Multimodal Feature Fusion Method for Unbalanced Sample Data in Social Network Public Opinion

Jian Zhao 1,2,3, Wenhua Dong 1,2,3, Lijuan Shi 3,4, Wenqian Qiang 1, Zhejun Kuang 2,3,*, Dawei Xu 1 and Tianbo An 1

1 School of Cyber Security, Changchun University, Changchun 130022, China; zhaojian@ccu.edu.cn (J.Z.); daniel_daren@outlook.com (W.D.); qiangayo@163.com (W.Q.); xudw@ccu.edu.cn (D.X.); tambo@163.com (T.A.)
2 School of Computer Science and Technology, Changchun University, Changchun 130022, China
3 Jilin Provincial Key Laboratory of Human Health Status Identification and Function Enhancement, Changchun 130022, China; shilj@ccu.edu.cn
4 School of Electronic Information Engineering, Changchun University, Changchun 130022, China
* Correspondence: kuangzhejun@ccu.edu.cn

Abstract: With the wide application of social media, public opinion analysis in social networks has been unable to be met through text alone because the existing public opinion information includes data information of various modalities, such as voice, text, and facial expressions. Therefore multimodal emotion analysis is the current focus of public opinion analysis. In addition, multi-modal emotion recognition of speech is an important factor restricting the multi-modal emotion analysis. In this paper, the emotion feature retrieval method for speech is firstly explored and the processing method of sample disequilibrium data is then analyzed. By comparing and studying the different feature fusion methods of text and speech, respectively, the multi-modal feature fusion method for sample disequilibrium data is proposed to realize multi-modal emotion recognition. Experiments are performed using two publicly available datasets (IEMOCAP and MELD), which shows that processing multi-modality data through this method can obtain good fine-grained emotion recognition results, laying a foundation for subsequent social public opinion analysis.

Keywords: multi-modality; fine-grained; model fusion; emotion features; feature generation

1. Introduction

Online public opinion gathers public views on public social events and has a huge impact on associated objects [1], such as the melamine event in 2008 [2], the Wei Zexi event in 2016 [3], and “COVID-19” in 2020 [4]. The exposure of these events has a heavy blow on relevant objects, revealing many problems existing behind the event and causing a significant impact on people’s life safety. People’s emotions have undergone great changes with the development and spread of public opinion, which has caused much trouble for public opinion management, including the World Health Organization and governments of different countries [5]. In addition, online public opinion has caused great distress for people to analyze online public opinion because of its attributes of complex content, the coexistence of truth and falsehood, and easy dissemination [6]. As an important part of public opinion analysis, emotion recognition plays an important role in the field of artificial intelligence, which still remains a challenging task even with the development of deep learning and natural language processing. The main reason is that there are many ways and characteristics of expressing emotion, such as implicit emotion, dialogue emotion, and so on. We can capture emotional features by studying different ways, such as speech features, video features, facial features, EEG features [6–9], etc. Therefore, fine-grained multi-modal emotion recognition has become one of the current hotspots in the field of emotion analysis. Among them, for the selection of multi-modal data, speech is the easiest to obtain, and it
is also the most widely used mode in people’s daily communication and there are also a large number of emotional features in the speech signal, so emotion analysis is the core and hotspot in the field of multi-modal research [10–12].

In the field of multi-modal emotion analysis, there are many research models for different modalities. Wollmer et al. [13] fused audio and video modalities for the first time and used bidirectional long and short-term memory neural networks to conduct multi-modal emotion analysis, whose experimental results showed that this multi-modality recognition effect was superior to the single-modality results. Morency et al. [14] conducted validation experiments for multi-modality, and the results demonstrated that a joint model integrating video, audio, and text features could be effectively used to identify emotions in online videos. Afterward, Poria et al. [15] proposed a new approach for multi-modal emotion analysis, which consisted of collecting emotion from web videos by presenting a model that used audio, video and text modalities as information sources. Soleymani et al. [16] proposed a multi-modal emotion data analysis framework to retrieve user opinions and sentiments from video content. Poria et al. [17] proposed a multi-modality analysis model based on LSTM that enables utterances to capture contextual information from their surroundings in the same video, thereby aiding emotion analysis. In 2018, Majumder et al. [18] proposed a novel feature fusion strategy, which is performed hierarchically by first fusing the two modes together and then fusing the third pattern to provide a new direction for multi-modal feature fusion. More and more people study different model fusion methods and optimize them continuously, but the problem of sample imbalance in the data set itself, as well as the emotion feature retrieval method that restricts a single modality in the process of modal fusion, are not analyzed.

