Research on Sentiment Analysis Technology Based on Multi-modal Data

Zheyu Qiu
Computer of Science and Technology
Nanjing University of Posts and Telecommunications
Jiangsu, Nanjing, China
1023040814

***Abstract*—This thesis proposes a sentiment analysis model framework based on multimodal data, which includes four components: text processing network, audio vision processing network, feature fusion network, and sentiment analysis network. The BERT model embedded in the knowledge map is used to process the text, and the LSTM network is used to analyze the audio vision. The feature fusion is processed by direct splicing, and several linear change layers and Relu activation functions are used to analyze the emotion. In terms of model testing, multiple indicators of the model were tested, and the results of each indicator were discussed and evaluated, as well as the overall model of the system was tested, using the network built by pytorch, trained on the mosi data set and as a result of the test, the accuracy rate of the two classifications reached 84.91%, the F1_score reached 0.8487, and the accuracy rate of the five classifications reached 51.6%. The experimental results show that the model in this thesis has good performance, and can achieve a certain accuracy rate in the sentiment analysis of multi-modal data, and has a relatively ideal accuracy.**

***Index Terms*—Multimodal, Sentiment Analysis, Knowledge Graph, BERT, LSTM**

I. Introduction

Sentiment analysis is a natural language processing technique used to determine the sentiment, such as positive or negative, of a given content in a text. However, traditional sentiment analysis techniques often only use textual data for analysis, which may ignore other available sources of important information. In recent years, as large amounts of multimodal data (such as text, audio, and video) have become easier to obtain and store, research on sentiment analysis based on multimodal data has become an area of great concern.

Multimodal sentiment analysis technology aims to improve the accuracy of sentiment analysis by integrating text with other types of data. For example, in a video, elements such as a character's expressions, body language, and voice can convey their emotional state very well. This additional information can enhance traditional sentiment analysis that only targets text data and improve its accuracy.

II. Data Preprocessing

In this section, we will process the data and gain a deeper understanding of the dataset we have. Multimodal Sentiment Analysis datasets like MOSI, SIHS, and MOSEI include various types of data, primarily aimed at analyzing and understanding human emotional expression.

A. Dataset Origin

The MOSI (Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos) dataset is a resource for studying sentiment intensity and subjectivity analysis in online opinion videos. The text data in the MOSI dataset is sourced from comments on YouTube videos, covering a wide range of topics and emotions, including positive, negative, and neutral sentiments. The audio data is extracted from YouTube videos, including spoken voices related to the comments. The video data consists of video clips related to the comments, capturing facial expressions, body language, and other visual features of the speakers. The MOSEI dataset is a larger MOSI dataset with more than tens of thousands of multi-modal data. And SIHS is a Chinese multi-modal dataset.

TABLE I
Multimodal dataset

Dataset	Language	Quantity
CH-SIMS	Ch	2281
CMU-MOSI	Eng	2199
CMU-MOSEI	Eng	More than 23500

B. Feature Extraction

For different modalities of data, I use different methods for preliminary processing.

For text data, I use the BERT model for processing. The BERT model is based on the Transformer architecture, which can effectively capture the context of text in both directions and has very good results in processing text content.

For audio data, I use the OpenSmile tool for feature extraction. This is a popular audio feature extraction tool that can generate various feature vectors for different types of audio data such as speech, music, and ambient sounds.

For video data, I use the OpenFace tool for processing. This is an open-source tool for face recognition and facial expression analysis based on deep learning.

After processing, I can get the vector dimensions of the three modes respectively. Text, audio, and video are (457, 39, 768), (457, 400, 33), (457, 55, 709) respectively.

III. Model Building

The large scale of multimodal big data makes sentiment analysis difficult. This paper proposes a multi-modal sentiment

analysis model to address the need to consider multiple data sources simultaneously in practical applications. This model relies on deep learning technology and integrates input from three data sources: text, voice, and video. It can perform certain emotional analysis and judgment on multi-modal data. The model framework is shown in Figure 1 below:

