# Master's Thesis

# Noise-Tolerant Multimodal Sentiment Analysis using Modal-Independent Classifiers

（モーダル独立な分類器を用いたノイズに
頑健なマルチモーダル感情分析)

48-216431

**Nattapong Tiyajamorn**

Department of Information and Communication Engineering
Graduate School of Information Science and Technology
The University of Tokyo

*Supervisor*

Associate Professor Naoki Yoshinaga

January 26, 2023

*A thesis submitted in fulfillment of the requirements for the degree of Master*
*in*
*Graduate School of Information Science and Technology*
*The University of Tokyo*

# Abstract

The COVID-19 pandemic has led to an increase of online multi-modal conversation via video conferencing platforms such as Zoom and WebEx. Although these multi-modal communications are expected to be enhanced by natural language processing (NLP), various types of noises such as errors of automatic speech recognition (ASR), low-resolution data due to limited communication bandwidth, and absence of some modalities due to privacy concerns makes it difficult to apply existing multimodal NLP. In this study, aiming to address the noises in online multimodal data in multi-modal sentiment classification, we propose a robust method using modal independent classifiers which can retain the model performance across different data quality. The key component of our model is a gate network that determines the reliability of each mono-modal input when accumulating outputs of mono-modal classifiers. Experimental results on an online dialogue dataset, Hazumi, confirmed that our model outperform baselines with noisy data.

# Contents

# List of Figures

# Tables

# Chapter 1

# Introduction

## 1.1 Background

The COVID-19 pandemic has resulted in a significant growth in online communication. Platforms such as Zoom, Google Hangouts, and Skype have seen a sudden increase in usage as people look to stay connected with friends, family, and colleagues. This increase on online communication has been driven by the need to socially distance, with many people now working from home. The pandemic has also led to the creation of new online communities, as people come together to discuss the challenges of living in a time of crisis.

As the world becomes more and more digitized, the need for Sentiment Analysis (SA) [1] is growing. Sentiment analysis is a process of computationally identifying and categorizing opinions expressed in a piece of text, especially with the increase in online communication data, in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral. This is important because it can help businesses or individuals gauge public opinion on a given topic, and make decisions accordingly. For example, if a company sees that sentiment towards their product is mostly negative, they may take steps to improve it. Similarly, an individual may use sentiment analysis to get a sense of how people feel about a potential investment.

However, the majority of research in the field of SA has focused on textual data. With the advent of social media, there is an ever-increasing amount of multimodal data available, which has the potential to provide valuable insights into the sentiment of a text. Multi-modal machine learning [2] is a powerful tool that can be used to improve the accuracy of predictions and improve the performance of machine learning models. By using multiple data sources, machine learning models can learn from a variety of data types and improve the accuracy of their predictions.

Multimodal SA takes into account not only the text of a document, but also the images, videos, and other media that are associated with it. This can be used to identify the sentiment of a text that may be otherwise difficult to interpret, such as when the text is sarcastic or ironic. Additionally, multimodal SA can be used to identify the sentiment of a document when the text is unavailable, such as in a video or image. To overcome this limitation, some multimodal sentiment analysis models are tailored to the specific domain and context. For example, a sentiment analysis model trained on tweets should be different from one trained on Facebook posts.

Although these multi-modal communications are expected to be enhanced by natural language processing (NLP), various types of noises such as errors of automatic speech recognition (ASR) [3] such as, wav2vec [4],or low-resolution data due to limited communication bandwidth, and absence of some modalities due to privacy concerns makes it difficult to apply existing multimodal NLP. The model should be able to recognize the differences in language and media used in the two contexts. In addition, models should be able to recognize and analyze sentiment in different media, such as images, videos, and audio. This will allow the model to better understand the sentiment of a text, even when the text is unavailable.

To improve the performance of sentiment analysis models in different domains and contexts, researchers are also exploring the use of transfer learning. This involves training a model on a large dataset of one domain or context, and then fine-tuning it on a smaller dataset of another domain or context. This allows the model to leverage the knowledge it has already learned on the first dataset, and quickly adapt to the new dataset.

Another approach to overcome the limitation of current sentiment analysis models is to use ensemble models. Ensemble models involve combining the predictions of multiple models, each trained on a different dataset or using a different algorithm. This can help to improve the overall performance of the model by reducing the impact of any errors or biases present in individual models.

The problem of current sentiment analysis models is that they are limited to a specific domain or context. To overcome this limitation, researchers are exploring various approaches such as tailoring models to specific domains and contexts, using transfer learning and ensemble models. These approaches are expected to improve the performance of sentiment analysis models in different domains and contexts and make it more accurate and reliable.

In this study, aiming to address the noises in online multi-modal data in multimodal sentiment classification, we propose a robust method using modal independent classifiers which can retain the model performance across different data quality. The key component of our model is a gate network that determines the reliability of each mono-modal input when accumulating outputs of mono-modal classifiers.

To evaluate the proposed method, we conducted experiments on the Hazumi dataset [5], which is a dialogue corpus with multimodal conversation data with annotation. Our experiments show that the proposed method improved the performance of the model by mitigating the effect of the data noise. This study proposed a robust multimodal sentiment classification method that can effectively use multi-modal data in various noisy environments, and showed that our method can retain the model performance when data quality of one or more modalities is low. Our proposed method is expected to be used in various applications that involve multimodal sentiment analysis. In addition, our proposed method can help reduce the costs associated with data collection and annotation, as it allows models to learn from a variety of noisy data sources, including social media and public conversation data.

In summary, the proposed method in this study addresses the problem of current multimodal sentiment analysis models by using a gate network to determine the reliability of each mono-modal input and by using modal independent classifiers. This approach helps to mitigate the effect of data noise and improve the performance of the model. The proposed method was evaluated on the Hazumi dataset [5] and showed promising results in terms of performance.

## 1.2 Contributions of this paper

The contributions of this thesis are as follows:

- Introducing the multimodal sentiment analysis task
- Suggesting the problem in current multimodal sentiment analysis model
- Proposing a model for multimodal sentiment analysis by using a gate network and modal independent classifiers
- Analyzing the dataset for multimodal sentiment analysis
- Evaluating the proposed method and related methods for multimodal sentiment analysis

## 1.3 Contents in this paper

This thesis consists of following parts

**Chapter 2** Introducing basic knowledge related to this study.

**Chapter 3** Explaining the related works in multimodal sentiment analysis.

**Chapter 4** Describing the architecture and details of proposed method .

**Chapter 5** Explaining the dataset used in experiments and the result.

**Chapter 6** Concluding the study and suggesting future improvement.

# Chapter 2

## Preliminary Knowledge

In this chapter, we will discuss the basic knowledge necessary to understand the topic at hand and use it to provide insight into the topic.

