university of groningen

faculty of science and engineering

Multimodal sarcasm recognition based on different feature fusion methods

Youyang Cai

university of groningen

faculty of science
and engineering

University of Groningen

Multimodal sarcasm recognition based on different feature fusion methods

Master's Thesis

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)

Youyang Cai (s5746019)

June 11, 2024

Contents

Acknowledgments

Firstly, I would like to extend my deepest gratitude to Xiyuan and Matt for their invaluable guidance throughout my project. Their expertise in the field and patience with my learning process were crucial in the completion of this work. Xiyuan's detailed feedback on my methodology and Matt's insights during the analysis phase helped refine my approach and significantly enhanced the quality of my research.

Secondly, I am really thankful to the instructors of the MSc Voice Technology program. Their dedication to teaching and their willingness to support us outside of regular class hours have been beneficial. The practical workshops and the one-on-one sessions they offered not only deepened my understanding of complex concepts but also instilled a robust foundation for my professional skills in voice technology.

Lastly, I want to thank my classmates for their collaboration and camaraderie during this journey. Learning alongside such dedicated individuals has been an inspiring experience. Their diverse perspectives challenged me to think more critically and creatively, enriching my learning process. The study groups and late-night discussion sessions were invaluable, providing both support and motivation throughout this challenging yet rewarding program.

Abstract

In the current digital age where virtual assistants are widespread and sarcasm is frequently used online, the detection of sarcasm has become a critical challenge. Traditional sarcasm detection methods have largely focused on textual data, but the emergence of multimodal analysis has introduced new dynamics into the field, reflecting the complex nature of human communication where sarcasm often involves discrepancies across different modalities. For example, a sarcastic remark might be made with a positive tone but accompanied by a sarcastic facial expression.

This thesis introduces the Contrastive-Attention-based (ConAtt) model, designed to enhance sarcasm detection by leveraging a cross-modal contrastive attention mechanism. This model effectively captures and analyzes inconsistencies between modalities—such as textual praise accompanied by a complaining tone—by extracting several contrastive features from the discourse. Experimental validations on the Multimodal Sarcasm Detection Dataset (MUStARD) dataset demonstrate the ConAtt model's effectiveness, marking a significant advancement over traditional sarcasm detection approaches.

The ConAtt model exhibits substantial improvements in key performance metrics including Precision, Recall and F-Score, highlighting the benefits of integrating multimodal data for detecting sarcasm. This study not only emphasizes the importance of multimodal integration but also validates the efficacy of the contrastive attention mechanism in parsing complex and subtle cues across various communication channels. The capabilities of the ConAtt model are pivotal for the accurate identification and interpretation of sarcasm, offering substantial potential for applications that require a nuanced understanding of human interactions.

These findings lay a robust foundation for future research in multimodal sarcasm detection, opening new avenues for exploring more intricate and nuanced forms of communication. This work underscores the growing relevance of advanced multimodal analysis techniques in the broader context of natural language processing and human-computer interaction.

Key Words: Sarcasm detection, Multimodalities, Contrastive attention

1 Introduction

The rapid evolution of instant messaging technologies has significantly transformed the landscape of communication, introducing new platforms like Twitter where individuals share their daily lives and express views using figurative language such as irony and sarcasm. Unlike traditional communication, which predominantly utilized textual content, modern interactions increasingly employ a mix of modalities including text, audio, and video. This multimodal data surge provides a rich reservoir for understanding complex communicative phenomena like sarcasm. Sarcasm involves a complex interplay of sentiments where the expressed sentiment is often the opposite of the literal meaning conveyed by the text. For example, a phrase such as "that's really funny" that is accompanied by a sarcastic tone or laughter may actually indicate sarcasm rather than genuine amusement. This inherent characteristic makes sarcasm detection a particularly challenging domain within natural language processing.

Sarcasm detection, also known as sarcasm recognition, is a traditional subfield of sentiment analysis and natural language processing. Traditionally, this area focused primarily on textual content, employing analysis of special characters, emoticons, and unique syntactic and semantic patterns to identify sarcastic expressions [1] [2]. However, text-based methods have limitations, as sarcasm can manifest in forms that extend beyond mere text, necessitating broader analytical approaches. With technological advancements, multimodal sarcasm detection has emerged as a promising field. This approach integrates text, audio, and visual cues to enhance detection accuracy. Early initiatives, such as those by Morency et al. [3], demonstrated the effectiveness of combining visual, auditory, and textual features for more accurate sarcasm detection. This has been substantiated by subsequent research indicating that acoustic features like pitch variability and speech rate are indicative of sarcasm in audio modes [4]. The inclusion of audiovisual multimodal datasets has reduced the relative error rates in sarcasm detection by 12.9%, highlighting the superiority of multimodal methods over traditional single-text modalities [5].

The main challenge in multimodal sarcasm detection is accurately discerning the underlying intent behind the text, which often contradicts its apparent meaning. This task requires a nuanced understanding of both the content and the context in which it is presented. Sarcasm detection techniques have included rule-based methods, statistical methods utilizing specific features and algorithms, and, with the advent of more complex technology, deep learning methods. Deep learning methods have advanced to account for the nuanced cues embedded across textual, auditory, and visual signals, extract features from different modalities, and establish cross-modal relationships to enhance these features [6] [7]. Despite these advancements, effectively fusing cross-modal features remains a significant challenge. Most current approaches rely on concatenating features from various modalities, assigning equal importance to each. This method may not always accurately reflect true communicative intent. Additionally, effectively capturing modal inconsistencies—which play a pivotal role in detecting sarcasm—remains an underexplored area in multimodal analysis.

Human communication exhibits dual dynamics, such as inter-modal and intra-modal dynamics [8]. Inter-modal dynamics pertain to how different communication modalities interact, such as the combination of textual content and vocal tone in a sarcastic remark. Intra-modal dynamics, on the other hand, focus on variations within a single modality, such as changes in tone or facial expressions. These dynamics are crucial for detecting sarcasm, as they often exploit inconsistencies that can be subtle and highly context-dependent. For example, a sarcastic remark might be delivered with a cheerful tone and a smile, which contradicts the negative sentiment of the spoken words.

To address the intricacies of multimodal sarcasm detection, we propose the ConAtt model. This model uses a contrastive attention fusion strategy that is adept at capturing both intra-modal and inter-modal inconsistencies [9]. By focusing on differences within and between modalities, the ConAtt model effectively improves the detection of sarcasm. The advantage of this model lies in its ability to discern contrasts between linguistic content and accompanying audiovisual signals, which are key factors in sarcasm detection. For example, when the textual components of an utterance appear neutral or positive, but the corresponding audio and visual cues—such as a complaining tone or a somber facial expression—indicate negativity, this discrepancy signals sarcasm. The ConAtt model utilizes a contrastive attention mechanism to target these inter-modal inconsistencies. It generates opposing attention weights for interactions between modalities, effectively highlighting contrasts between them. This approach not only underscores inconsistencies across different modal expressions—such as when audio and visual cues are consistent but textual content differs—but also enhances sarcasm detection in scenarios where traditional monomodal approaches may fail. This focused analysis across modalities enables ConAtt to effectively detect both subtle and overt cues of sarcasm.

The organization of this thesis is as follows: The document is divided into five main sections. The introduction is followed by a literature review of the paper. This review examines traditional rule-based methods for sarcasm detection (Section 2.1.1), machine learning approaches (Section 2.1.2), deep learning approaches (Section 2.1.3), and multimodal sarcasm detection (Section 2.2), which leads to the formulation of research questions and hypotheses. The methodology is detailed in the third section, which includes an overview of the dataset structure, the architecture of the model, and specific details of the experimental design. The fourth section presents the results, comparing the performance of the baseline models with the ConAtt model. Additionally, an ablation study is conducted to explore the impact of different components on multimodal sarcasm recognition. By incrementally adding various modalities and applying the contrastive attention mechanism, the influence of each component on sarcasm detection is examined. These findings are then compared with previous research results. Discussions on the model's structure highlight the limitations of the current study and suggest directions for future research. Finally, the fifth chapter provides conclusions and perspectives on future research directions.

2 Literature Review

The field of sarcasm detection has seen significant transformations, evolving from simple text-based analyses to sophisticated multimodal approaches. Initially, rule-based methods were prominent, relying heavily on textual indicators and semantic inconsistencies to identify sarcasm. However, these approaches often struggled with accuracy and adaptability across different domains. The advent of machine learning brought about a shift towards statistical models, enhancing sarcasm detection capabilities through supervised learning and manual feature extraction. As technology advanced, deep learning methods, particularly neural networks, became pivotal in sarcasm detection, offering the ability to automatically learn complex patterns and handle high-dimensional data. The limitations of text-only analysis led to the integration of multimodal methods, combining textual, visual, and auditory data to capture the nuanced expressions of sarcasm more effectively.

2.1 Text-based Sarcasm Detection

2.1.1 Rule-based Approaches

In the field of sarcasm detection, early methods primarily relied on rule-based approaches, which focused on identifying inconsistencies in sentiment polarity within the text to detect sarcasm. Such methods involved recognizing specific indicators or patterns within sentences that are commonly associated with sarcasm, including the analysis of punctuation, hashtags, and syntactic structures [10] [11] [12]. For instance, Suhaimin et al. decomposed the corpus into words and symbols, utilized Malay and English dictionaries to correct spelling errors, performed stop-word removal to eliminate meaningless words, extracted lexical features in n-gram form, and converted all tokens to lowercase to ensure consistency [13]. In addition to lexical features, syntactic features focused on the construction and structure of sentences were analyzed, where Suhaimin et al. tagged the translated corpus with POS, including 36 different tags representing various grammatical categories, mapped to correspondence groups (NOUN, VERB, ADJECTIVE, ADVERB), with non-conforming words removed to emphasize word-tag pairs representing syntactic features [13].

