Interspeech 2018
2-6 September 2018, Hyderabad

Imbalance Learning-based Framework for Fear Recognition in the MediaEval Emotional Impact of Movies Task

Xiaotong Zhang, Xingliang Cheng, Mingxing Xu, Thomas Fang Zheng

Department of Computer Science and Technology
Center for Speech and Language Technologies, Research Institute of Information Technology
Tsinghua University, Beijing, China
{zhang-xt16,chengxl16}@mails.tsinghua.edu.cn, {xumx,fzheng}@tsinghua.edu.cn

Abstract

Fear recognition, which aims at predicting whether a movie segment can induce fear or not, is a promising area in movie emotion recognition. Research in this area, however, has reached a bottleneck. Difficulties may partly result from the imbalanced database. In this paper, we propose an imbalance learning-based framework for movie fear recognition. A data rebalance module is adopted before classification. Several sampling methods, including the proposed softsampling and hardsampling which combine the merits of both undersampling and oversampling, are explored in this module. Experiments are conducted on the MediaEval 2017 Emotional Impact of Movies Task. Compared with the current state-of-the-art, we achieve an improvement of 8.94% on F_1, proving the effectiveness of proposed framework.

Index Terms: fear recognition, imbalance learning, sampling methods, softsampling, hardsampling

1. Introduction

Automatically recognizing fear that is induced by movie segments is a challenging task in movie emotion recognition. Many applications can be found in this field, such as detecting horror movie and protecting children from potentially harmful video contents [1], etc..

In recent years, movie emotion recognition has made great progress. Penet et al. presented a violent shots detection system that studied several methods for introducing temporal and multimodal information in the framework [2]. In [3], Mixture of Experts (MoE)-based fusion model was proposed by Goyal et al. to combine multi-modalities for predicting the emotion evoked in movies.

The frameworks mentioned above, however, are not suitable for movie fear recognition since they do not have the ability to work on the extremely imbalanced database.

Imbalance data learning is an important challenge in movie fear recognition. On the one hand, the horror movie is only a branch of film genre with limited sources in the movie market. On the other hand, the number of fear segments should be controlled in a reasonable range even in a horror movie, according to the movie theory [4]. Under this circumstances, the total number of movie segments that can induce fear is far less than not-fear segments. That is to say, the original data we can obtain from the movie market is extremely imbalanced in this binary classification task (i.e. fear vs. not-fear), making the classifier hard to be trained.

In this paper, we propose an imbalance learning-based framework to solve the above-mentioned challenge. A data rebalance module, which combines conditional data sampling methods, is applied before classification. We also integrate multimodal features, including audio features, visual features, and emotion-space features, at the feature level. Posterior probabilities predicted by the classifiers are fused using soft voting at the decision level.

The remainder of the paper is organized as follows. The related work is introduced in Section 2. Section 3 describes the imbalance learning-based framework in detail. Experiments are conducted in Section 4 and results are presented in Section 5. In section 6, we discuss the results from different perspectives. Finally, the conclusion is drawn in Section 7.

2. Related Work

Much work focusing on movie emotion recognition has been studied in the past few decades. The complex interplay between multi-modalities, such as audio and video, makes movie emotion recognition a more challenging task compared with speech emotion recognition.

Srivastava et al. proposed a bimodal framework for movie emotion recognition [5]. In this framework, they combined facial expression recognition with lexical analysis of dialogues in movies to recognize emotions of characters in movies. Midlevel concept feature, which is based on detectable movie shot concepts, was proposed to bridge the “affective gap” by Ellis et al. [6].

Traditional classifiers that are often adopted in movie emotion recognition include Support Vector Machine (SVM) [7] and Random Forest (RF) [8]. In recent years, deep learning has become a new classifier in many emotion recognition tasks. Nguyen et al. introduce a novel approach using 3-dimensional convolutional neural networks (C3Ds) and multimodal deep-belief networks (DBNs) to improve the performance of multimodal emotion recognition [9].

Data imbalance is an important factor that may largely influence the capability of the classifier. Imbalance data learning can be categorized into two groups, i.e. sampling methods and cost-sensitive learning [10].

Undersampling and oversampling are two traditional sampling methods [10]. Undersampling, such as EasyEnsemble and BalanceCascade [11], randomly removes some data from the set of the majority class. In contrast, oversampling replicate samples for the set of the minority class. Algorithms such as SMOTE [12], Borderline-SMOTE [13], Adaptive Synthetic Sampling [14], and MWMOTE [15] are all classic methods of oversampling.