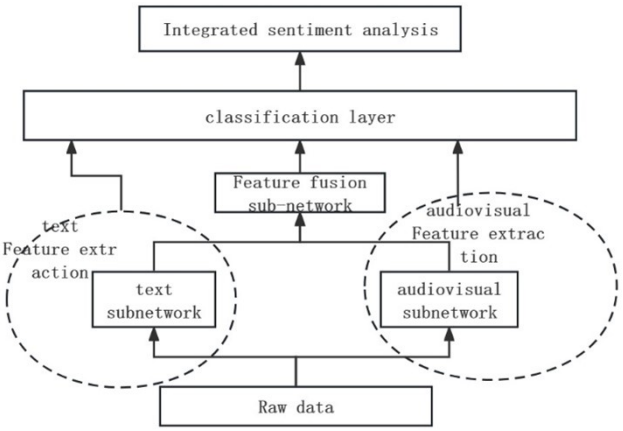

Fig. 1. Model Framework

A. Text Subnetwork

This network is mainly responsible for processing text content and converting text into numerical values for subsequent processing. Since the BERT model is built based on Transformer, it mainly uses the self-attention mechanism:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{1}$$

Convert the input text data into three categories: word vectors, position vectors, and word vectors, and then complete the processing of text data through the trained BERT model.

B. Audiovisual Subnetwork

For the audio and video data that have been processed using OpenSmile and OpenFace, input it into the LSTM network for the next step of processing, and finally get the dimensional vector results of audio and video.

C. Feature Fusion and Emotion Classification

In multimodal deep learning, the fusion layer aims to fuse features from different modalities to produce a more comprehensive and accurate joint representation. Simple concatenation operations are often used to connect features from different modalities. The classification layers in this experiment are composed of three linear transformation layers and two Relu activation functions. The formula of Relu activation function is as follows:

$$f(x) = \max(0, x) \tag{2}$$

IV. Sentiment Classification Result Analysis

This experiment trained and tested the model on the MOSI and SIMS data sets. Both MOSI and SIMS data sets are divided into three parts: Train, Val, and Test, with the ratios of 6:1:3 and 6:2:2. During the model training process, some indicator parameters were tested, including two-class classification accuracy, F1_score score, and five-class classification accuracy.

A. Model Indicator Test

This experiment designed a multi-modal sentiment analysis model. During the training process of the model on the MOSI English dataset, there are some indicators that change with the training process, such as two-class accuracy, five-class accuracy, and F1_score.

1) Secondary Classification Accuracy: The model introduced in this article can be trained to achieve binary classification of emotions, that is, to judge multi-modal data as positive or negative. Accuracy is a very critical metric because it tells us how well the model can correctly predict sentiment tendencies. Therefore, it is very important to evaluate the sentiment analysis capabilities of the model through accuracy.

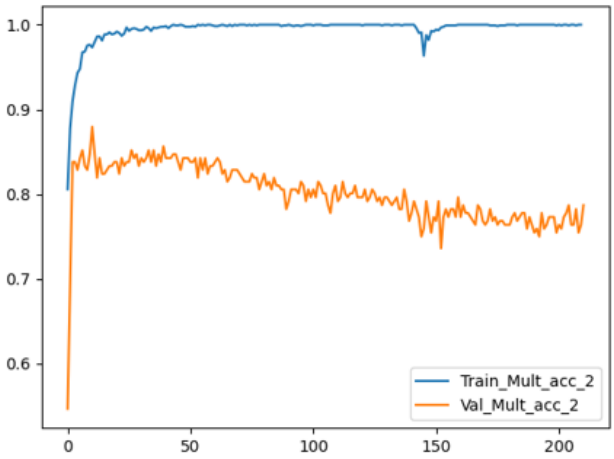

Fig. 2. Secondary Classification Accuracy

2) F1_Score: F1 Score is a widely used indicator to measure the effectiveness of a classification model. It combines the precision and recall of the classifier. The value of F1 Score ranges from 0 to 1. The higher the value, the better the classifier is. Specifically, F1 Score is the harmonic mean of precision and recall, which can be a good measure of the performance of a classifier when dealing with imbalanced data.

3) Five Classification Accuracy: Emotional classification is very difficult because human emotions are very complex, and the same text may trigger different emotional reactions in different contexts. Therefore, for a good five-category emotion model, it needs to have a very powerful feature extraction ability and classification ability. Through continuous improvement and training of the model, the five-category classification accuracy of the model has reached a relatively high level. The

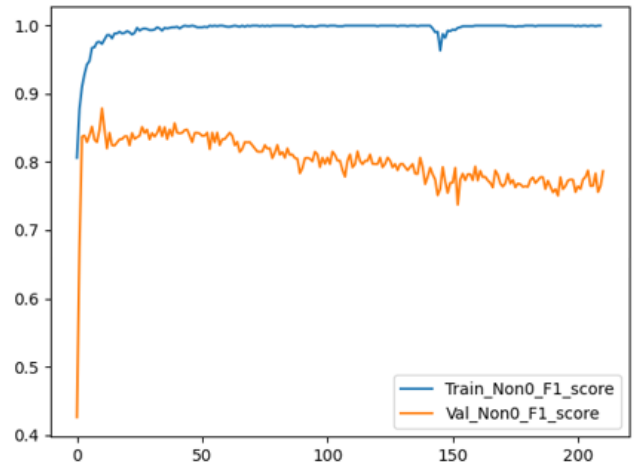

Fig. 3. F1_Score

following is the accuracy of five classifications after model training.