Semantic properties also play a critical role, as they involve the meaning of language and individual words, which can vary depending on context. Semantic methods are widely used due to their effectiveness in identifying sarcasm, for example, Bharti et al. employed a semantic-based approach where a sentence containing a negative phrase within a positive context was considered sarcastic [14] [15]. Moreover, early research in sentiment analysis often treated the task as an unsupervised learning problem, where researchers developed sentiment dictionaries and judgment rules for specific domains through manual curation. These lexicon-based methods involved data preprocessing, including cleaning and noise reduction, tokenization, and computing sentiment scores and classification results based on the emotional polarity of words [16].

In summary, while these rule-based approaches are intuitive and straightforward, they have significant limitations such as an over-reliance on the construction of dictionaries and rules, which not only affects accuracy but also increases manual labor costs and struggles to adapt to texts from different domains. Furthermore, these methods often fail to effectively handle ambiguous or context-dependent sarcastic expressions, highlighting the need for more sophisticated techniques in sarcasm detection.

2.1.2 Machine Learning Approaches

Supervised learning involves analyzing a large volume of sample data to build classifiers or regressors for predicting new data. Common classification algorithms include Naive Bayes (NB), Support Vector Machine (SVM) [17], and Maximum Entropy (ME). For instance, Pang et al. [18] demonstrated the superiority of machine learning methods over traditional sentiment dictionary approaches. Xin Guo et al. [19] used WordNet for feature extraction and vector representation, building SVM models for sarcasm detection.

However, machine learning-based methods, despite their robust capabilities in detecting sarcasm, typically require manual feature extraction, including lexical polarity, text labels, grammatical structures, and sarcastic expressions. For example, Ptáček et al. [20] collected tweets labeled with "#sarcasm" using the Twitter API to mark sarcastic sentiments and designed manual features based on lexical frequency and sarcasm indicators such as emoticons, punctuation, quotes, and capital letters, evaluated using SVM and ME classifiers. Eke et al. [21] extracted a variety of features from texts, including lexical features, text length, grammatical features, syntactic features, emoticon features, tag features, and semantic features, and employed a two-stage classification process using decision trees, SVM, K-Nearest Neighbors (K-NN), logistic regression trees, and random forests; the first stage used Bag-of-Words and Term Frequency-Inverse Document Frequency (TF-IDF) to process lexical features into a dictionary, and the second stage utilized this dictionary along with other features for sarcasm detection. Choosing the appropriate classifier is crucial, as different classifier combinations can impact the analysis outcomes.

2.1.3 Deep Learning Approaches

With the advent of deep learning technology, neural network-based methods have significantly advanced the field of natural language processing, particularly in sarcasm detection. Techniques such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) [22], and Long Short-Term Memory networks (LSTMs) have become fundamental due to their capability to automatically learn features and patterns from data, enhancing their ability to handle high-dimensional data and understand text sequences. Originally developed for image recognition, CNNs have shown remarkable performance in text processing tasks, effectively extracting local features from input texts using multiple convolutional kernels, which enhances classification performance [23] [24]. Wang et al. [25] introduced a densely connected CNN model that incorporates a multi-scale feature attention mechanism to improve accuracy in classification tasks. On the other hand, RNNs and LSTMs are particularly well-suited for processing sequential data [26], such as text or speech, due to their ability to capture temporal and sequential information which is crucial for understanding language structures. However, RNNs are prone to issues like gradient vanishing or exploding, which can limit their performance and stability. To address these issues, Hochreiter et al. [27] introduced LSTM, which includes design elements such as input gates, forget gates, and output gates to effectively mitigate gradient problems and handle long-term dependencies.

In sarcasm detection, Jain et al. investigated the structural characteristics of sarcastic sentences using a hybrid model combining LSTM and CNN to capture contrasts in sentences at different memory levels [28]. Similarly, Ghosh and Veale [29] employed a method for sarcasm detection through social media by collecting tweets tagged with #sarcasm and expanding the dataset using an LSA-based approach to include additional indicative tags. They created a balanced training dataset and an annotated test set, and developed a deep neural network model. This model integrates contextual cues such as tags, profile references, and emoticons, utilizing the Stanford constituency parser to enhance the extraction of relevant features from the data. The network architecture includes an LSTM layer followed by a fully connected Deep Neural Network (DNN) layer, which utilizes the outputs from the LSTM to generate a high-order feature set that aids in effective classification. These models showcase the effectiveness of deep learning in identifying nuanced expressions of sarcasm across various data forms.

Although single-modality methods are effective in text-based domains, such as processing Twitter data, their limitations become apparent in real-life communication scenarios. In these settings, communication often extends beyond mere text to include tone of voice, facial expressions, and body language. The absence of this multimodal information hinders the effectiveness of traditional sarcasm detection methods. Consequently, there is a growing emphasis in current research on developing multimodal detection methods. These methods aim to integrate data from various modalities—textual, auditory, and visual—to enhance the accuracy and efficiency of sarcasm detection. By capturing the complexity of human communication more fully, multimodal approaches are better able to identify and analyze sarcasm in different contexts.

2.2 Multimodal Sarcasm Detection

Due to the limitations of text-based unimodal sarcasm detection, which cannot utilize features from different modalities, the integration of multimodal methods has become increasingly important. Multimodal sarcasm detection integrates textual, visual, and auditory information to enhance detection performance by capturing the nuanced expressions of sarcasm that unimodal data often miss [30]. This approach leverages the rich, complementary information available across different modalities, helping to overcome the ambiguity that often accompanies text-only analysis.

2.2.1 Multimodal Fusion Methods

The fusion techniques employed in multimodal sarcasm detection typically include early fusion, late fusion, and hybrid fusion. Early fusion, also known as feature fusion, involves integrating features from all modalities at an early stage before any model training occurs. This method can capitalize on the high correlation between modalities to enhance model training efficiency. However, managing high dimensionality and preserving relevant information without introducing redundancy pose significant challenges [31]. On this basis, Busso et al. [32] focused on the interactions between facial features such as lips, eyebrows, and head movements, and vocal features, significantly advancing the development of multimodal feature fusion. Morency et al. [3] utilized Hidden Markov Models to integrate textual, visual, and auditory features for tri-modal emotion analysis. They constructed a YouTube video emotion dataset to support their experimental data.

Late fusion, or decision fusion, involves training separate models for each modality and then combining their outputs towards the final decision-making process. This method is less susceptible to the complexities of direct feature integration and can effectively manage the uncorrelated errors from different modal outputs [33]. Common late fusion techniques include weighted averaging, concatenation, and summation, using advanced models like CNNs and RNNs to integrate results effectively. For instance, Wu et al. [34] proposed a method that extracts prominent emotional features from voice and inputs them to an SVM and Multi-Layer Perceptron (MLP) for localized recognition results, which are then fused using a linear weighting method to enhance performance. Ellis et al. [35] utilized three separate SVM to classify textual, video, and audio features respectively. They then applied a linear weighting strategy for decision fusion, demonstrating the significance of multimodal integration.

Hybrid fusion combines the strengths of both early and late fusion, integrating some features early on and leaving other aspects of decision-making to be combined after individual modal analyses. This method potentially yields a more robust model by leveraging both feature-level and decision-level insights, albeit at the cost of increased model complexity and training challenges [36]. Poria et al. [37] used a Gaussian kernel-based feature fusion for integrating text, visual, and voice features, and designed a decision fusion method based on weighted rules, choosing an Extreme Learning Machine (ELM) as the classifier for sentiment analysis. Rozgic et al. [38] initially utilized overlapping segments of electroencephalogram (EEG) signal sequences. Subsequently, they merged segment-level features to obtain response-level feature vectors. Finally, they employed decision fusion to achieve emotional classification results. Sayedelahl [39] proposed a hybrid fusion approach, combining feature and decision modalities, to enhance the assessment of emotions from spontaneous spoken dialogues.

2.2.2 Constrastive Attention

Advancements in multimodal fusion techniques are continuously addressing the challenges of effectively modeling the dynamics between different modalities. For instance, Morency et al. [3] utilized Hidden Markov Models to integrate text, visual, and audio features for tri-modal emotion analysis, creating a YouTube video emotion dataset to support their experimental data. Similarly, Sebe et al. [40] tackled the challenges of visual and audio emotion recognition by proposing the use of probabilistic graphical models, such as Bayesian networks, to probabilistically merge facial and vocal features, enhancing the accuracy of emotion detection. However, these approaches often focus mainly on the integration of modalities and tend to overlook the inconsistencies between them.