Cost-sensitive learning mainly considers the costs of the misclassified samples [16]. This method has been used in many

10.21437/Interspeech.2018-1744

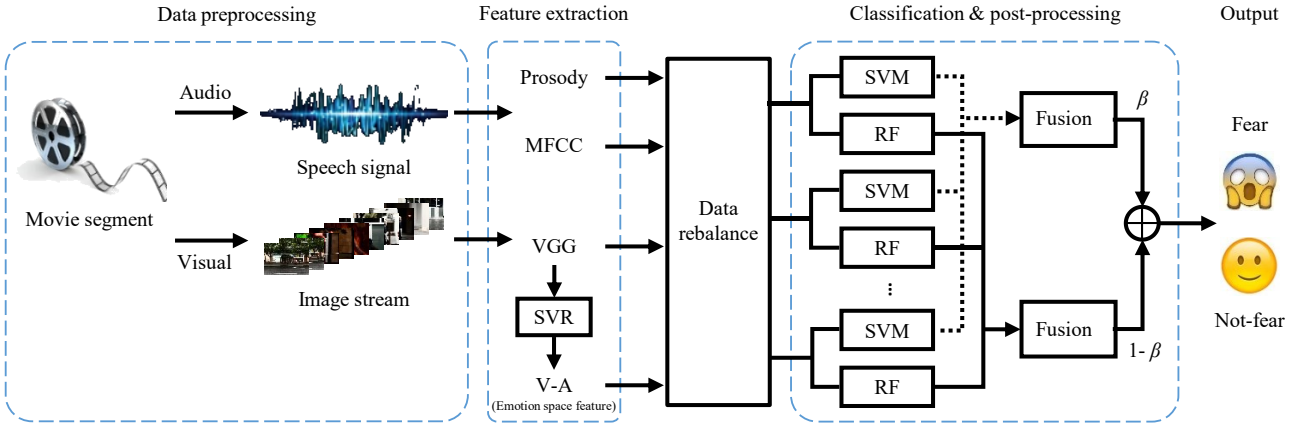

Figure 1: *Imbalance learning-based framework for movie fear recognition*

classification systems, including boosting [17], decision trees [18], and neural networks [19].

3. Imbalance Learning-based Framework

The general workflow of imbalance learning-based framework is illustrated in Fig.1. There are three main modules, i.e. feature extraction, data rebalance, and classification & post-processing. The sampling methods in data rebalance module and the fusion strategy in classification & post-processing module are described with more details in the following subsections.

3.1. Problem definition

Given N training movies M_1, M_2, …, M_N, each of them is segmented using a 10s-window with 5s shift, so $M_n=\{m_{n1}, m_{n2}, \ldots, m_{nk_n}\}$, k_n is the number of segments in M_n. The training set M can be defined by all segments, i.e. $M=\{m_{11}, m_{12}, \ldots, m_{1k_1}, m_{21}, \ldots, m_{Nk_N}\}$. Each segment m_{ij} has a label l_{ij} to indicate whether m_{ij} can induce fear (l_{ij}=1) or not (l_{ij}=0).

Then, the fear recognition task can be considered as a binary classification problem, aiming at finding a mapping function Φ to predict labels for given movie segments, as shown in Equation (1):

$$\hat{l}_{ij} = \Phi(m_{ij}) \tag{1}$$

3.2. Sampling methods

Undersampling and oversampling are two traditional sampling methods that are widely used for data rebalancing. However, undersampling gets true balanced data at the cost of discarding useful training samples while the samples generated by oversampling may be unreliable. Taken the complementarity of both undersampling and oversampling into consideration, we propose two data rebalance methods which combine the traditional sampling methods to enhance their strengths.

The combination methods are illustrated in Fig.2. One is what we call hardsampling. It applies undersampling before oversampling, while the other one, which is called softsampling, applies undersampling after oversampling.

Given an imbalanced dataset, where the size of the majority class is A and the size of the minority is B. A is extremely bigger than B.

