university of groningen / campus fryslân

Chinese Multi-Model Sarcasm Detection

Based on Contrastive Attention Residual Late Fusion

Zhengkun Mei

university of groningen / campus fryslân

Chinese Multi-Model Sarcasm Detection Based on Contrastive Attention Residual Late Fusion

Master's Thesis

To fulfill the requirements for the degree of
Master of Science in Voice Technology
at University of Groningen under the supervision of
Assoc. Prof. M. Coler (Voice Technology, University of Groningen)
and
X. Gao (Voice Technology, University of Groningen)

Zhengkun Mei (s5075580)

ABSTRACT

Sarcasm detection has become increasingly crucial in the era of virtual assistants and the widespread use of sarcasm on the internet. While previous research has focused on text, the significance of other modalities, such as audio and image, has gained prominence. However, the detection of sarcasm in the Chinese language remains limited to text-based approaches due to the lack of multimodal datasets. This paper addresses this research gap by introducing a novel Chinese multimodal sarcasm dataset that combines text and audio modalities. To effectively capture the rich information from different modalities, I propose a late fusion contrastive attention residual model. This model leverages convolutional fusion layers to fuse raw and high-dimensional features from multiple modalities, and the two modalities' information is organically combined through the mechanism of contrastive attention, facilitating the detection of sarcasm. The experimental results demonstrate the effectiveness of the proposed approach, showcasing improved performance in Chinese sarcasm detection. The detail result you can check in my paper. This research and newly created Chinese multi-model dataset contributes to advancing multimodal sarcasm detection techniques and lays the foundation for future studies in this domain. The full dataset is publicly available for use and academic research, and the Github link is https://github.com/ZhengkunMei/Chinese-multimodel-sarcasm-dataset

Key Words: Sarcasm detection, Multimodalities, Late fusion, Chinese sarcasm dataset, Contrastive attention

CONTENTS

1

INTRODUCTION

In the digital age, where social interactions are increasingly taking place on online platforms, the accurate interpretation of nuanced linguistic expressions is important. Sarcasm has emerged as a particularly challenging aspect of natural language understanding. Sarcasm is a linguistic phenomenon in which individuals employ metaphors and exaggerations to expose and criticize individuals or social events, mostly employing positive words to convey their underlying negative sentiments (Potamias et al., 2020). For example:

"我告诉你,我是有尊严的人知道吗?" (I'm telling you, I'm a man of dignity, you know?)

"那是, 您这尊严都在脸上呢." (That's right, your so-called 'dignity' is merely superficially displayed on your face.)

In this dialogue, the second speaker, "B" responds to the first speaker's statement of being a person with dignity by affirming that their dignity is, in fact, displayed on face. This can be interpreted as a way of mocking or dismissing the first speaker's assertion. It implies that the first speaker's claim of having dignity is evident only superficially, without genuine substance or credibility.

Previous work mainly focused on meaning as analysed through text (Joshi et al., 2017). Sarcasm texts often contain special characters or emoticons that distinguish them from ordinary texts (Carvalho et al., 2009), and have unique organisational rules, where sarcasm can be identified through syntactic and semantic features (Suhaimin et al., 2017). However, sarcasm is also expressed through different modalities. Acoustic features as another modality contain a wealth of information relevant to sarcasm detection. Sarcasm has a overall reductions in mean f0, decreases in f0 variation (standard deviation), and changes in HNR (Cheang & Pell, 2008). Rakov and Rosenberg (2013) find that acoustic features like mean f0, standard deviation of f0, f0 range, mean amplitude, amplitude range, speech rate, harmonics-to-noise ratio (HNR), and one-third octave spectral values (as a measure of nasality) can be used to detect sarcasm in audio.

Accordingly, a multimodal approach which takes into account different modalities alongside text holds significant promise. Compared to the single text modality, multimodel cues could improve sarcasm detection (Gu et al., 2018). Yuanyuan and Jing (2022) find that combining data from multiple modalities such as text or images contains more information than using only single modality data. Castro et al. (2019) create a new multimodel dataset using audio features extracted by librosa. The inclusion of audio-visual multi-modalities dataset reduce sarcasm detection relative error rate by 12.9%.

Multimodal approaches have demonstrated potential in enhancing sarcasm detection capabilities. However, the advancement of Chinese sarcasm recognition is impeded by the dearth of appropriate multimodal databases. Presently, Chinese sarcasm detection predominantly relies on text-based recognition due to the absence of a Chinese multimodal sarcasm database. To bridge this research gap, this paper aims to address the issue by curating a Chinese multimodal sarcasm dataset that encompasses both text and audio modalities. Additionally, a novel contrastive attention residual late fusion model is proposed for sarcasm detection on this dataset. This paper creates a Chinese multimodal sarcasm dataset containing 1500 utterances and 1500 context based on both text and audio modalities by collecting recordings from deyunshe. The contrastive attention residual late fusion model uses the idea of residuals to continuously fuse original features at the back end as well as high-dimensional features after fusion by several convolutional fusion layers. Two modalities are fused based on the contrastive attention mechanism to balance two modalities' information. This model is experimented on this newly created dataset and the results show that the proposed multimodal architecture outperforms existing methods that use single modality or use early fusion method.

Accordingly, this work is organised as follows: The paper is divided into four parts. After the introduction follows a comprehensive literature review, which examines the sarcasm based on traditional rule-based and lexicon-based method (2.1.1), sarcasm recognition based on machine learning and deep learning (2.1.2),sarcasm recognition based on multimodality (2.2) and research question and hypothesis generated from the gap between existing Chinese sarcasm detection methods and the state-of-art (2.3). The subsequent focuses on the methodology, encompassing the description of the dataset establishment, including its basic structure, innovative aspects, and specific details regarding the model structure and experimental design. In chapter 4, the results are presented, comparing the performance of baseline models(model built on single modality and model using early fusion) with the contrastive attention residual late fusion model. Additionally, I conducted an ablation study to explore the impact of different components on multi-modal sarcasm recognition. By gradually adding different modalities, introducing residuals and convolutional fusion layers, applying the contrastive attention mechanism and incorporating speaker and context information, I investigated the effects of each part on sarcasm recognition. These findings were then compared with previous research results. Comparison and discussion of the modeling structure leads to the limitations of the study as well as directions for future research. Finally, in chapter 5, I provide a conclusion and an outlook for future research directions.

2

LITERATURE REVIEW

This chapter provides a comprehensive literature review on sarcasm detection, focusing on both single-modality and multi-modal approaches. The chapter is organized as follows: Section 2.1 discusses various approaches to sarcasm detection, including rule-based, lexicon-based, machine learning-based, and deep learning-based methods. Within this section, subsection 2.1.1 delves into the rule- and lexicon-based approaches, exploring how specific indicators and patterns are used to identify sarcasm. Subsection 2.1.2 delves into machine learning and deep learning approaches, highlighting the use of traditional algorithms as well as CNN and RNN models for improved performance. Section 2.2 explores the realm of multi-modal sarcasm detection, considering the limitations of text-based approaches and the importance of integrating multiple modalities. Subsections examine the early fusion and late fusion methods, respectively, showcasing their benefits and drawbacks. Section 2.3 emphasizes the need for multi-modal approaches in Chinese sarcasm detection and proposes research question and hypothesis based on the gap between existing method in Chinese sarcasm and the contrastive attention residual late fusion model. By structuring the literature review in this way, I provide a comprehensive analysis of existing research and set the foundation for my proposed multi-modal approach in the subsequent sections.

2.1. TEXT-BASED SARCASM DETECTION

The task of sarcasm recognition was first regarded as a text classification task. For text classification tasks, there are four main categories of approaches for detecting sarcasm: rule-based, lexicon-based, machine learning-based, and deep learning-based (Aboobaker et al., 2020).