### 2.1 Neural Network

A neural network [6] is a type of machine learning algorithm. It is a type of artificial intelligence that imitates the operation of a human brain. A neural network is composed of layers of interconnected neurons that communicate with each other. Each neuron is a unit that processes data and transmits the data to other neurons in the network.

Neural networks are used in a wide range of applications, from pattern recognition and image classification to speech and natural language processing. They are also used for autonomous systems, such as self-driving cars and robots.

The operation of a neural network relies on the use of weights and biases. Weights are numerical values that represent the strength of the connection between neurons, while biases are numerical values that control the direction and magnitude of the output from neurons. The weights and biases are optimized through an iterative process known as backpropagation [7]. This process adjusts the weights and biases to reduce the error between the predicted output and the actual output.

Neural networks are also capable of learning from their mistakes. This means that they can adjust the weights and bias values in order to make more accurate predictions. This allows neural networks to become more accurate over time as they learn from their mistakes.

### 2.1.1 Multilayer perceptron

Multilayer Perceptron (MLP) [8] is a type of Neural Network composed of multiple layers of neurons. MLPs are commonly used for classification and regression tasks.

The basic structure of a MLP consists of an input layer, one or more hidden layers, and an output layer. The input layer takes in the input data and passes it to the hidden layers for processing. The hidden layers process the input data and pass it to the output layer. The output layer provides the output from the MLP.

The hidden layers are the most important part of an MLP. Each hidden layer is composed of neurons that are connected to the neurons in the preceding and following layers. The neurons in the hidden layers are responsible for learning to recognize patterns in the input data and to produce the desired output.

The MLP learns by adjusting the weights of the connections between neurons. The weights are adjusted based on the input data and the desired output. Over time, the MLP learns to produce the desired output for a given input.

MLPs are a powerful tool for solving complex tasks. They have been used to solve tasks such as image recognition, language translation, and text classification. MLPs can also be used to predict future events or trends. As MLPs continue to evolve and improve, they will become more capable of solving complex tasks.

### 2.1.2 Transformer

The Transformer model [9] is a powerful machine learning model that has revolutionized the field of natural language processing (NLP). It is based on an encoder-decoder architecture, which means it uses an encoder to read an input sentence, and a decoder to generate an output sentence based on the input.

The Transformer model is based on the concept of self-attention. Self-attention is a mechanism in which the model pays attention to all parts of the input sentence, rather than just the words or phrases at the beginning or end of the sentence. This allows the model to capture long-range dependencies and understand the relationship between words, even in complex sentences. This allows the model to produce more accurate translations and other NLP tasks. The Transformer model also uses a multi-headed self-attention mechanism which allows the model to process different aspects of the input sentence in parallel. This increases the speed and accuracy of the model, and helps it to understand the context of a sentence better.

The Transformer model in NLP is trained with a special type of training called “masked language modeling”. This means that the model is given an input sentence where some of the words are hidden, and it is asked to predict what the missing words are. This helps the model to learn more about the context of sentences and to better understand the relationships between words.

Overall, the Transformer model has revolutionized the field of NLP and is used in many applications today. It is a powerful, accurate, and efficient model that has enabled NLP tasks to be done much faster and more accurately than ever before.

## 2.2 Multitask Learning

Multitask learning [10] is a machine learning technique in which multiple tasks are solved simultaneously. It is based on the idea that multiple related tasks can be better solved when they are trained together instead of training each separately. This allows the model to use shared information between tasks, making it more accurate and efficient. For example, a multitask learning model might be used to simultaneously predict the sentiment of a movie review and the genre of the movie. It might be able to make more accurate predictions by utilizing the relationship between the sentiment and the genre.

Multitask learning is often used for natural language processing tasks, such as language translation, text classification, and question answering. It can also be used for computer vision tasks, such as object detection, image segmentation, and facial recognition. By utilizing the shared information between tasks, multitask learning models can improve the accuracy and efficiency of machine learning models.

### 2.2.1 Gate Layer

Gate layer in multi-task learning [11] and mixture of experts [12] is a method of regularization that helps to improve the performance of a model trained on multiple tasks. It works by assigning weights to each task, allowing the model to focus more on certain tasks over others. This helps to reduce the risk of overfitting, as the model can learn from a more balanced dataset and is less likely to overfit on one task. The gate also helps to avoid the interference between tasks and encourages the model to learn from each task independently. This not only improves the overall accuracy of the model, but also helps to prevent catastrophic forgetting, which is when the model forgets the previously learned knowledge when training on a new task. Furthermore, gate in multi-task learning can improve the generalization performance of a model by allowing it to learn from multiple tasks simultaneously. This can be done by assigning different weights to different tasks, allowing the model to focus more on certain tasks while ignoring others. By doing this, the model can learn from a more diverse dataset and can better generalize its knowledge to new tasks.

Additionally, the gate layer in multi-task learning can also be used to improve the interpretability of the model. By assigning different weights to different tasks, the model's focus can be better understood, and it can be easier to identify which tasks the model is particularly good at, and which tasks it struggles with. This can be particularly useful in industries such as healthcare, where the interpretability of a model is crucial for understanding its performance and making decisions based on its predictions.

Another advantage of using the gate layer in multi-task learning is its ability to adapt to changes in the data. As the data used in different tasks can change over time, the gate layer can be adjusted to reflect these changes, allowing the model to continue to learn and improve its performance. This is particularly useful in industries such as finance, where the data used to make predictions can change rapidly.

Overall, the gate layer in multi-task learning is a powerful tool that can be used to improve the performance, interpretability, and adaptability of a model. Its ability to assign weights to different tasks allows the model to focus on certain tasks while ignoring others, which can help to reduce the risk of overfitting and improve its generalization performance. This makes it a valuable tool for businesses and industries looking to make data-driven decisions and stay ahead of the competition.

# Chapter 3

## Related Work

This chapter will provide an overview of the existing research in the field and the most recent developments in the area.

### 3.1 Sentiment Analysis

Sentiment analysis [13] is a form of natural language processing (NLP) that is used to identify and classify opinions expressed within text. The goal of sentiment analysis is to identify the opinion of the person or source associated with the text, which can be positive, negative, or neutral. This is done by using a combination of machine learning algorithms and dictionaries of known sentiment-bearing words. Many algorithms are trained on a corpus of text known to contain sentiment-bearing words, which are then used to identify sentiment in new text.