Based on this problem, the main contributions of this paper are as follows:

1. In the text and speech models, the paper first analyzes the speech emotion feature retrieval method, which restricts the modal fusion in the existing modal fusion method, analyzes the emotion features in the speech signal, and then puts forward the MA2PE speech feature retrieval method.

2. As for the problem of sample disequilibrium, several common data processing methods for sample disequilibrium are analyzed, and SOM oversampling method are finally proposed.

3. A fine-grained emotion recognition method for both text and speech modalities of sample disequilibrium data is proposed and validated on IEMOCAP and MELD datasets, proving that this method outperforms the existing models.

2. Data Preprocessing and Model Architecture

2.1. Audio Feature Extraction Method—MA2PE

The traditional audio feature retrieval method MFCC obtains audio features with loss in treble or bass more or less. Meanwhile, the length of the generated audio features is also different due to the different lengths of each audio segment in the dataset; therefore, its length needs to be processed in a cutting or filling manner before the retrieved audio features are sent into the model for analysis, which to a large extent makes audio features have different deletions or redundancies [19–21].

Based on this, a new feature retrieval method, MA2PE, is proposed, which adopts a series of operations to retrieve and convert all speech data of different lengths into 8-dimensional feature vectors. The specific operation methods are as follows:

First, a 44.1 kHz sampling rate is used to read the time series of each audio, and the mean and standard deviation of the absolute value of each audio time series are obtained as two of the features, which not only retain other features but also play a role in integrating all length vectors, and the mean and standard deviation can also reflect certain feature differences.

Second, because the energy of the speech signal is related to its tone, thus it can be used to detect certain elevated emotions, such as anger, excitement, etc., as the tone and loudness of voices emitted by people in the case of anger will be different. Therefore, the
The following formula is used to calculate the root mean square of each audio spectral feature frame by frame and calculate the corresponding standard deviation to learn and represent the tone and loudness features of speech signals.

\[
E = \sqrt{\frac{1}{n} \sum_{i=1}^{n} y[i]^2}
\]  

(1)

Here \( n \) denotes the length of the time series, and \( y[i] \) denotes the input signal.

Then, we obtained the ratio of mute state to the total state in the audio signal as the fifth feature because of the fact that our speech rate will also vary in different emotional states, which leads to different proportions of mute state and it is significant for us to learn emotion features. Besides, we also calculate the harmonic energy in the time-frequency signal as the sixth feature. When people change their emotions, there will be changes in harmonics, which provides a reference for us to do fine-grained emotion analysis.

Finally: we all know that the waveform generated by articulation changes with our mood, which is the treble signal. There are many commonly used treble detection algorithms, including the modified autocorrelation function method (MACF) and normalized cross-correlation function method (NCCF), for example. This paper adopts the autocorrelation algorithm based on the center clipping frame, which is calculated by the following formula:

\[
res[n] = \begin{cases} 
  y[n] - C_l, & y[n] \geq C_l \\
  0, & |y[n]| < C_l \\
  y[n] + C_l, & y[n] \leq C_l 
\end{cases}
\]

(2)

where \( y[n] \) is the input signal and \( C_l \) is half of the mean value of the input signal. We also put the calculated autocorrelation coefficient \( res[n] \) and the normalized value into the features of speech emotion.

Through the above methods, the 8-dimensional feature retrieval method of MA2PE can be obtained.