2.1.1. RULE- AND LEXICON-BASED SARCASM DETECTION

The rule-based approach to sarcasm detection relies on identifying specific indicators or patterns in sentences that are associated with sarcasm. These indicators can be based on various properties of the language used, such as the semantic, syntactic, pragmatic and

lexical of the text. In the lexical property, different lexical features are considered in classifying sarcasm. Suhaimin et al. (2017) broke the corpus into words and symbols, including punctuation and hashtags. Due to noisy social media data containing spelling errors and non-standard words, Malay and English dictionaries were utilized to correct misspelled words. Stopword removal was also performed to eliminate meaningless words. Subsequently they extracted the lexical features in n-gram form, excluding single characters like 'n', 't', and 'b'. All tokens were then converted to lowercase for consistency, which can be used as a common feature in NLP. In syntactic property, the syntactic information is mainly concerned with the formation and structure of a sentence or the rules how a sentence is organized. Suhaimin et al. (2017) tagged translated corpus with POS, which consists of 36 different tags representing various grammatical categories. Each POS tag was mapped into one of the four correspondence groups(NOUN, VERB, ADJECTIVE, ADVERB) and all other words that did not belong to these groups were removed. The authors used the word-tag pair to represent the syntactic feature. Semantic property contains information of a language's meaning and the single word's meaning too. A same single word can have different meaning based on context. The semantic method is one of the commonly using rule-based approaches due to its efficiency in identifying sarcasm (Ilavarasan et al., 2020). Bharti et al. (2015) used a semantic-based approach for their research work. If a negative phrase appears inside a positive sentence, the sentence is judged to be sarcasm. Pragmatic features encompass symbolic or figurative elements found in text, such as smilies, emoticons, and replies. These elements serve as expressive cues to convey emotions, attitudes, or additional contextual information within the text (Bharti et al., 2015). Kreuz and Roberts (1995) first introduced the concept of pragmatic features. Rajadesingan et al. (2015) employed a systematic approach to detect sarcasm in tweets and utilized psychological and behavioral pragmatic features of users along with their past and present tweets for analysis.

For the lexicon-based approach, the fundamental idea behind the lexicon-based approach is to utilize various words expressing opinions to indicate different sentiments. A lexicon, which refers to either a set of words or a dictionary of a language, is the foundation of this approach. The lexicon-based approach can be categorized into two primary types: the dictionary-based approach and the corpus-based approach. The lexicon-based approach is useful in identifying opinion words that are contextual and dependent on syntactic patterns. Researchers, such as Riloff et al. (2013), have utilized a corpus-based approach to identify sarcasm by analyzing instances where positive sentiment contradicts with the situation. For example, "absolutely adore it when my bus is late" positive word "adore" followed by mostly negative word "late". This inconsistency can reflect sarcastic utterances, and semantic-based databases can realize sarcasm recognition. The dictionary-based approach involves the manual compilation of opinion words. This set of opinion words can be expanded by incorporating their synonyms and antonyms, which are context-dependent.

2.1.2. Machine learning and deep learning approach

Traditional machine learning algorithms were mainly used initially to detect sarcasm in text modality, including support vector machine algorithm (Vapnik & Chervonenkis, 1964), naive bayes algorithm (Wang & Manning, 2012), K-Nearest neighbor algorithm

(Cover, 1953), decision tree algorithm (Breiman, 2017) etc. The machine learning based approach to sarcasm detection can be divided into two main steps, feature engineering and classification algorithm selection (Gong, 2020). Typical features include lexical features, TF-IDF features, etc. González-Ibánez et al. (2011) employed unigrams and dictionary-based features as the lexical features, and three pragmatic features: i) positive emoticons such as smileys, ii) negative emoticons such as frowning faces, and iii) ToUser, which indicates whether a tweet is a reply to another tweet as input into SVM and logical regression to detect sarcasm.

However, the machine learning algorithms ignore the relationship between words and sentences, and the processing and generalization capabilities of high-dimensional data are relatively poor. In order to solve this problem, deep learning method(CNN, RNN, LSTM) was introduced in the text field. While CNN models were originally developed for computer vision tasks, their effectiveness has been demonstrated in natural language processing (NLP) as well. In fact, they have achieved impressive results in various NLP tasks, including semantic parsing as shown by Yih et al. (2014). This highlights the versatility and success of CNN in handling not only visual data but also textual information. Kim (2014) introduced CNN into the text classification task to obtain more complex high-dimensional data, improve the performance of the model in text classification tasks, and have higher accuracy than traditional machine learning algorithms. Compared with CNN, which is good at extracting features with invariant positions, the RNN model and LSTM (Baziotis et al., 2018) is better at modeling in text order. Subsequent work is based on CNN, RNN, and LSTM network to improve. Aiming at the structural characteristics of satirical sentences, Jain et al. (2020) created a dataset of 30,000 tweets contains 12,000 satirical text and 18,000 non-satirical Hindi-English mixed language text. They proposed a multi-level memory network based on a hybrid model (applying LSTM and CNN to text data) to capture the contrast of sentences at different memory levels to detect sarcasm. Ghosh and Veale (2016) collected the dataset for sarcasm detection from Twitter by using the hashtag #sarcasm as an initial retrieval cue and crawling the Twittersphere. To expand the dataset, they used an LSA-based approach to include additional indicative hashtags. They also harvested tweets from user profiles with a bias towards sincerity or sarcasm. The training dataset consisted of 39,000 tweets, equally balanced between sarcastic and non-sarcastic data. They created a test set of 2,000 annotated tweets. Additionally, they considered elements like hashtags, profile references, and emoticons to extract contextual information. They used the Stanford constituency parser for parsing tweets and applied pre-processing and post-processing techniques to extract relevant information from the parse tree. They proposed a network structure consisting of an LSTM layer followed by a fully connected DNN layer. The LSTM layer processes the input data and produces a sequence of outputs. These outputs are then fed into the DNN layer, which generates a higher order feature set based on the LSTM output, making it easily separable for the desired number of classes. Finally, a softmax layer is added on top of the DNN layer to produce the final classification probabilities. However, detecting single modalities alone is insufficient for effectively solving sarcasm detection in real-life situations due to the absence of other modalities' information. Therefore, the future direction of research lies in developing multi-modal detection methods.

2.2. MULTI-MODEL SARCASM DETECTION

Due to the limitations of text-based single-modality sarcasm detection, which cannot leverage the features of different modalities, and the lack of multi-modal datasets in the field of Chinese sarcasm detection, it has become increasingly important to incorporate multi-modal approaches to integrate both text and speech information to enhance the performance of Chinese sarcasm detection. Multimodal sarcasm detection refers to the use of computer techniques to identify satirical language expressions in multiple modalities (e.g. text, audio, etc.). Sarcasm is a special form of utterance which can express sentiment. Sentiment involves subjective experiences, physiological responses, and behavioral responses. Multimodal recognition is to identify and predict sentiment through these physiological responses and behavioral responses. In 1997, Duc et al. (1997) first proposed the concept of "Multi-modal", they improved person authentication by combining results from multiple biometric modalities(speech audio data and face images). Multimodal fusion methods for sarcasm detection mainly include early fusion (Williams et al., 2018) and late fusion (Cambria et al., 2018).

Early fusion directly extract the features from single modalities and pass them to the classifier for classification. Schifanella et al. (2016) curated a dataset by gathering images and texts from three social platforms. They employed SVM to pioneer the exploration of multi-modal sarcasm detection. Castro et al. (2019) introduced a novel dataset MUStARD, consisting of sarcastic and non-sarcastic videos drawn from different sources with text, image and audio three modalities. They also did a simple experiment based on their new dataset. Text features were extracted using BERT and the authors take the last four transformer layers of the first token ([CLS]) in the utterance and average them to get a unique utterance representation. Audio features were extracted using Librosa to get the MFCC, melspectrogram and spectral centroid features and visual features are extracted for each frame in the utterance video using a pre-trained ResNet-152 image classification model. Then three modalities features were early fused and fed into a rather simple SVM model. The results were shown to significantly outperform their unimodal counterparts, with relative error rate reductions of up to 12.9%. However, this method overlooks the inconsistencies between the different modes. Direct cascading in different semantic spaces leads to information loss and fails to effectively address the issues of information redundancy and complementarity (Ding et al., 2022).

