the high word error rate of ASR. Fortunately, ASR technique has become more and more accurate. It is an effective approach to decode the speech signals into text and utilize the text-based methods for semantic level speech emotion recognition and contribute to the final fused system. However, the practical dilemma is that we don't have large scale emotional speech or transcript database, especially for languages other than English.

In this paper, we propose an effective method to automatically construct emotional spoken language text corpus from written dialogs in fictions. With the proposed method, large scale emotional spoken language text database can be built and adopted as an auxiliary dataset for the emotion recognition tasks on ASR generated texts.

We have utilized this method to build the Chinese Emo-

[‡]. The corresponding author.

978-1-5386-0563-9/17/$31.00 ©2017 IEEE

tional Written Dialogs (CEWD) database, available at Chinese Linguistic Data Consortium [www.chineseldc.org]. We also provide the results of text-based emotion recognition baselines on CEWD dataset to show its effectiveness.

The rest of this paper is organized as follows. Our proposed method and the released dataset are explained in Section 2. The emotion recognition baseline methods applied in experiments are described in Section 3. The experimental results and analysis are presented in Section 4. And the conclusion and future works are provided in Section 5.

2. The Chinese emotional written dialogs dataset

In this section, we first explain the proposed method and the construction of the Chinese Emotional Written Dialogs (CEWD) dataset from thousands of fictions, and then present the distribution of 32 annotated emotion states in the corpus.

2.1. Creating the emotional dictionary

In fictions, a conversation often goes with descriptions of the corresponding speakers, specifying the expressions on the face, the speaker's moods or the speaking manners that reflect the speaker's feelings. Thus, we can estimate the emotion tendency from the descriptions of a speaker. A strategy is to check the emotional words with a dictionary.

To construct the emotional dictionary, the first step is to collect a set of words, which can be verbs, adjectives or adverbs, that have strong emotion tendencies. Next, these words are properly classified into emotional categories. We refer to the existing Chinese emotional dictionary, NTUSD [19], and enrich the collection with supplemented Chinese terms. To reduce the ambiguity in the dictionary construction, 5 people voted to decide the category of each emotional word. Finally, we obtain 581 words in total and classify these words manually into 32 emotions categories.

The classification of emotional words is based on semantic knowledge. Synonyms and the words having similar emotion tendency are categorized into one class. For example, the words *glad, pleasure, happy, joyful, cheerful* are categorized as *happy* state. Emotion states such as *excited, angry, happy, depressed, serious, disgusted* and *sad*, account for the majority in emotional dictionary.

Instead of constructing the entire large scale supervised emotional transcript corpus manually, it is more efficient to build a small scale emotion dictionary, which will be used to search emotional dialogs heuristically in fictions and classify them automatically.

2.2. Extracting the emotional short conversations

We download tens of thousands of ebooks from the Internet. The book sources contain the Chinese native works and the translated works from non-Chinese languages. The categories of the novels mainly include literary classics, scientific fictions, romantic fictions, mystery novels and essays. First, we search all the conversations from the novels with regular expressions and obtain millions of conversations. Second, we split a conversation into two parts according to sentence structure, one is the description of speaker, another is the what the speaker said. Then, if there are emotion words in the description part, we choose the corresponding adjectives as the emotion tag. If not, the conversation is supposed to be unlabeled. Note that, some conversations have no description of speakers, then these sentences are unlabeled. Third, the conversations with emotion tags can be automatically classified into 32 emotion categories according to the emotional dictionary described in subsection 2.1. For instance, in an extracted conversation as follows,

> *He said excitedly, "I am going on vacation in December!"*

we have the emotion word *excitedly* in the description of the speaker, then, the corresponding emotion tag is *excited*. In the context of natural dialogs, similar dialogs may have different emotion tags. When learning word embeddings, it's a good practice to leave the description parts in the training corpus to capture more emtional information of words via co-occurence statistics.

TABLE 1. COMPONENTS OF THE CEWD DATASET

emotional words	581	for dialogs searching
emotion categories	32	for dialogs labeling
labeled dialogs	101093	for supervised learning
unlabeled dialogs	2772514	for word2vec training

We only check the emotion tags from syntactic instead of the semantic meanings in speakers' descriptions. Therefore, there is a small part of conversations having inconsistent emotion tags while the most dialogs are correctly labeled.