As attention mechanisms become increasingly prevalent in the field of deep learning, researchers are integrating these mechanisms into multimodal feature fusion frameworks more extensively. For example, Bedi et al. [41] developed a Hindi-English mixed code dataset (MaSaC) for detecting sarcasm in conversations, proposing a multimodal hierarchical attention network structure. This structure captures the correlations between modalities, thus effectively uncovering sarcastic sentiments. Extending this approach, Pan et al. [42] proposed a multimodal sarcasm framework based on images, text, and post tags that not only integrates features across modalities but also uses attention mechanisms to simulate inconsistencies between them. Furthermore, Wang et al. [43] considered using a pretrained language model to model multimodal information by concatenating image and text features and inputting them into a BERT model, allowing the model to focus on inconsistencies between modalities using attention mechanisms.

To further enhance the detection of sarcasm, researchers implement the contrastive attention mechanism [44] [45], which is specifically designed to extract inter-modality incongruent information. This innovative approach leverages the discrepancies between different types of data to identify sarcasm more effectively, emphasizing the importance of detecting contrasting cues that are often pivotal in understanding sarcastic expressions.

Despite these advances, as research progresses, the field of multimodal fusion methods continues to evolve significantly. However, the effective integration of multimodal features for accurate modeling remains a primary challenge in sarcasm detection.

2.3 Research Question and Hypothesis

In the rapidly evolving field of digital communication, the integration of diverse modalities such as text, audio, and visual cues has significantly enriched the complexity of interpreting communicative cues, particularly sarcasm [9]. Despite the considerable advances in multimodal sarcasm detection, effectively integrating these modal features to capture nuanced incongruities between different modalities remains a formidable challenge. These incongruities are often pivotal in sarcasm detection, as sarcasm frequently exploits discrepancies, such as a mismatch between the tone of voice and the text content, to convey meanings contrary to the literal interpretation of words [46].

Building on the complexities and challenges identified in the field of multimodal sarcasm detection, we propose the following research question: How can different feature fusion methods utilizing multiple modalities affect the accuracy of sarcasm recognition? Based on the research question exploring how different feature fusion methods affect sarcasm recognition accuracy, this thesis hypothesizes that the ConAtt model, which utilizes a contrastive attention mechanism [47], will surpass early fusion methods in terms of Precision, Recall and F-Score. The ConAtt model is designed to specifically enhance the recognition of sarcasm by highlighting and effectively utilizing modal discrepancies. This approach is expected to provide a more nuanced and accurate detection of sarcasm, particularly by capturing and emphasizing the incongruities between different modalities, such as textual, auditory, and visual cues. Additionally, the hypothesis underscores the role of multimodal data in improving sarcasm detection performance, positing that the integration of multimodal data not only enriches the context available for analysis but also significantly refines the detection capabilities of sarcasm, a complex communicative phenomenon often embedded in subtle cues and contradictions [48].

3 Methodology

In this chapter, we introduce the early fusion method as Experiment 1, which serves as a comparative experiment. Experiment 2 employs the ConAtt model, designed to validate that the contrastive attention strategy enhances model performance by capturing discrepancies between modalities. The organization of this chapter is as follows: initially, we discuss the structure of the dataset and the process of feature extraction. Subsequently, we detail the model architecture, including Experiment 1 for comparison, and provide a comprehensive description of the structure and fusion process of Experiment 2—the ConAtt model. Lastly, we outline the experimental procedures, including the evaluation measures used and the cross-validation methods employed. This structure provides a clear framework for understanding the dataset creation, feature extraction, and experimental design aspects of our research.

3.1 Dataset and Feature Extraction

In this section, we will introduce the MUStARD [47] and the methodology for extracting textual, auditory, and visual features to be utilized in the experiment. Three vectors are used to represent each utterance: a textual feature (T) for linguistic content (text), an acoustic feature (A) for acoustic characteristics (audio), and a visual feature (V) for visual information.

3.1.1 Dataset

We use the MUStARD from GitHub[1] as the benchmark for detecting sarcasm. This dataset consists of video clips from four popular TV shows: Friends, The Big Bang Theory, The Golden Girls, and Sarcasmaholics Anonymous. The dataset includes a total of 690 videos, with an equal distribution of sarcastic and non-sarcastic samples.

The MUStARD dataset is designed to capture sarcasm by incorporating complementary information from multiple modalities—vision, audio, and text, along with speaker identification. Each sample in the dataset includes these three modalities and the identity of the speaker, providing a comprehensive context for sarcasm detection.

The dataset is divided into two main parts: the final utterance and the preceding sentences of the final utterance, which provide the context. The context provides additional information that is crucial for understanding whether a statement is sarcastic or not. Statistical details of the dataset can be found in Table 1 [5].

3.1.2 Audio Preprocessing

The preprocessing steps involve the reduction of background noise from all audio files, followed by resampling to a uniform 22,050 Hz and conversion to a mono channel to ensure consistency in audio quality across the dataset.

[1] `https://github.com/soujanyaporia/MUStARD`

Table 1: Dataset statistics by utterance and context.

Statistics	**Utterance**	**Context**
Unique words	1991	3205
Avg. utterance length (tokens)	14	10
Max. utterance length (tokens)	73	71
Avg. duration (seconds)	5.22	13.95

3.1.3 Feature Extraction

Text Features: Textual feature (T) is generated using a pre-trained BERT model [49], and finally a 768-dimensional ($d_t = 768$) vector representation is extracted for each target sentence. The procedure starts with tokenizing each sentence using the BERT tokenizer, which converts text into a sequence of tokens. To standardize input lengths, shorter sentences are padded with [PAD] tokens, while longer sentences are truncated to the maximum length of 512 tokens. This ensures all sequences are uniformly processed. The core of the feature extraction involves initially computing the average of the hidden states from the last four transformer layers of the first [CLS] token, as these layers typically capture a richer contextual representation of the sentence. However, subsequent empirical findings highlighted in [50] suggest that a more accurate representation is obtained by using the hidden state of the last [CLS] token. Thus, in later stages, this approach is adopted.

Speech Features: In our research, we employ the Librosa library [51] to extract low-level acoustic features from each utterance within our dataset. Feature extraction is carried out by segmenting each audio signal into d_w non-overlapping windows. This segmentation enables the extraction of local features such as Mel-frequency cepstral coefficients (MFCCs), mel-spectrograms, spectral centroids, and their respective temporal derivatives (delta features). Here is an overview of each feature and its relevance to our analysis:

MFCCs: These capture the spectral characteristics and energy distribution within short-time frames, proving essential for distinguishing between sarcastic and non-sarcastic utterances based on acoustic differences.

Mel-spectrogram: This feature offers a visual representation of the frequency content over time, emphasizing regions of higher energy. It is instrumental in analyzing spectral patterns and acoustic cues within both sarcastic and non-sarcastic speech.

Spectral centroid: It characterizes the “center of mass” of the frequency distribution, providing insights into pitch and tonal differences between different types of speech.

Temporal derivatives: Representing the rate of change of acoustic features over time, these derivatives capture dynamic changes in speech delivery, which are crucial for identifying sarcasm due to its unique temporal variations.

To handle the variable lengths of audio sources in our dataset, we compute the average of these acoustic features for each focused utterance, referred to as acoustic feature A, with a dimensionality of 298 ($d_a = 298$), to create a consistent framework for further analysis.

Video Features: To accurately extrate the visual features, we adopt a structured approach using the pre-trained ResNet-152 model, which is based on the ImageNet dataset. Specifically, we focus on the pool5 layer of ResNet-152 [52], recognized for its ability to encapsulate high-level visual descriptors. For each frame in the dataset, designated as the f frame, we use this layer's output as a feature vector. The pool5 layer generates a 2048-dimensional vector, capturing a rich set of visual data points that represent key aspects of the image content effectively. To create a uniform feature set for each utterance, regardless of the number of frames it contains, we compute the average of these vectors across all frames within a focused utterance. This averaging process smooths out anomalies and provides a stable, representative visual feature set. The resultant average vector, denoted as visual feature (V), retains the 2048-dimensional structure ($d_v = 2048$).

3.2 Multimodal Fusion

In our study, we explore multimodal frameworks which are essential for integrating and correlating multiple types of data simultaneously, specifically acoustic, visual, and textual information [53]. These frameworks employ three principal fusion strategies: early, late, and hybrid fusion, each catering to different aspects of data integration and processing.

Early Fusion [54] involves integrating different sources of data into a single, comprehensive feature vector before any classification process. This vector is then fed into a single classifier. In contrast, Late Fusion [55] refers to the aggregation of outputs from multiple classifiers, each trained independently on different modalities. The final decision is typically made by averaging these outputs or through more complex decision-making mechanisms like voting. Hybrid Fusion [56] employs an intermediate shared representation layer where data from various modality-specific sources are jointly learned. This shared layer facilitates a deeper integration of modalities, allowing the model to leverage both inter-modal and intra-modal interactions more effectively.

In the context of our research, we first employ early fusion methods in Experiment 1 to demonstrate how integrating multimodal data can enhance sarcasm detection performance. This establishes a foundational comparative framework for evaluating the effectiveness of more advanced fusion techniques. Subsequently, we focus on the application of the ConAtt model as Experiment 2, which serves as a prime example of the contrastive attention mechanism. The ConAtt model was specifically selected due to its innovative use of contrastive attention mechanisms. These mechanisms significantly heighten the model's sensitivity to the subtle nuances typically found in sarcastic expressions. Experiment 2 aims to validate the hypothesis that the contrastive attention mechanism can improve model performance further by adeptly capturing discrepancies between modalities.

The ensuing sections will detail the experimental setup, the specific architectures employed for each fusion strategy, and the results obtained from our comparative study.