While late fusion combines the results of multiple classifiers trained on various modalities (Cambria et al., 2017). This step effectively resolves the inconsistency between the various modes. However, some useful information will inevitably be lost in the respective fusion process (Ding et al., 2022). Cai et al. (2019) created a dataset for sarcasm detection that includes both text and graphics. They also introduced attribute features into the task of recognizing sarcasm and developed a multimodal hierarchical fusion model (MHFM) to analyze the data. Ding et al. (2022) concatenate the internal features of the subsequent fusion with the original one to comprehensively consider the collaborative fusion of the original modal semantic space and the unified semantic space. This late-fusion model, which was almost like a residual neural network, not only effectively fusesd the different modalities on high-dimensional features, but also avoided the problem of overfitting by fusing them with the previous features and was able to utilize the information from the original features. Their results shows that the multi-modality has

a 4.85% improvement over the single-modality, and the Error rate reduction has an improvement of 11.8%.

2.3. RESEARCH QUESTION AND HYPOTHESIS

Compared to the multi-model sarcasm detection based on different multi-modalities English dataset, Chinese sarcasm detection still mainly relied on single-modality(text-modality). Chinese sarcasm dataset is mostly based on Chinese social media platforms such as weibo (Gong, 2020). Tang and Chen (2014) collected 950 sarcasm texts from Weibo based on the use of emoticons, linguistic features and emotional polarity. The linguistic structure of irony and the elements of irony were also explored and summarised. Liu et al. (2014) constructed three unbalanced datasets on sarcasm, with data from Sina Weibo, Tencent Weibo and Netease Forum. The Sina Weibo dataset contained 238 satirical texts and 3621 non-satirical texts, while the Tencent Weibo dataset contained 359 satirical texts and 5128 non-satirical texts. The authors proposed an integrated multi-strategy learning approach to solve the data imbalance problem. Based on their work, Sun et al. (2016) added 1030 satirical corpus and non-sarcastic corpus from Sina Weibo and blogs to experiment convolutional neural networks and LSTM sequential neural network models for the task of sarcasm recognition.

Due to the lack of a multimodal sarcasm database for Chinese, especially the lack of use of sarcastic audios, sarcasm detection in Chinese is limited to text-based recognition, ignoring the important role of audio in sarcasm recognition and the improvement of sarcasm recognition with modality fusion. Whether a late-fusion model using multiple modalities can utilize information from different modalities to improve the performance of Chinese sarcasm recognition has become my research question. Therefore this paper intends to make a Chinese multi-model sarcasm dataset with audio and text modality by collecting sarcasm and non-sarcasm from deyunshe's recordings. Based on Ding et al. (2022) work, and contrastive attention mechanism proposed by Zhang et al. (2021), this paper proposed a multimodal late fusion model based on the idea of residual and contrastive attention mechanism. The model extracts the features of the respective modalities through the subnet and uses the contrastive attention mechanism to fuse them at the back-end through a convolutional fusion layer, with each fusion adding the original features to exploit the residual information. The model will be experimented on my newly created multimodal Chinese sarcasm dataset and compared with the experimental results of other models on this dataset to demonstrate the improvement of the model for Chinese sarcasm recognition. In this thesis, it is hypothesized that this late-fusion model using the skip-connection residual mechanism(Ding et al., 2022) and contrastive attention mechanism(Zhang et al., 2021) can organically combine the information of the two modalities and be applied to the field of Chinese sarcasm recognition to improve the performance of recognition.

3

METHODOLOGY

In this chapter, I present the methodology used to answer 'How a contrastive attention residual late fusion model improve sarcasm detection performance based on Chinese corpus' and to test the hypothesis that the late fusion model based on the residuals ,convolutional fusion layers and contrastive attention mechanism which effectively combines text and audio modalities could outperform other models using single-modality and simple early fusion method. To those ends, the chapter is organized as follows: Firstly, I discuss the process of dataset collection and feature extraction, including the requirements for dataset selection, the source chosen, the dataset structure, and the reasons why specific features are used in the experiments. Next, I introduce the experimental setup, including the baseline models employed for comparison, such as Majority, Random, SVM, Fully-connection network, LSTM, Late-fusion, and Residual Convolutional Late-fusion. I describe the structure and fusion processes of the contrastive attention late-fusion model in detail. Finally, I outline the experimental procedures, including the different recognition approaches considered, the evaluation measures used, and the cross- validation methodology employed. This organization provides a clear framework for understanding the dataset creation, feature extraction, and experimental design aspects of the research.

3.1. DATASET AND FEATURE EXTRACTION

For this part, I will introduce the requirements of collection, the source I selected, the structure of this dataset and which audio features will be used in experiment.

DATA COLLECTION

To build a Chinese multi-modalities sarcasm dataset, the first step is to select proper sarcasm source. Considering openness of the data, efficiency of creation, and amount of information contained, dataset collection need to meet following requirements:

1. The data source must be open and easy to collect. The data should be accessible for academic use and could be collected through public platform.

2. The data should contain at least two modalities. Previous Chinese sarcasm recognition was focused on the textual domain, and data was mostly collected from text communication platforms such as Weibo, where the sarcasm exists only in the form of text. In creating multi-modalities sarcasm dataset, the sarcasm source should have different forms.

3. The data should have a relatively high proportion of sarcasm within the whole corpus. Sarcasm often exhibits aggression and is context-dependent, making it less prevalent in general context (Wallace et al., 2014). When selecting the source of dataset, it should be a context prone to sarcasm.

To meet the requirements of multi-modalities sarcasm dataset, I collected sarcasm and non-sarcasm from the recordings of Deyunshe and manually transcribed them. Deyunshe is one of the very few established Chinese crosstalk companies in China dedicated to revitalizing the crosstalk art form (Hou & Lim, 2021). The crosstalk is a comprehensive performing art consisting of many forms of humor, such as homophone, hyperbole, sarcasm and so on (Huang et al., 2022).

The database is mainly Deyunshe's voice data obtained through the Himaraya app. The app claims that users are not allowed to use the platform for commercial purposes without the written permission of Himalaya, signing a separate agreement or specifying specific commercial services available to users; No any part or all of the voice on the platform is commercially exploited, reproduced, copied, sold, surveyed, or advertised. However, since the voice information collected by this platform is completely open, there are no restrictions on non-commercial downloading and reprinting. In addition, Deyunshe itself has always opened the recordings of its programs to the outside world and can be downloaded and used freely. Under the following circumstances, a work may be used without the permission of the copyright owner, and no remuneration shall be paid to it, but the name of the author and the title of the work shall be specified, and other rights enjoyed by the copyright owner in accordance with this Law shall not be violated: (1) Using other people's published works for personal study, research or appreciation; (2) For school classroom teaching or scientific research, translate or reproduce a small amount of published works for the use of teaching or scientific research personnel. Therefore, for collecting and using Deyunshe's voice on Himaraya app, I only need to mention the ownership of deyunshe's voice.

As the primary purpose of crosstalk is to entertain the audience and evoke laughter. Unlike normal speech, crosstalk heavily relies on comedic elements, including sarcasm, puns, exaggeration, and clever wordplay. As a result, the proportion of sarcasm and other comedic devices is much higher in crosstalk performances compared to typical everyday conversations or formal speeches.

Crosstalk performances are typically recorded in the form of audio or video. This characteristic makes crosstalk data multimodal, as it contains modalities beyond just text. AI models can benefit from this aspect by utilizing and extracting features from different modalities.

Crosstalk perfectly aligns with these three requirements, as it exhibits multimodal features, contains a wealth of sarcastic expressions, is easily clippable, and can be sourced openly and legally. This provides an excellent foundation for creating a multimodal sar-

casm database.

3.1.1. DATASET STRUCTURE

The dataset follows the data structure of MUStARD (Castro et al., 2019). MUStARD is specifically compiled from well-known television shows such as Friends and The Big Bang Theory, containing audiovisual utterances that have been meticulously labeled with sarcasm annotations. Each target utterance within this dataset is accompanied by preceding dialogues and speaker labels, serving as contextual information that plays an important role in comprehending the underlying sarcastic remarks. The key elements are shown in Table 3.1.

Key	**Value**
utterance	The text of the target utterance to classify.
speaker	Speaker of the target utterance.
context	List of utterances (in chronological order) preceding the target utterance.
context_speakers	Respective speakers of the context utterances.
sarcasm	Binary label for sarcasm tag.