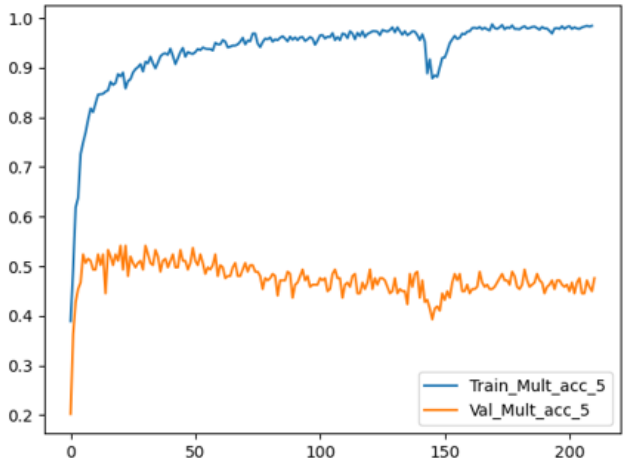

Fig. 4. Five Classification Accuracy

V. Conclusion and Future Work

This paper proposes a multimodal sentiment analysis model framework. The model has shown good performance on the MOSI dataset. However, there are still many areas for improvement in the model. For example, the model can be further optimized in terms of data preprocessing, feature extraction, model design, and parameter tuning. In future research, we will continue to explore how to improve the performance of the model, as well as how to apply the model to more practical application scenarios.

A. Data testing

For the designed model, the model performance was tested after training on the Chinese data set SIMS to verify the stability and reliability of the model. The experiment displays all the measured data visually in tabular form so that we can intuitively observe all data testing conditions. At the same time, the histogram is used to count the number of each emotion for the two-category and five-category categories, and finally compared with the standard labels to obtain the actual accuracy rate.

序号	音频编号	二分类结果	五分类结果	原始文本
1	video_0003$_$0006	消极	弱消极	这事结婚前咱俩不是说好了？
2	video_0010$_$0038	消极	消极	你妈说来磊儿来了影响方一几学习。
3	video_0023$_$0007	消极	消极	你看我行吗？
4	video_0021$_$0005	积极	积极	那也让我见识下你的本事。
5	video_0045$_$0009	消极	中性	以我对你的判断我觉得你应该会先派黄福福
6	video_0027$_$0039	消极	弱消极	镇长，您把心放宽，多保重。
7	video_0019$_$0002	消极	弱消极	无所谓，乱世。
8	video_0024$_$0031	积极	中性	老马，平时在里边儿都爱喝啤啊
9	video_0016$_$0022	积极	弱积极	现在好了，汽水不冰了，可我的心却是冰冰的
10	video_0024$_$0054	消极	弱消极	倒是也行，你是达人，你看着办。
11	video_0051$_$0034	消极	弱消极	娘
12	video_0010$_$0036	积极	积极	你看他发挥多稳定啊，说明他心理素质特别好。
13	video_0016$_$0033	消极	消极	夏涵夏涵，我赢了，夏涵你知道吗，我为了打败你这套连招我练了多少
14	video_0011$_$0004	消极	消极	皇上怕是也没心情见二位小主了。
15	video_0028$_$0023	消极	消极	要赶他还在拘留病房抓他出来。
16	video_0008$_$0011	积极	弱积极	花多少钱我都可以的啦。
17	video_0058$_$0002	消极	消极	你确实还差点意思
18	video_0041$_$0024	积极	中性	赵钱孙李，周吴郑王，冯陈褚卫，我叫孟鹤堂
19	video_0016$_$0034	消极	消极	大春我真城护你，能像个傻逼似的。
20	video_0050$_$0010	消极	弱消极	在我还没有变成老大太之前，留下我一生最美丽的瞬间
21	video_0013$_$0026	消极	弱消极	把本来已经平息了的事件再次推到了风口浪尖上。
22	video_0027$_$0008	消极	弱消极	要不你先搬着啊，老太太我给你照顾两天，等你安顿好了再接过去。
23	video_0015$_$0035	消极	弱消极	就算这本日记是真的，你也不能指证我是凶手。
24	video_0015$_$0028	积极	弱积极	我想我完成一次完美的犯罪。
25	video_0043$_$0010	积极	中性	有很多很多演员确实在他那学习过以后发生了翻天覆地的改变
26	video_0036$_$0004	消极	弱消极	没有光，没有声音