2.2. Oversampling Processing Method Based on SOM

The traditional sample disequilibrium processing method includes an oversampling method, undersampling method, and reweighting method [22–24], among which the traditional oversampling and undersampling method is to copy and delete the original sample. However, this method of obtaining text features causes repetition and waste of data features and cannot play a great role in the learning of fine-grained emotion features, while the multi-modal reweighting method also has problems more or less [25–27]. Therefore, this paper proposes a SOM oversampling method - an oversampling method based on TF-IDF synonymous substitution. It uses the commonly used text segmentation method to segment the text, obtains different keywords through TF-IDF, expands data according to the needs of the text, finds multiple alternative topics from the big data thesaurus, and replaces them to generate new sample data. In addition, through the text generated according to the sample and the voice data corresponding to the text of the original data, the package in the moviepy library is used to modify the audio content without changing audio information such as pitch, pitch, etc. The characteristics of the new samples generated by this method are different from those generated by the previous method, and they are diverse. In this way, if they are sent to the model for learning, more characteristics can be obtained, which is very useful for the analysis of fine-grained emotion features.

2.3. Modal Fusion Method

For the modal fusion method, the fusion methods based on the decision layer and the feature layer are selected respectively for the analysis and verification of the comparative experiments. The following two fusion models are designed respectively, as shown in Figures 1 and 2.
Among the fusion methods based on the decision layer, there are several common combination methods, including the voting method, fuzzy integration method and D-S evidence reasoning method. However, this classification method cannot allocate appropriate weights based on the features of each modality sufficiently [28]. Hence, this paper proposes a dynamic weight allocation method. According to the classification results of speech and text, the same weight is given first, and the weight ratio is adjusted continuously by comparing the correct results with the linear weighting plan, and finally, the suitable weight is obtained.

2.4. Model Framework

The multi-modal fine-grained emotion analysis model based on feature layer fusion designed in this paper is mainly composed of four modules: processing module of few shot data, text feature retrieval module, speech feature retrieval module, and multi-modal feature layer fusion module. As shown in Figure 3, the model processes text and speech modalities separately.

According to the text modality, firstly, the text data expansion method is used to amplify the few shot data in the sample disequilibrium data, and then the amplified text is feature retrieved.

For the speech modality, the corresponding audio data generation method is used to generate the corresponding text, and then the audio data feature is retrieved. Because the audio feature dimension proposed in this paper is only 8-dimensional, it is easy to complicate the features by selecting dot multiplication or dot product operation. Thus, finally, the feature layer is fused by simple stitching and then sent into the model for analysis to obtain the required results.
3. Parameter Setting and Result Analysis

3.1. Datasets

Due to the lack of emotional data of existing network discourse, we mainly use two multi-modal dialogue emotional data sets, MELD and IEMOCAP, for experiments. The MELD dataset comes from about 13,000 utterances from 1433 dialogues from the TV series Friends. It is divided into seven categories in total, as shown in Table 1. The IEMOCAP dataset contains about 12 h of audio-visual data, including video, voice, facial motion capture, and text content. Contains anger, happiness, sadness, fear, surprise, and neutral six emotions; this study focuses on two modes, the corresponding amount of data (text and audio) shown in Table 1.

Table 1. IEMOCAP dataset and MELD dataset data volume.

| Emotion Category | IEMOCAP Total Data Volume | MELD Total Data Volume |
|------------------|---------------------------|------------------------|
| ang              | 1103                      | 1607                   |
| Hap/joy          | 595                       | 2308                   |
| Sad/sadness      | 1041                      | 1002                   |
| fear             | 1084                      | 358                    |
| surprise         | 1849                      | 1636                   |
| neutral          | 1708                      | 6436                   |
| disgust          | —                         | 361                    |
3.2. Experiment Procedures

(1) For the few shot data, the oversampling method is used to generate text and speech data corresponding to the results, respectively. The specific generation results are shown in Figure 4.

(2) Text dialogues are divided into tokens and each word becomes lowercase.

(3) As for the text processing scheme, TF-IDF is used to obtain the weight matrix of the text.

(4) Set the audio sampling rate to 44,100 Hz.

(5) Audio features were retrieved by using the audio feature retrieval method of MA2PE.

(6) A new two-modality feature is generated by simply stitching the generated speech and text features.

(7) The model is trained by using the algorithm of Random Forest and tested on the test set, and the results of each modality are calculated.