Sentiment analysis is used in a variety of industries, from marketing to customer service. It can be used to identify customer sentiment about a product or service, or

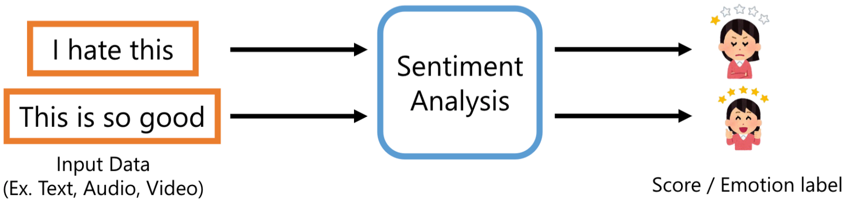

Figure 1: Overview of Sentiment Analysis

to analyze the sentiment of a particular author or publication. It can also be used to identify political sentiment in news articles, or to detect the overall sentiment of a conversation on social media. Sentiment analysis can be used to gain valuable insights into customer sentiment and opinion, which can help businesses make better decisions and improve their customer experience.

Additionally, sentiment analysis can be used to track brand reputation and monitor customer satisfaction over time. By consistently analyzing customer sentiment, businesses can make adjustments and improvements to their products or services to better meet customer needs and expectations.

Another key application of sentiment analysis is in the financial industry. By analyzing news articles and social media posts related to a particular company or industry, investors and traders can gain valuable insights into market sentiment and make more informed investment decisions. This is particularly useful for identifying early warning signs of market changes or potential risks.

Overall, sentiment analysis is a powerful tool that can be used to gain insights and make data-driven decisions in a variety of industries. Its ability to identify and classify opinions expressed in text can provide valuable insights that can help businesses improve their products, services, and customer experience. With the growing amount of text data available, sentiment analysis is becoming an increasingly important tool for businesses to stay ahead of the competition and make better decisions.

There are many possible ways to categorize Sentiment Analysis. In this paper Sentiment Analysis can be divided into three distinct types: Binary Sentiment Analysis, Fine-grained Sentiment Analysis, and Emotion Detection. [1]

Binary Sentiment Analysis or Positive/Negative Sentiment Analysis is a type of natural language processing (NLP) task used to assess sentiment in text. It is a form of sentiment analysis that seeks to classify a piece of text as either positive or negative. Binary sentiment analysis is typically used to quickly identify the sentiment of a large volume of text and can be used to inform decisions in fields such as marketing and customer service. The process starts by collecting data

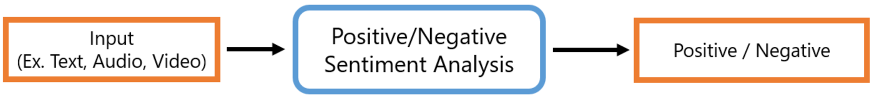

Figure 2: Overview of Positive/Negative Sentiment Analysis

from various sources, such as reviews, comments, and social media posts. This data is then preprocessed to remove irrelevant words, such as those that do not have a direct impact on the sentiment of the text. After the preprocessing step, the text is split into individual sentences that can be analyzed for sentiment. The next step is to employ an algorithm, such as a machine learning model, to classify each sentence as either positive or negative. The output of the algorithm is then used to calculate an overall sentiment score for the text. Binary sentiment analysis provides an efficient and cost-effective way to quickly analyze the sentiment of a large volume of text and can be used to inform decisions in fields such as marketing and customer service.

Fine-grained sentiment analysis [14] is a form of sentiment analysis that goes beyond the binary classification of sentiment (i.e. positive or negative). It allows for a more nuanced understanding of sentiment by going beyond just two categories and instead allowing for a range of sentiment to be determined. For example, fine-grained sentiment analysis might classify sentiment on a five point scale, from strongly negative to strongly positive. In contrast, binary sentiment analysis simply categorizes sentiment as either positive or negative.

Fine-grained sentiment analysis is more complex and requires more sophisticated algorithms than binary sentiment analysis. Because of this, it is better suited for situations where a more nuanced understanding of sentiment is needed, such as in customer feedback or opinion mining. It can also be used to determine the intensity of sentiment, for example whether a comment is slightly positive or strongly positive. This type of analysis can be used to better understand customer

Input
(Ex. Text, Audio, Video)
Fine-grained
Sentiment Analysis
Score or Rating

Figure 3: Overview of Fine-grained Sentiment Analysis

sentiment and make more informed decisions.

Another advantage of fine-grained sentiment analysis is that it is more accurate than binary sentiment analysis. This is because it is able to capture subtle differences in sentiment that may be missed by binary classification. For instance, it can distinguish between a comment that is merely positive and one that is enthusiastic.

Emotion detection [15] is a type of sentiment analysis that goes beyond simply determining whether something is positive or negative. Unlike regular sentiment analysis, which focuses primarily on the sentiment expressed in the text, emotion detection takes into account the emotion behind the sentiment. It attempts to identify the emotions that are felt or implied by a text, such as joy, fear, anger, disgust, or surprise. By analyzing the sentiment and emotion of a text, emotion detection can provide a more complete picture of how people are feeling and thinking about something. This can be useful in many scenarios, such as determining how customers feel about a product or service, or understanding how people are reacting to a particular event or news story. Emotion detection can also be used to detect the emotions of social media users, allowing businesses to better understand the needs and opinions of their target audience. As emotion detection technology continues to improve, we can expect to see more applications for it in various fields.

## 3.2 Multimodal NLP

Multimodal NLP is an area of Natural Language Processing that focuses on the analysis of multiple forms of data. It combines multiple forms of information such as text, images, audio, and video to interpret and understand the content. The

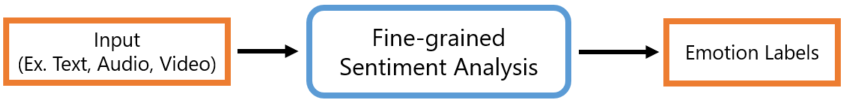

Figure 4: Overview of Emotion Detection

goal of multimodal NLP is to derive insights from the data and to make better decisions. This type of NLP is particularly useful in areas such as computer vision and speech recognition.

Multimodal NLP builds on the existing Natural Language Processing techniques such as sentiment analysis, entity extraction, and topic modeling. It incorporates additional sources of data such as images, audio, and video to better interpret the content. This allows the machine to better understand the context of the text, images, and audio. For example, with speech recognition, it can better understand the context of the words and phrases used in order to accurately interpret the speech.