Fig. 5. Data testing

The table is divided into several titles: serial number, audio number, two-category results, five-category results, and original text. A total of 457 pieces of test data are recorded. Part of the interface is shown in Figure 5 above. At the same time, use histograms to draw histograms for the results of two categories and five categories respectively, as shown in Figure 6 and Figure 7 below:

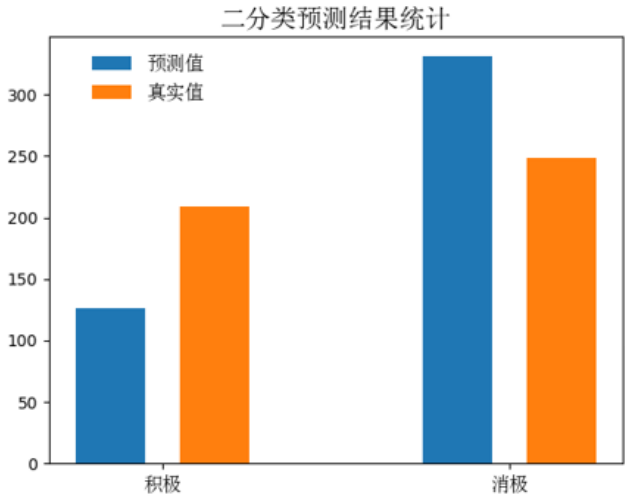

Fig. 6. Two-category histogram

Through a comprehensive analysis of the performance of the model in this article on the Chinese data set SIMS, it can be seen that the model has high two-classification and five-classification capabilities, and can achieve a good accuracy

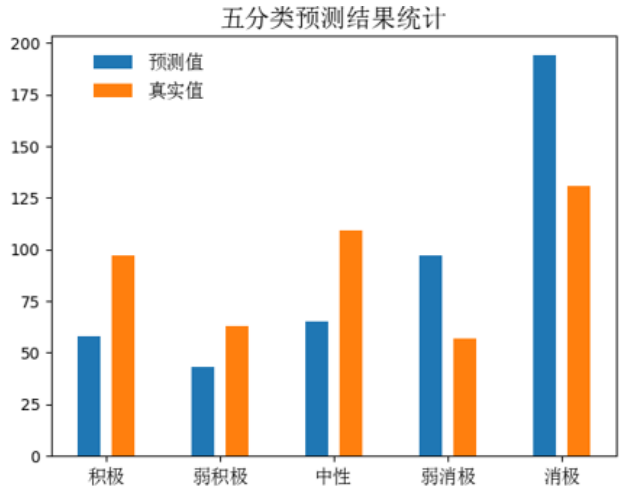

Fig. 7. Five-category bar chart

rate. It shows that the model in this paper has strong prediction and analysis capabilities on multi-modal emotional data and can be put into practical use very well.

VI. DISCUSSION

Multimodal sentiment analysis is a field that studies data involving multiple perceptual modalities (such as text, audio, video, etc.) to understand and infer emotions. The development in this field represents an important progress in the field of sentiment analysis, which has a wide range of application prospects and challenges. The analysis of emotions here in this article only includes negativity, neutrality, and positivity. However, emotions in reality are richer and more changeable. People can be excited and happy, or they can cry with joy; similarly, for sadness, it can be caused by illness. It can come as painful sadness, or it can be the devastation of a breakup between lovers. In reality, people's emotions are very complex. It is very difficult to accurately analyze emotional tendencies, and there is still a long way to go.

VII. CONCLUSION

This article mainly studies multi-modal sentiment analysis technology. By analyzing and evaluating the models proposed by predecessors, we propose our own model and test it. The specific work is as follows:

A. Improvements in Data Preprocessing

Summarized the development process of multi-modal emotion analysis models and established a scientific and reasonable theoretical basis for multi-modal emotion analysis; learned the previous processing methods of text, audio and video and multi-modal fusion. Analytical knowledge.

B. Advanced Feature Extraction Techniques

Summarize and analyze the multi-modal emotion analysis models proposed by previous people. By summarizing the characteristics, advantages and disadvantages of these models, a preliminary study is conducted on the current multi-modal emotion analysis models, and at the same time, the model of this article is proposed. Provides references and guidance;

C. Enhanced Model Architecture

The architecture of the multimodal sentiment analysis model can be enhanced by integrating more advanced neural network layers and attention mechanisms. Additionally, exploring different fusion strategies for combining features from various modalities may yield better results.