Table 3.1: MUStARD elements

Following the json files' structure of MUStARD, the Deyunshe crosstalk shows were carefully analyzed scene by scene. Sarcasm and non-sarcasm segments were meticulously intercepted from the audio files, and corresponding audio clips were exported. These segments were then transcribed manually based on what was heard, and the speaker of each utterance was recorded. Additionally, preceding sentences before the target utterance were intercepted, manually transcribed, and the speaker information was noted. Speaker and context information are essential in building a multimodal crosstalk sarcasm database. Speaker data helps understand individual styles and influences sarcasm use, while context aids in detecting sarcasm and grasping sarcasm dynamics during performances. In the experiment, context and speaker information will be added gradually to the baseline model. The example of json's format is shown in Figure 3.1.

To ensure a balanced dataset, a total of 1500 discourse items were collected, with approximately half of them containing sarcasm and the other half being non-sarcastic utterances. During the interception of the discourse items, efforts were made to use sentences of comparable length, thereby maintaining relative balance in the timing of

the discourses. The aim was to achieve consistency in the number of sarcasm and non-sarcasm instances, as demonstrated in Table 3.2.

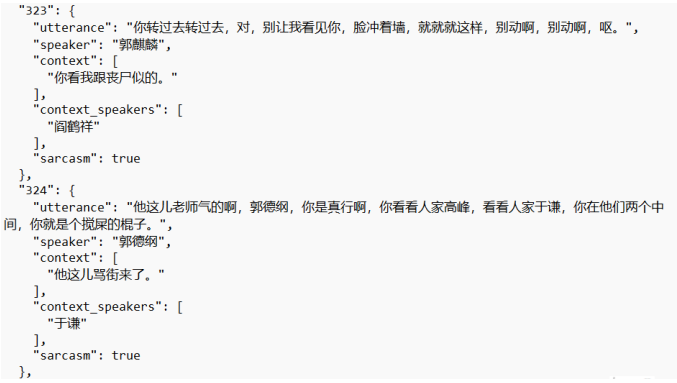

Figure 3.1: JSON format

In order to ensure a relative balance between sarcasm and non-sarcasm, half of the 1500 discourse items collected were sarcasm and half were non-sarcasm, and sentences of comparable length were used in the interception of the discourse items to ensure a relative balance in the timing of the discourse. As Table 3.2 shows, sarcasm and non-sarcasm remain relatively consistent in terms of number and length.

Statistics	**Utterance**	**Context**
Number(sarcasm)	750	750
Number(non-sarcasm)	750	750
Average_duration (sarcasm)	4.9	6.1
Average_duration(non-sarcasm)	5.5	5.8

Table 3.2: Dataset statistics by utterance and context

3.1.2. FEATURE EXTRACTION

As the database contains both speech and text modalities, I will extract the features for both speech and text. The process followed to extract each of them is described below:

Text features: Cui et al. (2019) trained 3-layer RoBERTa-wwm-ext-large model(Chinese BERT with Whole Word Masking) for further accelerating Chinese natural language processing. The text features extraction used this pre-trained Chinese model(chinese rbtl3 pytorch). This pre-trained model improves the bert baseline on different kinds of text

datasets. In tests with the sentiment analysis database ChnSentiCorp, the model achieved an accuracy of 95.8%, an improvement of 0.8% over the bert model. I utilized the RoBERTa-wwm-ext-large model to obtain 1024-dimensional token representations for each utterance. To create a compact and meaningful representation for each utterance, I summed the token vectors and then calculated the average. As a result, each utterance was transformed into a 1024-dimensional utterance representation. This process allowed me to capture the essential information from the tokens while reducing and unifying the overall dimensionality. The averaged vector of this token is more responsive to the characteristics of a sentence as a whole than a single sentence vector, which is more suitable for a sentence representation.

Speech features: Following the work of Castro et al. (2019), I use the Librosa (McFee et al., 2015) to extract low-level features from the audio data stream for each utterance in the dataset. Firstly, the audio files are loaded and the sample rate is set as 22050. Next, I employ a heuristic vocal-extraction method to eliminate background noise from the audio signal. This method isolates the vocal component of the audio and remove any unwanted noise or interference from the recording. Then the Librosa extract audio features from each windows, and the features includes MFCC, melspectrogram, spectral centroid and their associated temporal derivatives (delta).

MFCC: Captures spectral characteristics and energy distribution in short-time frames. Useful for distinguishing sarcasm and non-sarcasm based on acoustic differences.

Mel-spectrogram: Provides a visual representation of the frequency content over time, highlighting regions of higher energy. Helps analyze spectral patterns and acoustic cues in sarcastic and non-sarcastic speech.

Spectral centroid: Characterizes the center of mass of the frequency distribution. Gives insights into pitch and tonal differences.

Temporal derivatives: Represents the rate of change of acoustic features over time. Includes dynamic changes, helpful for distinguishing sarcasm as it involves specific temporal variations in speech delivery.

For different sentence has different length, I took the mean features of window features in one utterance. In order to test which feature combinations are better for the task of sarcasm detection and for fusing with text features, several sets of comparison experiments are performed. The comparison group includes, MFCC only(with 40 dimensions), MFCC and melspectrogram and their delta(336 dimensions) and MFCC, melspectrogram and their delta and spectral centroid information(337 dimensions). These features are fed into the SVM individually or stitched together with text features for comparison experiments. After adjustment of the parameters, I set penalty coefficient to 0.1, 1, 10, 100, 300 and use rbf, linear as the kernal to test results then get 100 as the coefficient and rbf as the kernal which has the best results.

As shown in the Figure 2, when only using audio features for detection, the 336-dimensional features (MFCC+melspectrogram+delta) achieve the highest score in all three indicators. The average score reaches 74.59, which is the highest compared to the other two. When considering the fusion of text features with audio features, the 336-dimensional features still get the best result. When adding the spectral centroid information, the performance of sarcasm recognition witness a clear drop. Therefore, I ex-

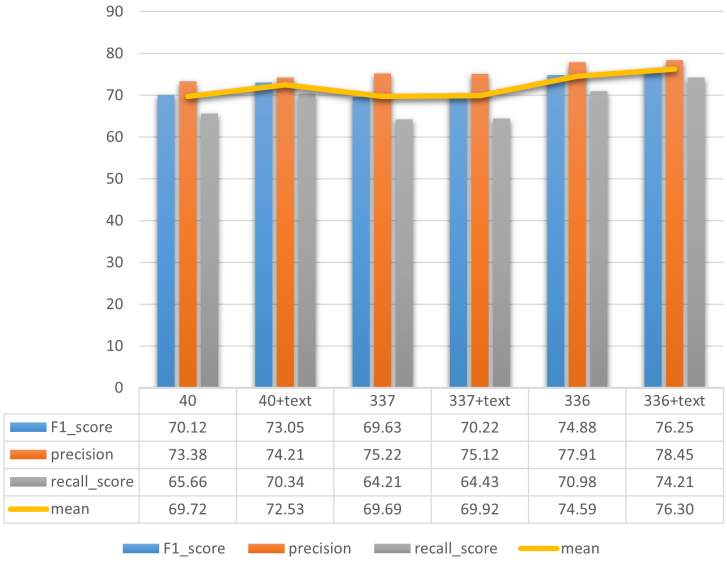

Figure 3.2: Feature selection

tract MFCC, melspectrogram and their delta(336 dimensions) as my experiment audio feature in the end.

3.2. EXPERIMENT

This part introduces the structure of the baseline model used for the newly created multimodal Chinese sarcasm dataset. It also proposes a multimodal contrastive attention residual late fusion model and provides a detailed description of the model along with an explanation of the experimental procedure.

3.2.1. BASELINES

Majority: Following the idea of Castro et al. (2019), all predictions are labeled with the majority of label types. This process only requires a label file and does not involve a feature file.

Random: Following the work of Castro et al. (2019), a list of all possible categories (sarcasm and non-sarcasm) is created, and random predictions are made for each instance, which are then evaluated with the actual labels.

SVM: Following the work of Castro et al. (2019), the speech and text features are separately input into SVM(Pedregosa et al., 2011) for training and evaluation. Then, based on this, the audio and text features are concatenated, early-fused, and trained using the SVM.

Fully-connection network: Similar to SVM, a three-layer linear network is constructed, with each layer activated using ReLU. The text and speech features are fed into the network separately, and the early-fusion features are passed through this network for training.

