Journal of Advances in Mathematics and Computer Science

36(7): 30-37, 2021; Article no.JAMCS.72724
ISSN: 2456-9968
(Past name: British Journal of Mathematics & Computer Science, Past ISSN: 2231-0851)

Multimodal Video Sentiment Analysis Using Audio and Text Data

Yanyan Wang[1*]

[1]School of Science, Zhongyuan University of Science, No.41 Zhongyuan Rd, Zhengzhou, China.

Author's contribution

The sole author designed, analysed, interpreted and prepared the manuscript.

Article Information

DOI: 10.9734/JAMCS/2021/v36i730381
Editor(s):
(1) Dr. Leo Willyanto Santoso, Petra Christian University, Indonesia.
Reviewers:
(1) Naruboina Srilatha, India.
(2) Saroj Kumar Dash, India.
Complete Peer review History: https://www.sdiarticle4.com/review-history/72724

Original Research Article

Received 15 June 2021
Accepted 19 August 2021
Published 25 August 2021

Abstract

Nowadays, video sharing websites are becoming more and more popular, such as YouTube, Tiktok. A good way to analyze a video's sentiment would greatly improve the user experience and would help with designing better ranking and recommendation systems [1,2]. In this project, we used both acoustic information of a video to predict its sentiment levels. For audio data, we leverage transfer learning technique and use a pre-trained VGGish model as a features extractor to analyze abstract audio embeddings [6]. We then used MOSI dataset [5] to further fine-tune the VGGish model and achieved a test accuracy of 90% for binary classification. For text data, we compared traditional bag-of-word model to LSTM model. We found that LSTM model with word2vec outperformed bag-of-word model and achieved a test accuracy of 84% for binary classification.

Keywords: Video sentiment analysis; multimodal data; transfer learning; abstract feature extraction; text mining.

1 Introduction

Online video is becoming the dominant media on the internet. People upload and share a huge amount of video clips in video sharing website, such as YouTube. Besides, video plays an important role in social networks, such

**Corresponding author: Email: wangyanyanno1@163.com;*

as Facebook. With all graphical, acoustic and text information, video provides tons of information for customers. In the meanwhile, all these information gives business owners great potential to design and improve current web applications.

Sentiment level is an important factor for design user-friendly web product. Knowing customer's sentiment level can improve the experience of new feeds system, video recommendation system, etc. Currently, sentiment level prediction based on text data is well studied [1,2]. However, there are very limited studies that used acoustic information to analyze the sentimental levels [3,4]. In this work, we use both the acoustic and text data to train deep neural networks for predicting sentiment levels.

2 Methods

2.1 Dataset

We used Multimodal Opinion level Sentiment Intensity (MOSI) dataset to perform experiments [5]. The MOSI dataset contains 2199 video clips collected from YouTube. For each video clips, both audio recording and transcript is provided in separate files for each sentence. The sentiment level of each video clip is evaluated by five workers on a scale from -3 to 3.

2.2 Data pre-processing and label generation

We pre-processed the original MOSI dataset to generate labels for training. We generated 4 types of labels: 7-class, 3-class, binary and filtered binary class. We first calculated the mean value of the 5 human labeled scores. For 7-class, we rounded it to fall in one of the categories: -3, -2, -1, 0, 1, 2, 3. Then we add a constant 3 to each value to obtain the 7-class labels with all positive values. For 3-class labels, we set the label to be 0 if the mean value is less than 0, 1 if the value equal to 0, and 2 if the value is greater than 0. For binary labels, we set the label to be 0 if the mean value is less than 0, and to be 1 if greater than 0. We discarded the data whose mean score is equal to 0.

We added another step to remove the bias to the binary labels to obtain the filtered binary labels. Specifically, we calculated the standard deviation of the scores, and only kept the subset of samples whose standard deviation was less than 1. With the 7-class, 3-class and binary labels, we can better evaluate our models. With the comparison between the standard binary labels and the filtered binary labels, we can evaluate the effect of potential bias that exists in the MOSI dataset.

2.3 VGGish network and audio feature extraction

VGGish is a VGG-like pre-trained audio classification model which is released by Google [6,7,8]. This model is trained over 2 million human-labeled 10-second video soundtracks for acoustic scene classification tasks. The architecture of the VGGish model is shown in Fig. 1. The pre-trained VGGish network use VGG-like CNN architecture, and can generate embeddings of each 960ms audio window from its fully connected layers. In this work, we leverage this model to perform abstract features extraction on the MOSI dataset.