D. Application in Real-world Scenarios

Based on the study of previous work, this paper first proposes to use the knowledge graph to embed the text processing network BERT model, and then use OpenSmile to process audio data and OpenFace to process video data. Then after this preprocessing work, the three models are fused, and then the three modal vectors and the fused vectors are analyzed to obtain the results.

E. Integration with Other Technologies

Finally, after designing the multi-modal sentiment analysis model, this article tested the model training process and summarized the three parameters of the two-class classification accuracy, F1-score and five-class classification accuracy during the model training process. Overall testing was conducted on the SIMS data set. At the same time, the performance of the existing models and this paper's model on these three parameters is also compared, so as to conduct a certain evaluation and analysis of the advantages and disadvantages of this paper's model.

Although the research in this article has achieved certain results, there are still deficiencies in some aspects and require further strengthening and improvement:

F. Cross-lingual and Cross-cultural Analysis

Due to the limitations of the knowledge graph, the BERT model embedded in the knowledge graph in this article does not have a certain training direction. If you can make a suitable knowledge graph according to specific needs, text feature extraction will achieve higher efficiency.

G. Cross-lingual and Cross-cultural Analysis

The data set used in this article is a universal standard data set. If there are specific needs, making a data set that matches the needs for model training can achieve better results.

H. Cross-lingual and Cross-cultural Analysis

For some tasks with obvious single-modal emotions, single-modal emotion classification may be better than multi-modal, so single-modal emotion analysis can be performed for such tasks. Therefore, when performing multi-modal fusion, you can choose to judge whether to perform single-modal or multi-modal emotional analysis.

Finally, there is a summary of the course. One semester of big data analysis, from PCA to Mining patterns, from Music Classification to Cluster Analysis, learning various data

processing methods not only gave me a deeper understanding of data results, models, etc., but most importantly What is more, I have a preliminary framework for an idea of data processing, which will undoubtedly be of great help to me in various subsequent deep learning experiments, and will undoubtedly be of great benefit.

REFERENCES

[1] Poria, Soujanya, et al. "Multimodal sentiment analysis: Addressing key issues and setting up the baselines." IEEE Intelligent Systems 33.6 (2018): 17-25.

[2] Zadeh, Amir, et al. "CMU-MOSI: A multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos." arXiv preprint arXiv:1606.06259 (2016).

[3] Devlin, Jacob, et al. "BERT: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[4] Schuller, Björn W., et al. "The INTERSPEECH 2010 paralinguistic challenge." Eleventh Annual Conference of the International Speech Communication Association. 2010.

[5] Baltrušaitis, Tadas, Amir Zadeh, and Louis-Philippe Morency. "Open-Face 2.0: Facial behavior analysis toolkit." 2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018). IEEE, 2018.

[6] Baltrušaitis, Tadas, Chaitanya Ahuja, and Louis-Philippe Morency. "Multimodal machine learning: A survey and taxonomy." IEEE Transactions on Pattern Analysis and Machine Intelligence 41.2 (2018): 423-443.

Latex 代码：

```
\documentclass[conference]{IEEEtran}
\usepackage{graphicx}
\usepackage{amsmath}

\begin{document}

\title{Research on Sentiment Analysis Technology Based on Multi-modal Data}

\author{
    \IEEEauthorblockN{Zheyu Qiu}
    \IEEEauthorblockA{
        \textit{Computer of Science and Technology} \\
        \textit{Nanjing University of Posts and Telecommunications} \\
        Jiangsu, Nanjing, China \\
        1023040814
    }
}

\maketitle

\begin{abstract}
This thesis proposes a sentiment analysis model framework based on multimodal data, which includes four components: text processing network, audio vision processing network, feature fusion network, and sentiment analysis network. The BERT model embedded in the knowledge map is used to process the text, and the LSTM network is used to analyze the audio vision. The feature fusion is processed by direct splicing, and several linear change layers and Relu activation functions are used to analyze the emotion. In terms of model testing, multiple indicators of the model were tested, and the results of each indicator were discussed and evaluated, as well as the overall model of the system was tested, using the network built by pytorch, trained on the mosi data set and as a result of the test, the accuracy rate of the two classifications reached 84.91\%, the F1\_score reached 0.8487, and the accuracy rate of the five classifications reached 51.6\%. The experimental results show that the model in this thesis has good performance, and can achieve a certain accuracy rate in the sentiment analysis of multi-modal data, and has a relatively ideal accuracy.
\end{abstract}

\begin{IEEEkeywords}
Multimodal, Sentiment Analysis, Knowledge Graph, BERT, LSTM
\end{IEEEkeywords}
```

\section{Introduction}
Sentiment analysis is a natural language processing technique used to determine the sentiment, such as positive or negative, of a given content in a text. However, traditional sentiment analysis techniques often only use textual data for analysis, which may ignore other available sources of important information. In recent years, as large amounts of multimodal data (such as text, audio, and video) have become easier to obtain and store, research on sentiment analysis based on multimodal data has become an area of great concern.