The components and purposes of the released CEWD database are shown in Table 1. we extract 2873607 conversations in total, 101093 dialogs are labeled with emotion tags while the other 2772514 ones are unlabeled. A large amount of potential emotion dialogs lie in the massive unlabeled dialogs. On one hand, this corpus is in large scale enabling us to build the word or paragraph representations for a variety of emotion recognition and sentiment analysis applications. On the other hand, the corpus can be used as an auxiliary data to build distributed representations for small text dataset. Particularly, the ASR generated texts from thousands of speech utterances are far away from sufficient to apply the word embedding methods. Since, the linguistics of dialogs in corpus are very similar to the spoken languages in speech, we obtain better distributed representations and enhance the emotion recognition system with the CEWD corpus as a supplementary material. The experimental results in Section 4 also show the effectiveness of the CEWD corpus itself.

2.3. Distribution of emotion states

In this corpus, 101093 dialogs are classified into 32 emotion categories. The 32 emotion translations are listed

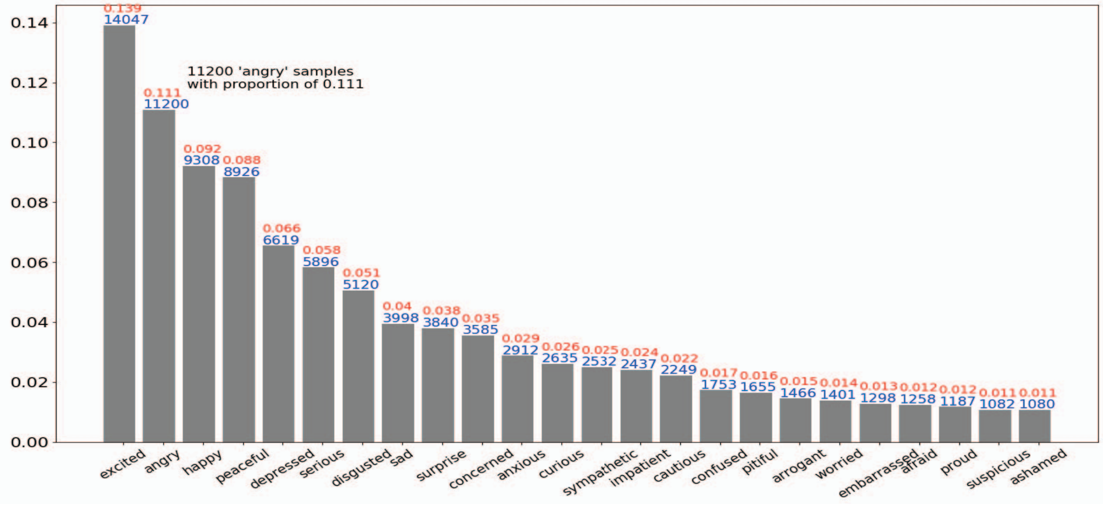

Figure 1. The distribution of the first 24 emotion states

TABLE 2. THE PROPORTIONS OF THE LAST 8 EMOTIONS

emotion state	count	proportion
respectful	882	0.87%
critical	717	0.71%
deceptive	478	0.47%
expectant	448	0.44%
thankful	412	0.41%
shy	353	0.35%
regretful	186	0.18%
hostile	133	0.13%

in the descending order of proportions as follows: *excited, angry, happy, peaceful, depressed, serious, disgusted, sad, surprise, concerned, anxious, curious, sympathetic, impatient, cautious, hesitating, worried, pitiful, arrogant, embarrassed, suspicious, afraid, ashamed, proud, respectful, critical, deceptive, expectant, thankful, shy, regretful* and *hostile*.

The proportions and the numbers of corresponding labeled samples of emotions are shown in Figure 1 and the Table 2. Figure 1 shows the details of the first 24 emotion states. The integer on the top of each bar is the number of samples and the fractional value means its proportion in dataset. The Table 2 presents a tiny fraction in corpus that the last 8 ones take. The first 8 emotions account for 64.41% while the last 8 emotions only take proportion of 3.56% in total. This distribution of emotions is consistent with the Zipf's law in our real life. In experiments, we will focus on those major emotion states.

