Design of a Multimodal Intelligent Public Opinion Analysis System

Xinyu Long*, Jiaqing Huang, Yinying Li, Peng Ai, Yang Zhang
College of Electrical Engineering, Southwest Minzu University, Chengdu, China

Abstract: As an emerging research field, multimodal sentiment analysis tasks aim to identify the speaker's emotions by combining information from different modalities. In recent years, it has been increasingly used in areas such as public opinion analysis, intelligent dialogue and user profiling. This article systematically sorts out the development context of multimodal intelligent public opinion analysis technology, and focuses on the application status of minority language processing technology in the field of public opinion monitoring. By analyzing the technical bottlenecks of the current public opinion analysis system, a multimodal fusion solution based on deep learning is proposed, and the effectiveness of the solution in social governance in ethnic minority areas is verified by combining actual cases. The research results show that a multimodal system integrating speech recognition, machine translation and sentiment analysis can significantly improve the efficiency of processing public opinion in minority languages and provide technical support for maintaining social stability in ethnic regions.

Keywords: Multimodal Analysis; Deep Learning; Public Opinion Analysis; Machine Translation; Sentiment Analysis

1. Introduction

1.1 Research Background and Significance

With the development of online public opinion, public opinion involving ethnic groups, religions, etc. has shown an increasing trend year by year in recent years. Taking Yunnan Province as an example, there are many problems. From 2017 to 2021, there are more than 100 cases of public opinion involving ethnic groups and religions every year, and they are increasing year by year. The public opinion monitoring and analysis systems currently on the market have defects such as high cost of use and non-specificity. There is no entity system specifically involving public opinion monitoring and analysis of minority languages. The multimodal intelligent public opinion analysis and monitoring system proposed in this paper can timely grasp the online public opinion dynamics in various regions including ethnic areas, conduct effective monitoring, thereby reducing the incidence of incidents endangering public safety and maintaining social stability in ethnic areas. It has higher market value and application value.

1.2 Analysis of Current Research Status at Home and Abroad

1.2.1 Current status of research on minority speech recognition (status of machine translation, status of public opinion methods)

China is a multi-ethnic country, and each ethnic group has its own unique language. Researching multilingual and cross-linguistic phenomena in a multi-ethnic context is of great significance to promoting language communication and integration, building cultural harmony, and maintaining national stability. Relying on information technology can greatly improve the quality of government social services, among which speech recognition technology is of great significance as a key link in breaking down language barriers.

1.2.2 Research Status of Sentiment Classification Methods

At present, the research on text sentiment analysis methods can be roughly divided into the following three types: the first is based on sentiment dictionaries; the second is based on machine learning algorithms; and the third is based on deep learning (which is widely used). Therefore, the research progress in text sentiment analysis is analyzed and discussed at home and abroad [1].

Most of the work of constructing sentiment dictionaries is done manually, and it is necessary to constantly integrate new words and combine different specific fields for detailed analysis. Therefore, the technology of automatically constructing sentiment dictionaries has attracted attention. In addition, the commonly used English sentiment knowledge bases include WordNet, GI, etc., and the Chinese ones include HowNet [2] and NTUSD [3] dictionary [4]. With the development of artificial intelligence and machine learning applications, people have begun to use machine learning algorithms to try to analyze the sentiment polarity of long and short texts [5]. Before the rise of deep learning, the method of analyzing sentiment tendency based on machine learning became a major research method. Commonly used machine learning algorithms for this method include: support vector machine, random forest, logistic regression, maximum entropy, K-Means, etc. [6].