LSTM: Following the work of Ding et al. (2022), train multimodal features separately and use early fusion features on a three-layer LSTM network.

Late-fusion: Audio and text features are fused at the back-end with high-dimensional features through a three-layer LSTM network respectively.

Residual Convolutional Late-fusion: Following and fine-tuning the work of Ding et al. (2022), the high-dimensional text and audio features extracted by LSTM are fused at the back-end. The original modalities features are fused with the two high-dimensional features at the first fusion layer. The spatial features at the first fusion layer are extracted using CNN and linear layers. The features after the Subnet are then combined with the features after the first fusion layer at the second fusion layer. This process incorporates the concept of residuals by continuously adding previous features to the fusion layer for feature extraction.

Contrastive Attention Residual Late-fusion: Based on the work of Ding et al. (2022), a contrastive attention mechanism is added to balance two modalites's information. The contrastive attention mechanism follows the structure of Zhang et al. (2021). With contrastive attention, the opponent attention weights are learned. Next, the opponent attention weights are applied to another modality to generate the corresponding contrastive vectors for representing the differences in that modality. Through the contrastive attention mechanism, the information of the two modalities no longer exists independently, but is organically fused together to accomplish the task of sarcasm recognition.

3.2.2. Contrastive Attention Residual Late-fusion Model

Figure 3.3 illustrates the fundamental structure of the contrastive attention residual late-fusion model. The model takes text data and audio data as two modal inputs. Librosa (McFee et al., 2015) is used to extract 336 dimensions of audio features, including MFCC, melspectogram, and their associated temporal derivatives. Bert is employed to extract 1024 dimensions of text features.

Both modal features undergo processing through a Subnet layer, as depicted in Figure 3.4(a). The Subnet layer consists of three fully-connected layers, with each layer activated using the ReLU function. By directly extracting the original features of audio and text, the two-layer LSTM is capable of effectively utilizing the original semantic space information of both modalities. The Subnet layer transforms the audio and text modal features into a unified 128-dimensional high-dimensional feature representation.

The high-dimensional features of the two modalities will be organically fused through an contrastive attention mechanism before being fed into the first fusion layer. If the two dimensional features are simply spliced together at the back-end, it is not conducive to utilizing the information generated by the interaction of the two modalities. Therefore, I refer to Zhang et al. (2021) article, in which such an contrastive attention mechanism

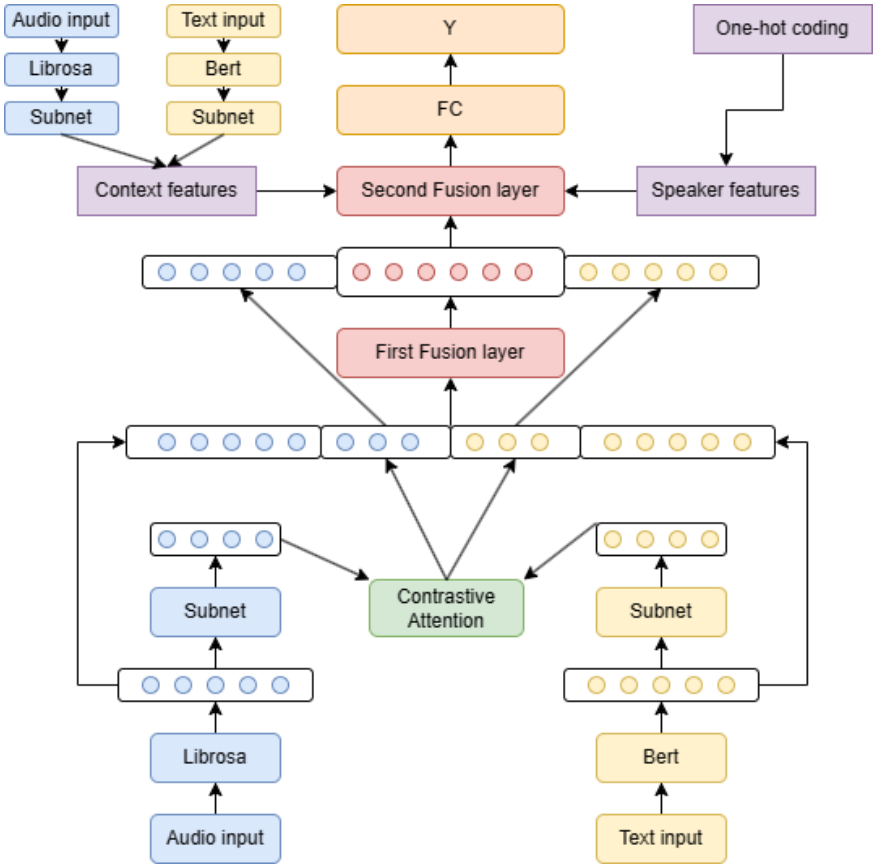

Figure 3.3: Contrastive Attention Residual Late-fusion

is used to deal with the features between cross-modalities. If we consider the features extracted by the audio Subnet as Q and correspond it to a weight matrix K, and d is the dimension of Q, the attention weight matrix for the audio modality (y1) can be represented as:

$$\text{y1} = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)$$

In this case, I didn't use the same method to add attention weights to the text modality. Instead, I applied a method of opponent attention weights, calculating the attention weight matrix for the text modality as:

$$\text{y2} = \text{softmax}(1 - \text{y1})$$

Compared to the traditional attention weight matrix, this method allows both modalities to establish associations and complement each other, enabling one modality to ef-

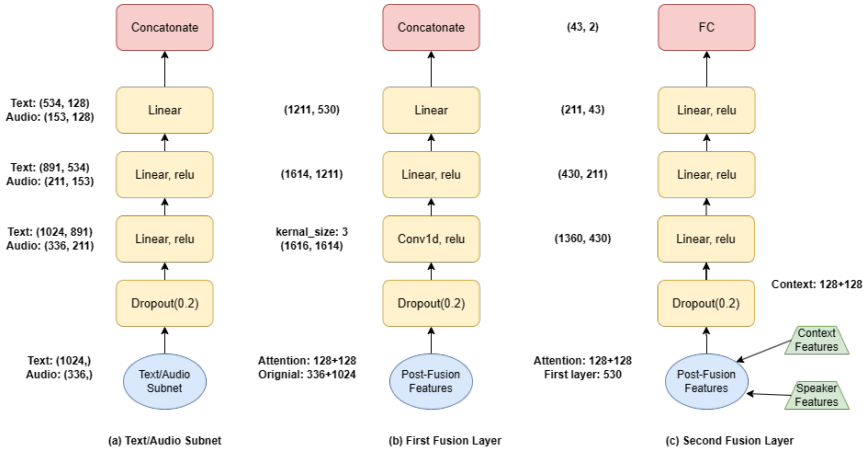

Figure 3.4: Subnet and Fusion layer

fectively utilize information that the other modality has not utilized, thus generating interactions and improving recognition performance. Finally, the high-dimensional features from text and speech are multiplied by their respective associated attention weight matrices to obtain the output under the attention mechanism. The output dimension is consistent with the extracted high-dimensional features.

Then I refer to the structure proposed by Ding et al. (2022) in their model, which adopts a residual-network-like structure and fuses the features of the original 1024-dimensional text features and 336-dimensional audio features in the first fusion layer. As shown in Figure 3.4 (b), unlike Ding et al. (2022)'s structure, I use a Conv1d fusion layer and two fully connected layer with Relu activation in the first fusion layer, to obtain 530-dimensional features in the first fusion layer. This fusion process allows the original features to contribute to the semantic space, while the high-dimensional features obtained from the Subnet capture non-linear semantic information.

Since the Subnet utilizes fully-connected layers for extracting orignial semantic space features, the Conv1d layer in the first-level fusion network focuses on extracting spatial features from the high-dimensional information. After the first fusion layer, the fused features in 530 dimensions are combined with the attention output features caculated by the attention weight matrix by the features extracted by Librosa and Bert (128+128 dimensions), and then fed into the second fusion layer. The second fusion layer consists of three fully connected network layers, each activated by Relu, and transforms the input features into 43-dimensional features. Compared to the orignial structure used by Ding et al. (2022), this design aims to mitigate the overfitting problem caused by an increased number of layers in the model.