Multimodal sentiment analysis technology aims to improve the accuracy of sentiment analysis by integrating text with other types of data. For example, in a video, elements such as a character's expressions, body language, and voice can convey their emotional state very well. This additional information can enhance traditional sentiment analysis that only targets text data and improve its accuracy.

\section{Data Preprocessing}
In this section, we will process the data and gain a deeper understanding of the dataset we have. Multimodal Sentiment Analysis datasets like MOSI, SIHS, and MOSEI include various types of data, primarily aimed at analyzing and understanding human emotional expression.

\subsection{Dataset Origin}
The MOSI (Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos) dataset is a resource for studying sentiment intensity and subjectivity analysis in online opinion videos. The text data in the MOSI dataset is sourced from comments on YouTube videos, covering a wide range of topics and emotions, including positive, negative, and neutral sentiments. The audio data is extracted from YouTube videos, including spoken voices related to the comments. The video data consists of video clips related to the comments, capturing facial expressions, body language, and other visual features of the speakers. The MOSEI dataset is a larger MOSI dataset with more than tens of thousands of multi-modal data. And SIHS is a Chinese multi-modal dataset.

```
\begin{table}[htbp]
    \caption{Multimodal dataset}
    \begin{center}
        \begin{tabular}{|c|c|c|}
            \hline
            Dataset & Language & Quantity \\
            \hline
            CH-SIMS & Ch & 2281 \\
            CMU-MOSI & Eng & 2199 \\
            CMU-MOSEI & Eng & More than 23500 \\
            \hline
        \end{tabular}
    \end{center}
\end{table}
```

\subsection{Feature Extraction}
For different modalities of data, I use different methods for preliminary processing.

For text data, I use the BERT model for processing. The BERT model is based on the Transformer architecture, which can effectively capture the context of text in both directions and has very good results in processing text content.

For audio data, I use the OpenSmile tool for feature extraction. This is a popular audio feature extraction tool that can generate various feature vectors for different types of audio data such as speech, music, and ambient sounds.

For video data, I use the OpenFace tool for processing. This is an open-source tool for face recognition and facial expression analysis based on deep learning.

After processing, I can get the vector dimensions of the three modes respectively. Text, audio, and video are (457, 39, 768), (457, 400, 33), (457, 55, 709) respectively.

\section{Model Building}
The large scale of multimodal big data makes sentiment analysis difficult. This paper proposes a multi-modal sentiment analysis model to address the need to consider multiple data sources simultaneously in practical applications. This model relies on deep learning technology and integrates input from three data sources: text, voice, and video. It can perform certain emotional analysis and judgment on multi-modal data. The model framework is shown in Figure \ref{fig:model_framework} below:

\begin{figure}[htbp]
    \centering
    \includegraphics[width=\linewidth]{model_framework.png}
    \caption{Model Framework}
    \label{fig:model_framework}
\end{figure}

\subsection{Text Subnetwork}
This network is mainly responsible for processing text content and converting text into numerical values for subsequent processing. Since the BERT model is built based on Transformer, it mainly uses the self-attention mechanism:

\begin{equation}
\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)
\end{equation}

Convert the input text data into three categories: word vectors, position vectors, and word vectors, and then complete the processing of text data through the trained BERT model.

\subsection{Audiovisual Subnetwork}
For the audio and video data that have been processed using OpenSmile and OpenFace, input it into the LSTM network for the next step of processing, and finally get the dimensional vector results of audio and video.

\subsection{Feature Fusion and Emotion Classification}
In multimodal deep learning, the fusion layer aims to fuse features from different modalities to produce a more comprehensive and accurate joint representation. Simple concatenation operations are often used to connect features from different modalities. The classification layers in this experiment are composed of three linear transformation layers and two Relu activation functions. The formula of Relu activation function is as follows:

\begin{equation}
f(x) = \max(0, x)
\end{equation}

\section{Sentiment Classification Result Analysis}
This experiment trained and tested the model on the MOSI and SIMS data sets. Both MOSI and SIMS data sets are divided into three parts: Train, Val, and Test, with the ratios of 6:1:3 and 6:2:2. During the model training process, some indicator parameters were tested, including two-class classification accuracy, F1_score score, and five-class classification accuracy.