The applications of multimodal NLP are vast and range from healthcare to autonomous vehicles. In healthcare, it can be used to detect diseases and other medical conditions from patient speech and images. In autonomous vehicles, it can be used to detect objects in the environment and to interpret navigation instructions.

Multimodal NLP is also being used in the field of education. For example, it can be used to create more engaging and interactive learning experiences. It can be used to analyze student responses and to tailor educational materials to their needs. This can help students learn more quickly and effectively.

Furthermore, multimodal NLP can be used in the field of marketing. It can be used to analyze customer feedback and to better understand customer preferences. This can help companies create more effective marketing campaigns and tailor their products and services to the needs of their customers.

Multimodal NLP is an exciting field of research that has the potential to revolutionize the way humans interact with machines. It has the potential to make machines more intelligent and to enable them to better understand and respond to human language. As this field of research continues to develop, it is likely to have a profound impact on the way humans interact with technology.

Multimodal NLP is becoming increasingly important in the emerging field of intelligent systems. It has the potential to enable machines to interact with humans in more natural and intuitive ways. For example, it could be used to build conversational agents that understand the user’ s intent from a combination of text, images, and audio. It could also be used to create virtual assistants that can interpret a user’ s voice or facial expressions.

In addition, multimodal NLP could be used to build systems that make smarter decisions. For example, it could be used to create autonomous robots that can interpret the environment and make decisions based on a combination of text, images, and audio. This would enable robots to better understand the world around them and make decisions more intelligently.

In conclusion, multimodal NLP is an exciting and rapidly growing field of research that has the potential to revolutionize the way humans interact with machines. It has the potential to enable machines to better understand and respond to human language and to make smarter decisions. It is likely that in the near future, multimodal NLP will become increasingly important in the development of intelligent systems.

### 3.3 Multimodal Sentiment Analysis

Multimodal sentiment analysis is an emerging field of research that deals with the analysis of sentiment in multimedia data. It combines the use of natural language processing (NLP), computer vision, and machine learning techniques to identify and classify sentiment in multimedia content such as images, videos, and audio. This field of research has gained a lot of attention in recent years, as it is considered to be able to provide more comprehensive insights into sentiment than relying on a single modality. This paper aims to provide a comprehensive review of the state-of-the-art methods in multimodal sentiment analysis, which includes both traditional and deep learning approaches. In order to provide a thorough overview of the literature in this field, a variety of related works from the fields of NLP, computer vision, and machine learning are included in this chapter.

### 3.3.1 Transformer-based Joint Encoding

Recognizing emotions from multimedia is a difficult challenge. Historically, emotion recognition has been done using separate signals, such as audio, video or text. However, with the implementation of Deep Learning, they can create new frameworks which use multiple signals simultaneously, making use of the joint data gathered from the different sources. The paper proposes a solution based on principles of Machine Translation and Visual Question Answering. Not only is this method extremely efficient, but it is also a viable option for Sentiment Analysis and Emotion Recognition. The results demonstrate that they can compete with, and even exceed, the current best results on the CMU-MOSEI dataset.

The model is based on the Transformer architecture [9], which eliminates the need for recurrent connections and instead uses attention-based global dependencies and Feed-Forward Neural Networks (FFNs) [16] to encode sequences. This results in a much more parallelizable model than the Recurrent Neural Network (RNN) [17]. The monomodal encoder is composed of a stack of identical blocks, each with its own set of parameters. Each block has two sub-layers with a residual connection and layer normalization around each of them. The attention mechanism involves combining a Query with a Key to generate a Context-aware attention map. For the multimodal transformer approach, The authors add a dedicated transformer for each modality. The authors also propose three additional ideas to enhance it: joint-encoding, modular co-attention, and a glimpse layer at the end of each block.

### 3.3.2 Multilogue-Net

Multilogue-Net [18] presents a new approach for analyzing emotions and sentiment in conversations. The authors propose a new model called Multilogue-Net, which is a context-aware recurrent neural network (RNN) that can analyze both text and audio data in conversations.

The authors begin by discussing the limitations of traditional emotion detection and sentiment analysis methods, which have focused on analyzing individual sentences or utterances, rather than taking into account the context of the entire conversation. They argue that this approach is not well-suited for understanding emotions and sentiment in real-world conversations, which are often more complex and nuanced than single sentences.

To address this issue, the authors propose Multilogue-Net, which is designed to analyze both text and audio data in a conversation and take into account the context of the entire conversation. The model is trained on a dataset of conversations that includes both text and audio data, as well as labels for the emotions and sentiments expressed in each conversation.

The authors evaluate Multilogue-Net on a number of benchmarks and find that it outperforms existing approaches for emotion detection and sentiment analysis in conversations. They also show that the model is able to accurately detect emotions and sentiments in real-world conversations, demonstrating its potential for practical applications. The authors also mention that their proposed model is the first one to combine audio and text modality in the same model.

The authors also point out that the model can be used in a variety of applications such as customer service interactions, virtual assistants, and human-computer interactions. The model can also be used in other applications such as medical interviews, legal proceedings, and mental health assessments.

In conclusion, this paper presents a new approach for analyzing emotions and sentiment in conversations that takes into account the context of the entire conversation, and demonstrates that this approach can improve the accuracy of emotion detection and sentiment analysis in real-world conversations. The proposed model can be used in a variety of applications, and it's the first one to combine audio and text modality in the same model, making it more powerful than other models in the field.

## 3.4 Automatic Speech Recognition

Automatic Speech Recognition (ASR) [3] is a technology that allows the recognition and translation of spoken language into written or computer-readable form. It is used to assist in a variety of tasks, ranging from voice-controlled interactive systems to automated customer service systems. Its applications can be found in industries ranging from medical and legal to consumer electronics and automotive.

ASR systems are typically composed of a speech recognition engine, which is responsible for recognizing the spoken words, and a language model, which is used to help the speech recognition engine to interpret the words correctly. The speech recognition engine is usually based on a hidden Markov model [19] which is a statistical model that can predict the likely sequence of words given a set of audio input. This model is trained using a large corpus of data to accurately model how humans speak. The language model is used to help the engine interpret the audio input, by providing context to the words that are spoken.

In order to achieve accurate speech recognition, ASR systems require large amounts of data and the use of powerful computing resources. This data is used for the training of the speech recognition engine and language model. The data used for training is generally collected from real-world environments, and may be collected through user studies, audio recordings, or crowdsourcing [20]. Once the data is collected, it must be processed and transcribed into a usable format for the speech recognition engine.