To evaluate the impact of context and speaker information on the model, I conduct a comparison experiment where I integrate the context and speaker features with the fusion features and pass them to the second fusion layer. The context features consist of speech features and text features, both extracted using Librosa and Bert, respectively.

The speech features are 336-dimensional, while the text features are 1024-dimensional. These features are then fed into the Subnet to extract high-dimensional features, resulting in two 128-dimensional features in the time domain.

As for the speaker features, they are mainly encoded using one-hot encoding. Since the dataset contains a fixed number of speakers, I count the categories of all speakers, where each category corresponds to one dimension. Therefore, the dimensionality of the speaker features is equal to the total number of speaker categories. This integration of context and speaker information aims to capture more comprehensive and contextualized representations for improved model performance and to test the effect of adding this component compared to the baselines.

3.2.3. EXPERIMENT SETUP

According to the baseline settings, several sets of controlled experiments are conducted. These experiments fall into four main categories based on the number of modalities and the approach to multimodal fusion: unimodal recognition, multimodal early fusion recognition, multimodal late fusion recognition, and fine-tuned multimodal late fusion recognition with the inclusion of context and speaker information.

To begin, experiments are carried out without any features using Majority and Random methods. In the Majority experiment, all predictions are based on the majority of values among the possible prediction types (sarcasm and non-sarcasm), such as all-sarcasm or all-non-sarcasm. In the random experiment, predicted values are randomly generated using the random function.

For unimodal detection and multimodal early fusion sarcasm detection, SVM, fully-connected networks, and LSTM are utilized as baseline tests. The same experimental steps are followed for all three models. Audio and text features are extracted using Librosa and BERT, respectively. The features from both modalities are separately fed into the three models for training, and a simple early fusion of the two modalities is performed by concatenating the features and feeding them into the model as a control group.

In the experiments on multimodal late fusion sarcasm detection, a relatively simple and fundamental late-fusion model is first built, as illustrated in Figure 3.4(a). The extracted text and audio features pass through a Subnet network comprising two three fully connected layesr to extract the semantic space information of the features. The two high-dimensional features are then fused in the late fusion layer and fed into the fully connected layer for training. Subsequently, the late-fusion model proposed by Ding et al. (2022) is adopted, approximating a residual neural network. The first fusion layer is modified to incorporate a combined one-dimensional convolutional layer and several fully connected layers to better capture high-dimensional spatial information. Then following the idea of Zhang et al. (2021), a contrastive attention mechanism is applied to better fuse two modalities' features. Finally, context and speaker information are added to the fused features to assess their impact on sarcasm recognition.

All experiments are evaluated based on precision, recall, and F-Scores. The dataset is evenly divided into five groups using cross-validation, and the average results from each fold are taken as the final indicators.

4

RESULT AND DISCUSSION

In this section, I compare the experimental results of the model proposed in this paper (contrastive attention late-fusion model) with other baseline models (SVM, LSTM, FC, LateFC, ResConv) on the Chinese multimodal sarcasm dataset, as well as conduct detailed ablation experiments to explore the effect of each component on the experimental results by gradually adding different modalities, adding residual convolution fusion layers, adding contrastive attention mechanism and adding context and speaker information. In the end, through the comparison of results and detailed ablation study, I discuss the limitations of this research and propose a future direction.

4.1. COMPARISON WITH BASELINES

Table 4.1 presents the results of speaker and context independent experiments, comparing the contrastive attention residual late-fusion model with other baseline models on the Chinese multimodal sarcasm dataset (created based on Deyunshe's text and audios). The baselines follow the approach of Castro et al. (2019), which initially use two random algorithms, Majority and Random, for sarcasm detection on the Chinese multimodal sarcasm dataset. Both baselines perform poorly, with the Random prediction achieving the lowest performance, with all three indicators hovering around 50%.

Subsequent experiments demonstrate a significant improvement in detection efficiency, whether using unimodal or multimodal features. In the early fusion experimental group, multimodal sarcasm recognition, whether utilizing SVM, LSTM, or fully connected neural networks, proves more effective than either unimodal text or speech recognition. Upon fusing the two modalities, SVM achieves precision, recall, and F-Scores of 78.45%, 74.21%, and 76.25%, respectively, which is 1 to 2 percentage points higher than the unimodal recognition of text or audio. Deep neural networks exhibit a greater capability to capture non-linear relationships between different modal features, as evidenced by the experimental results.

In the early fusion models of the two tested sets of deep neural networks, the LSTM layer outperforms the fully connected layer. This is attributed to the fact that both the

text and speech features of sarcasm exhibit significant temporal characteristics, and the LSTM is better equipped to capture the temporal dynamics of sarcasm compared to the fully connected layer. After fusing the two modal features, the LSTM achieves the best results in the early fusion experiments, with precision, recall, and F-Score reaching 79.47%, 80.53%, and 81.63%, respectively. The three indicators attain their highest values in the early fusion models.

Fusion type	**Modality**	**Algorithm**	**Precision**	**Recall**	**F-Score**
-	-	Majority	50.41	67.03	50.25
-	-	Random	49.49	51.3	53.26
Unimodel	T	SVM	76.14	74.14	74.14
	A	SVM	77.91	70.98	74.88
	T	FC	78.32	77.24	76.19
	A	FC	78.94	75	71.42
-	T	LSTM	74.28	72.47	70.74
-	A	LSTM	78.94	75	71.42
Early	T+A	SVM	78.45	74.21	76.25
	T+A	FC	78	78.78	79.59
	T+A	LSTM	79.47	80.53	81.63
Late	T+A	LateFC	80.13	79.86	79.59
Late	T+A	ResConv	82.13	82.86	84.59
Late	T+A	**AttResConv**	**85.33**	**86.19**	**87.07**
Δearly	-	ΔSVM	$\uparrow$6.88%	$\uparrow$11.98%	$\uparrow$10.82%
		ΔFC	$\uparrow$7.33%	$\uparrow$7.41%	$\uparrow$7.48%
		ΔLSTM	$\uparrow$5.86%	$\uparrow$5.66%	$\uparrow$5.44%
Δlate		ΔLateFC	$\uparrow$5.2%	$\uparrow$6.33%	$\uparrow$7.48%
		ΔResConv	$\uparrow$3.2%	$\uparrow$3.33%	$\uparrow$2.48%

Table 4.1: Comparison with Baselines

Instead of simply concatenating different modal features at the front end, the late fusion approach proves more effective in capturing the inherent distinctions between modalities and improving recognition results. The concept underlying late fusion is to extract high-dimensional features from each modality and fuse them in the back end.

Initially, a relatively straightforward late fusion structure is employed, where the audio and text features are passed through three fully connected layers before being fused in the back end. The experimental results closely align with the best early fusion model using LSTM, exhibiting slightly better precision values while the other two indicators are slightly lower.

The contrastive attention residual late-fusion model, inspired by the structure proposed by Ding et al. (2022) and the attention mechanism propoesed by Zhang et al. (2021) but with improvements in the fusion layer, outperforms all other models in terms of results. It demonstrates significant enhancements in all three indicators, with precision improving by nearly 6% compared to the SVM early fusion model, and recall and F-Score improving by more than 10%. Moreover, this model achieves a 5% improvement over the best early fusion model (LSTM). In comparison to the late fusion model that does not incorporate the idea of residual convolution, the model structure proposed in this paper leads to an overall enhancement of approximately 5% in all three indicators.

4.2. Ablation Study

This part will provide a detailed analysis of the experiment, examining the specific impacts of incorporating various modalities, implementing residual and convolutional fusion layers, adding contrastive attention mechanism and introducing context and speaker information on the experimental outcomes.

4.2.1. Role of multi-modalities

Table 4.2 presents the experimental results of the ablation experiments by incorporating modalities, indicating that multiple modalities enhance the detection of Chinese sarcasm compared to a single modality. In the unimodal experiments, audio and text exhibit similar performance in Chinese sarcasm detection. When utilizing models with SVM and fully connected neural networks, text shows slightly better experimental performance than audio. However, the LSTM model demonstrates significantly superior performance with audio compared to text, owing to the fact that sarcasm, being an emotional form of speech, exhibits richer temporal variations. In the case of audio unimodality, regardless of the model used, precision, recall, and F-Score consistently exceed 70%, with an average index of around 75%. Particularly, under the LSTM model, precision, recall, and F-Score reach 78.61%, 77.46%, and 80.9%, respectively. Precision attains the second-highest value for unimodal detection, while recall and F-Score achieve the highest values, affirming the importance of the audio modality in Chinese sarcasm recognition.