3. Emotion recognition methods

In this section, we introduce the text-based emotion recognition methods we adopt in the experiments.

3.1. Word embedding

The idea of unsupervised word representations was originally introduced by Bengio [20]. In 2013, two more efficient models, the continuous skip-gram and continuous bag-of-words (CBOW), were proposed by Mikolov [5] [6]. In 2014, a different approach named global vector for word representation (GloVe) was provided by Pennington [21].

The word embedding can be viewed as a parameterized function mapping,

$$F : words \to \mathbb{R}^n, \qquad (1)$$

which maps words to high-dimension vectors. The word embedding implicitly encodes the semantic and syntactic regularities of linguistics and represents words in continuous vector space. Semantically and syntactically similar words obtain similar vectors. The word vectors are accepted as standard features in a variety of NLP applications including sentiment analysis and emotion recognition [21].

Based on the word vectors, sentences, paragraphs and even documents can be encoded into distributed representations [17]. The paragraph vectors implicitly capture the sentence-level or paragraph-level features including semantic, syntactic and the ordering information of words. In our work, we build word vectors and paragraph vectors using the CEWD database in multiple different experimental setups.

3.2. Paragraph vectors and DNN networks

Since the semantic, syntactic and the ordering information of words have been encoded into paragraph vectors, it is not necessary to fit the paragraph vectors to recurrent neural networks (RNN). Actually, a neural network with one or multiple hidden layers and a softmax output layer is capable to predict the multi-categorical emotion label. The simplest

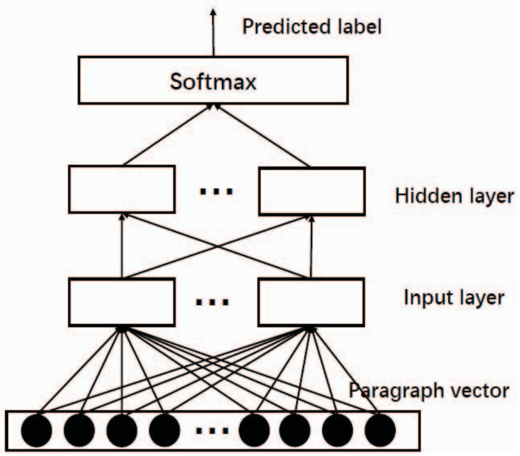

Figure 2. Emotion recognition with paragraph vectors

model with only one hidden layers is shown as Figure 2. As reported by Mesnil [22], the distributed representations capture rich semantics of words and paragraphs via co-occurence statistics. The fixed size paragraph vectors can be used as the feature representations of documents and then input to DNN networks for emotion recognition. This is the baseline system we adopt to recognize the emotion states.

For the training of DNN on imbalanced dataset, minority over-sampling [23] is an effective technique to improve the system performance. We over-sample the minority classes with duplications when training and tuning the baseline system. For the tasks on small datasets, an auxiliary dataset becomes very useful. Using the auxiliary dataset to pre-train the model would result in a more robust and effective system.

4. Experimental results and analysis

In this section, we present and analyze the experimental results of speech emotion recognition on the newest version of CHEAVD dataset as well as sentences emotion classification on the proposed CEWD dataset.

4.1. Speech emotion recognition on CHEAVD

The newest version of Chinese Natural Emotional Audio-Visual Database (CHEAVD) contains 6760 emotional segments extracted from Chinese movies and TV shows. 8 major emotions are *neutral, angry, happy, sad, worried, anxious, disgusted* and *surprise*. Each sample has an emotion label, a video clip and a corresponding speech wave. The emotion states distribution in this dataset is quite imbalanced. Thus, macro average precision is the primary metric while accuracy takes the second on emotion recognition task. More information is provided in [24] [25] [26].

4.1.1. ASR on CHEAVD speech. In the experiments, we conduct the emotion recognition on the speech part of CHEAVD dataset. We apply our in-house Mandarin speech recognition system to recognize CHEAVD speech files. Our ASR system is based on the KALDI toolbox [27]. The acoustic model is trained from 1000 hours of Mandarin speech with the chain model setup in [27]. We use the CEWD dataset and about fifteen million sentences of short film subtitles to build the language model to further reduce the word error rate on the CHEAVD database.