1.2.3 Current Status of Research on The Application of Public Opinion Systems in China

Public opinion is the sum of various emotions and opinions related to the interests of individuals and social groups within a certain time and space. At present, the existing public opinion analysis companies in China still have certain technical gaps in information collection and analysis in ethnic minority areas. In many cases, manual classification and organization are still required, which greatly restricts the efficiency of public opinion information processing. Through the application of NLP semantic segmentation, machine learning and other technologies, automated analysis of public opinion data processing can be achieved [7]. In order to solve the public opinion problem, some companies in China have purchased professional public opinion analysis equipment. However, due to the uncontrollability of public opinion, in order to ensure that they can grasp the public opinion dynamics related to them in real time, the network negative public opinion monitoring system, such as the Eagle Eye Speed Reading Network System, can help government and enterprise units to achieve 24-hour monitoring of public opinion related to the entire network and various regions, automatically push relevant information, and intelligently warn of negative sensitive messages. Intelligent investigation, in addition, for the investigation of negative public opinion risks in various regions, modern intelligent system software tools can be used, such as the big data public opinion monitoring and early warning system of Yifang Software [8], which supports all-weather coverage monitoring of public opinion risk points across the entire network [9], automatically extracts and identifies sensitive messages, and helps them to detect risks early and deal with them in a timely manner [10] .

1.2.4 Ethnic Speech Recognition and Translation

The main translation model used is the Transformer XL dual encoder-decoder model. The Transformer model is a classic NLP model launched by Google in 2017 (Bert uses the Transformer). In machine translation tasks, the Transformer outperforms RNN and CNN [11]. It only requires an encoder/decoder to achieve good results and can be efficiently parallelized.

In the future, the Transformer model will focus on the improvement of the self-attention mechanism of the new Transformer model and compare these models. In addition, there are also the latest applications of the model in various fields such as NLP [12], computer vision and reinforcement learning.

1.2.5 Intelligent Public Opinion Analysis Platform

The intelligent public opinion analysis platform is mainly based on the deep learning algorithm - CNN text sentiment recognition, public opinion warning and real-time visualization display. CNN, also known as convolutional neural network, is part of the most influential innovation in the field of computer vision. CNN has achieved very good results in image processing because its convolution and pooling operations can capture local features of the image. Similarly, CNN can also capture local information in text processing. Refer to Yoon Kim's paper on TextCNN

published in 2014. The model structure is shown in Figure 1.

Most of the object recognition algorithms based on deep learning technology are based on CNN and designed with

different structures to achieve the recognition effect. Based on AlexNet, these algorithms mainly develop along the following trends:

(1) Deepening

Depth is one of the core elements of CNN structure. ResNet makes the CNN network structure deeper and easier to converge, further improving the classification accuracy of deep learning algorithms and providing new ideas for CNN structure design.

The network width was adjusted based on ResNet, achieving higher classification accuracy with a shallow and wide network structure.

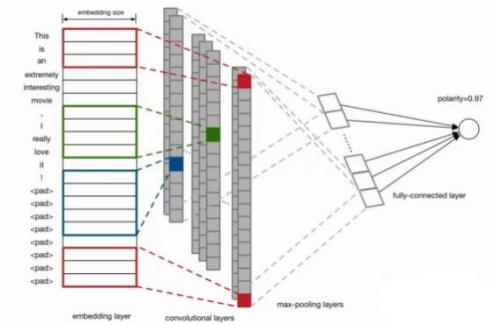

Figure 1. TextCNN Model Structure

(2) Enhanced convolution module functions

By enhancing the functions of the convolution module, CNN can be given stronger perception capabilities. For example, Deformable ConvNets breaks the restrictions of traditional CNN on the shape of the convolution kernel, allowing the convolution kernel to perform affine transformations such as translation, scaling, and rotation, generating flexible and non-fixed convolution kernel shapes, thereby improving the perception capabilities of the convolution kernel.

(3) Design new functional units, loss functions, etc.

The network performance can be improved by designing new functional units or loss functions. After Center Loss, the features extracted by the CNN network are visually reduced in dimension, and it can be found that the intra-class distance is reduced and the model's discriminative power is improved. Fisher Loss also adopts a similar idea, adding the inter-class distance into the loss function to improve the network classification performance.

(4) Radiation detection task

Correct classification is the basis for target detection. The development of deep learning in target recognition has also promoted the progress of target detection, giving rise to a series of new convolutional neural network structures, such as regional convolution.