ASR technology is becoming increasingly popular in a variety of industries and applications. It is used in voice-controlled interactive systems, such as virtual assistants, automated customer service systems, and speech-based search engines.

It is also used in medical and legal applications, such as transcription of medical records and legal documents.

In addition to its current applications, ASR technologies are being used to develop new and innovative solutions. For example, ASR can be used to develop natural language interfaces, which are designed to interact with a user in a more natural fashion. This type of interface could be used to facilitate interactions with virtual assistants, or to develop interactive robots. ASR can also be used to develop more accurate automated translations, allowing for easier communication between different languages.

ASR technology has the potential to revolutionize the way we interact with computers and the world around us. As ASR technologies continue to advance and become more affordable, they will become increasingly more accessible and available to the general public. As the technology becomes more widespread, it is expected to have a profound impact on the way we work, communicate, and interact with one another.

Finally, ASR technology is also used in consumer electronics, such as voice-controlled televisions and voice-enabled car navigation systems. ASR technology is continually evolving, and is expected to become increasingly ubiquitous in our day-to-day lives.

### 3.4.1 wave2vec

Wave2vec [4] is an innovative natural language processing (NLP) technique that uses a neural network to understand and generate context from audio signals. It uses an encoder-decoder architecture [21] to process audio data and extract meaningful information from it. Wave2vec works by first encoding audio signals in a vector form and then decoding the vectors to generate meaningful context. This process is a form of unsupervised learning that allows Wave2vec to learn and generate context from the raw audio data it receives.

The encoder component of Wave2vec takes audio data as input and converts it into a vector representation. This vector representation is then used by the decoder to generate meaningful context from the audio signals. The encoder is designed to be able to understand and recognize features in the audio data such as pitch, volume, timbre and other acoustic properties. It then converts these features into a vector representation that can be used by the decoder.

The decoder component of Wave2vec is responsible for generating context from the vector representation created by the encoder. It uses a recurrent neural network(RNN) [17] to process the vector data and generate meaningful context. The decoder is trained on a large corpus of audio data so that it can understand the context of the audio signals it receives.

Overall, Wave2vec is an innovative NLP technique that can be used to generate meaningful context from audio signals. It is a powerful tool that can be used to create natural language processing applications and can be used to improve the accuracy of speech recognition systems.

Wav2vec 2.0 [22] is the latest version of the wav2vec speech representation learning algorithm from Facebook AI Research. It is an unsupervised approach to learning speech representations from raw audio signals. Wav2vec 2.0 is different from its predecessor wav2vec in several ways.

Firstly, Wav2vec 2.0 uses a new tokenization strategy that results in higher quality representations. It uses a self-attention based approach that allows the model to learn representations in a more natural way. This means that the model can learn to represent longer phrases without sacrificing accuracy.

Secondly, Wav2vec 2.0 uses a larger training dataset. This makes it more robust and allows the model to better capture the nuances of speech. Additionally, the larger training dataset also makes it easier to train the model since more data points are available.

Thirdly, Wav2vec 2.0 has a much larger model. This allows it to learn a much richer representation of speech, which results in higher quality representations.

Finally, Wav2vec 2.0 also uses a new optimizer. This allows the model to be trained more efficiently and accurately.

Overall, Wav2vec 2.0 is an improved version of the original wav2vec algorithm.

It has a larger training dataset, a larger model, and a better tokenization strategy that results in higher quality representations.

### 3.4.2 Whisper

Whisper [23] is an text generation model developed by OpenAI. The model is based on a transformer-based architecture which consists of a sequence of self-attention layers and feed-forward layers. The self-attention layers allow the model to attend to different parts of the input sequence and extract important features from it. This is done by computing a matrix of attention scores, which are then used to determine which parts of the input the model should focus on. The feed-forward layers are used to further process the extracted features and make predictions. It utilizes a deep learning approach to generate natural-sounding speech from a given input text. The model is designed to generate speech that sounds more human-like than traditional text-to-speech systems, and can generate both spoken and written text. OpenAI Whisper uses a combination of both self-attention and feed-forward layers to enable it to handle complex NLP tasks. It has also been shown to work well on a variety of tasks, including question answering and language modeling.

The OpenAI Whisper model is trained on a massive corpus of natural language data, allowing it to learn the nuances of language and the patterns of human communication. The model is capable of producing both single-sentence and multi-sentence responses to various prompts. It can also generate responses to questions, interpret sentiment, and generate personalized responses to queries.

The OpenAI Whisper model is particularly useful for chatbot development and natural language processing applications. Its advanced language capabilities and flexibility make it an ideal tool for conversational AI applications. Additionally, it can be used to generate personalized responses to customer queries, allowing companies to provide customers with more personalized customer service.

OpenAI Whisper is a powerful tool in the development of conversational AI applications, and its flexibility and advanced language capabilities make it an excellent choice for businesses looking to provide personalised customer service. The model is easy to use and is able to generate natural-sounding speech in a variety of contexts. With its advanced language capabilities and flexibility, OpenAI Whisper is set to become a powerful tool in the development of conversational AI applications.

# Chapter 4

## Noise-Tolerant Multimodal Sentiment Analysis

Multimodal sentiment analysis is a challenging task that requires the integration of different modalities in order to accurately assess the sentiment of a piece of text, audio, or visual. In this paper, we present a novel method for multimodal sentiment analysis that combines linguistic, acoustic, and visual features to capture sentiment in audio-visual media. In the following section, we will introduce the proposed method in detail and discuss its merits.

In this chapter, we will first explain the method for feature extraction that we used in order to retrieve feature vectors from the data in each modality. Then we will explain the overview for each part of our proposed method which consist of two main part: monomodal model and multimodal model.

### 4.1 Feature Extraction

In this paper, we used three types of feature extraction methods or tools for each modality which will be explained in the following subsections.

Raw data
Feature extraction
Feature vector

Figure 5: Overview of Feature Extraction

### 4.1.1 Linguistic features

Glove embedding [24] is a type of word embedding technique that uses a pre-trained model to create numerical representations of words. It is based on the co-occurrence matrix of words in large corpora. Glove embedding has gained popularity as a powerful technique for learning word representations from data. It is especially useful for natural language processing tasks due to its ability to capture the context of words. Glove embedding uses a technique called matrix factorization, which is similar to principal component analysis. This technique allows it to effectively capture the semantic relationships between words as well as the syntactic relationships. Glove embedding is also useful for understanding natural language since it captures the meaning of words in their context. It has been used to improve the performance of many natural language processing applications such as sentiment analysis, question answering, and text classification.