When combining the two modalities for recognition, the recognition results consistently outperform those of either single modality for audio and text. Table 4.2 also provides a detailed overview of the results comparing multimodal and unimodal approaches under the same model. It is evident that, regardless of the model, the multimodal approach consistently achieves superior experimental results compared to the unimodal approach, with the multimodal indicators surpassing the unimodal ones by a margin of 0.5% to 4%. Specifically, under the LSTM early fusion model, after fusing audio features, the multimodal features exhibit an 8.06% improvement in recall and a 10.89%

Modality	**Algorithm**	**Precision**	**Recall**	**F-Score**
T	SVM	76.14	74.14	74.14
A		77.91	70.98	74.88
T+A		78.45	74.21	76.25
Δmulti-uni model		↑2.04% ↑0.54%	↑0.07% ↑3.23%	↑2.11% ↑1.37%
T	FC	78.32	77.24	76.19
A		78.94	75	71.42
T+A		78	78.78	79.59
Δmulti-uni model		↓0.32% ↓0.92%	↑1.54% ↑3.78%	↑3.4% ↑8.17%
T	LSTM	74.28	72.47	70.74
A		78.61	77.46	80.99
T+A		79.47	80.53	81.63
Δmulti-uni model		↑5.19% ↑0.86%	↑8.06% ↑3.07%	↑10.89% ↑0.64%
T+A	Late-fusion	80.13	79.86	79.59
T+A	ResConv	82.13	82.86	84.59
T+A	AttResConv	85.33	86.19	87.07

Table 4.2: Multi-model comparison

improvement in F-Score compared to the single text features. This substantial enhancement greatly contributes to the performance of the model, highlighting the significance of fusing audio features in multimodal representations for Chinese sarcasm recognition.

4.2.2. Role of Resnet Convolution Fusion

Table 4.3 presents the enhancements to the model through the incorporation of residuals and convolutional layers into the late fusion model's first fusion layer. In the initial late fusion model, all three indicators reach approximately 80%, yielding results comparable to those of early fusion using the LSTM neural network. Upon introducing residuals into the network, precision increases by 2.64%, recall by 0.26%, and F-Score by 1.76%, with marginal improvements across all three indicators. Subsequently, the first fusion layer is replaced with a convolutional neural network to extract spatial features on top of the residual late fusion, resulting in significant increases in all three values. Specifically, precision improves by 5.3%, recall by 6.26%, and F-Score by 7.48%. In summary, the incorporation of residuals and convolutional layers in the late fusion model's first fusion layer leads to notable improvements in precision, recall, and F-Score, demonstrating the effectiveness of these enhancements in enhancing the performance of the model.

Algorithm	**Precision**	**Recall**	**F-Score**
FClate	80.13	79.86	79.59
ResFC	82.77	80.12	81.35
ResConv	85.33	86.19	87.07
ΔRes	↑2.64%	↑0.26%	↑1.76%
ΔConv	↑5.3%	↑6.26%	↑7.48%

Table 4.3: skip-connection comparison

4.2.3. Role of Contrastive Attention

As can be seen from Table 4.1, with the introduction of the contrastive attention mechanism, the performance of the new model is again substantially improved over the previous model of residual convolutional model. All three evaluation metrics rose by about three percent, confirming the previous hypothesis that the contrastive attention mechanism is effective in solving the problem of the lack of correlation between different modalities independently of each other, and that by using the opponent weighting matrix, two modalities are correlated with each other and are able to pay more attention to the features that are not present in another modality.

4.2.4. Role of Context and Speaker

Table 4.4 provides a detailed overview of how the context and speaker features were integrated into the contrastive attention residual late fusion model to evaluate their impact on sarcasm detection. Initially, the text and audio features of the context are separately

inputted into the fully-connected Subnet to extract high-dimensional features. These features are then combined together with other features before being passed into the second fusion layer. However, this fusion approach yielded poor performance, with a decrease in precision by 9.98% and in other two evaluation indicators more than 10% compared to the original model that did not incorporate context features.

To address this issue, I hypothesize that the timing of incorporating context features into the fusion layer was incorrect. Considering that the first fusion layer already include a convolutional neural network known to significantly enhance the model's capability, I improve the experiment by fusing the context features from Subnet with other features prior to the first fusion layer. This modification resulted in a notable improvement, as all three indicators surpass the 80% . However, the performance remained noticeably lower than the original model that did not incorporate context features.

Subsequently, I conduct experiments by incorporating speaker information into the model. The results show only marginal changes, with a slight increase in precision and a slight decrease in recall and F-Score compared to the original model.

Finally, I integrate both context features and speaker features into the model, with the context features fused prior to the first fusion layer, as it demonstrated better performance. The final results still fall slightly below those of the original model, with only a marginal advantage of 0.05% in recall, while scoring lower than the original model in the other two indicators.

Algorithm	**Precision**	**Recall**	**F-Score**
AttResConv	85.33	86.19	87.07
+Context(second)	75.35 ↓9.98%	74.04 ↓12.15%	72.78 ↓14.29%
+Context(first)	81.04 ↓4.29%	82.66 ↓3.53%	84.35 ↓2.72%
+Speaker	85.61 ↑0.28%	85.32 ↓0.87%	85.03 ↓2.04%
+Context +Speaker	84.77 ↓0.56%	86.24 ↑0.05%	83.11 ↓3.96%

Table 4.4: Multi-model comparison

4.3. LIMITATIONS AND FUTURE RESEARCH

However, limitations exist in the dataset's scale(only in the crosstalk form), domain specificity, and a rather small size, which may affect generalizability. Moreover, since the sentence vectors are obtained by averaging the word vectors during feature extraction, the special word vectors in the satirical text and the variations embedded in the satirical utterances in the speech lose their significance. Also, when using the contrastive attention mechanism, due to the lack of temporal variation in the input features, it is not possible to supervise the features within the sentence, but only in the weights of the two modalities, which undoubtedly ignores the correlation information within the satirical text and speech.

Future research should focus on expanding the dataset size, diversifying the sources,

and exploring additional modalities such as facial expressions or gesture recognition. In feature extraction, a more subtle approach is needed to preserve the temporal information of speech and text and to include other useful features, such as the pitch of speech, for a better grasp of sarcasm, on the basis of which each temporal feature can be supervised within the sentence using transformer's multi-head attention mechanism. Advanced techniques like deep reinforcement learning or attention mechanisms can be employed on the context features to further enhance the model's performance by adding the context text and audio features. Investigation into cross-lingual sarcasm detection and adapting the model for other languages would contribute to its broader applicability.

5

CONCLUSION

The newly created Chinese multimodal sarcasm database fills the database gap in the field of Chinese sarcasm recognition. The proposed contrastive attention residual late-fusion model demonstrates superior performance compared to baseline models in Chinese multimodal sarcasm detection. The model outperforms unimodal and early fusion models, achieving higher precision, recall, and F-Scores. Incorporating residuals, convolutional layers and contrastive attention mechanism in the late fusion model enhances precision, recall, and F-Score, highlighting the effectiveness of these enhancements. The addition of context and speaker information shows mixed results, with context integration yielding limited improvement and speaker information having minimal impact. The study confirms the importance of multimodal features in Chinese sarcasm recognition, with organic integration of audio features and text features playing a crucial role in capturing temporal dynamics and information from different modalities. The model's performance showcases its potential for practical applications in various domains requiring sarcasm detection. Overall, the new dataset plays a crucial rule in Chinese sarcasm field and the proposed model demonstrates promising results in Chinese multimodal sarcasm detection, highlighting the significance of multimodal fusion, contrastive attention and the potential for further advancements in the field.