TABLE 3. WORD ERROR RATES COMPARISON IN CHEAVD DATABASE

baseline system + ASR	word error rate	MAP	accuracy
iFLYTEK	22.87%	36.54%	32.82%
Our ASR system	24.65%	37.08%	31.13%

To measure the quality of the ASR generated text, we manually transcribe 200 speech clips of CHEAVD dataset to form the ASR evaluation data, then, we apply the ASR application interface (API) of iFLYTEK [28] to get the recognized text for comparison. The word error rates (WER) of ASR generated texts are shown in Table 3.

The WER of iFLYTEK Mandarin API and our in-house system is 22.87% and 24.65%, respectively. Because some audios in CHEAVD have background sounds or noise, the word error rates are much higher than 10%. In the emotion recognition task on CHEAVD dataset, the small differences in word error rates seem to have no significant effect on the performance of emotion recognition. The averages of MAP and accuracy of baseline system (in 4.1.2) are shown in Table 3. It is clear that the differences of emotion recognition results on ASR generated texts, corresponding to ASR engine of iFLYTEK and ours respectively, are nearly negligible. Therefore, we choose our ASR system to recognize the speech into text.

4.1.2. Three emotion recognition systems. Because some audio streams are too short or in mute state, the ASR generated texts from CHEAVD dataset only contain 6023 short sentences, which is insufficient to build balanced word vectors or paragraph vectors. One option is to train the word vectors on the Wikipedia Chinese corpus and construct the sentence vectors with word vectors. Alternatively, using CEWD corpus as an auxiliary data, we directly generate the word vectors and paragraph vectors for the ASR generated texts. In the experiments, every written dialog is represented by one paragraph vector with dimension of 128.

We obtain paragraph vectors through the aforementioned methods. Since the newest version of the CHEAVD dataset has not yet provided a test set [26], we split the 6023 samples into training set and test set randomly. The training set contains 5037 samples while the test set has 986 samples. We adopt a baseline system, an over-sampling system and an optimized system to recognize emotion states. The experimental results are shown in Table 4, all metrics are presented with the averages of 10 test cases.

The baseline system. In the baseline system of emotion recognition on CHEAVD dataset, we use the unlabeled dialogs to learn the paragraph vectors with dimension of

TABLE 4. Emotion recognition on CHEAVD dataset

system	MAP	accuracy	f1-score
baseline system	37.08%	31.13%	29.89%
over-sampling system	41.94%	40.99%	39.04%
optimized system	**47.50%**	**43.91%**	**42.36%**

TABLE 5. Confusion matrix of the optimized system

	hap	**ang**	**sur**	**dis**	**neu**	**wor**	**anx**	**sad**
hap	**94**	12	3	0	15	11	2	1
ang	17	**109**	2	2	19	14	4	0
sur	17	9	**10**	3	11	7	0	1
dis	17	19	3	**14**	15	8	2	1
neu	43	63	9	2	**128**	12	11	3
wor	14	18	2	0	21	**45**	3	1
anx	23	14	1	0	16	10	**27**	0
sad	13	20	2	0	23	9	3	**8**

[1] { **hap** : happy, **ang** : angry, **sur** : surprise, **dis** : disgust, **neu** : neutral, **wor** : worried, **anx** : anxious, **sad** : sad }.

TABLE 6. Results of the first N emotion states

N emotion states	MAP	accuracy	f1-score
6	**57.65%**	**60.41%**	**56.89%**
8	43.40%	49.27%	41.23%
12	33.13%	37.28%	31.19%

TABLE 7. Confusion matrix of the 6 emotion states

	dep	**ang**	**pea**	**ser**	**exi**	**hap**
dep	**499**	233	207	115	176	116
ang	108	**1468**	160	116	294	113
pea	153	195	**1000**	121	136	158
ser	122	187	149	**481**	101	121
exi	85	220	111	59	**2137**	149
hap	84	160	133	93	258	**1182**

[1] { **dep** : depressed, **ang** : angry, **pea** : peaceful, **ser** : serious, **exi** : excited, **hap** : happy }.