3.2.1 Experiment 1

In Experiment 1, we utilize the early fusion method, also known as feature-level fusion. This method involves combining features from different modalities—namely, textual, visual, and auditory—into a single multimodal feature vector. The primary advantage of this approach is its ability to capture the interrelationships among features from different modalities, which can significantly enhance performance by leveraging these integrated features for more robust predictions.

Our methodological approach begins with the extraction of features specific to each modality: textual(T), visual(V), and acoustic(A) features are obtained using the aforementioned methods 3.1.3. Once these features are extracted, they are concatenated to form a unified feature vector that represents all modalities.

This integrated multimodal feature vector, as depicted in the Feature Fusion phase of Figure 1, is then fed into a classifier for training and prediction. The selection of robust classifiers capable of handling large, complex feature sets is crucial. Suitable classifiers include deep learning models such as MLP, LSTM, and SVM.

MLP: MLP is well-suited for processing high-dimensional feature vectors derived from early fusion. It utilizes its multilayer structure to learn complex patterns within the input data. MLP's effectiveness lies in its ability to integrate and classify various input features through nonlinear activation functions, such as ReLU or Sigmoid, which are pivotal for understanding and interpreting the implicit correlations present in multimodal data.

LSTM: LSTM is particularly applicable to the analysis of multimodal data with temporal dynamics, such as joint textual and auditory cues [26]. LSTM's gating mechanisms help the model process and remember long-term dependencies without losing information over time, which is crucial for understanding the subtle linguistic and tonal variations inherent in sarcasm.

SVM: SVM [57] is a powerful supervised learning algorithm commonly used for both classification and regression tasks. SVMs operate by finding a hyperplane in the feature space that maximizes the margin between different data point categories. This model is particularly well-suited for handling high-dimensional data, effectively performing even when the number of data points is less than the number of features. SVMs enhance their adaptability and flexibility by employing kernel techniques, such as linear, polynomial, and radial basis function kernels, to manage data that is not linearly separable. This capability makes SVMs perform excellently on small datasets and particularly suitable for intricate pattern recognition challenges, such as distinguishing between sarcastic and non-sarcastic expressions.

SVMs were selected as the baseline classifier for Experiment 1 due to their proven effectiveness in efficiently classifying small to medium-sized datasets [57]. In subsequent ablation studies, we will conduct comparative analyses using MLP and LSTM to further evaluate the performance across different models. The detailed structure is shown in Figure 1. Here, V, T, and A represent the visual features, textual features, and acoustic features, respectively. Additionally, the early fusion model's source code is open and publicly available on GitHub[2].

[2] https://github.com/soujanyaporia/MUStARD

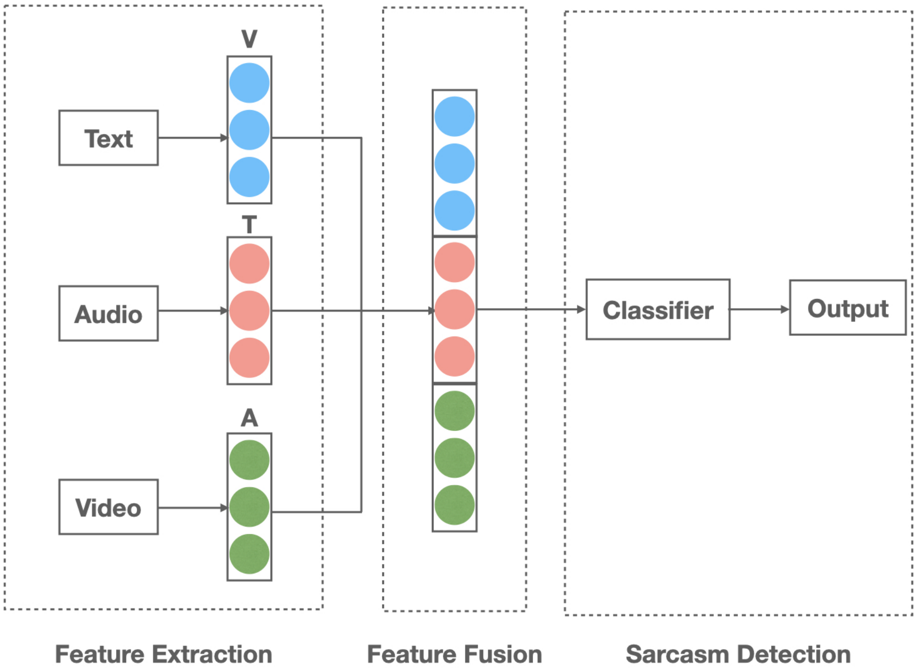

Figure 1: The Proposed Early Fusion Method

3.2.2 Experiment 2

In this section, we detail the architecture and functional components of the ConAtt model [47], which is designed for multimodal sarcasm detection. Our model effectively evaluates sarcasm within a conversational context, providing a nuanced analysis that incorporates historical interactions. The ConAtt model is open source code for public use and available on github[3].

The multimodal sarcasm detection task involves analyzing an utterance sequence $S_i = \{u_1, u_2, \ldots, u_i\}$ within a conversation that has proceeded for t turns. The focus is on the i-th utterance u_i, which is the subject of testing, while the preceding utterances serve as its historical context. The objective of the ConAtt model is to assign a binary label to u_i (1 for sarcasm and 0 for non-sarcasm), conditioned on both the utterance itself and its interaction with previous utterances. Here, S_i denotes the sequence of utterances in the conversation up to the i-th utterance, u_i points to the i-th utterance in the sequence.

[3] https://github.com/xiaoqiangzhang203/Multi-modal-Sarcasm-Detection-ConAttSD

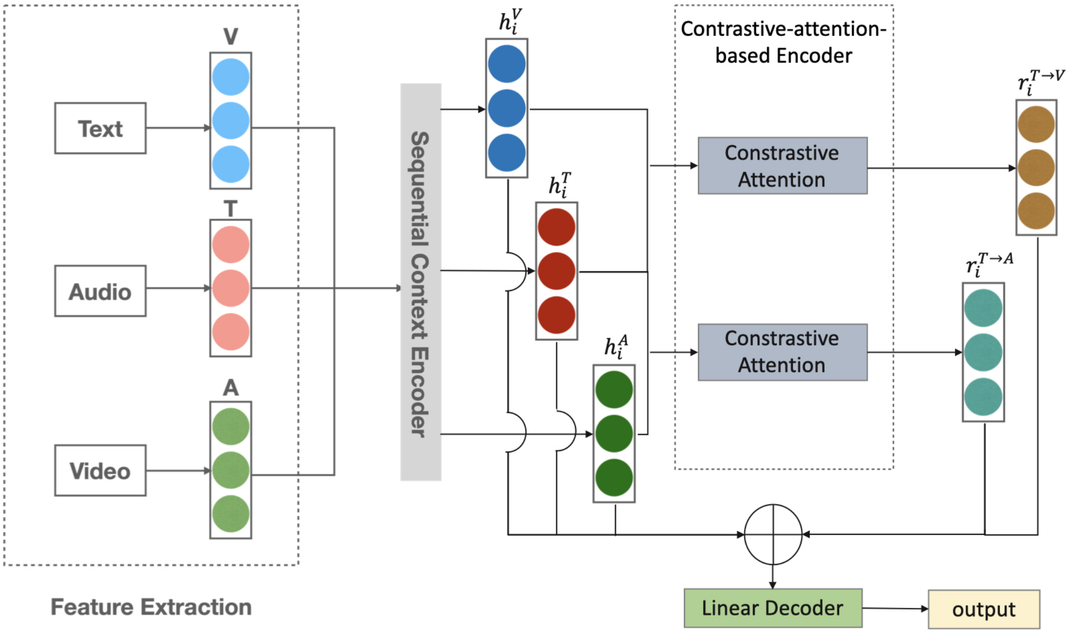

Figure 2: The Proposed ConAtt Model Construction

The ConAtt model, as illustrated in the Figure 2 [47], consists of four main components: Feature Extraction, the Sequential Context Encoder, the Contrastive-Attention-based Encoder, and the Linear Decoder. Feature Extraction employs the methods discussed above to extract features from text, audio, and video. Subsequently, the Sequential Context Encoder dynamically captures the intra-modal influences that are transmitted throughout the conversation, effectively processing the sequential flow of dialogues. Then, the Contrastive-Attention-based Encoder, utilizing an inter-modal contrastive attention mechanism, extracts inconsistencies between different modalities within the conversation, highlighting discrepant cues that are indicative of sarcasm. Finally, the Linear Decoder assigns a sarcasm label to the utterance by synthesizing the insights gathered from both encoders, thus determining the presence or absence of sarcasm based on the integrated feature representation. This structured approach allows for nuanced detection of sarcasm across varied communicative cues.

Sequential Context Encoder: In this study, we introduce a sequential context encoder designed to analyze the sequence and context of conversations, focusing on how utterances influence each other across time. As illustrated in Figure 3, this encoder comprises two sub-encoders: one based on Gated Recurrent Units (GRU) and another based on the Transformer architecture.