![Figure 4](image)

**Figure 4.** The representation of sample data volume after over-sampling in MELD dataset (where 0–6 in label represents six emotional labels: ‘anger’, ‘disgust’, ‘fear’, ‘joy’, ‘neutral’, ‘sadness’, ‘surprise’).

3.3. Experiment Results and Analysis

In this part, the results of previous experiments are presented and analyzed. Tables 2 and 3 present the results of the MELD data set and IEMOCAP data set in the case of text, speech, and multi-modal fusion, respectively. Because there are differences in the data volume between MELD and IEMOCAP, and there are differences in the sample disequilibrium proportion, the MELD dataset and IEMOCAP dataset will be analyzed separately in this paper.

3.3.1. Experiment Results and Analysis of MELD Data Set

This paper first conducts an evaluation experiment of model comparison on the MELD data set and the experimental results are shown in Table 2. The heat map of the results on the multi-modal fine-grained emotion analysis structure based on feature layer fusion is shown in Figure 5. The benchmark models of this experiment are the Text-CNN, HiGRU-sf [29], cMKL, bLSTM [17], and DialogueRNN [15] models shown in Table 2, all of which are the optimal results of multi-modal fine-grained analysis of the MELD data set. Other models, including SVC, LR, etc., are all models used after over-sampling the samples. According to the results in the table and the following analysis results, we can draw the following conclusions:
As for text data: we can observe that the accuracy rate of almost all emotions is higher than that of audio, which may be due to the fact that the features obtained from text are richer than those of audio. As shown in Figure 6, compared with the benchmark model, we
can find that almost all oversampling processing methods are superior to other methods, while the multi-modal fine-grained emotion analysis method based on feature layer fusion proposed in this paper can bring better results than other methods, and the classification effect is also the best. The accuracy rate of all emotion categories is about 80%, whether acc or w_avg is used as the evaluation index. It is the best category compared with other methods because the classification effect reaches 84.50%.

Figure 6. Fine-grained Emotion Accuracy Rate of Different Methods of MELD Data Set (Text).

As for audio data: Compared with the original audio feature retrieval method, the current one is more general and with a better result, especially for the problem of sample disequilibrium, which can effectively prevent the occurrence of the overfitting phenomenon. As shown in Figure 7, we can observe that the results of logistic regression, multi-layer perceptron and support vector machines are relatively low, while the results of ours are the highest, whose overall classification effect reaches 88.88%, and its accuracy rate of aversive emotion reaches 97.95% at the highest. In addition, it is also found that this method has high accuracy in judging negative emotions. Because of the characteristics of audio signals, it is not easy to learn but very easy to overfit. The commonly used speech feature retrieval method, such as MFCC, retrieves features in various lengths due to the problem of audio length. For the unity of the model, we need to cut or fill, which leads to redundancy or loss of feature information. While our feature retrieval method can effectively solve this problem and also retain the audio features. Therefore, it can be found that the accuracy rate of audio features retrieved on MA2PE is higher than that of text.

For multi-modal data sets: we can find the results of modal fusion, whether feature layer fusion or decision layer fusion is superior to a single modality. In the figure, text + audio (back) refers to decision layer fusion, while text + audio refers to feature layer fusion method. Besides, the best method improves the accuracy rate by almost 10%. Compared with the weighted fusion method based on the decision layer, the method of feature layer fusion is similar in effect, but the method of decision layer fusion requires reasonable learning of the weights of different modalities, which is very time-consuming. While the feature layer fusion method of stitching not only retains the features of the two modalities of text and speech but also does not cause great time consumption of the fused features because the speech features are only 8-dimensional. Therefore, it is a more appropriate choice. Compared with other basic models, our model has better results in any kind of fine-grained emotional data set, and the final classification effect reaches 94.13%, which is 33.88% higher than the best DialogueRNN model of the benchmark model. From Figure 8,
we can also find that the overall results of all the models processed by the oversampling method are superior to those of the benchmark model.

Figure 7. Fine-grained Emotion Accuracy Rate of Different Methods of MELD Data Set (speech).

Figure 8. Fine-grained Emotion Accuracy Rate of Different Methods of MELD Data Set (Multimodalities).