Neural network (Regions with CNN Features, RCNN), Single Shot MultiBox Detector (SSD) and other algorithms are representative. Based on the RCNN series of networks,

YoLo (You Only Look Once) and single-shot multi-frame detectors were born one after another. Currently, SSD can achieve a detection speed of 58FPS, which can basically meet the real-time requirements.

The detailed development process is shown in Figure 2.

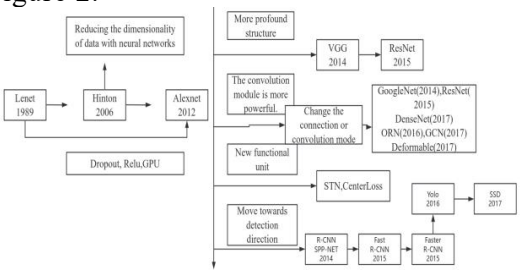

Figure 2. Development Trend of Deep Learning Algorithms Based on Convolutional Neural Networks

1.3 Main Research Contents of This Paper

1.3.1 Multimodal voice information collection

The voice signal is acquired through audio collection, wired monitoring or wireless detection, and the voice input in different modes is classified into three types: Chinese, minority language and text, and processed into Chinese text format and transferred to the server for processing.

1.3.2 Minority speech recognition and translation

In response to the problems of minority speech and text translation in the collection process, and taking into account the characteristics of minority languages, we plan to use deep learning methods to establish a word segmentation model to improve the accuracy of translation and recognition; LSTM and CNN network models will be used for text translation.

1.3.3 Public opinion analysis system design

(1) Text preprocessing and feature extraction

First, the number of texts with different sentiment categories in the original text set is counted. Then, the collected texts are randomly selected to form a training set and a

test set, and the text sentiment labels are added. Finally, the doc2vec program is used to train, extract and regularize the text features.

(2) Feature classification

The labels of the training set are used for CNN network training, and the labels of the test set are used to evaluate the performance. The trained CNN network model and the text features in the test set are used to perform sentiment classification on the text in the test set.

(3) Judgment on the results of public opinion analysis

The sentiment classification and stem extraction results are normalized and divided into three levels: rumors, dangerous speeches, and reactionary speeches, and the probability of meeting these levels is calculated.

1.3.4 Public opinion alert and visualization platform

The visualization platform constantly displays the current monitoring status of the system, such as viewing monitoring, voice monitoring, wireless channel monitoring, and network text monitoring. When public opinion is judged to be reactionary speech, the current time mark is returned, and the corresponding time is marked in combination with the camera monitoring, and the public opinion information is displayed on the visualization platform, including the location, time, and information.

2 Working Principle of Multimodal Intelligent Public Opinion Analysis System

2.1 Multimodal Fusion Method

How to effectively fuse the feature information of different modalities is the key to multimodal sentiment analysis. There are three general methods of multimodal fusion (As shown in Figure 3): feature layer-based fusion method, decision layer-based fusion method, and consistency regression-based fusion method.