As for the hardsampling (see Fig.2(a)), $X(X \leqslant B)$ samples and αX (α >1) samples are randomly chosen from the minority class and the majority class, respectively. After this imbalanced undersampling, oversampling is used to generate another (α-1)*X samples for the minority class. Therefore, a total of αX samples are obtained for both the minority class and the majority class, forming a $2\alpha X$-sample subset. This process is repeated T times and we get T subsets with $2\alpha X$ training samples in each subset.

Softsamping (see Fig.2(b)) uses oversampling in the first place to balance the overall training samples. Undersampling is then performed to generate Y samples for the majority class and Y samples for the minority class randomly.

The ratio of generated minority samples and real minority samples in each subset maintains to be α-1 (see the dashed line in Fig.2(a)) in hardsampling method, while in softsampling, the percentage in each subset is at random.

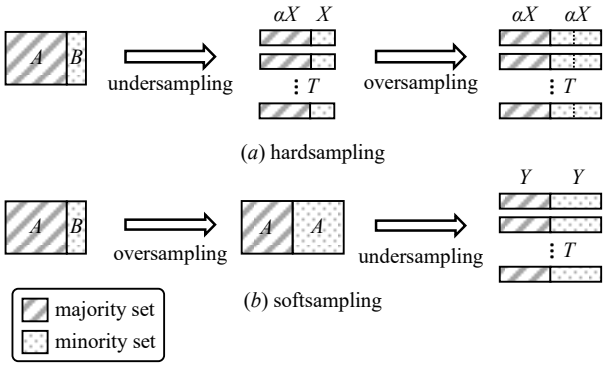

Figure 2: *Sampling methods*

3.3. Post-processing strategy

Late fusion is carried out at the decision level in the post-processing module. Two traditional classifiers, Support Vector Machine (SVM) and Random Forest (RF), are adopted as the classifiers for the rebalanced T subsets.

Soft voting, which is based on posterior probabilities, is used to fuse the predictions of SVMs and RFs. The fusion results are distinguished and voted using different weights, as presented in Equation (2):

$$P = \beta P_{SVM} + (1-\beta) P_{RF} \tag{2}$$

where P_x is the predicted value output by classifier x. x refers to SVM or RF. β presents the weight. A sample will be classified as fear if the voting result P is larger than a threshold t, which is determined experimentally.

4. Experiments

4.1. Database

The experiments are conducted on the LIRIS-ACCEDE database [20], which is provided in MediaEval 2017 Emotional Impact of Movies Task [1]. The development set consists of 30 movies with 442.08 minutes in total length. The test set consists of a selection of 14 movies with 477.22 minutes in total length.

Movies in the development set and the test set are fragmented into consecutive 10-second segments sliding with a shift of 5-second in the whole file. For each segment, valence and arousal values for consecutive 10-second are provided, as well as the indication whether this segment is able to induce fear (value 1) or not (value 0).

The first 6 movies in the development set are chosen as the validation set for parameter tuning. The training set contains the remaining movies. Data distribution is shown in Table 1.

Table 1: *Data distribution*

	# Movie	# Fear Segments	# Not-fear Segments	# Total Segments
Training Set	24	230	4062	4292
Validation Set	6	53	929	982
Test Set	14	204	5506	5710

4.2. Feature extraction

Total 4172-d multimodal features were extracted including 74-d audio features, 4096-d visual features, and 2-d V-A features.

Audio features. Fear segments often contain terrifying background sounds such as unexpected screams, irregular tones, and low grumble. To depict this characteristic, we extract the 35-d prosody features (see Table 2, 5*7=35) and the 39-d Mel-Frequency Cepstral Coefficients (MFCC) features by using openSMILE toolbox [21] with configuration files named "prosodyShsViterbiLoudness.conf" and "MFCC12 _E_D_A_ Z.conf", respectively.

Table 2: *Prosody features*

Low Level Descriptor	Functional
1. F0	1. Standard deviation
2. log(F0)	2. Mean
3. Voicing final unclipped	3. Linear regression
4. Harmonics to noise ratio	4. Centroid
5. Loudness	5. Percentile 10.0
	6. Percentile 90.0
	7. Percentile range 0-1

Visual features. As for visual features, we capture images from movie segments every one second. CNN has proven to be useful in learning image features. We adopt the deep features extracted by VGGNet (pre-trained VGG16 [22] fc6 layer). Ten images are captured in a 10-second movie segment and the segment-level features are the average of each image features.