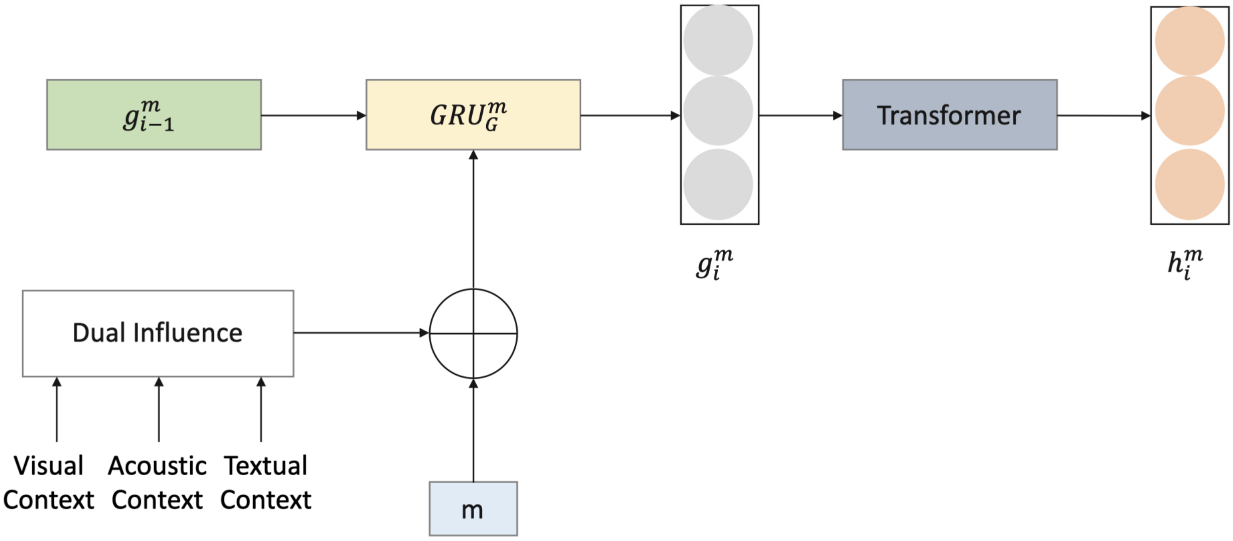

Figure 3: Sequential Context Encoder

The GRU-based encoder extracts the sequential context information. Specifically, it uses a GRU mechanism to update the global state for each utterance, which represents the sequential context [47] as follows:

$$g_i^m = \mathrm{GRU}_G^m\left((u_i^m \oplus c_i^m), g_{i-1}^m\right) \tag{1}$$

Here, $m \in \{T, A, V\}$ denotes the modality, g_i^m is the global state for the i-th utterance in modality, u_i^m is the i-th utterance, and c_i^m is the context, which captures both intra-modal and inter-modal influences. The global state g_i^m is updated by concatenating u_i^m and c_i^m using the concatenation symbol $\oplus$, expressed as: $g_i^m = u_i^m \oplus c_i^m$.

In the Transformer-based sub-encoder, we enhance the capability to capture extensive sequential context by utilizing the Transformer model, known for its efficiency in handling long-range dependencies. The Transformer operates over the global states, applying multiple layers of self-attention and feed-forward neural networks to produce a refined sequential context vector h_i^m for each utterance.

Contrastive-Attention-based Encoder: In order to discern incongruent information across multiple modalities for effective sarcasm recognition, following the work of [47], we introduce the inter-modality contrastive attention mechanism. This mechanism utilizes contrastive attention to process three sequential context vectors—h_i^T (textual), h_i^A (audio), and h_i^V (visual)—derived from the sequential context encoder. The contrastive attention mechanism, initially developed for person re-identification in computer vision and later adapted for natural language processing tasks such as text summarization, extends from the traditional self-attention mechanism utilized in Transformers. The process involves three primary steps: calculation of conventional attention weights a_c using Eq. 2, derivation of opponent attention weights a_o via an opponent function followed by a softmax operation as defined in Eq. 3, and the formation of a contrastive vector through the weighted summation of the elements of V, using opponent attention weights as specified in Eq. 4. Here, Q, K, and V represent the queries, keys, and values, respectively, with d_k denoting the dimension of the key vector K.

$$a_c = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{2}$$

$$a_o = \text{softmax}\left(1 - a_c\right) \tag{3}$$

$$r = a_o V \tag{4}$$

This contrastive attention framework aims to highlight less relevant or irrelevant aspects between two vectors. Specifically, in the context of this research, the text modality serves as the anchor. Directed bi-modal variants, such as T → A and T → V, are formulated by designating the textual vector h_i^T as Q and the respective audio and visual vectors as K. The vectors h_i^A and h_i^V are then used as V to generate the inter-modality contrastive vectors, namely $r_i^{T\to A}$ and $r_i^{T\to V}$. These vectors represent the incongruence within their corresponding bi-modal interactions, effectively isolating discrepancies between text and the other modalities.

Linear Decoder: As illustrated in Figure 2, the classification process relies on the integration of output vectors from distinct encoding stages. Specifically, three sequential context vectors—h_i^T (textual), h_i^A (audio), and h_i^V (visual), generated by the sequential context encoder—are merged with the two inter-modal contrastive vectors—$r_i^{T\to A}$ and $r_i^{T\to V}$, produced by the Contrastive-Attention-based Encoder. The combined vector is then fed into a softmax classifier. This classifier is responsible for determining the sarcasm label of the utterance u_i. It assesses whether the utterance expresses sarcasm by analyzing the complex interplay of multimodal signals encapsulated within the combined vector. This method highlights the importance of integrating diverse modal inputs to enhance the precision of sarcasm detection in intricate communicative contexts.

3.3 Experiment

3.3.1 Baselines

In our exploration of sarcasm recognition using the MUStARD dataset, we conduct two key experiments. Experiment 1 is designed to validate whether integrating multimodal data enhances sarcasm detection performance. In Experiment 2, we test the effectiveness of the ConAtt model. This approach is hypothesized to further improve model performance by effectively capturing discrepancies between modalities, thereby demonstrating the comparative advantages of our proposed method over Experiment 1. The code for my model is currently available on github[4].

Experiment 1: Following the methodology outlined in [5], we implement an early fusion strategy to integrate features from three modalities—speech, text, and audio—into a single feature vector. This integrated vector is then input into a SVM classifier for training and evaluation.

Experiment 2: Based on the work referenced in [47], our ConAtt model integrates four key components: Feature Extraction, the Sequential Context Encoder, the Contrastive-Attention-based Encoder, and the Linear Decoder. This integration enables nuanced detection of sarcasm across diverse communicative cues. The architecture of this model is depicted in Figure 2.

[4] https://github.com/youyang-cai/Multimodal-sarcasm-recognition

Table 2: Hyperparameters for fine-tuning ConAtt

Optimizer	Batch size	Learning rate	Dropout rate
Adam	64	0.0001	0.5

3.3.2 Setup

Following the work of [47], there are two setups available for experiments, we have opted for a speaker-dependent setup, as it simplifies the processing requirements. We implement a five-fold cross-validation method in our experiments, ensuring each fold is randomly generated in a stratified manner to maintain label balance across the dataset. During each iteration of the validation process, one fold is designated as the test set, while the others are used for training. Validation subsets are systematically derived from the training folds during each iteration of the cross-validation process. Specifically, 10% of each training set is further partitioned to serve as the validation subset. To evaluate the effectiveness of our sarcasm detection models, our study conducted Experiment 1 and Experiment 2. Experiment 1 was divided into unimodal and multimodal tests, employing SVM to process features from respective modalities. This was done to validate the importance of multimodal data over unimodal data when it comes to sarcasm detection and to establish a baseline for Experiment 2.

In Experiment 2, we employed the contrastive attention mechanism, which integrates data from multiple modalities and focuses on capturing discrepancies between these different modalities. This strategy aligns with our hypothesis that such an approach would yield superior model performance by effectively highlighting and utilizing these discrepancies. To deepen our analysis, we conducted ablation studies in Experiment 1 by varying the type of classifier used. We replaced the SVM with a MLP and a LSTM network in separate trials to test the robustness of our findings and to further validate the significance of multimodal data in sarcasm detection. Experiment 2 explored the efficacy of the contrastive attention mechanism by conducting trials with and without the Contrastive-Attention-based Encoder. This part of the study was crucial for demonstrating how the contrastive attention mechanism aids in recognizing inconsistencies across modalities, which is essential for accurately detecting sarcasm.

Additionally, due to the randomized nature of fold creation which led to an overlap of speakers across training and testing sets, we classified our approach as speaker-dependent. This classification was important for interpreting the performance metrics of our models. We used weighted metrics for Precision, Recall, and F-Score, where weights were determined based on the relative frequencies of sarcasm and non-sarcasm instances within the dataset, ensuring that our evaluation metrics reflected the distribution of the classes accurately.

During the training phase, we employ the Adam optimizer [58] to fine-tune all hyperparameters. A detailed overview of the hyperparameters is given in Table 2. Each training batch consists of 64 conversations, with a learning rate meticulously set to 0.0001 in order to optimize convergence speed and stability. In the GRU component of our model, a dropout rate of 0.5 is implemented to reduce overfitting, enhancing the model's generalization capabilities across unseen data. The hidden representations within the GRU are configured to a dimensionality of 150, providing a balance between model complexity and computational efficiency. Furthermore, for the Transformer-based components of our architecture, we configure three blocks (B=3) and set the number of attention heads to six. This configuration is chosen to effectively capture the intricate dependencies within the data while maintaining manageable computational demands. These settings collectively ensure that our model is robust and capable of handling the complexities inherent in multimodal sarcasm detection.