\subsection{Model Indicator Test}
This experiment designed a multi-modal sentiment analysis model. During the training process of the model on the MOSI English dataset, there are some indicators that change with the training process, such as two-class accuracy, five-class accuracy, and F1_score.

\subsubsection{Secondary Classification Accuracy}
The model introduced in this article can be trained to achieve binary classification of emotions, that is, to judge multi-modal data as positive or negative. Accuracy is a very critical metric because it tells us how well the model can correctly predict sentiment tendencies. Therefore, it is very important to evaluate the sentiment analysis capabilities of the model through accuracy.

\begin{figure}[htbp]
    \centering
    \includegraphics[width=\linewidth]{secondary_classification_accuracy.png}
    \caption{Secondary Classification Accuracy}
    \label{fig:secondary_classification_accuracy}
\end{figure}

\subsubsection{F1_Score}

F1 Score is a widely used indicator to measure the effectiveness of a classification model. It combines the precision and recall of the classifier. The value of F1 Score ranges from 0 to 1. The higher the value, the better the classifier is. Specifically, F1 Score is the harmonic mean of precision and recall, which can be a good measure of the performance of a classifier when dealing with imbalanced data.

\begin{figure}[htbp]
    \centering
    \includegraphics[width=\linewidth]{f1_score.png}
    \caption{F1_Score}
    \label{fig:f1_score}
\end{figure}

\subsubsection{Five Classification Accuracy}
Emotional classification is very difficult because human emotions are very complex, and the same text may trigger different emotional reactions in different contexts. Therefore, for a good five-category emotion model, it needs to have a very powerful feature extraction ability and classification ability. Through continuous improvement and training of the model, the five-category classification accuracy of the model has reached a relatively high level. The following is the accuracy of five classifications after model training.

\begin{figure}[htbp]
    \centering
    \includegraphics[width=\linewidth]{five_classification_accuracy.png}
    \caption{Five Classification Accuracy}
    \label{fig:five_classification_accuracy}
\end{figure}

\section{Conclusion and Future Work}
This paper proposes a multimodal sentiment analysis model framework. The model has shown good performance on the MOSI dataset. However, there are still many areas for improvement in the model. For example, the model can be further optimized in terms of data preprocessing, feature extraction, model design, and parameter tuning. In future research, we will continue to explore how to improve the performance of the model, as well as how to apply the model to more practical application scenarios.

\subsection{Data testing}
For the designed model, the model performance was tested after training on the Chinese data set SIMS to verify the stability and reliability of the model. The experiment displays all the measured data visually in tabular form so that we can intuitively observe all data testing conditions. At the same time, the histogram is used to count the number of each emotion for the two-category and five-category categories, and finally compared with the standard labels

to obtain the actual accuracy rate.

```
\begin{figure}[htbp]
    \centering
    \includegraphics[width=\linewidth]{data testing.png}
    \caption{Data testing}
    \label{fig:data testing}
\end{figure}
```

\text{}
The table is divided into several titles: serial number, audio number, two-category results, five-category results, and original text. A total of 457 pieces of test data are recorded. Part of the interface is shown in Figure 5 above.
\text{}
At the same time, use histograms to draw histograms for the results of two categories and five categories respectively, as shown in Figure 6 and Figure 7 below:

```
\begin{figure}[htbp]
    \centering
    \includegraphics[width=\linewidth]{Two-category histogram.png}
    \caption{Two-category histogram}
    \label{fig:Two-category histogram}
\end{figure}

\begin{figure}[htbp]
    \centering
    \includegraphics[width=\linewidth]{Five-category bar chart.png}
    \caption{Five-category bar chart}
    \label{fig:Five-category bar chart}
\end{figure}
```

\text{}
Through a comprehensive analysis of the performance of the model in this article on the Chinese data set SIMS, it can be seen that the model has high two-classification and five-classification capabilities, and can achieve a good accuracy rate. It shows that the model in this paper has strong prediction and analysis capabilities on multi-modal emotional data and can be put into practical use very well.

\section{DISCUSSION}
Multimodal sentiment analysis is a field that studies data involving multiple perceptual modalities (such as text, audio, video, etc.) to understand and infer emotions. The development in this field represents an important progress in the field of sentiment analysis, which has a wide range of application prospects and challenges.
The analysis of emotions here in this article only includes negativity, neutrality, and positivity. However, emotions in reality are richer and more changeable. People can be excited and happy, or they can cry with joy; similarly, for sadness, it can be caused by illness. It can come as painful sadness, or it can be the devastation of a breakup between lovers. In reality, people's emotions are very complex. It is very difficult to accurately analyze emotional tendencies, and there is still a long way to go.