128. We train the DNN networks with two hidden full connection layers, which have 256 and 128 cells inside respectively, to predict the emotion states. Dropout layers are adopted to avoid over-fitting. This baseline system is less robust and the predictions are slightly unstable. The averages of MAP, accuracy and f1-score are 37.08%, 31.13% and 29.89% respectively.

The over-sampling system. The emotion states ”disgusted” and ”surprise” are the minority classes in CHEAVD dataset. Both the two classes have less than 200 samples. We over-sample the two classes with duplications for 3 times in the training set. Then, we train and tune the baseline model with the over-sampled training set. By adopting the over-sampling trick, the averages of MAP, accuracy and f1-score are 41.94%, 40.99% and 39.04% respectively, which improves the performance significantly by comparison to the results of the baseline system.

The Optimized system. In the optimized system, we use the labeled samples of CEWD dataset to pre-train the DNN networks. First, we select the emotion states of ”*angry, happy, sad, worried, anxious, disgusted, surprise*” and extract the corresponding samples out from CEWD dataset. Besides, we select 10000 unlabeled dialogs as the ”*neutral*” components. Then, we obtain 47671 labeled samples and pre-train the DNN networks for 100 epochs. After the pre-training, we fine tune the model with the over-sampled training set. As a result, a more robust system and improved performance are obtained. The averages of MAP, accuracy and f1-score are 47.50%, 43.91% and 42.36% respectively. The confusion matrix of a test case is shown in the Table 5. The optimized system achieves the best performance.

As mentioned in [25] [26] [29], the CHEAVD is an imbalanced dataset and takes the macro average precision as the primary metric. It is a challenge to achieve a promising result in this eight emotion states recognition task. The baseline MAP and accuracy of the last version of CHEAVD 1.0 dataset are 30.34% and 21.18% provided in [25]. And the best MAP and accuracy of CHEAVD 1.0 dataset are 54.43% and 32.17% in terms of video-audio based method [30]. However, the CHEAVD dataset has been updated to the version of 2.0 and the version of 1.0 has been not accessible. Our experiments are carried out on the version 2.0 of CHEAVD. It is inappropriate to compare our experimental results directly with the previous works on the version 1.0 of CHEAVD. But, all these works prove that the text-based method on ASR generated text is effective in the speech emotion recognition, and our proposed dataset CEWD is valuable. By fully exploring the utterances in speech, text-based method can be an enhancement approach for audio-based method in spoken language emotion recognition.

4.2. Emotion recognition on CEWD corpus

We select the first 12 emotion states, which are *excited, angry, happy, peaceful, depressed, serious, disgusted, sad, surprise, concerned, anxious* and *curious*, to form a closed set classification task. We carry out 3 experiment cases for the first 6, 8 and 12 emotion states recognition tasks. We extract the corresponding samples and split them into training set, validation set and testing set in the proportion of 0.7 : 0.1 : 0.2.

We adopt the paragraph vectors and the baseline DNN networks explained in subsection 3.2. The emotion recognition results on the first N emotions are shown in Table 6. For the prediction of the first 6 emotion states, the MAP, accuracy and f1-score are 57.65%, 60.41% and 56.89% respectively. The confusion matrix of the first 6 emotion states prediction is shown in Table 7. The promising results prove the effectiveness of the CEWD dataset.

5. Conclusion

In this paper, we introduce an effective and efficient method for automatically constructing large scale emotional spoken language text corpus from fictions, and we share the Chinese emotional written dialogs (CEWD) dataset constructed with the proposed method. Through the experiments, we show that the method is practical and the CEWD dataset is valuable for emotion recognition on spoken language.

We adopt the text-based method to obtain a comparable result in the emotion recognition task on CHEAVD dataset. By using the text-based method as a complementary approach for the video-audio based methods, it is hopeful to build a more robust and effective system for multi-modal emotion recognition.

Acknowledgment

This research was funded in part by the National Natural Science Foundation of China (61401524, 61773413), Natural Science Foundation of Guangdong Province (2014A030313123), Natural Science Foundation of Guangzhou City (201707010363), the Fundamental Research Funds for the Central Universities(15lgjc12), Science and Technology Development Foundation of Guangdong Province, National Key Research and Development Program (2016YFC0103905) and IBM Faculty Award.