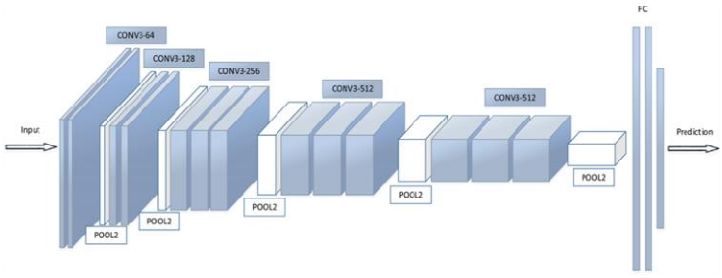

Fig. 1. Architecture of pre-trained VGGish model for abstract acoustic feature extraction

We experimented with extracting abstract features from the last three fully connected (FC) layers. The last FC layer has a dimension of 128, and the second and third last FC layers all have a dimension of 4096. Since the VGGish model takes a 960ms audio as input, we divide each audio file into several 960ms windows and extract an audio embedding for each window. Then we train a Support Vector Machine (SVM) classifier using these audio embeddings to predict its sentiment level out of the given labels. Apart from evaluating the train and test accuracy for each audio window, we also evaluated the "per file accuracy", which is the accuracy after applying majority vote to all 960ms window predictions of the same audio file.

To further adapt the VGGish model to the MOSI dataset, we built an end-to-end system by adding a Softmax layer at the end of the last fully connected layer to fine-tune the VGGish model. The dimension of the Softmax label is decided by the number of label categories. For each experiment, we trained 300 iterations. Then we extract the feature embeddings from the fine-tuned VGGish model and evaluated the accuracy of SVM classifier on these embeddings after fine-tuning.

2.4 LSTM and text feature generation

Recurrent neural network like LSTM captures the sequential characteristics of text which is more suitable for text sentiment analysis [9,10,11]. In this project, we first use Word2Vec to translate each word to a 300-dimension vector. Then we feed the word embedding to the LSTM network for sentiment classification. For our model, the window size of LSTM is 20 words [12].

3 Resutls and Discussion

3.1 Audio: Transfer learning using VGGish

We first evaluated the performance of abstract audio embeddings with different dimensions, which are extracted directly from pre-trained VGGish model. The Accuracy for each 960ms window, as well as for each file are reported. The comparison result is shown in Fig. 2.

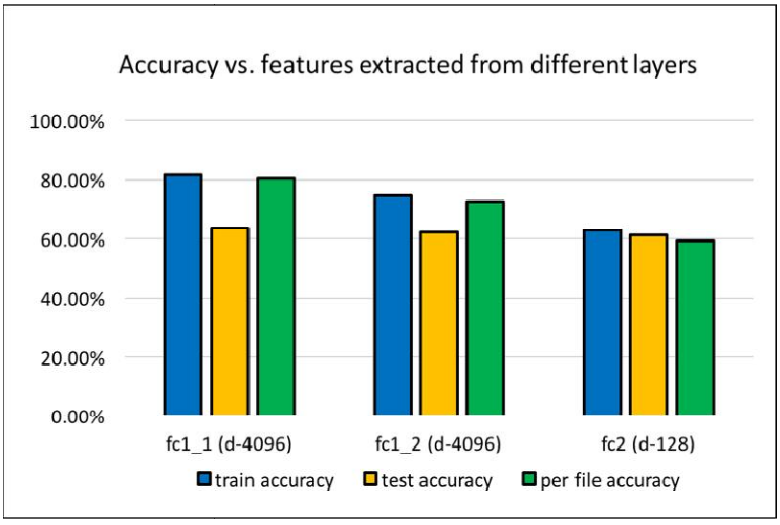

Fig. 2. Sentiment classification accuracy of feature embeddings extract from different fully connected layers, using pre-trained VGGish model

The data shows that the features extracted from the third last 4096-dimensional layer perform the best. It gives a train accuracy of 81.65%, a test accuracy of 63.74% and a per file accuracy of 80.50%. The performance of the second last 4096-dimension layer is slightly worse than the first one. The last 128-dimension layer only has a train accuracy of 62.95%. Therefore, in the following experiment, without explicit description, we use the first 4096-dimensional fully connected layer for feature extraction, since the high-dimension feature vector is able to

capture more information. The accuracy for each window and each file is similar, indicating that further increase the length of the audio clip will not improve the classifier performance.