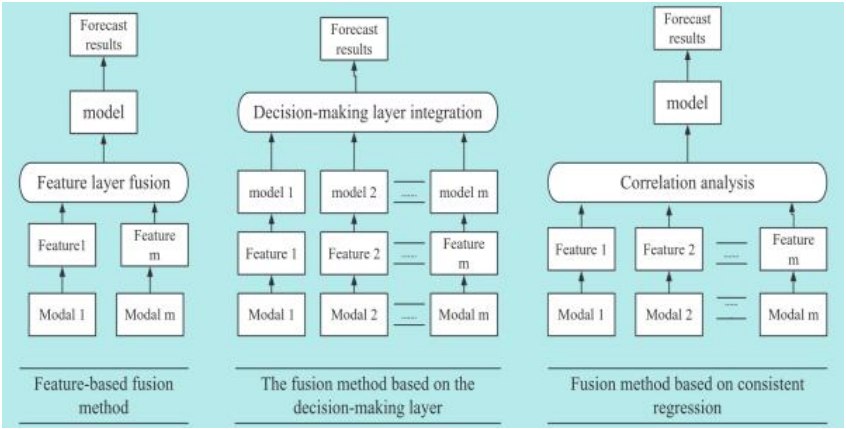

Figure 3. Multimodal Emotion Fusion Method

The fusion method based on the feature layer is also called the early fusion method. By extracting the modal features of various modalities and fusing them at the input, a joint representation is constructed, and sentiment classification is performed on this basis. The implementation of this method is relatively simple. It relies on a general model to learn the sentiment information of different modalities without building a specific model. However, the fusion method based on the feature layer cannot learn the correlation between different modalities and is prone to overfitting of the data. The fusion method based on the decision layer is also called the late fusion method. First, an independent sentiment analysis model is established based on the comment data of different modalities, and then the results of sentiment analysis of different models are fused through a specific decision method. The decision methods include: average [13], majority voting [14], weighted sum [15] or learning model. Since this method has an integrated model, the model parameters are usually relatively small and can adapt well to changes in the number of modalities. However, since an independent sentiment analysis model is established for each modality, the interaction relationship between modalities cannot be established. The fusion method based on consistency regression assumes that the emotions conveyed by different modal information in the comment object are consistent. First, the

feature information of different modalities is extracted, and the correlation weights are learned through the correlation learning algorithm to achieve the fusion of different modalities. Consistency regression fusion can learn the interaction between different modalities well, but this method cannot learn the complementarity between different modal data.

2.2 Evaluation Indicators

For the task of analyzing netizens' emotions in emergencies, four indicators, namely, accuracy, precision, recall, and F1 value, are usually used to evaluate the constructed model. Accuracy can simply and directly reflect the performance of the netizens' emotion analysis model in emergencies. For the situation of unbalanced sample data, accuracy cannot objectively reflect the difference between models. Therefore, in the process of analyzing netizens' emotions in emergencies, more attention is paid to the F1 value. The calculation of accuracy (ACC), precision (P), recall (R), and F1 value is shown in the following formula.

$$P = \frac{TP}{TP+FP} \quad (1)$$

$$R = \frac{TP}{TP+FN} \quad (2)$$

$$Fl = \frac{2*P*R}{P+R} \quad (3)$$

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

Among them, TP is the number of positive comments predicted as positive comments, FP is the number of negative comments predicted as positive comments, TN is the number of positive comments predicted as negative comments, and FN is the number of negative comments predicted as negative comments.

2.3 Deep Neural Networks

A deep neural network [16] consists of multiple layers of fully connected neural networks. Each layer only receives connections from the previous layer and provides connections to the next layer in the hidden layer. Figure 4 describes the architecture of the deep neural network model in detail. The input consists of the connection between the input features and the hidden layer. The input layer can be constructed by TF-IDF, word embedding or other methods. The output layer is the predicted sentiment probability. The deep neural network is a discriminative model that is trained using the back-propagation algorithm.

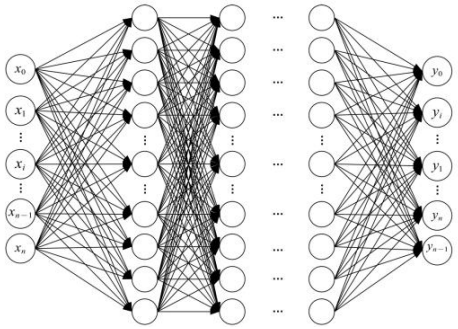

Figure 4. Deep Neural Network Model

For the input feature x, the hidden layer unit is calculated as shown in the following formula.