\section{CONCLUSION}
This article mainly studies multi-modal sentiment analysis technology. By analyzing and evaluating the models proposed by predecessors, we propose our own model and test it. The specific work is as follows:
\subsection{Improvements in Data Preprocessing}
Summarized the development process of multi-modal emotion analysis models and established a scientific and reasonable theoretical basis for multi-modal emotion analysis; learned the previous processing methods of text, audio and video and multi-modal fusion. Analytical knowledge.

\subsection{Advanced Feature Extraction Techniques}
Summarize and analyze the multi-modal emotion analysis models proposed by previous people. By summarizing the characteristics, advantages and disadvantages of these models, a preliminary study is conducted on the current multi-modal emotion analysis models, and at the same time, the model of this article is proposed. Provides references and guidance;

\subsection{Enhanced Model Architecture}
The architecture of the multimodal sentiment analysis model can be enhanced by integrating more advanced neural network layers and attention mechanisms. Additionally, exploring different fusion strategies for combining features from various modalities may yield better results.

\subsection{Application in Real-world Scenarios}
Based on the study of previous work, this paper first proposes to use the knowledge graph to embed the text processing network BERT model, and then use OpenSmile to process audio data and OpenFace to process video data. Then after this preprocessing work, the three models are fused, and then the three modal vectors and the fused vectors are analyzed to obtain the results.

\subsection{Integration with Other Technologies}
Finally, after designing the multi-modal sentiment analysis model, this article tested the model training process and summarized the three parameters of the two-class classification accuracy,

F1-score and five-class classification accuracy during the model training process. Overall testing was conducted on the SIMS data set. At the same time, the performance of the existing models and this paper's model on these three parameters is also compared, so as to conduct a certain evaluation and analysis of the advantages and disadvantages of this paper's model.

\text{}
Although the research in this article has achieved certain results, there are still deficiencies in some aspects and require further strengthening and improvement:

\subsection{Cross-lingual and Cross-cultural Analysis}
Due to the limitations of the knowledge graph, the BERT model embedded in the knowledge graph in this article does not have a certain training direction. If you can make a suitable knowledge graph according to specific needs, text feature extraction will achieve higher efficiency.

\subsection{Cross-lingual and Cross-cultural Analysis}
The data set used in this article is a universal standard data set. If there are specific needs, making a data set that matches the needs for model training can achieve better results.

\subsection{Cross-lingual and Cross-cultural Analysis}
For some tasks with obvious single-modal emotions, single-modal emotion classification may be better than multi-modal, so single-modal emotion analysis can be performed for such tasks. Therefore, when performing multi-modal fusion, you can choose to judge whether to perform single-modal or multi-modal emotional analysis.

\text{}
Finally, there is a summary of the course. One semester of big data analysis, from PCA to Mining patterns, from Music Classification to Cluster Analysis, learning various data processing methods not only gave me a deeper understanding of data results, models, etc., but most importantly What is more, I have a preliminary framework for an idea of data processing, which will undoubtedly be of great help to me in various subsequent deep learning experiments, and will undoubtedly be of great benefit.

\begin{thebibliography}{00}
\bibitem{b1} Poria, Soujanya, et al. "Multimodal sentiment analysis: Addressing key issues and setting up the baselines." IEEE Intelligent Systems 33.6 (2018): 17-25.
\bibitem{b2} Zadeh, Amir, et al. "CMU-MOSI: A multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos." arXiv preprint arXiv:1606.06259 (2016).
\bibitem{b3} Devlin, Jacob, et al. "BERT: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).
\bibitem{b4} Schuller, Björn W., et al. "The INTERSPEECH 2010 paralinguistic challenge." Eleventh Annual Conference of the International Speech Communication Association. 2010.

\bibitem{b5} Baltrušaitis, Tadas, Amir Zadeh, and Louis-Philippe Morency. "OpenFace 2.0: Facial behavior analysis toolkit." 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). IEEE, 2018.
\bibitem{b6} Baltrušaitis, Tadas, Chaitanya Ahuja, and Louis-Philippe Morency. "Multimodal machine learning: A survey and taxonomy." IEEE Transactions on Pattern Analysis and Machine Intelligence 41.2 (2018): 423-443.
\end{thebibliography}

\end{document}