We then conducted the experiment on other three types of labels. The result is shown in Fig. 3. The data shows that the filtered binary labels has an improved performance on training accuracy. However, the testing and per file accuracy are pretty close to the normal binary labels. The performance of 3-class and 7-class labels are worse than binary labels. However, all four types of labels perform better than random guess.

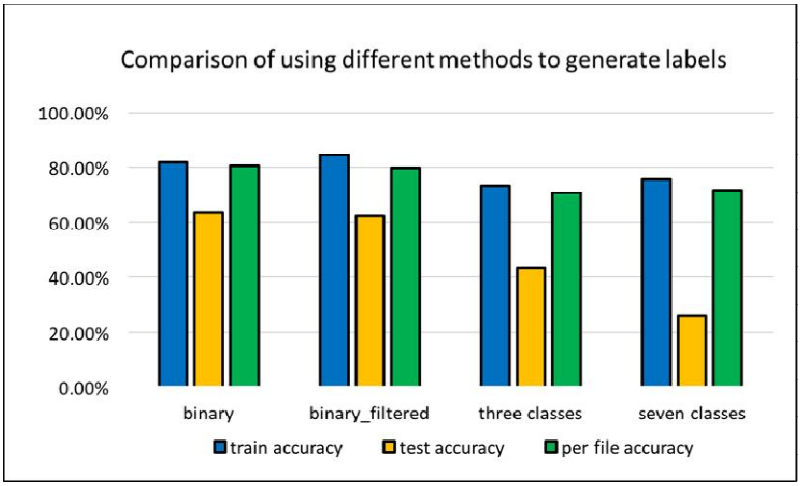

Fig. 3. Comparison of using different methods to generate labels for video sentiment level

To further adapt the VGGish model to the MOSI dataset, we built an end-to-end system by adding a softmax layer in the end to fine tune the VGGish model. The learning history of this end-to-end model is showed in Fig. 4. The data shows a clear rising curve for train and test accuracy. For all four types of labels, test accuracies start from the random guess value, and rise to a plateau value. The final result from the end-to-end system is very close to that from using SVM to classified directly extracted features. For binary, filtered binary, 3-class, and 7-class classification, the test accuracy is around 63%, 63%, 43%, 26%.

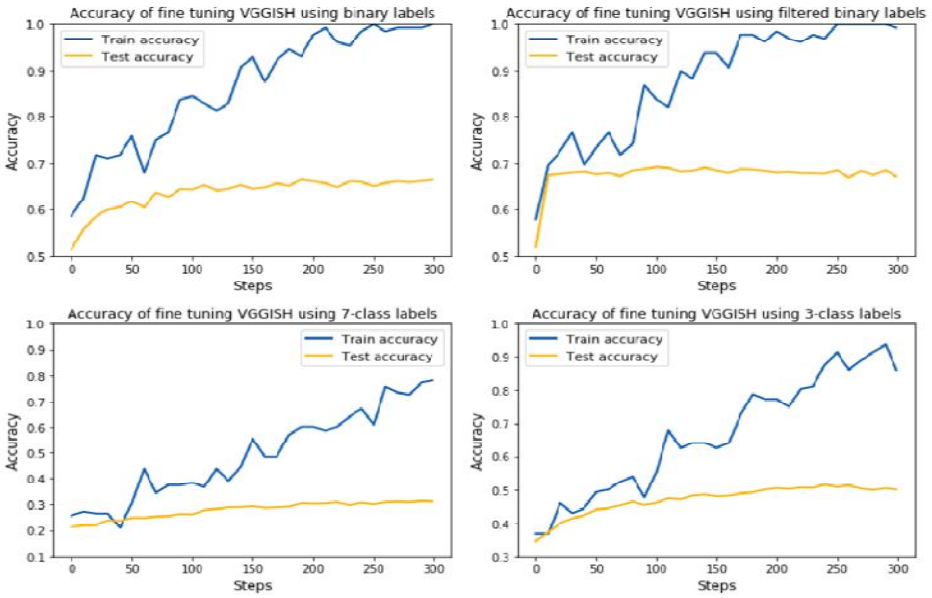

Fig. 4. Learning history of fine-tuning VGGish model with MOSI audio data

Using abstract feature vectors from this fine-tuned model from the first 4096-dimension layer, the classification accuracy improves by a large margin compared with pre-trained model. The results are shown in Fig. 5. Notably, the SVM training accuracy for both binary labels are greater than 97%, and the training accuracy for 3-class and 7-class are 95% and 98%, respectively. The test accuracy gets improved significantly as well. The test accuracy of filtered binary labels is 90.28%. The per-file-accuracy for all four types of labels are greater than 93%, which are significantly increased.