$$h = \sigma\left(\sum_i w_i x_i + b_i\right) \quad (5)$$

Among them, α is a nonlinear activation function. Commonly used activation functions include Sigmoid function, Tanh function and RELU function:

$$\text{simgoid}\,(a) = \frac{1}{1+e^{-a}} \quad (6)$$

$$\text{Tan}h\,(a) = \frac{e^{2a}-1}{e^{2a}+1} \quad (7)$$

$$\text{RELU}\,(a) = \max(0, a) \quad (8)$$

2.4 Feature Extraction

Feature extraction is an important step in the analysis of netizens' emotions in sudden incidents. The quality of emotion prediction results depends on the quality of extracted features. In the early days, feature extraction methods were mainly based on manual and rule-based methods. This method can only obtain shallow text emotion features and requires a lot of manpower and professional knowledge to formulate rules. In recent years, more and more researchers have found that improving deep learning models can obtain deeper text emotion features, making deep learning models have great potential in sentiment analysis tasks. Text feature extraction methods based on deep learning include: Deep Neural Networks (DNN) [16], RNN [17], LSTM [18], TextCNN [19], attention mechanism [20], etc.

3. Implementation of Multimodal Intelligent Public Opinion Analysis System

3.1 Research Methods

3.1.1 Literature Analysis

This book uses literature analysis to summarize the research results of domestic

and foreign scholars on language translation and public opinion analysis, to understand the relevant concepts and theories, to carry out model research on the basis of clarifying the research context and conducting systematic analysis, and to provide theoretical support for model construction.

3.1.2 Experimental Method

First, the professional knowledge involved in this article is classified into hardware and software parts. The software part includes data acquisition, model building, and model testing. The hardware part includes embedded design. This article will conduct experiments and breakthroughs in each part.

3.1.3 Deep Learning

This paper uses deep learning methods to implement two NLP models: language translation and public opinion analysis. See Figure 5 for details. By building a dual encoder-decoder model of Transformer XL for text translation, and then building a public opinion analysis model that integrates CNN-BiLSTM, the optimal model is selected and the model parameters are determined through validation set verification.

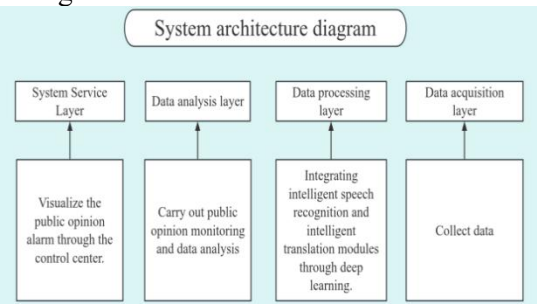

Figure 5. System Architecture Diagram

3.2 Specific Implementation Process

3.2.1 Multimodal Data Collection Platform

(1) Video surveillance data

The system proposed in this paper can be connected to the camera in the monitoring system for real-time monitoring and voice pickup. If it is determined that the voice at a certain moment contains illegal information, the monitoring of that time period will be immediately marked or saved for subsequent processing.

(2) Voice monitoring data

The voice data of this system can come from a camera that can pick up voice or use a microphone alone to monitor the voice of a physical scene. It can also monitor calls or intercept wireless channels as a wireless monitoring method.

(3) Text data

The text data of this system can come from text directly obtained from the Internet, or it can be text obtained after converting voice data into text, which are all used as the text data of the system.

3.2.2 Processing of Minority Languages

(1) Ethnic voice to text conversion

The voice collection part of a specific camera collects voice signals with time information or all voice signals in communication monitoring in real time, and converts voice signals into corresponding ethnic characters by calling iFLYTEK's API for converting minority languages into corresponding ethnic characters.

(2) Text preprocessing

The main work of text preprocessing is to segment the text translated from minority languages into Chinese words and extract semantic stems. In view of the problem of word segmentation ambiguity in the traditional dictionary word segmentation method, this paper intends to adopt a word segmentation method that combines statistics and dictionaries, establish a language model based on the dictionary, count and select the word segmentation result with the highest probability, and then use the word segmentation model based on LSTM EDM to correct the result.