References

[1] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[2] Y. Zhang, R. Jin, and Z.-H. Zhou, "Understanding bag-of-words model: a statistical framework," *International Journal of Machine Learning and Cybernetics*, vol. 1, no. 1-4, pp. 43–52, 2010.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[4] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning, "Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora," in *Proceedings of EMNLP*, 2009, pp. 248–256.

[5] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *Proceeding of ICLR*, 2013.

[6] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[7] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin, "Learning sentiment-specific word embedding for twitter sentiment classification," in *Proceeding of ACL*, 2014.

[8] O. Irsoy and C. Cardie, "Opinion mining with deep recurrent neural networks." in *Proceeding of EMNLP*, 2014, pp. 720–728.

[9] R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proceedings of EMNLP*, 2013.

[10] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[11] X. Wang, Y. Liu, C. Sun, B. Wang, and X. Wang, "Predicting polarities of tweets by composing word embeddings with long short-term memory," in *Proceedings of ACL*, vol. 1, 2015, pp. 1343–1353.

[12] D. Tang, B. Qin, and T. Liu, "Document modeling with gated recurrent neural network for sentiment classification," in *Proceeding of EMNLP*, 2015.

[13] Y. Kim, "Convolutional neural networks for sentence classification," *Proceeding of EMNLP*, 2014.

[14] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," in *Proceedings of ACL*, 2014.

[15] J. Wang, L.-C. Yu, K. R. Lai, and X. jie Zhang, "Dimensional sentiment analysis using a regional cnn-lstm model," in *Proceeding of ACL*, 2016.

[16] K. S. Tai, R. Socher, and C. D. Manning, "Improved semantic representations from tree-structured long short-term memory networks," in *Proceeding of ACL*, 2015.

[17] Q. V. Le and T. Mikolov, "Distributed representations of sentences and documents." in *Proceedings of ICML*, vol. 14, 2014, pp. 1188–1196.

[18] F. Metze, A. Batliner, F. Eyben, T. Polzehl, B. Schuller, and S. Steidl, "Emotion recognition using imperfect speech recognition," in *Proceedings of INTERSPEECH*. ISCA, 2010.

[19] L.-W. Ku, Y.-T. Liang, H.-H. Chen *et al.*, "Opinion extraction, summarization and tracking in news and blog corpora." in *AAAI spring symposium: Computational approaches to analyzing weblogs*, vol. 100107, 2006.

[20] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A neural probabilistic language model," *journal of machine learning research*, vol. 3, no. Feb, pp. 1137–1155, 2003.

[21] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation." in *Proceedings of EMNLP*, vol. 14, 2014, pp. 1532–43.

[22] G. Mesnil, T. Mikolov, M. Ranzato, and Y. Bengio, "Ensemble of generative and discriminative techniques for sentiment analysis of movie reviews," *arXiv preprint arXiv:1412.5335*, 2014.

[23] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, no. 1, pp. 321–357, 2002.

[24] W. Bao, Y. Li, M. Gu, M. Yang, H. Li, L. Chao, and J. Tao, "Building a chinese natural emotional audio-visual database," in *Proceedings of ICSP*. IEEE, 2014, pp. 583–587.

[25] Y. Li, J. Tao, B. Schuller, S. Shan, D. Jiang, and J. Jia, "Mec 2016: the multimodal emotion recognition challenge of ccpr 2016," in *Chinese Conference on Pattern Recognition*. Springer, 2016, pp. 667–678.

[26] "Multimodal emotion recognition challenge (mec 2017)," accessed: 2017-4-25. [Online]. Available: http://www.chineseldc.org/htdocsEn/emotion.html

[27] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *Proceedings of ASRU*, 2011.

[28] "Automatic speech recognition service of iFLYTEK," accessed: 2017-4-3. [Online]. Available: http://www.xfyun.cn/services/voicedictation

[29] Y. Li, J. Tao, L. Chao, W. Bao, and Y. Liu, "Cheavd: a chinese natural emotional audiovisual database," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–12, 2016.

[30] J. Deng, N. Cummins, J. Han, X. Xu, Z. Ren, V. Pandit, Z. Zhang, and B. Schuller, "The university of passau open emotion recognition system for the multimodal emotion challenge," in *Chinese Conference on Pattern Recognition*. Springer, 2016, pp. 652–666.