Valence-Arousal features. We consider that audio and visual features are not well-designed to describe induced emotions since they also contain some redundant information that is irrelevant to affects. In order to learn features that are directly related to emotion, features in the image space are embedded into the Valence-Arousal (V-A) emotion space to extract emotion-related features, which we call V-A features. According to the labeled V-A of 10-second movie segment on the training set, we train a Support Vector Regression (SVR) model using VGG features and then embed the VGG features of the training set, the validation set and the test set in the emotion space using the pre-trained SVR.

4.3. Evaluation metrics

Accuracy, precision, recall, and F_1 are evaluation metrics that are often used to assess the performance of binary classification systems, defined as

$$accuracy = \frac{tp + tn}{tp + tn + fp + fn}$$
$$precision = \frac{tp}{tp + fp}, \qquad recall = \frac{tp}{tp + fn} \tag{3}$$
$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

where *tp*, *tn*, *fp*, and *fn* represent true positive, true negative, false positive, and false negative, respectively. Among all these metrics, F_1 is the most comprehensive one, especially when the dataset is imbalanced.

The official metrics used in the MediaEval 2017 Emotional Impact of Movies Task are all calculated at the movie-level. To be more specific, all the metrics are firstly calculated on each movie separately and then averaged, defined as

$$R = \frac{1}{N}\sum_{i=1}^{N} R(M_i) \tag{4}$$

where *R* refers to the results of accuracy, precision, recall, and F_1. *N* is the number of movies. M_i represents the *i*-th movie.

4.4. Experimental setup

In the data rebalance module, α and *X* is fixed to 2 and 200 respectively, and *Y* is fixed to 230.

In the classification & post-processing module, SVM and RF are trained using the scikit-toolkit [23]. RBF kernel is used for SVM. The number of trees in RF is 100 and the maximum number of leaf nodes is 50. Decision threshold *t* and weight β are determined simultaneously by grid search from 0 to 1 with a step of 0.05. All these parameters are tuned on the validation set based on movie-level F_1.

5. Results

5.1. Contribution of sampling methods

We choose the framework without data rebalance module as the baseline system. Traditional undersampling and oversampling methods we choose are random sampling and SMOTE, respectively. Random sampling generates 20 subsets using all fear samples (230, see Table 1) and the same number of not-fear samples. SMOTE generates fear samples until the numbers of fear samples and not-fear samples are the same (4062, see Table 1). In softsampling and hardsampling, the undersampling and oversampling methods are also random sampling and SMOTE, respectively. The features used in these systems are all the integration of audio, visual, and Valence-Arousal.

Table 3 presents the results of different sampling methods. The results on F_1 demonstrate that softsampling outperforms traditional sampling methods. The framework without data rebalance module has the worst performance on most of the metrics, especially on F_1.

It seems that the performance of hardsampling is worse than simply undersampling or oversampling. One possible reason is that the number of samples used for SMOTE in hardsampling is much less than in softsampling. Many incorrect samples are generated in hardsampling during this process, making the classification unreliable.

Table 3: *Contribution of sampling methods*

Sampling Method	Accuracy	Precision	Recall	F_1
Without Rebalance	0.7408	0.2586	0.2278	0.1863
SMOTE	0.7258	0.2625	**0.5379**	0.3056
Random Sampling	**0.8317**	**0.3518**	0.3435	0.2961
Hardsampling	0.8201	0.2594	0.3895	0.2843
Softsampling	0.7745	0.3171	0.4969	**0.3246**

5.2. Contribution of modalities

According to the conclusion drawn in section 5.1, we apply softsampling in the data rebalance module to present the contributions of modalities, shown in Table 4.

The results show that V-A plays an important role in the multimodal features. Visual is the best single modality that contributes most to the framework. Considering dimensions, the performance indicating that 2-d V-A features are extremely emotion-related and pure, comparing with 4096-d visual features and 74-d audio features.

Table 4: *Contribution of modalities*

Modality	Accuracy	Precision	Recall	F_1
Audio+Visual+V-A	0.7745	0.3171	**0.4969**	**0.3246**
Audio+Visual	**0.7844**	**0.3205**	0.4510	0.3090
V-A	0.6589	0.2267	0.3402	0.2412
Audio	0.6449	0.1825	**0.4787**	0.2353
Visual	**0.7735**	**0.3064**	0.4346	**0.2907**

5.3. Comparison with other research groups

Table 5 compares the best performance of proposed framework with other groups in the MediaEval 2017 Emotion Impact of Movie Task. The performance of the proposed framework achieves a qualitative leap on F_1, up to 8.94% increase compared with the best group THUHCSI [27].