3.2.3 Transformer XL-based Dual Encoder-decoder Translation Model

The encoder and decoder of the translation model are a pair from Transformer XL, the other sub-encoder is a bidirectional LSTM, and the sub-decoder is an LSTM combined with an attention mechanism. At the same time, in order to better capture word order information, the encoder side introduces a complex-valued word embedding method. In view of the fact that some languages do not have parallel corpora, but there are some mutually translated corpora based on ci and ancient poems, a weakly supervised learning method is used to initialize the translation model.

3.2.4 Text Sentiment Recognition Based on CNN

The doc2vec program is used to train, extract and regularize text features. After regularization, the obtained feature vector is input into the CNN network for network

parameter training. Then, the trained model is used to perform sentiment classification on the text to obtain the classification results.

3.2.5 Public Opinion Alert and Visualization Platform

(1) Public opinion alert

If the public opinion analysis part determines that the input information contains illegal information, the public opinion alarm stage will be carried out. The input content can be displayed through the visualization platform, and the corresponding monitoring information can be selected to view the direct source of the data and make the necessary processing.

(2) Real-time display of analysis results

This system contains a visualization platform to display the current monitoring status of the system at all times, such as viewing monitoring, voice monitoring, wireless channel monitoring, and network text monitoring. Information that needs to be alerted is displayed at all times, showing the area, time, and direct content of the acquisition.

The specific process is shown in Figure 6.

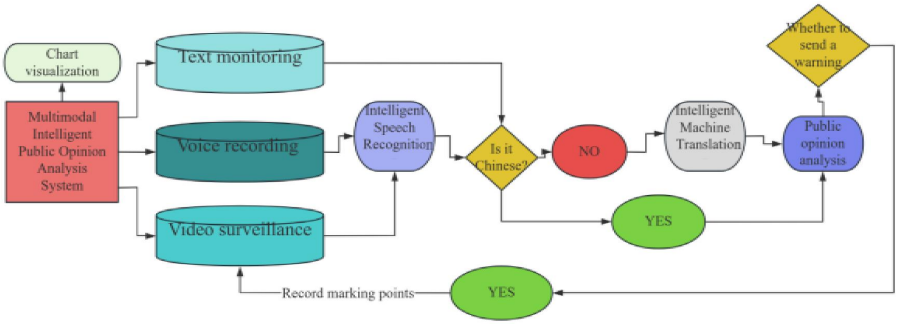

Figure 6. The Implementation Process of the Public Opinion Analysis System

4. Experimental Design and Results Analysis

4.1 Experimental Environment and Dataset

The experiment uses a high-performance computing platform (GPU/multi-core CPU/64GB memory) and PyTorch 1.10 framework, integrates Transformers, Librosa, Jieba and other tools for multimodal processing, and implements distributed training through Horovod. The dataset consists of three parts: 1) 100,000 minority language speech data (five languages including Tibetan/Yi/Uyghur), including sentiment annotation; 2) 200,000 Chinese texts and 50,000 Chinese-Minority bilingual parallel corpora; 3) Multimodal verification set (speech-text-geographic triple data). Preprocessing uses spectral denoising and MFCC feature extraction, uses Wav2Vec 2.0 for speech transcription, and text processing combines BERT-wwm-ext and Transformer word segmentation models, and enhances low-resource language data through back translation and masking strategies.

4.2 Experimental Design and Model Architecture

The experimental system strictly follows the multimodal fusion framework and is divided into four modules: speech recognition and translation, feature fusion, sentiment classification, and visual warning. The speech recognition module uses the Wav2Vec 2.0 model, which is fine-tuned on minority speech data and outputs phoneme sequences; the machine translation module builds a Chinese-Minority bilingual encoder based on Transformer-XL, and achieves cross-language alignment through the attention mechanism. During the training process, the initial learning rate is set to 1e-4, and label smoothing technology is used to alleviate the impact of data noise.