Glove embedding has also been used in many other applications, such as machine translation, information retrieval, and recommendation systems. Glove embedding is a computationally efficient technique, requiring only a few minutes to train a model on a standard laptop. This makes it a popular choice for many natural language processing tasks, as it does not require a large amount of computing power. Glove embedding is also known for its ability to capture the meaning of words in their context, which is important for understanding natural language. Furthermore, it is highly flexible and can be used to represent a variety of languages and text types.

We uses Glove embedding for our feature extraction in linguistic modality. We embed each word in a vector of 300 dimensions using GloVe. If a word from is not presented in the vocabulary list, we replaced it with the unknown token.

### 4.1.2 Acoustic features

OpenSmile [25] is an open source software library for audio signal processing. It is the most widely used open source library for audio feature extraction and provides a comprehensive set of tools for extracting audio features from speech and music recordings. OpenSmile can also be used for speech recognition, natural language processing, audio segmentation, and sound event detection.

OpenSmile has a modular design, allowing users to customize the feature extraction process. It provides a range of pre-defined feature extraction algorithms, including low-level descriptors (e.g., energy, pitch, and spectral shape), higher-level descriptors (e.g., formant, glottal, and prosodic features), and machine learning-based features (e.g., phoneme recognition, emotion recognition, and speaker recognition). The library also includes tools for acoustic modeling, audio segmentation, and acoustic scene classification. OpenSmile is a powerful and versatile tool that has been successfully used in several research projects and commercial applications. It has been applied to speech recognition, natural language processing, speaker recognition, music classification, and emotion recognition, among many other tasks. It is also used in the development of commercial products such as voice-enabled virtual assistants, automated customer service systems, and music recommendation systems.

We used OpenSmile for acoustic feature extraction. The 384-dimension features were extracted with the same method as they were in INTERSPEECH 2009 Emotion Challenge feature set (IS09) [26].

### 4.1.3 Visual features

OpenFace [27] is a facial recognition and analysis library. OpenFace provides state-of-the-art facial analysis algorithms, allowing developers to easily create applications to detect and recognize faces in images and video streams.

OpenFace works by first extracting facial features from an image or video stream. It then uses a deep learning neural network to compare those features to a large database of known faces. Once it has identified the faces, OpenFace can track them in real-time and provide rich information about the facial expressions and movements of the person in the image or video. It can also be used to create 3D models of a person's face, as well as to analyze facial expressions and recognize

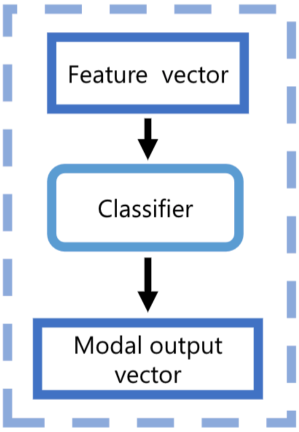

Figure 6: Monomodal Model

emotions.

This library is a powerful tool that can be used to create a wide variety of applications, from facial recognition to emotion detection. It is being used in a variety of industries, from security and surveillance to healthcare and marketing. OpenFace can be used to create applications such as facial authentication systems and virtual assistants that can recognize and respond to facial expressions and emotions.

OpenFace is an efficient and powerful tool that is revolutionizing the way facial recognition and analysis is being used in the modern world. With its advanced facial recognition algorithms, OpenFace enables developers to create powerful applications that can be used in a variety of industries.

The 66-dimension features extracted by OpenFace are consist of 48-dimension landmark features and 18-dimension action unit features.

## 4.2 Monomodal Classification Model

The monomodal model is composed of a stack of multiple blocks of identical classification model [28] corresponding to the maximum number of modalities presented in the data. In this paper, the data mainly consists of three modalities which are linguistic, acoustic, and visual.

The model are based on a simple multilayer perceptron (MLP) classification model. [8] Each block of monomodal model contains two hidden layer and one output layer. Both hidden layers are applied ReLU activation [29] and the output layer has Softmax activation [30].

The output vector will be adjusted based on the number of categories or labels presented in the classification dataset.

## 4.3 Multimodal Model

Our idea of multimodal model are based on the way to combine the result of each output vector from the monomodal model for each modality.

In this paper, we considered trying to implement multiple ways of combining the output vectors from each modality. We implemented three methods in order to combine the results: concatenation, averaging, and gate layer. We will further explain the methods in following subsection.

### 4.3.1 Concatenation Model

The concatenation model or concat model combines the result from each monomodal model by simply concatenating all the vector together. Concatenation of vectors is a process of joining two or more vectors together. This can be done by adding the elements of one vector to the end of another vector. This process can be used to combine multiple vectors into a single unit. The result of the concatenation is a new vector with the combined elements of all the original vectors. However, in order to be able to separate each output vector, we also

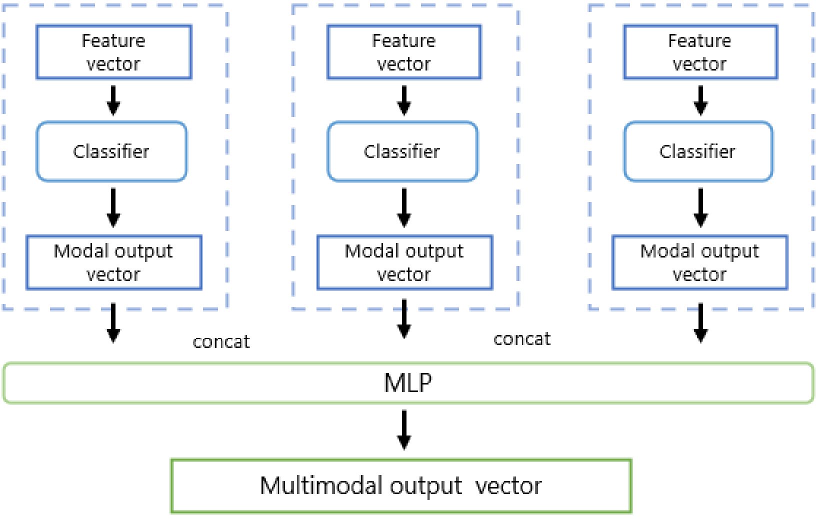

Figure 7: Concat Model

append the blank token at the beginning and the end of each output vector from our monomodal model. Thus, the final MLP model and recognize that there are three separated vectors from each modality: linguistic, acoustic, and visual.