We notice that the first three groups [24-26] did not pay attention to the imbalance of the database. This makes their frameworks tend to predict more segments as not-fear, which may improve accuracy and precision. However, recall is pretty low in this case, making F_1 declines in general.

Table 5: *Comparison with other groups*

Group Name	Accuracy	Precision	Recall	F1
HKBU [24]	0.7630	0.1688	0.0657	0.0786
MIC-TJU [25]	**0.8623**	**0.3756**	0.0991	0.1424
TCNJ-CS [26]	0.7296	0.2553	0.1922	0.1740
THUHCSI [27]	0.8153	0.2318	0.2781	0.2352
Proposed method	0.7745	0.3171	**0.4969**	**0.3246**

6. Discussion

There are 14 movies in the test set, 4 of which do not contain fear segments. Therefore, even if the system can perfectly classify each test segment, precision is only 0.7143 (i.e. (14-4)/14) at the most, so do recall and F_1. Moreover, some movies have only few fear segments, making recall sensitive to the algorithm tuning. The imbalanced distribution of fear segments between movies may also result in the unreliability of movie-level evaluation metrics.

Therefore, to further assess the performance of our system, we explore the evaluation metrics from another two perspectives. (1) movie-level metrics are recalculated on the sub-test set which only contains movies that have at least one fear segment. (2) segment-level metrics on the whole test set are computed. Segment-level ignores the differences between movies and treat all the segments equally. When segment-level metrics is adopted to evaluate our system, we tune parameters at the segment-level correspondingly.

Table 6 demonstrates that our framework can find 69.57% fear segments on the sub-test set with 44.39% precision. From this point of view, our framework can be used as a reference in the horror movie ranking or detection. However, precision at the segment-level is only 20.24% with 58.33% recall, indicating that there is still a far distance before application. Moreover, accuracy is 90.30% in this case, which suggests that it is not a suitable evaluation metric on imbalanced database.

Table 6: *Different evaluation perspectives*

Perspective	Accuracy	Precision	Recall	F_1
Movie-level	0.7745	0.3171	0.4969	0.3246
Movie-level (sub)*	0.7628	0.4439	0.6957	0.4544
Segment-level	0.9030	0.2024	0.5833	0.3005

* It refers to a subset of the test set, in which each movie contains at least one fear segment.

7. Conclusion

In this paper, we propose a novel imbalance learning-based framework for movie fear recognition. Several sampling methods are applied in the data rebalance module. Softsampling and hardsampling are proposed to combine the advantages of oversampling and undersampling. Multimodal features are extracted and integrated at the feature level. Posterior probabilities predicted by the classifiers are fused using soft voting at the decision level. The results of our framework reach a state-of-the-art performance on recall and F_1 in the MediaEval 2017 Emotion Impact of Movie Task.

Although this paper has provided a promising baseline for movie fear recognition, there are still several potential improvements remain to be investigated in the future. Firstly, considering the temporal structure of movies, contextual features are going to be learned using LSTM. Secondly, softsampling and hardsampling will be explored in detail to further assess their performance.

8. Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (61433018, 61171116) and the National High Technology Research and Development Program of China (863 program) (2015AA016305).

9. References

[1] E. Dellandréa, M. Huigsloot, and L. Chen, et al., "The Mediaeval 2017 Emotional Impact of Movies Task," *MediaEval 2017 Multimedia Benchmark Workshop Working Notes Proceedings of the MediaEval 2017 Workshop*, 2017.

[2] C. Penet, C.H. Demarty, and G. Gravier, et al., "Multimodal information fusion and temporal integration for violence detection in movies," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2393-2396, 2012.

[3] A. Goyal, N. Kumar, T. and Guha, et al., "A multimodal mixture-of-experts model for dynamic emotion prediction in movies," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2822-2826, 2016.

[4] Y. L. Wang, *Research on American Horror Movies since 2000*. Jiangsu: Nanjing Normal University, 2011.

[5] R. Srivastava, S. Yan, and T. Sim, et al., "Recognizing emotions of characters in movies," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp.993-996, 2012.