3.3.2. Experiment Results and Analysis of IEMOCAP Data Set

A comparison and evaluation experiment of the model is carried out on the IEMOCAP data set and the experimental results are shown in Table 3. The heat map of the results on the multi-modal fine-grained emotion analysis structure based on feature layer fusion is shown in Figures 9 and 10. As shown in Table 3, the benchmark models of this experiment are HiGRU [29], HiGRU-sf [29], mement [30], cLSTM [17], TFN [31], MFN [32], CMU [33], and ICON [10], all of which are the optimal results for the multi-modal fine-grained analysis of the IEMOCAP data set. Other models, including SVC, LR, etc., are all models used after over-sampling the samples. According to the results in the table and the following analysis results, we can reach the following conclusions:
Table 3. IEMOCAP dataset analysis result table.

| Method | Model     | ang | hap | exc | sad | fru | neu | acc | w_avg |
|--------|-----------|-----|-----|-----|-----|-----|-----|-----|------|
| HiGRU  | text      | 64.12 | 39.86 | 62.21 | 78.37 | 60.37 | 62.50 | —   | 62.52 |
| HiGRU-sf | text     | 71.18 | 51.75 | 62.88 | 70.20 | 61.68 | 64.84 | —   | 64.06 |
| LR     | audio     | 40.07 | 24.36 | 16.67 | 42.85 | 15.38 | 0   | 32.44 | 24.47 |
|        | text      | 72.71 | 69.36 | 69.89 | 61.78 | 49.03 | 45.37 | 64.01 | 62.94 |
|        | text + audio | 73.18 | 73.66 | 73.35 | 68.77 | 51.87 | 51.65 | 67.82 | 66.71 |
| MLP    | audio     | 40.22 | 25.91 | 28.86 | 42.20 | 17.41 | 46.66 | 34.10 | 33.07 |
|        | text      | 80.32 | 88.61 | 77.19 | 60.53 | 55.37 | 50.60 | 71.43 | 71.37 |
|        | text + audio | 77.66 | 89.64 | 84.92 | 82.12 | 58.41 | 54.33 | 77.94 | 76.48 |
| MNB    | audio     | 69.76 | 23.24 | 0   | 0   | 0   | 0   | 24.78 | 17.28 |
|        | text      | 73.58 | 64.17 | 65.35 | 57.35 | 49.64 | 46.24 | 61.47 | 60.59 |
|        | text + audio | 70.97 | 60.27 | 69.16 | 67.54 | 50.64 | 55.87 | 63.05 | 62.79 |
| SVC    | audio     | 39.66 | 24.73 | 100 | 39.37 | 11.36 | 0   | 32.67 | 36.57 |
|        | text      | 75.29 | 75.33 | 69.02 | 68.60 | 50.00 | 46.69 | 67.62 | 66.08 |
|        | text + audio | 73.22 | 83.18 | 77.02 | 68.89 | 50.85 | 49.61 | 70.51 | 69.07 |
| XGB    | audio     | 59.81 | 62.93 | 55.34 | 67.40 | 29.60 | 39.72 | 58.96 | 54.13 |
|        | text      | 70.16 | 42.78 | 71.30 | 59.61 | 47.66 | 45.80 | 51.46 | 55.91 |
|        | text + audio | 55.20 | 58.89 | 73.67 | 63.31 | 47.39 | 66.40 | 60.16 | 60.68 |
| Menten | Text + audio + video | 67.1 | 24.4 | 65.2 | 60.4 | 68.4 | 56.8 | — | 59.9 |
|        | Text + audio + video | 70.0 | 25.5 | 58.8 | 58.6 | 67.4 | 56.5 | — | 59.8 |
|        | Text + audio + video | 69.1 | 23.2 | 63.1 | 58.0 | 65.5 | 56.6 | — | 58.8 |
|        | Text + audio + video | 72.3 | 24.0 | 64.3 | 65.6 | 67.9 | 55.5 | — | 60.1 |
| CMU    | Text + audio + video | 67.6 | 25.7 | 69.9 | 66.5 | 71.7 | 53.9 | — | 61.9 |
| ICON   | Text + audio + video | 68.2 | 23.6 | 72.2 | 70.6 | 71.9 | 59.9 | — | 64.0 |
| Our model | audio      | 57.97 | 71.67 | 59.89 | 66.30 | 38.55 | 41.76 | 61.62 | 57.88 |
|         | text      | 77.46 | 84.78 | 77.35 | 55.81 | 56.45 | 53.72 | 69.89 | 69.77 |
|         | text + audio | 72.76 | 92.38 | 80.72 | 73.73 | 53.21 | 58.13 | 75.87 | 73.93 |