4 Results

This chapter presents a detailed analysis of the impact of multimodal data integration on sarcasm detection. Through a series of methodically designed experiments, we explore how the combination of various modalities—text, audio, and visual—enhances the performance of sarcasm detection systems. The results from these experiments provide empirical evidence supporting the hypothesis that the ConAtt model, which utilizes a contrastive attention mechanism to highlight modal discrepancies, enhances sarcasm detection accuracy compared to early fusion methods. This underscores the benefits of leveraging diverse datasets and the contrastive attention mechanism in capturing the subtleties of sarcastic expressions. This chapter is structured to first discuss the comparative performance of different modal setups, followed by insights from ablation studies, and concluding with the limitations of the current study and directions for future research.

4.1 Comparison With Baselines

In an effort to explore the impact of multimodal data on sarcasm detection performance, we conducted an analysis using various modality configurations across two distinct experiments. The results of these experiments are summarized in Table 3. Here, the Error Rate Reduction (ERR) quantifies the decrease in relative error, calculated as the difference between these two values divided by the absolute error of the unimode, exemplified by the formula $3.8/(100-68.1)$.

Experiment 1 evaluated the effectiveness of unimodal versus multimodal configurations in sarcasm detection, demonstrating a significant superiority of multimodal setups. The integration of all three modalities—text, audio, and visuals—proved to be the most effective, achieving Precision, Recall, and F-Score rates of approximately 71.7%. These metrics indicate an upward trend compared to the highest-performing single modality, visuals, which scored around 67.4%. This configuration also resulted in an ERR of approximately 11.91%, which further indicates that a single feature cannot fully express sarcasm, and more information is needed to assist sarcasm detection.

In Experiment 2, the implementation of the ConAtt model led to further enhancements in sarcasm detection performance by effectively capturing discrepancies between modalities. Using multimodal data (T+A+V) and the ConAtt model produced improved Precision, Recall, and F-score values of 75.3%, 75.1%, and 75.0%, respectively. With gains of 3.4% in Precision and Recall as well as a 3.3% improvement in the F-score over the best-performing model from Experiment 1, these measures show improvements in all parameters. This yielded the ERR of up to 11.66%. The performance boost provided by the ConAtt model underscores the efficacy of contrastive attention fusion strategy in highlighting modal inconsistencies, illustrating the robustness of the ConAtt model and the importance of multimodal data integration.

Further exploration into the influence of the three modalities on the ConAtt model involved evaluating its performance using both dual-modal and tri-modal variants. Among the dual-modal setups, the combination of text and audio (T+A) displayed superior performance, achieving Precision at 71.8%, Recall at 71.7%, and F-Score at 71.7%. Configurations excluding the acoustic feature (T+V) performed the worst, emphasizing the critical role of audio features in capturing speaker-specific details such as pitch and intonation, which are essential in detecting sarcasm. The tri-modal variant (T+A+V), however, achieved the highest performance across all metrics, suggesting that each modality contains unique and complementary information that contributes significantly to the overall

Table 3: Comparison With Baselines

Experiment	**Modality**	**Precision**	**Recall**	**F-Score**
Experiment 1	T	65.1	64.6	64.6
	A	65.9	64.8	64.8
	V	**68.1**	**67.4**	**67.4**
	T+A	66.7	66.7	66.6
	T+V	71.8	71.6	71.6
	A+V	67.4	66.5	66.6
	T+A+V	**71.9**	**71.7**	**71.7**
	Δmulti-unimodal	↑3.8%	↑4.3%	↑4.3%
	ERR	↑11.91%	↑13.19%	↑13.19%
Experiment 2	T+A	71.8	71.7	71.7
	T+V	70.5	70.3	70.3
	A+V	70.9	70.6	70.5
	T+A+V	**75.3**	**75.1**	**75.0**
	ΔExperiment 1	↑3.4%	↑3.4%	↑3.3%
	ERR	↑12.09%	↑12.01%	↑11.66%
	Δmulti-unimodal	↑7.2%	↑7.7%	↑7.6%
	ERR	↑22.57%	↑23.62%	↑23.31%

$^{\Delta\text{multi-unimodal}}$ represents the discrepancy between the peak value observed in multi-mode and the maximum value achieved by any unimode.

$^{\text{ERR}}$ quantifies the decrease in relative error, calculated as the difference between these two values divided by the absolute error of the unimode, exemplified by the formula $3.8/(100-68.1)$.

$^{\Delta\text{Experiment 1}}$ refers to the difference obtained by comparing the optimal model from Experiment 2 with the model from Experiment 1 that has the same input data. The corresponding Error rate reduction is calculated as the difference between these two values divided by the absolute error of Experiment 1, exemplified by the calculation $3.4/(100-71.9)$.

efficacy of sarcasm detection. Sarcasm is often conveyed through discordance between modalities, and a contrastive attention mechanism, designed to extract several contrasting features of discourse, can significantly enhance performance.

4.2 Ablation Study

4.2.1 Role of Multi-modalities

The ablation study reported in Experiment 1 offers a clear illustration of the effectiveness of multimodal data integration in sarcasm detection, highlighting the importance of selecting the appropriate classifier based on the type of data and specific detection needs. Table 4 presents detailed outcomes using various classifiers, including SVM, MLP, and LSTM.

For the MLP classifier, the integration of text, audio, and visuals (T+A+V) achieved the highest scores with Precision, Recall, and F-score values of 69.0%, 68.8%, and 68.9%, respectively. These results represent improvements of 1.1% in Precision, 1.6% in Recall, and 1.7% in F-score compared to the highest-performing unimodal configuration, which was the visual modality. Similarly, the LSTM classifier demonstrated its strongest performance when combining the audio and visual modalities (A+V), achieving Precision, Recall, and F-score values of 72.6%, 72.3%, and 72.2%, respectively. When compared to the most effective unimodal approach using only visuals, which recorded Precision, Recall, and F-score of 68.6%, 67.2%, and 66.5%, these results represent significant improvements. Specifically, there was an increase of 4.0% in Precision, 5.1% in Recall, and 5.7% in F-score. These enhancements underscore the substantial advantages of multimodal approaches over unimodal approaches, particularly in capturing the complexities of sarcasm detection more effectively.

Moreover, the LSTM classifier demonstrated a clear advantage, consistently outperforming both unimodal and multimodal configurations implemented with SVM and MLP classifiers. This suggests that LSTM networks are particularly adept at handling the complexities of multimodal data integration in sarcasm detection, likely due to their ability to effectively process sequential and contextual information [26]. In contrast, the SVM classifier exhibited superior performance compared to the MLP, which indicates its appropriateness for managing smaller datasets [57], such as the MUStARD dataset used in this study, where the risk of overfitting is a significant concern. This highlights the critical importance of choosing the appropriate classifier based on both the nature of the dataset and the specific requirements of the task.

These findings demonstrate the significance of multimodal data in sarcasm detection. Across SVM, MLP and LSTM classifiers, configurations utilizing multimodal data consistently outperform those relying on a single modality. This underscores the vital role that multimodal approaches play in enhancing the accuracy and depth of sarcasm detection in digital communications. By leveraging diverse data streams and strategically chosen classifiers, these systems are better equipped to decode the complexities of communicative phenomena, thus advancing our understanding of sarcasm in a nuanced and effective manner.

4.2.2 Role of Contrastive Attention

In Experiment 2, the effectiveness of the ConAtt model was thoroughly validated, demonstrating substantial performance enhancements when integrated with a multimodal fusion strategy. This model's adeptness at capturing discrepancies across different modalities aligns seamlessly with our initial

Table 4: Ablation Study Results

Fusion Type	**Modality**	**Algorithm**	**Precision**	**Recall**	**F-Score**
Experiment 1	T	SVM	65.1	64.6	64.6
	A		65.9	64.8	64.8
	V		**68.1**	**67.4**	**67.4**
	T+A		66.7	66.7	66.6
	T+V		71.8	71.6	71.6
	A+V		67.4	66.7	66.6
	T+A+V		**71.9**	**71.7**	**71.7**
	Δmulti-unimodal		↑3.8%	↑4.3%	↑4.3%
	T	MLP	61.6	61.2	61.1
	A		60.0	58.1	57.1
	V		**67.9**	**67.2**	**67.2**
	T+A		63.5	63.2	63.2
	T+V		68.5	68.1	68.2
	A+V		63.2	62.9	62.9
	T+A+V		**69.0**	**68.8**	**68.9**
	Δmulti-unimodal		↑1.1%	↑1.6%	↑1.7%
	T	LSTM	64.9	63.8	63.9
	A		67.0	65.9	65.1
	V		**68.6**	**67.2**	**66.5**
	T+A		70.2	69.6	69.2
	T+V		71.2	67.2	66.2
	A+V		**72.6**	**72.3**	**72.2**
	T+A+V		71.4	70.5	70.3
	Δmulti-unimodal		↑4.0%	↑5.1%	↑5.7%
Experiment 2	T+A+V	ConAtt w/o CA	70.7	70.4	70.4
	T+A+V	ConAtt	**75.3**	**75.1**	**75.0**
	ΔConAtt w/o CA		↑4.6%	↑4.7%	↑4.6%

ΔConAtt w/o CA This represents the discrepancy between the models with and without the Contrastive-Attention-based Encoder.

hypothesis. Equipped with a contrastive attention mechanism, the ConAtt model significantly outperformed its counterpart without this mechanism across all measured metrics. This represents an improvement in precision, recall, and F-Score of approximately 4.6% over the model without the contrastive attention mechanism, which achieved scores of 70.7%, 70.4%, and 70.4% respectively.