BIBLIOGRAPHY

Aboobaker, R. A., Potamias, Siolas, G., Stafylopatis, A.-G., & Jihad. (2020). A survey on sarcasm detection approaches. *Indian Journal of Computer Science and Engineering, 11*, 751–771. https://doi.org/10.21817/indjcse/2020/v11i6/201106048

Baziotis, C., Athanasiou, N., Papalampidi, P., Kolovou, A., Paraskevopoulos, G., Ellinas, N., & Potamianos, A. (2018). NTUA-SLP at semeval-2018 task 3: Tracking ironic tweets using ensembles of word and character level attentive rnns. *CoRR, abs/1804.06659*. http://arxiv.org/abs/1804.06659

Bharti, S. K., Babu, K. S., & Jena, S. K. (2015). Parsing-based sarcasm sentiment recognition in twitter data. *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, 1373–1380.

Breiman, L. (2017). *Classification and regression trees*. Routledge.

Cai, Y., Cai, H., & Wan, X. (2019). Multi-modal sarcasm detection in twitter with hierarchical fusion model. *Proceedings of the 57th annual meeting of the association for computational linguistics*, 2506–2515.

Cambria, E., Das, D., Bandyopadhyay, S., & Feraco, A. (2017). Affective computing and sentiment analysis. *A practical guide to sentiment analysis*, 1–10.

Cambria, E., Hazarika, D., Poria, S., Hussain, A., & Subramanyam, R. (2018). Benchmarking multimodal sentiment analysis. *Computational Linguistics and Intelligent Text Processing: 18th International Conference, CICLing 2017, Budapest, Hungary, April 17–23, 2017, Revised Selected Papers, Part II 18*, 166–179.

Carvalho, P., Sarmento, L., Silva, M. J., & De Oliveira, E. (2009). Clues for detecting irony in user-generated contents: Oh...!! it's" so easy";-. *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, 53–56.

Castro, S., Hazarika, D., Pérez-Rosas, V., Zimmermann, R., Mihalcea, R., & Poria, S. (2019). Towards multimodal sarcasm detection (an _obviously_ perfect paper). *arXiv preprint arXiv:1906.01815*.

Cheang, H. S., & Pell, M. D. (2008). The sound of sarcasm. *Speech communication*, *50*(5), 366–381.

Cover, H. (1953). Cover tm, hart pe. *Nearest neighbor pattern classification, IEEE Trans. Inf. Theory*, *13*(1), 21–27.

Cui, Y., Che, W., Liu, T., Qin, B., Yang, Z., Wang, S., & Hu, G. (2019). Pre-training with whole word masking for chinese bert. *arXiv preprint arXiv:1906.08101*.

Ding, N., Tian, S.-w., & Yu, L. (2022). A multimodal fusion method for sarcasm detection based on late fusion. *Multimedia Tools and Applications*, *81*(6), 8597–8616.

Duc, B., Bigün, E. S., Bigün, J., Maıtre, G., & Fischer, S. (1997). Fusion of audio and video information for multi modal person authentication. *Pattern Recognition Letters*, *18*(9), 835–843.

Ghosh, A., & Veale, T. (2016). Fracking sarcasm using neural network. *Proceedings of the 7th workshop on computational approaches to subjectivity, sentiment and social media analysis*, 161–169.

Gong. (2020). *Study of chinese text irony recognition methods* (Master's thesis). Harbin Institute of Technology.

González-Ibánez, R., Muresan, S., & Wacholder, N. (2011). Identifying sarcasm in twitter: A closer look. *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, 581–586.

Gu, Y., Yang, K., Fu, S., Chen, S., Li, X., & Marsic, I. (2018). Multimodal affective analysis using hierarchical attention strategy with word-level alignment. *Proceedings of the conference. Association for Computational Linguistics. Meeting, 2018*, 2225.

Hou, P., & Lim, B. (2021). Commercialization of traditional performing arts in mainland china: A case study of deyunshe. *Journal of Management, Economics, and Industrial Organization*, *5*(1), 86–99.

Huang, B., Du, S., & Wan, X. (2022). Crossdial: An entertaining dialogue dataset of chinese crosstalk. *arXiv preprint arXiv:2209.01370*.

Ilavarasan, E., et al. (2020). A survey on sarcasm detection and challenges. *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, 1234–1240.

Jain, D., Kumar, A., & Garg, G. (2020). Sarcasm detection in mash-up language using soft-attention based bi-directional lstm and feature-rich cnn. *Applied Soft Computing*, *91*, 106198.

Joshi, A., Bhattacharyya, P., & Carman, M. J. (2017). Automatic sarcasm detection: A survey. *ACM Computing Surveys (CSUR)*, *50*(5), 1–22.

Kim, Y. (2014). Convolutional neural networks for sentence classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1746–1751. https://doi.org/10.3115/v1/D14-1181

Kreuz, R. J., & Roberts, R. M. (1995). Two cues for verbal irony: Hyperbole and the ironic tone of voice. *Metaphor and symbol*, *10*(1), 21–31.

Liu, P., Chen, W., Ou, G., Wang, T., Yang, D., & Lei, K. (2014). Sarcasm detection in social media based on imbalanced classification. *Web-Age Information Management: 15th International Conference, WAIM 2014, Macau, China, June 16-18, 2014. Proceedings 15*, 459–471.

McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., & Nieto, O. (2015). Librosa: Audio and music signal analysis in python. *Proceedings of the 14th python in science conference*, *8*, 18–25.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, *12*, 2825–2830.

Potamias, R. A., Siolas, G., & Stafylopatis, A.-G. (2020). A transformer-based approach to irony and sarcasm detection. *Neural Computing and Applications*, *32*, 17309–17320.

Rajadesingan, A., Zafarani, R., & Liu, H. (2015). Sarcasm detection on twitter: A behavioral modeling approach. *Proceedings of the eighth ACM international conference on web search and data mining*, 97–106.

Rakov, R., & Rosenberg, A. (2013). " sure, i did the right thing": A system for sarcasm detection in speech. *Interspeech*, 842–846.

Riloff, E., Qadir, A., Surve, P., De Silva, L., Gilbert, N., & Huang, R. (2013). Sarcasm as contrast between a positive sentiment and negative situation. *Proceedings of the 2013 conference on empirical methods in natural language processing*, 704–714.

Schifanella, R., De Juan, P., Tetreault, J., & Cao, L. (2016). Detecting sarcasm in multimodal social platforms. *Proceedings of the 24th ACM international conference on Multimedia*, 1136–1145.

Suhaimin, M. S. M., Hijazi, M. H. A., Alfred, R., & Coenen, F. (2017). Natural language processing based features for sarcasm detection: An investigation using bilingual social media texts. *2017 8th International conference on information technology (ICIT)*, 703–709.

Sun, X., He, J., & Ren, F. (2016). Pragmatic analysis of irony based on hybrid neural network model with multi-feature. *Journal of Chinese Information Processing*, *30*(6), 215.

Tang, Y.-j., & Chen, H.-H. (2014). Chinese irony corpus construction and ironic structure analysis. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 1269–1278.

Vapnik, V. N., & Chervonenkis, A. Y. (1964). On a perceptron class. *Automation and Remote Control*, *25*, 112–120.

Wallace, B. C., Kertz, L., Charniak, E., et al. (2014). Humans require context to infer ironic intent (so computers probably do, too). *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 512–516.

Wang, S. I., & Manning, C. D. (2012). Baselines and bigrams: Simple, good sentiment and topic classification. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 90–94.

Williams, J., Kleinegesse, S., Comanescu, R., & Radu, O. (2018). Recognizing emotions in video using multimodal DNN feature fusion. *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*, 11–19. https://doi.org/10.18653/v1/W18-3302

Yih, W.-t., He, X., & Meek, C. (2014). Semantic parsing for single-relation question answering. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 643–648.

Yuanyuan, C., & Jing, M. (2022). Detecting multimodal sarcasm based on sc-attention mechanism. *Data Analysis and Knowledge Discovery*, *6*(9), 40–51.

Zhang, X., Chen, Y., & Li, G. (2021). Multi-modal sarcasm detection based on contrastive attention mechanism. *Natural Language Processing and Chinese Computing: 10th CCF International Conference, NLPCC 2021, Qingdao, China, October 13–17, 2021, Proceedings, Part I 10*, 822–833.