Figure 9. Result Heat Map of Fusion Model of IEMOCAP Data Set (out model).
Analysis of the text: As shown in Figure 11, the method of oversampling and the method without oversampling are compared, and the results of the oversampling method are greatly improved. Compared with other models of oversampling, the emotion analysis model based on feature layer fusion proposed in this paper can achieve the classification effect at 67.77 due to the vast majority of models.

Analysis of the audio: Because there is no benchmark model for comparison, we compare different models of the same processing method. As shown in Figure 12, the SVC of the emotion analysis model based on feature layer fusion proposed in this paper is higher than that of other models. Relatively speaking, our model is superior to other models because of its better stability.
Figure 12. Fine-grained Emotion Accuracy Rate Analysis of Different MOCAP Methods (Audio).

Analysis of multi-modal fusion: As shown in Figure 13, compared with the other three modality fusion benchmark models, the fusion results of our two modalities are superior to the fusion results of the three modalities, showing good results in various emotional result features. Compared with other models in the same processing, the classification effect of our model is basically the same as other models.

Figure 13. Fine-grained Emotion Accuracy Rate Analysis of Different MOCAP Methods (Multimodalities).

However, by comparing the classification effect of IEMOCAP and MELD data, we find that this method has a great effect on the MELD data set and a good effect on the IEMOCAP data set. Relatively speaking, its classification effect is poor. Therefore, we analyze the differences between the two data sets. By comparing the results of the two data sets shown in Table 1, we observe that the MELD data set is very wide and with stronger data disequilibrium performance. After data expansion, the data volume is basically around 8000, as shown in Figure 4, while the data disequilibrium of IEMOCAP data itself is not
obvious. After data expansion, the data volume basically remains around 1500. Compared with the MELD data set, the data volume is much less and the samples are of relative equilibrium, so it can be concluded that the multi-modal fine-grained emotion analysis model based on feature layer fusion proposed in this paper has a better effect on the sample disequilibrium data.

4. Conclusions

This paper mainly aims at the limitations of traditional speech emotion features existing in multi-modal emotion recognition and the problem of accuracy rate decline caused by disequilibrium sample data of multi-modal research. Besides, it further studies speech emotion features and sample disequilibrium problems before a speech emotion feature retrieval method of MA2PE and an oversampling method of SOM are respectively proposed. The MA2PE method can express emotion features better, which can improve the accuracy rate of emotion expression by nearly 30%. SOM oversampling method is used to amplify samples and improve the utilization rate of data. In addition, it is verified in two public data sets—MELD and IEMOCAP, whose results show that it can achieve good results.

Based on the methods of MA2PE and SOM, we propose a multi-modal fine-grained emotion analysis model based on feature layer fusion. By combining the first two methods, the model fuses text features and speech features at the feature layer and puts them into the model for analysis. Through the analysis results of different methods of the two data sets, we can conclude that this method has a good effect on the study of multi-modal data based on sample disequilibrium, and the more unequal the data are, the better the effect is.

This paper mainly analyzes public emergencies in voice and text modes, but public opinion events also include other modalities, such as video, expression, etc., which also have important guidance for the event results. Therefore, in the subsequent research, more attention will be paid to the study of public opinion data of more modalities.