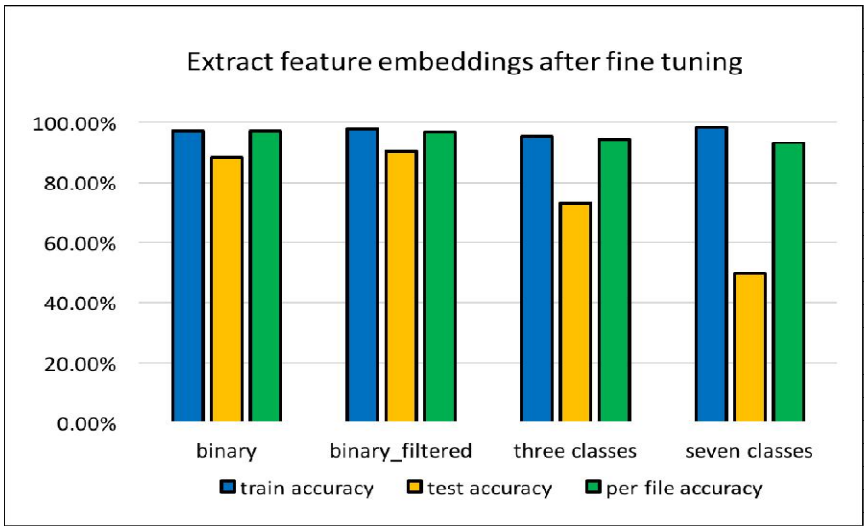

Fig. 5. Sentiment classification accuracy of feature embeddings extract from different fully connected layers, using fine-tuned VGGish model

3.2 Text: LSTM with word2vec

Before we implemented deep-learning methods, we tried bag-of-word model and other traditional machine learning methods (e.g., SVM with different kernels, random forest, k-neighbors). Best test accuracy came from Random Forest (78%). Using linear SVM, the model performed best in binary label. But in seven-class label, test accuracy is only 36%. The results are shown in Fig. 6 and Fig. 7, respectively.

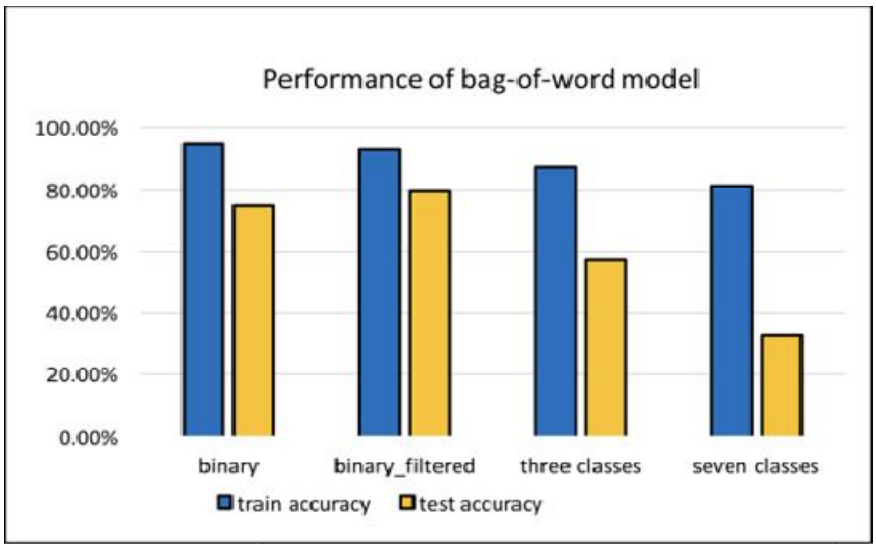

Fig. 6. Performance of bag-of-word model for sentiment classification using different labeling strategies