Multimodal fusion strategies are divided into feature layer fusion and decision layer fusion. Feature layer fusion concatenates the MFCC features of speech, the BERT vectors of text, and the translated Chinese word vectors into joint features, which are input into the fully connected layer for classification; decision layer fusion independently trains the TextCNN (text) and BiLSTM (speech) sub-models, and integrates the results through weighted voting, with the weights determined by the validation set grid search. The sentiment classification module adopts an improved TextCNN structure, introduces dilated convolution to enhance the ability to

capture long-distance dependencies, and outputs three-classification probabilities. Public opinion risk judgment is based on threshold rules, combining sentiment classification results with keyword matching (such as "violence" and "division") to determine the level of speech, and triggers a manual review process when multimodal features conflict.

The visual early warning platform is developed based on the Django framework and integrates ECharts to realize real-time display of public opinion heat maps, sentiment distribution and early warning information. It also links with surveillance cameras through GPS coordinates to present a three-dimensional visualization effect of "public opinion-location-time".

4.3 Comparative Experiments and Evaluation

The experimental design covers five sets of comparative tests to fully verify the system performance. The baseline models include TextCNN for text only, Wav2Vec 2.0 for speech only, and Transformer-XL for translation only; the multimodal fusion strategy compares the feature layer, decision layer, and hybrid fusion effects; the cross-language ablation experiment removes the translation module and directly classifies minority texts; the modality loss test simulates low-quality speech or text loss scenarios; and the real-time verification measures the end-to-end processing delay.

The evaluation indicators cover sentiment classification (accuracy, macro-average F1 value), speech recognition (word error rate, sentence error rate), machine translation (BLEU value, translation editing rate) and the overall system (warning accuracy, false alarm rate).

4.4 Experimental Results and Discussion

Experimental results show that the multimodal fusion model significantly outperforms the unimodal baseline in the sentiment classification task. The decision-layer fusion strategy effectively integrates complementary information between modalities through weighted voting, and achieves the best classification performance; the hybrid fusion strategy performs slightly worse due to overfitting caused by parameter redundancy. Cross-language ablation experiments show that the system performance decreases significantly after removing the translation module, verifying the necessity of cross-language alignment; the speech modality can still provide auxiliary information in noisy scenarios and enhance the robustness of the system.

The speech recognition module performs well on low-resource languages, but has a higher error rate in languages with more dialect variants (such as Yi); the machine translation module can effectively retain semantic information and support cross-language public opinion processing. Real-time tests show that the system's end-to-end latency meets actual deployment requirements and GPU resource usage is controllable.

The visual case demonstrates the system’s complete processing flow for minority public opinion: speech recognition and translation generate Chinese text, sentiment classification determines risk level, and the platform links the map to mark hot spots and push warnings.

5. Summary and Outlook

The multimodal fusion strategy significantly improves the accuracy and reliability of public opinion analysis by integrating text, voice and translation information. Decision-layer fusion performs particularly well in complex scenarios (such as emotional disguise), while feature-layer fusion has certain limitations due to insufficient learning of inter-modal correlations. The introduction of the cross-language translation module solves the problem that traditional systems cannot cover multiple languages, but dialect processing still needs to be optimized. The real-time performance of the system can meet most application scenarios, but further lightweight design is needed when deployed on edge devices.

The experiment also revealed some challenges: the recognition error rate of a few dialects is relatively high, and the regional corpus needs to be expanded; the parameter redundancy problem of the hybrid fusion strategy needs to be solved urgently. Overall, the system has demonstrated significant advantages in multimodal fusion, cross-language processing and real-time early warning, providing reliable technical support for public opinion

monitoring in ethnic areas.

The technical innovation of this study lies in the deep integration of speech recognition, machine translation and deep learning models, which fills the technical gap in public opinion monitoring in minority languages. Future work can be carried out from three aspects: multi-language expansion, model lightweighting and cross-modal alignment, such as supporting more minority languages, reducing computing overhead through knowledge distillation, or exploring the joint semantic understanding of speech, text and images. In actual deployment, it is necessary to further collect user feedback and optimize the system's usability and adaptability to cope with the complex and changing public opinion environment.