Author Contributions: W.D. and J.Z. conceived and put forward the research ideas and carried out the research. W.D., J.Z. and L.S. were in charge of the calculation and experimental data. W.D., J.Z., W.Q., Z.K. and D.X. collected information and wrote the paper. T.A. Suggest revisions and refinements to the thesis. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by [the Jilin Provincial Department of Science and Technology] grant number [No. 2019020195JC. 20200601004JC. 20200301054RQ. 20200404207YY], and funded by [Science and Technology Development Plan of Jilin Province] grant number [202004031205F], and funded by [Natural Science Foundation of Jilin Province] grant number [20210101477JC], and funded by [the National Natural Science Foundation of China for supporting the research in this article] grant number [No. 61502052].

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Cai, M.; Luo, H.; Meng, X.; Cui, Y. Topic-Emotion Propagation Mechanism of Public Emergencies in Social Networks. Sensors 2021, 21, 4516. [CrossRef] [PubMed]
2. China Dairy Products Found Tainted with Melamine. BBC News, 9 July 2010.
3. Death of Wei Zexi. Available online: https://en.wikipedia.org/w/index.php?title=Death_of_Wei_Zexi&oldid=1071405010 (accessed on 5 March 2022).
4. COVID-19. Available online: https://covid19.rs (accessed on 26 September 2021).
5. Fei, G.; Yue, C.; Yaxin, W.; Yueyi, L.; Zhiran, C. Emotional health status and social mentality of the Chinese general public during the 2019 novel coronavirus pneumonia pandemic. Sci. Technol. Rev. 2020, 38, 68–76.
6. Soleymani, M.; Pantic, M.; Pun, T. Multimodal emotion recognition in response to videos. IEEE Trans. Affect. Comput. 2011, 3, 211–223. [CrossRef]
7. Koromilas, P.; Giannakopoulos, T. Deep multimodal emotion recognition on human speech: A review. *Appl. Sci.* 2021, 11, 7962. [CrossRef]

8. Mittal, T.; Bhattacharyya, U.; Chandra, R.; Bera, A.; Manocha, D. M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. *Proc. AAAI Conf. Artif. Intell.* 2020, 34, 1359–1367. [CrossRef]

9. Zheng, W.L.; Dong, B.N.; Lu, B.L. Multimodal emotion recognition using EEG and eye tracking data. In Proceedings of the 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Chicago, IL, USA, 26–30 August 2014; pp. 5040–5043.

10. Hazarika, D.; Poria, S.; Mihalcea, R.; Cambria, E.; Zimmermann, R. ICON: Interactive Conversational Memory Network for Multimodal Emotion Detection. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; pp. 2394–2604. [CrossRef]

11. Jiang, Q.; Chen, L.; Xu, R.; Ao, X.; Yang, M. A Challenge Dataset and Effective Models for Aspect-Based Sentiment Analysis. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 6280–6285. [CrossRef]

12. Mai, S.; Xing, S.; Hu, H. Analyzing Multimodal Sentiment Via Acoustic- and Visual-LSTM With Channel-Aware Temporal Convolution Network. *IEEE/ACM Trans. Audio Speech Lang. Process.* 2021, 29, 1424–1437. [CrossRef]

13. Wöllmer, M.; Metallinou, A.; Eyben, F.; Schuller, B.; Narayanan, S. Context-Sensitive Multimodal Emotion Recognition from Speech and Facial Expression using Bidirectional LSTM Modeling. In Proceedings of the 51st Annual Conference of the International Speech Communication Association, Chiba, Japan, 26–30 September 2010.

14. Morency, L.P.; Mihalcea, R.; Doshi, P. Towards Sentiment Analysis: Harvesting Opinions from the Web. In Proceedings of the 13th International Conference on Multimodal Interfaces, Alicante, Spain, 14–18 November 2011; Association for Computing Machinery: Stroudsburg, PA, USA, 2011; pp. 169–176. [CrossRef]

15. Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; Mihalcea, R. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 527–536. [CrossRef]

16. Soleymani, M.; Garcia, D.; Jou, B.; Schuller, B.; Chang, S.F.; Pantic, M. A survey of multimodal sentiment analysis. *Image Vis. Comput.* 2017, 65, 3–14.