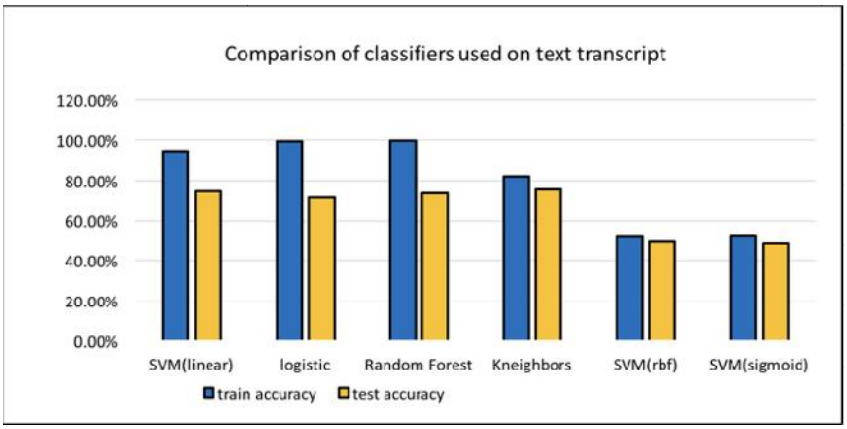

Fig. 7. Comparison of different classifiers for text sentiment classification

Then we trained LSTM models to classify sentiment label based on work2vec conversion of the raw text. The training history is shown in Fig. 8.

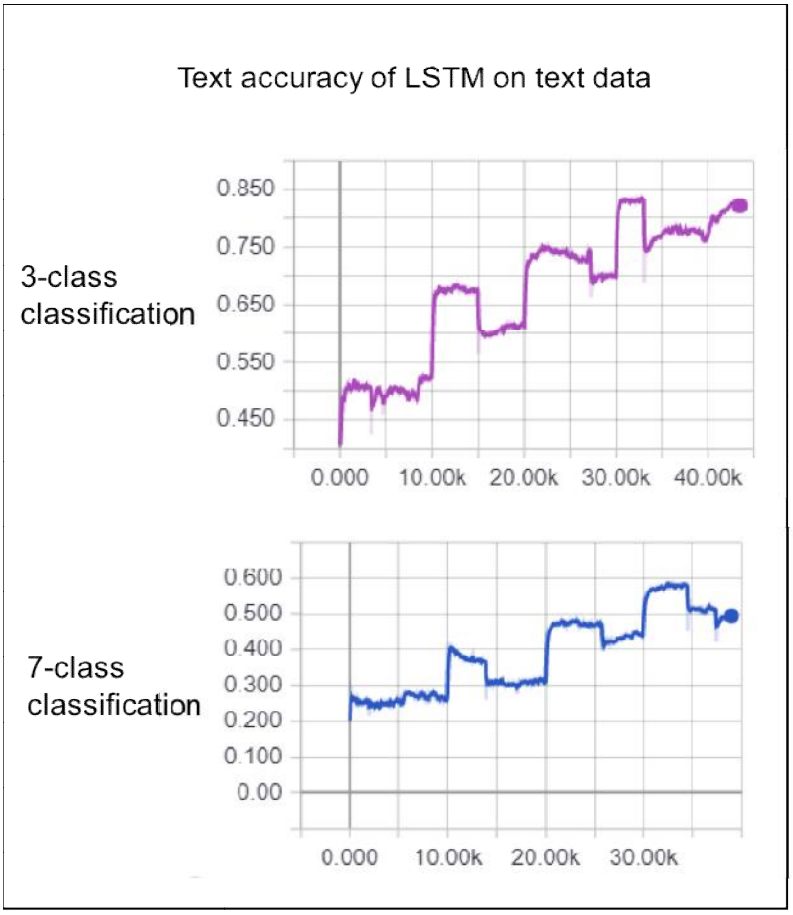

Fig. 8. Training history of LSTM model for sentiment classification from text data

The accuracy of the LSTM model using different labeling strategies is shown in Fig. 8. We can achieve a highest accuracy of 83% on binary filtered label.