### 4.3.2 Average Model

In this model, we combines the result from each monomodal model by using the average of output vectors from each modality. The average of vectors is a numerical measure of the middle point or central tendency of a set of vectors. It is calculated by taking the sum of all the elements of the vectors and then dividing it by the total number of elements. This will give the average value for all the elements in the vector. The average of vectors can be used to measure how close the elements of the vector are to each other, and it can also be used to determine how much variability there is in the set of vectors. Thus, the final MLP model won"t recognize that there are three separated vectors from each modality and only know the average value of them.

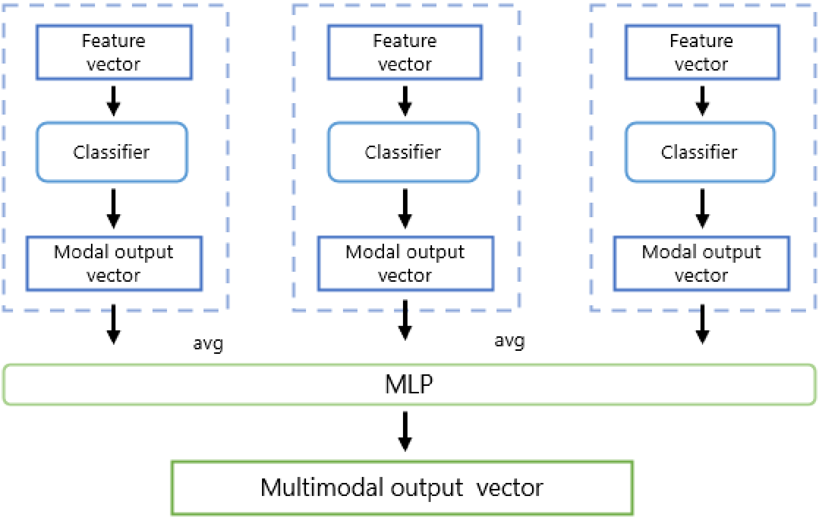

Figure 8: Avg Model

### 4.3.3 Gate Layer Model

A Gate Layer in a Mixture of Experts is a network component that uses a set of weights to decide which of the experts should be used for a given task. It works by weighing the output of each expert based on the input and then choosing the one with the highest weight. The weights are learned during training, and can be adjusted to better match the given task. The gate layer helps to ensure that the most appropriate expert is chosen for the task, which helps to improve the overall accuracy of the system.

In this paper, we instead applied the gate layer to control the flow of information going from monomodal model from each modality through the final multimodal MLP which produce the final classification output. The gate layer will be trained with all feature vectors with the objective to determine the quality and how trustworthy is for each output vector from the individual monomodal model for each modality.

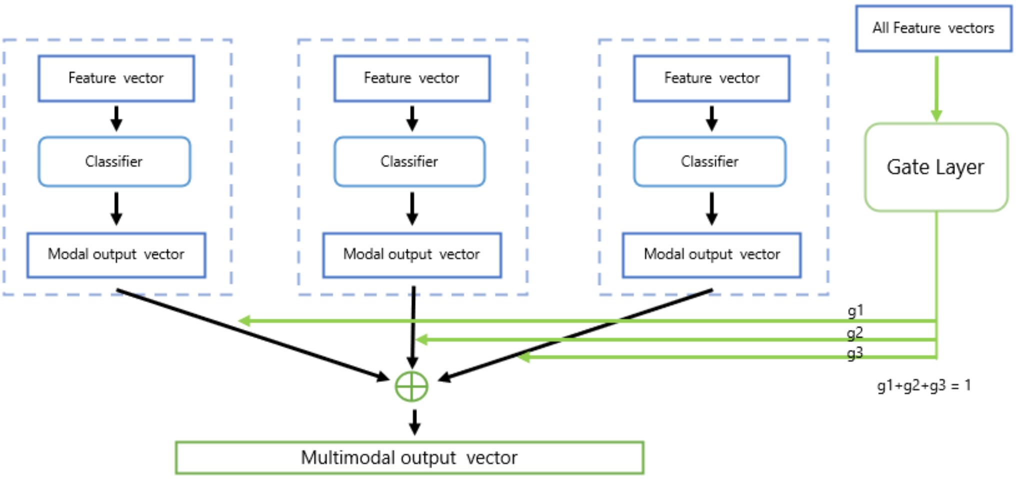

Figure 9: Gate Layer Model

# Chapter 5

## Evaluation

In this chapter, we will discuss the overview of dataset used in our research and showing the results and analysis for each experiments.

### 5.1 Dataset

In this section, we will be discussing the use of a specific dataset in order to analyze this particular research question about Multimodal Sentiment Analysis.

Hazumi dataset [5] proposes a novel multimodal human-agent dialogue corpus with annotations at both utterance and dialogue levels. The corpus consists of video recordings of interactions between participants and an agent in a simulated home environment. The recordings are annotated with a variety of linguistic, visual, and contextual features, providing a comprehensive dataset for researchers interested in studying multimodal dialogue.

The corpus is designed to facilitate research in a variety of dialogue-related tasks, including natural language understanding, dialogue management, and multimodal generation. The dataset contains over 6,000 utterances and 500 dialogues, each annotated with a variety of features such as dialogue act, discourse structure, and user emotions. Furthermore, the corpus includes a variety of visual features such as facial expressions, gestures, body language, and visual cues.

The authors have developed a web-based annotation platform to facilitate annotation of the corpus. This platform allows annotators to quickly and accurately annotate dialogues with the required features. The authors also provide a set of

| Dataset | Number of Experiments | Number of Dialogues |
|---|---|---|
| Hazumi2010 | 33 | 2798 |
| Hazumi2012 | 63 | 5334 |
| Hazumi2105 | 29 | 2235 |
| Total | 125 | 10367 |

Table 1: The number of dialogues in each Hazumi dataset (online)

tools to analyze the annotated data and to generate dialogues based on the annotations.

Overall, the authors have developed an extensive corpus for studying multimodal dialogue that provides researchers with valuable insights into human-agent interactions. The annotations at both the utterance and dialogue levels allow for a comprehensive analysis of human-agent dialogue, and the tools provided by the authors facilitate further research in this field.