[6] J. G. Ellis, W. S. Lin, and C. Y. Lin, et al., "Predicting evoked emotions in video," *IEEE International Symposium on Multimedia*, pp.287-294, 2014.

[7] Y. J. Liu, M. Yu, and G. Zhao, et al., "Real-time movie-induced discrete emotion recognition from EEG signals," *IEEE Transactions on Affective Computing*, 2017.

[8] M. Kotti and Y. Stylianou, "Effective emotion recognition in movie audio tracks," *IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE,* pp. 5120-5124, 2017.

[9] D. Nguyen, K. Nguyen, and S. Sridharan, et al., "Deep spatio-temporal features for multimodal emotion recognition," *Applications of Computer Vision*, pp. 1215-1223, 2017.

[10] H. He and E.A. Garcia. "Learning from imbalanced data." *IEEE Transactions on Knowledge and Data Engineering*, pp.1263-1284, 2009.

[11] X.Y. Liu, J. Wu, and Z.H. Zhou, "Exploratory under sampling for class imbalance learning," *Proc. Int'l Conf. Data Mining*, pp. 965-969, 2006.

[12] N.V. Chawla, K.W. Bowyer, and L.O. Hall, et al., "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, pp. 321-357, 2002.

[13] H. Han, W. Wang, and B. Mao, et al., "Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning," *International conference on intelligent computing*, pp. 878-887, 2005.

[14] H. He, Y. Bai, and E.A. Garcia, et al., "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," *International symposium on neural networks*, pp. 1322-1328, 2008.

[15] S. Barua, M.M. Islam, and X. Yao, et al., "MWMOTE--Majority weighted minority oversampling technique for imbalanced data set learning," *IEEE Transactions on Knowledge & Data Engineering*, pp. 405-425, 2013.

[16] C. Elkan, "The foundations of cost-sensitive learning," Proc. Int'l Joint Conf. Artificial Intelligence, pp. 973-978, 2001.

[17] Y. Sun, M.S. Kamel, and A.K.C. Wong, "Cost-sensitive boosting for classification of imbalanced data," *Pattern Recognition, vol. 40, no. 12*, pp. 3358-3378, 2007.

[18] M.A. Maloof, "Learning when data sets are imbalanced and when costs are unequal and unknown," *Proc. Int'l Conf. Machine Learning, Workshop Learning from Imbalanced Data Sets II*, 2003.

[19] M.Z. Kukar and I. Kononenko, "Cost-sensitive learning with neural networks," *Proc. European Conf. Artificial Intelligence*, pp. 445-449, 1998.

[20] Y. Baveye, E. Dellandrea, and C. Chamaret, et al., "LIRIS-ACCEDE: A video database for affective content analysis," *IEEE Transactions on Affective Computing*, pp. 43-55, 2015.

[21] F. Eyben, F. Weninger, and F. Gross, et al., "Recent developments in opensmile, the munich open-source multimedia feature extractor," *Proceedings of the 21st ACM international conference on Multimedia*, pp. 835-838, 2013.

[22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *international conference on learning representations*, 2015.

[23] F. Pedregosa, G. Varoquaux, and A. Gramfort, et al., "Scikit-learn: Machine learning in python." *Journal of Machine Learning Research*, pp. 2825-2830, 2011.

[24] Y. Liu, Z. Gu, and T.H.Ko, "HKBU at MediaEval 2017 emotional impact of movies task," *MediaEval 2017 Multimedia Benchmark Workshop Working Notes Proceedings of the MediaEval 2017 Workshop*, 2017.

[25] Y. Yi, H. Wang, and J. Wei, "MIC-TJU in MediaEval 2017 emotional impact of movies task," *MediaEval 2017 Multimedia Benchmark Workshop Working Notes Proceedings of the MediaEval 2017 Workshop*, 2017.

[26] S. Yoon, "TCNJ-CS @ MediaEval 2017 emotional impact of movie task," *MediaEval 2017 Multimedia Benchmark Workshop Working Notes Proceedings of the MediaEval 2017 Workshop*, 2017.

[27] Z. Jin, Y. Yao, and Y. Ma, et al., "THUHCSI in MediaEval 2017 emotional impact of movies task," *MediaEval 2017 Multimedia Benchmark Workshop Working Notes Proceedings of the MediaEval 2017 Workshop*, 2017.