Furthermore, the ConAtt model consistently outperforms all configurations tested in Experiment 1, regardless of the classifier used, demonstrating its robustness and effectiveness in detecting sarcasm. The results clearly demonstrate the effectiveness of the contrastive attention mechanism. Our ConAtt model adeptly manages both intra-modal and inter-modal variations, skillfully capturing the subtle discrepancies between modalities. Specifically, using the MLP classifier in a T+A+V setup in Experiment 1 resulted in an F-Score of only 68.9%, significantly lower than the 75.0% achieved by the ConAtt model in Experiment 2. Similarly, employing the LSTM classifier in the same modality configuration, the best F-Score from Experiment 1 was 70.3%, which pales in comparison to the 75.0% attained by the ConAtt model. These comparative results not only underscore the robustness of the ConAtt model but also its superior effectiveness in detecting sarcasm, illustrating the significant benefits of integrating the contrastive attention mechanism into multimodal fusion strategies.

4.3 Limitations and Future Research

Based on the insights and findings from our experiments, our limitations and directions for future research are as follows.

4.3.1 Limitations

The very limited size of the dataset utilized in this work is one of its primary weaknesses because it hinders our ability to thoroughly investigate more precise sarcasm detection. The restricted scope of the dataset could potentially restrict the applicability of our research findings, as a more extensive dataset would be able to more consistently confirm the model's efficacy across various circumstances and sarcastic kinds. To ensure that sarcasm detection algorithms in real-world contexts are both scalable and adaptable, a substantial amount of data is needed to train them to account for cultural and linguistic variations.

Furthermore, this study's primary reliance on attention-based multimodal fusion methods introduces another limitation. While these methods are proficient at integrating data from various modalities, enhancing the significance of relevant information, and minimizing noise, they may underperform when correlations between modalities are weak or complex. In scenarios where modal interactions are not straightforward or easily quantifiable, attention mechanisms might not fully capture the nuanced relationships necessary for accurate sarcasm detection.

4.3.2 Future Research

Future research should focus on expanding the dataset size and employing diverse datasets to further validate and refine the results. Increasing the dataset diversity would help in testing the models against a broader spectrum of sarcasm expressions, thus enhancing their applicability and reliability in various communication settings.

It is also essential to investigate and create more sophisticated modal fusion techniques. The primary fusion technique used in this work is a multi-modal approach based on attention mechanisms, which can fuse several data modes in a targeted manner to increase the weight of important information while suppressing noise and irrelevant data. However, the fusion approach based on attention mechanism may not be able to properly learn the information of the modes when the correlation between the modes is weak or difficult to define. In order to improve the model's accuracy and performance, more focus should be given in future work on the analysis and investigation of the correlation between modes in order to build an appropriate fusion method.

Moreover, the detection of sarcasm across languages and cultural backgrounds has become a significant area of research. Future research can gain from combining emotional computation and sarcasm recognition for more thorough multimodal assessments. Models' capacity to identify minute differences in sarcasm can be greatly improved by gaining a better grasp of the relationships between sarcastic and emotional cues. This is especially important in situations when sarcasm is significantly dependent on tone and a larger communication context, like in texts with intricate storylines or conversational talks. By using this method, scholars can process and evaluate texts from a variety of linguistic and cultural backgrounds more successfully, exposing the texts' deeper levels of sarcasm.

By addressing these areas, future research can enhance the effectiveness and applicability of sarcasm detection systems, ensuring that they are not only technically robust but also versatile in handling the complexities of human communication.

5 Conclusion

This study has elevated the field of sarcasm detection by introducing and validating the ConAtt model. This model has demonstrated superior performance over early fusion methods, achieving higher Precision, Recall and F-Score, particularly in managing dynamic multimodal information. At the core of the ConAtt model is a contrastive attention mechanism that expertly integrates and analyzes inconsistencies across different modalities—text, audio, and video. This capability is essential for the accurate identification and nuanced analysis of sarcasm, which is often subtly conveyed through complex multimodal interactions.

The ConAtt model's ability to precisely detect subtle variations in sarcastic expressions is critical for effective detection, showcasing the importance of multimodal features in sarcasm recognition. The organic integration of textual, auditory, and visual features captures the temporal dynamics and complex information flows of different modalities, significantly enhancing sarcasm detection capabilities. The superiority of multimodal approaches over unimodal ones enriches the diversity of information sources. Additionally, the selection of an appropriate classifier plays an important role in the outcomes of sarcasm detection.

To sum up, the ConAtt model has proven to be effective in detecting sarcasm and has confirmed the usefulness of contrastive attention mechanisms. Through its efficient integration and analysis of disparities among textual, audio, and video data, this model presents encouraging paths for future developments in the field.

Bibliography

[1] A. Joshi, P. Bhattacharyya, and M. Carman, “Automatic sarcasm detection: A survey,” *ACM Computing Surveys (CSUR)*, vol. 50, no. 5, pp. 1–22, 2017. DOI: 10.1145/3124420.

[2] P. Carvalho, L. Sarmento, M. Silva, and E. De Oliveira, “Clues for detecting irony in user-generated contents: Oh...!! it’s ”so easy”;-,” in *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, pp. 53–56, 2009. DOI: 10.1145/1651461.1651471.

[3] L.-P. Morency, R. Mihalcea, and P. Doshi, “Towards multimodal sentiment analysis: Harvesting opinions from the web,” in *Proceedings of the 13th international conference on multimodal interfaces*, pp. 169–176, 2011. DOI: 10.1145/2070481.2070509.

[4] R. Rakov and A. Rosenberg, ““sure, i did the right thing”: a system for sarcasm detection in speech,” in *Proc. Interspeech 2013*, pp. 842–846, 2013. DOI: 10.21437/Interspeech.2013-239.

[5] S. Castro, D. Hazarika, V. Pérez-Rosas, R. Zimmermann, R. Mihalcea, and S. Poria, “Towards multimodal sarcasm detection (an _Obviously_ perfect paper),” *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4619–4629, 2019. DOI: 10.18653/v1/P19-1455.

[6] Y. Cai, H. Cai, and X. Wan, “Multi-modal sarcasm detection in Twitter with hierarchical fusion model,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (A. Korhonen, D. Traum, and L. Màrquez, eds.), (Florence, Italy), pp. 2506–2515, Association for Computational Linguistics, July 2019. DOI: 10.18653/v1/P19-1239.

[7] S. Sangwan, M. S. Akhtar, P. Behera, and A. Ekbal, “I didn’t mean what i wrote! exploring multimodality for sarcasm detection,” in *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2020. DOI: 10.1109/IJCNN48605.2020.9206905.

[8] A. Zadeh, M. Chen, S. Poria, *et al.*, “Tensor fusion network for multimodal sentiment analysis,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1103–1114, 2017. DOI: 10.18653/v1/D17-1115.

[9] S. Pramanick, A. B. Roy, and V. M. Patel, “Multimodal learning using optimal transport for sarcasm and humor detection,” *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 546–556, 2021. DOI: 10.48550/arXiv.2110.10949.

[10] S. Hiai and K. Shimada, “A sarcasm extraction method based on patterns of evaluation expressions,” in *2016 5th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, pp. 31–36, 2016. DOI: 10.1109/IIAI-AAI.2016.198.

[11] M. Abulaish and A. Kamal, “Self-deprecating sarcasm detection: An amalgamation of rule-based and machine learning approach,” in *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pp. 574–579, 2018. DOI: 10.1109/WI.2018.00-35.

[12] A. Rajadesingan, R. Zafarani, and H. Liu, “Sarcasm detection on twitter: A behavioral modeling approach,” in *Proceedings of the eighth ACM international conference on web search and data mining*, pp. 97–106, 2015. DOI: 10.1145/2684822.2685316.

[13] M. S. M. Suhaimin, M. H. A. Hijazi, R. Alfred, and F. Coenen, “Natural language processing based features for sarcasm detection: An investigation using bilingual social media texts,” in *2017 8th International conference on information technology (ICIT)*, pp. 703–709, 2017. DOI: 10.1109/ICITECH.2017.8079931.

[14] S. K. Bharti, K. S. Babu, and S. K. Jena, “Parsing-based sarcasm sentiment recognition in twitter data,” in *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, pp. 1373–1380, 2015. DOI: 10.1145/2808797.2808910.

[15] E. Ilavarasan *et al.*, “A survey on sarcasm detection and challenges,” in *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pp. 1234–1240, 2020. DOI: 10.1109/ICACCS48705.2020.9074163.

[16] V. Hatzivassiloglou and K. R. McKeown, “Predicting the semantic orientation of adjectives,” in *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, (Madrid, Spain), pp. 174–181, Association for Computational Linguistics, 1997. DOI: 10.3115/976909.979640.

[17] C. C. Chang and C. J. Lin, “Libsvm: A library for support vector machines,” *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, p. Article 27, 2007. DOI: 10.1145/1961189.1961199.