17. Poria, S.; Cambria, E.; Hazarika, D.; Majumder, N.; Zadeh, A.; Morency, L.P. Context-Dependent Sentiment Analysis in User-Generated Videos. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver, BC, Canada, July 30–4 August 2017; Association for Computational Linguistics: Stroudsburg, PA, USA, 2017; pp. 873–883. [CrossRef]

18. Majumder, N.; Hazarika, D.; Gelbukh, A.; Cambria, E.; Poria, S. Multimodal sentiment analysis using hierarchical fusion with context modeling. *Knowl.-Based Syst.* 2018, 161, 124–133. [CrossRef]

19. Zhang, S.; Zhang, S.; Huang, T.; Gao, W. Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching. *IEEE Trans. Multimed.* 2018, 20, 1576–1590. [CrossRef]

20. Chi, P.H.; Chung, P.H.; Wu, T.H.; Hsieh, C.C.; Chen, Y.H.; Li, S.W.; Lee, H.Y. Audio Albert: A Lite Bert for Self-Supervised Learning of Audio Representation. In Proceedings of the 2021 IEEE Spoken Language Technology Workshop (SLT), Shenzhen, China, 19–22 January 2021; pp. 344–350. [CrossRef]

21. Hou, Y.; Yu, X.; Yang, J.; Ouyang, X.; Fan, D. Acoustic Sensor-Based Soundscene Analysis and Acoustic Assessment of Bird Species Richness in Shennongjia National Park, China. *Sensors* 2022, 22, 4117. [CrossRef] [PubMed]

22. Zhou, Y.; Xie, H.; Fang, S.; Wang, J.; Zha, Z.; Zhang, Y. TDI TextSpotter: Taking Data Imbalance into Account in Scene Text Spotting. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual, 20–24 October 2021; Association for Computing Machinery: New York, NY, USA, 2021; pp. 2510–2518.

23. Wang, Z.; Wu, C.; Zheng, K.; Niu, X.; Wang, X. SMOTE-Tomek-Based Resampling for Personality Recognition. *IEEE Access* 2019, 7, 129678–129689. [CrossRef]

24. Wasikowski, M.; Chen, X.W. Combating the Small Sample Class Imbalance Problem Using Feature Selection. *IEEE Trans. Knowl. Data Eng.* 2010, 22, 1388–1400. [CrossRef]

25. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized [BERT] Pretraining Approach. *arXiv* 2019, arXiv:1907.11692

26. Li, R.; Chen, H.; Feng, F.; Ma, Z.; Wang, X.; Hovy, E. Dual Graph Convolutional Networks for Aspect-based Sentiment Analysis. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Virtual, 1–6 August 2021; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 6319–6329. [CrossRef]

27. Aye, Y.M.; Aung, S.S. Sentiment analysis for reviews of restaurants in Myanmar text. In Proceedings of the 2017 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), Kanazawa, Japan, 26–28 June 2017; pp. 321–326. [CrossRef]
28. Song, X. Research on Multimodal Emotion Recognition Based on Text, Speech and Video. Master’s Thesis, Shan Dong University, Shandong, China, 2019.

29. Jiao, W.; Yang, H.; King, I.; Lyu, M.R. HiGRU: Hierarchical Gated Recurrent Units for Utterance-level Emotion Recognition. *arXiv* 2019, arXiv:1904.04446.

30. Sukhbaatar, S.; Szlam, A.; Weston, J.; Fergus, R. End-To-End Memory Networks. In *Proceedings of the Advances in Neural Information Processing Systems*; Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2015; Volume 28.

31. Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; Morency, L.P. Tensor Fusion Network for Multimodal Sentiment Analysis. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 7–11 September 2017; Association for Computational Linguistics: Stroudsburg, PA, USA, 2017; pp. 1103–1114. [CrossRef]

32. Zadeh, A.; Liang, P.P.; Mazumder, N.; Poria, S.; Cambria, E.; Morency, L.P. Memory Fusion Network for Multi-View Sequential Learning. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; AAAI Press: Palo Alto, CA, USA, 2018.

33. Hazarika, D.; Poria, S.; Zadeh, A.; Cambria, E.; Morency, L.P.; Zimmermann, R. Conversational Memory Network for Emotion Recognition in Dyadic Dialogue Videos. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), New Orleans, LA, USA, 1–6 June 2018; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; pp. 2122–2132. [CrossRef]