Acknowledgments

This paper was supported by Student innovation and entrepreneurship training program of Southwest Minzu University (D202411142203590006)

References

[1] Sun Hongkai, Huang Xing. Special topic research: Language identification. Language Strategy Research, 2018, (2):5-5.

[2] Deng L, Yu D. Deep leaning: Methods and applications. Foundations and Sipnal Processing, 2014, 7(3-4):197-387

[3] Liu Z, Zhang L, Tu CC, et al. Statistical and semantic analysis of rumors in Chinese social media. Scientia Sinica Informationis, 2015, 45(12):1536-1546.

[4] Bian T, Xiao X, Xu T, et al. Rumor detection on social media with bi-directional graph convolutional networks//Proceedings of the AAAI Conference on Artificial Intelligence.2020, 34(01):549-556.

[5] Maimaiti Ayifu, Wushouer, Paridan, Yang Wenzhong. Uyghur named entity recognition based on Bi LSTM-CNN-CRF model. Computer Engineering, 2018, 44(08):230-236.

[6] Chiu JPC, Nichols E. Named Entity Recognition with Bidirectional LSTM-CNNs. Computer Science, 2015.

[7] Yadav V, Bethard S. A Survey on Recent Advances in Named Entity Recognition from Deep Learning models//Proceedings of the 27th International Conference on Computational Linguistics. 2018, 2145-2158.

[8] Taboada M, Brooke J, Tofiloski M, et al. Lexicon-Based Methods for Sentiment Analysis. Computational Linguistics, 2011, 37(2):267-307.

[9] Ma Y, Peng H, Khan T, et al. Sentic LSTM: a hybrid network for targeted aspect-based sentiment analysis. Cognitive Computation, 2018, 10(4):639-650.

[10] Hyun D, Park C, Yang M C, et al. Target-aware convolutional neural network for target-level sentiment analysis. Information Sciences, 2019, 491:166-178.

[11] Qin Y, Wang Z, Zhou W, et al. Learning from explanations with neural module execution tree//International Conference on Learning Representations. 2020.

[12] Kuo Zunwang. Research on Several Key Theories and Applications of Network Public Opinion Analysis. Xinjiang University, 2021.

[13] Shutova E, Kiela D, Maillard J. Black holes and white rabbits: Metaphor identification with visual features//Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies. 2016, 160-170.

[14] Morvant E, Habrard A, Ayache S. Majority vote of diverse classifiers for late fusion//Structural, Syntactic, and Statistical Pattern Recognition: Joint IAPR International Workshop, S+ SSPR 2014, Joensuu, Finland, August 20-22, 2014. Proceedings. Springer Berlin Heidelberg, 2014, 153-162.

[15] Evangelopoulos G, Zlatintsi A, Potamianos A, et al. Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. IEEE Transactions on Multimedia, 2013, 15(7):1553-1568

[16] Wadawadagi R, Pagi V. Sentiment analysis with deep neural networks: comparative study and performance assessment. Artificial Intelligence Review, 2020, 53(8):6155-6195.

[17] Dharaniya R, Indumathi J, Uma G V. Automatic scene generation using

sentiment analysis and bidirectional recurrent neural network with multi-head attention. Neural Computing and Applications, 2022, 34(19):16945-16958.

[18]Swathi T, Kasiviswanath N, Rao A A. An optimal deep learning-based LSTM for stock price prediction using twitter sentiment analysis. Applied Intelligence, 2022, 52(12):13675-13688.

[19]Sun K, Shi X, Gao H, et al. Incorporating Pre-trained Transformer Models into TextCNN for Sentiment Analysis on Software Engineering Texts//Proceedings of the 13th Asia-Pacific Symposium on Internetware. 2022, 127-136.

[20]Wu H, Zhang Z, Shi S, et al. Phrase dependency relational graph attention network for Aspect-based Sentiment Analysis. Knowledge-Based Systems, 2022, 236:107736.