[18] B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up? sentiment classification using machine learning techniques,” in *Proceedings of the EMNLP*, 2002. DOI: 10.3115/1118693.1118704.

[19] X. Guo, G. Zhang, S. Wang, *et al.*, “Multi-way matching based fine-grained sentiment analysis for user reviews,” *Neural Computing and Applications*, vol. 32, no. 1-2, 2020. DOI: 10.1007/s00521-019-04686-9.

[20] T. Ptácek, I. Habernal, and J. Hong, “Sarcasm detection on czech and english twitter,” in *International Conference on Computational Linguistics*, 2014.

[21] C. Eke, A. Norman, and L. Shuib, “Multi-feature fusion framework for sarcasm identification on twitter data: A machine learning based approach,” *PLoS One*, vol. 16, p. e0252918, Jun 2021. DOI: 10.1371/journal.pone.0252918.

[22] J. L. Elman, “Finding structure in time,” *Cogn. Sci.*, vol. 14, pp. 179–211, 1990. DOI: 10.1207/s15516709cog1402$_1$.

[23] Y. Kim, “Convolutional neural networks for sentence classification,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1746–1751, 2014. DOI: 10.3115/v1/D14-1181.

[24] M. Pota, M. Esposito, G. D. Pietro, and H. Fujita, “Best practices of convolutional neural networks for question classification,” *Applied Sciences*, 2020. DOI: 10.3390/app10144710.

[25] S. Wang, M. Huang, and Z. Deng, “Densely connected cnn with multi-scale feature attention for text classification,” in *International Joint Conference on Artificial Intelligence*, 2018. DOI: 10.24963/ijcai.2018/621.

[26] C. Baziotis, N. Athanasiou, P. Papalampidi, A. Kolovou, G. Paraskevopoulos, N. Ellinas, and A. Potamianos, “Ntua-slp at semeval-2018 task 3: Tracking ironic tweets using ensembles of word and character level attentive rnns,” *CoRR*, vol. abs/1804.06659, 2018. DOI: ”10.18653/v1/S18-1100” .

[27] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, pp. 1735–80, 12 1997. DOI: 10.1162/neco.1997.9.8.1735.

[28] D. Jain, A. Kumar, and G. Garg, “Sarcasm detection in mash-up language using softattention based bi-directional lstm and feature-rich cnn,” *Applied Soft Computing*, vol. 91, p. 106198, 2020. DOI: 10.1016/j.asoc.2020.106198.

[29] A. Ghosh and T. Veale, “Fracking sarcasm using neural network,” in *Proceedings of the 7th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pp. 161–169, 2016. DOI: 10.18653/v1/W16-0425.

[30] J. Zhang, X. Wu, and C. Huang, “Adamow: Multimodal sentiment analysis based on adaptive modality-specific weight fusion network,” *IEEE Access*, vol. 11, pp. 48410–48420, 2023. DOI: 10.1109/ACCESS.2023.3276932.

[31] J. Gao, P. Li, Z. Chen, and J. Zhang, “A survey on deep learning for multimodal data fusion,” *Neural Computation*, vol. 32, pp. 829–864, 2020. DOI: 10.1162/neco$_{a0}$1273.

[32] C. Busso and S. S. Narayanan, “Interrelation between speech and facial gestures in emotional utterances: a single subject study,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2331–2347, 2007. DOI: 10.1109/TASL.2007.905145.

[33] E. Cambria, D. Hazarika, S. Poria, A. Hussain, and R. B. V. Subramanyam, “Benchmarking multimodal sentiment analysis,” *ArXiv*, vol. abs/1707.09538, 2017. DOI: 10.1007/978-3-319-77116-8$_1$3.

[34] C.-H. Wu and W.-B. Liang, “Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels,” *IEEE Transactions on Affective Computing*, vol. 2, no. 1, pp. 10–21, 2010. DOI: 10.1109/T-AFFC.2010.16.

[35] J. G. Ellis, B. Jou, and S.-F. Chang, “Why we watch the news: a dataset for exploring sentiment in broadcast video news,” in *Proceedings of the 16th international conference on multimodal interaction*, pp. 104–111, 2014. DOI: 10.1145/2663204.2663237.

[36] J. Williams, S. Kleinegesse, R. Comanescu, and O. Radu, “Recognizing emotions in video using multimodal dnn feature fusion,” in *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*, pp. 11–19, 2018. DOI: 10.18653/v1/W18-3302.

[37] S. Poria, A. Hussain, and E. Cambria, “Beyond text based sentiment analysis: Towards multimodal systems,” tech. rep, University of Stirling, Stirling FK9 4LA, UK, 2013.

[38] V. Rozgić, S. N. Vitaladevuni, and R. Prasad, “Robust eeg emotion classification using segment level decision fusion,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1286–1290, 2013. DOI: 10.1109/ICASSP.2013.6637858.

[39] A. Sayedelahl, R. Araujo, and M. S. Kamel, "Audio-visual feature-decision level fusion for spontaneous emotion estimation in speech conversations," in *2013 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pp. 1–6, 2013. DOI: 10.1109/ICMEW.2013.6618372.

[40] N. Sebe, I. Cohen, T. Gevers, *et al.*, "Emotion recognition based on joint visual and audio cues," in *18th international conference on pattern recognition (ICPR'06)*, pp. 1136–1139, 2006. DOI: 10.1109/ICPR.2006.489.

[41] M. Bedi, S. Kumar, M. S. Akhtar, *et al.*, "Multi-modal sarcasm detection and humor classification in code-mixed conversations," *IEEE Transactions on Affective Computing*, 2021. DOI: 10.1109/TAFFC.2021.3083522.

[42] H. Pan, Z. Lin, P. Fu, Y. Qi, and W. Wang, "Modeling intra and inter-modality incongruity for multi-modal sarcasm detection," in *Findings of the Association for Computational Linguistics: EMNLP 2020* (T. Cohn, Y. He, and Y. Liu, eds.), (Online), pp. 1383–1392, Association for Computational Linguistics, Nov. 2020. DOI: 10.18653/v1/2020.findings-emnlp.124.

[43] X. Wang, X. Sun, T. Yang, *et al.*, "Building a bridge: a method for image-text sarcasm detection without pretraining on image-text data," in *Proceedings of the 1st International Workshop on Natural Language Processing Beyond Text*, pp. 19–29, 2020. DOI: 10.18653/v1/2020.nlpbt-1.3.

[44] C. Song, Y. Huang, W. Ouyang, *et al.*, "Mask-guided contrastive attention model for person reidentification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1179–1188, 2018. DOI: 10.1109/CVPR.2018.00129.

[45] X. Duan, H. Yu, M. Yin, *et al.*, "Contrastive attention mechanism for abstractive sentence summarization," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 3044–3053, 2019. DOI: 10.18653/v1/D19-1301.

[46] B. Liang, C. Lou, X. Li, L. Gui, M. Yang, and R. Xu, "Multi-modal sarcasm detection with interactive in-modal and cross-modal graphs," *Proceedings of the 29th ACM International Conference on Multimedia*, 2021. DOI: 10.18653/v1/2022.acl-long.124.

[47] X. Zhang, Y. Chen, and G. Li, "Multi-modal sarcasm detection based on contrastive attention mechanism," *Natural Language Processing and Chinese Computing*, vol. 13028, pp. 822–833, 2021. DOI: 10.1007/978-3-030-88480-2$_{6}$6.

[48] Y. Zhang, J. Wang, Y. Liu, L. Rong, Q. Zheng, D. Song, P. Tiwari, and J. Qin, "A multitask learning model for multimodal sarcasm, sentiment and emotion recognition in conversations," *Information Fusion*, vol. 93, pp. 282–301, May 2023. DOI: 10.1016/j.inffus.2023.01.005.

[49] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *North American Chapter of the Association for Computational Linguistics*, 2019. DOI: 10.18653/v1/N19-1423.

[50] N. Ding, S. Tian, and L. Yu, "A multimodal fusion method for sarcasm detection based on late fusion," *Multimedia Tools and Applications*, vol. 81, pp. 8597–8616, 2022. DOI: 10.1007/s11042-022-13450-w.

[51] B. McFee, C. Raffel, D. Liang, D. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "Librosa: Audio and music signal analysis in python [j]//proceedings of the python in science conference," pp. 18–24, 2015. DOI: 10.25080/Majora-7b98e3ed-003.

[52] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. DOI: 10.1109/CVPR.2016.90.

[53] Y. K. Atri, S. Pramanick, V. Goyal, and T. Chakraborty, "See, hear, read: Leveraging multimodality with guided attention for abstractive text summarization," *Knowledge-Based Systems*, p. 107152, 2021. DOI: 10.48550/arXiv.2105.09601.

[54] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, "Convolutional mkl based multimodal emotion recognition and sentiment analysis," in *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pp. 439–448, 2016. DOI: 10.1109/ICDM.2016.0055.

[55] E. Cambria, "Affective computing and sentiment analysis," *IEEE Intelligent Systems*, vol. 31, no. 2, pp. 102–107, 2016. DOI: 10.1007/978-3-319-55394-8_1.

[56] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency, "Memory fusion network for multi-view sequential learning," *ArXiv*, vol. abs/1802.00927, 2018. DOI: 10.1609/aaai.v32i1.12021.

[57] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011. DOI: 10.48550/arXiv.1201.0490.

[58] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014. DOI: 10.48550/arXiv.1412.6980.