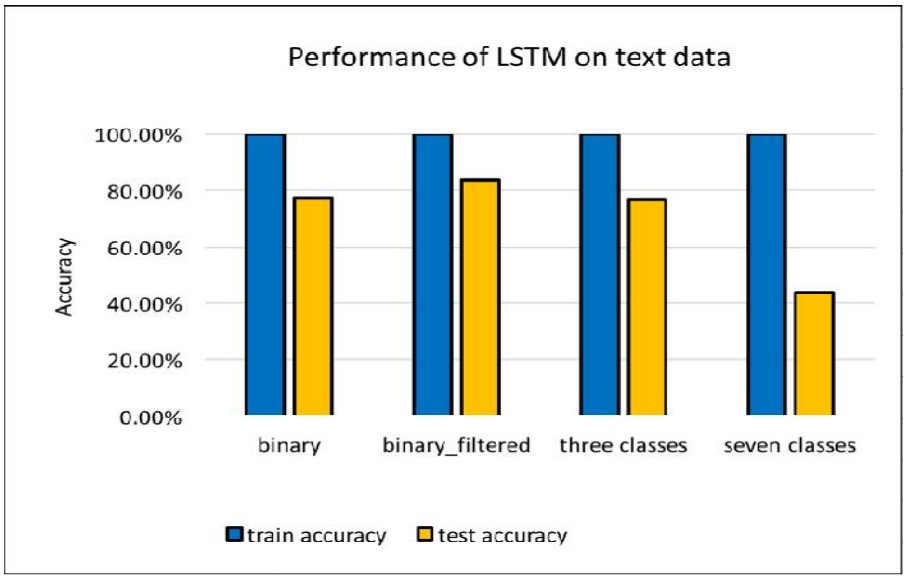

Fig. 9. The accuracy of the LSTM model using different labeling strategies

4 Conclusion

In this work, with MOSI dataset, we implemented LSTM and word2vec embedding in text data, and use pre-trained VGGish model to extract abstract features from audio data. With filtered binary labels, LSTM model can reach 83% test accuracy, which is much higher than traditional machine learning models, such as SVM. For filtered binary labels, using SVM on features directly extracted from the pre-trained VGGish model gives a test accuracy of 62\%. However, after fine tuning the VGGish model, the test accuracy is increased to 90%.

For future work, we will explore ways to directly concatenate audio and text embeddings together, to further improve the prediction accuracy for sentiment. We will also explore how to apply video sentiment information to existing video ranking and recommendation systems.

Competing Interests

Author has declared that no competing interests exist.

References

[1] Pang B, Lee L. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. InProceedings of the 42nd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics. 2004;271.

[2] Vinodhini G, Chandrasekaran RM. Sentiment analysis and opinion mining: a survey. International Journal. 2012;2(6):282-92.

[3] Williams J, Comanescu R, Radu O, Tian L. Dnn multimodal fusion techniques for predicting video sentiment. InProceedings of grand challenge and workshop on human multimodal language (Challenge-HML). 2018;64-72.

[4] Chen M, Wang S, Liang PP, Baltrušaitis T, Zadeh A, Morency LP. Multimodal sentiment analysis with word-level fusion and reinforcement learning. InProceedings of the 19th ACM International Conference on Multimodal Interaction. 2017;163-171.

[5] Zadeh A, Zellers R, Pincus E, Morency LP. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. IEEE Intelligent Systems. 2016;31(6):82-8.

[6] Hershey S, Chaudhuri S, Ellis DP, Gemmeke JF, Jansen A, Moore RC, et al. CNN architectures for large-scale audio classification. In2017 ieee international conference on acoustics, speech and signal processing (icassp). IEEE. 2017;131-135.

[7] Sengupta A, Ye Y, Wang R, Liu C, Roy K. Going deeper in spiking neural networks: VGG and residual architectures. Frontiers in neuroscience. 2019;13:95.

[8] Chen H, Xie W, Vedaldi A, Zisserman A. Vggsound: A large-scale audio-visual dataset. InICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE. 2020;721-725.

[9] Zhou C, Sun C, Liu Z, Lau F. A C-LSTM neural network for text classification. arXiv preprint arXiv:1511.08630; 2015.

[10] Wang J, Yu LC, Lai KR, Zhang X. Dimensional sentiment analysis using a regional CNN-LSTM model. InProceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2016;2:225-230.

[11] Wang J, Yu LC, Lai KR, Zhang X. Tree-structured regional CNN-LSTM model for dimensional sentiment analysis. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2019;28:581-91.

[12] Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781; 2013.

Peer-review history:
The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)
https://www.sdiarticle4.com/review-history/72724