Hazumi dataset is recorded by SANKEN, Osaka University. The data is collected from the dialogue between volunterrs and the agent in the form of Wizard of Oz (WoZ) [31] The experiments are recorded with the name Hazumi followed by year and month the the experiment was conducted, for example, Hazumi2010. In 2020 and 2021, all experiments were conducted online through an online conference tool due to the pandemic situation called COVID-19.

In this paper, we mainly focus on this version of Hazumi dataset due to its nature of online recording. The online dataset are consisted of 3 versions: Hazumi2010, Hazumi2012, and Hazumi2105. In these version, the volunteers and agent will have a conversation with the duration of approximately 15 minutes. Each conversation and interaction will be annotated by multiple annotators through a web-based online platform. However, there will be no recording of Kinect [32] and Body movement sensor due to the difficulty in an online setup. The number of Experiments and dialogues is shown in Table1.

| Dataset | Label | | | | | | |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| Hazumi2010 | 6 | 41 | 315 | 3279 | 3162 | 1312 | 166 |
| Hazumi2012 | 17 | 176 | 747 | 4208 | 5301 | 3732 | 1664 |
| Hazumi2105 | 4 | 34 | 252 | 1348 | 1989 | 2183 | 866 |
| Total | 27 | 251 | 1314 | 8835 | 10452 | 7227 | 2696 |

Table 2: The number of labels in each Hazumi dataset (online)

The data contains ELAN files [33] and experiment dump files. In this paper, we used the human transcript, time, and sentiment annotation recorded in ELAN files and the audio and video features provided in experiment dump files. The human transcription in ELAN files is transcribed by one annotator. The transcription are recorded without full stop mark and filler words are marked by symbol F. The sentiment annotations in ELAN files are annotated by the score in the range from 1 to 7, in which 1 is negative sentiment, and 7 is positive sentiment score.

As shown in Table 2, the data is more likely to be annotated within the score in the middle, approximately in the range of score 3 to 5. The annotator rarely uses an extreme score of 1 and 7. The sentiment annotations is in the range from 1 to 7 scoring, in which 1 is negative sentiment, and 7 is positive sentiment score. The lack of score of 1 may be cause by the emotional bias or guilt of rating conversation as extremely negative, while it is easier to rate for extremely positive in the form of score of 7.

In Machine Learning, train test split is a powerful tool for evaluating machine learning models. It is used to split a dataset into two parts: a training set and a testing set. The training set is used to train the model and the testing set is used to evaluate the performance of the model. The train test split helps prevent overfitting and allows for a more accurate evaluation of the model. It is an important step in any machine learning project and should be used whenever possible.

| Dataset | Number of Experiments | Number of Dialogues |
|---|---|---|
| Train | 101 | 8281 |
| Test | 12 | 1056 |
| Valid | 12 | 1030 |
| Total | 125 | 10367 |

Table 3: The number of dialogues in training/test/validation data from Hazumi dataset (online)

In this paper, we split the data into training data, test data, and validation data in a proportion of approximately 8:1:1. The data split by speakers is as shown in Table 3. The data split by speaker will have no same speaker across the different dataset, for example, there will be no same speaker in test data from the training data. Thus, all test data will always be somehow unseen to the model.

### 5.2 Experimental Results

In this section, we will explain the results from our experiment with our proposed method. We experimented with all of our three model: concat odel, avg model, and gatelayer model. We compared the results with the results from the Transformer joint-encoder model mentioned in Chapter 3.3.1. The experiments were done on both wav2vec and Whisper ASR model.

We trained the model with both human transcription and transcription genereated by ASR model mentioned above. All model are trained with the mixture of data contain all or some modalities. L represents linguistic modality from human transcription, while L' represents liguistic modality from transcription by ASR model. A represents acoustic modality, and V represent visual modality. The training data are consists of L+A+V, L+A, and L data with the proportion of 4:3:3. The results for each experiment will be shown and discussed in the following subsections.

| Model | Training data | Test Data | 3-label | 7-label |
|---|---|---|---|---|
| | | | Accuracy | Accuracy |
| Concat | Human Transcription | L’+A | 45.3 | 28.1 |
| | | L’+A+V | 49.6 | 30.7 |
| | ASR | L’+A | 45.8 | 28.8 |
| | | L’+A+V | 50.1 | 31.0 |
| Avg | Human Transcription | L’+A | 45.0 | 27.7 |
| | | L’+A+V | 49.0 | 30.1 |
| | ASR | L’+A | 45.2 | 28.0 |
| | | L’+A+V | 49.7 | 30.8 |
| Gate Layer | Human Transcription | L’+A | **51.8** | **30.5** |
| | | L’+A+V | **53.3** | **33.9** |
| | ASR | L’+A | **52.0** | **30.6** |
| | | L’+A+V | 53.7 | 33.9 |
| Joint Transformer | Human Transcription | L’+A | 49.7 | 29.0 |
| | | L’+A+V | 52.7 | 32.6 |
| | ASR | L’+A | 50.3 | 29.8 |
| | | L’+A+V | **54.4** | **34.2** |

Table 4: The wav2vec results evaluated on Hazumi dataset(online). Training data is split by speaker.

| Model | Training data | Test Data | 3-label | 7-label |
|---|---|---|---|---|
| | | | Accuracy | Accuracy |
| Concat | Human Transcription | L'+A | 46.3 | 30.9 |
| | | L'+A+V | 50.1 | 33.1 |
| | ASR | L'+A | 47.2 | 31.5 |
| | | L'+A+V | 51.0 | 33.9 |
| Avg | Human Transcription | L'+A | 47.1 | 31.4 |
| | | L'+A+V | 51.1 | 32.7 |
| | ASR | L'+A | 47.9 | 31.8 |
| | | L'+A+V | 52.0 | 33.1 |
| Gate Layer | Human Transcription | L'+A | **52.9** | **31.9** |
| | | L'+A+V | **54.1** | **34.2** |
| | ASR | L'+A | **53.2** | **32.1** |
| | | L'+A+V | 54.4 | 35.3 |
| Joint Transformer | Human Transcription | L'+A | 50.6 | 29.2 |
| | | L'+A+V | 53.9 | 34.1 |
| | ASR | L'+A | 51.5 | 30.2 |
| | | L'+A+V | **54.7** | **35.2** |

Table 5: The wav2vec results evaluated on Hazumi dataset(online). Training data is randomly split.

### 5.2.1 Results with wave2vec

In this experiment, we test our model on ASR transcription from wav2vec model. We experimented on both the data with regular train test split that randomly split the dialogue data, and the data using train test split considered the split by speakers so that there will be no same speaker data across train and test data.

The results shows that our proposed method slightly outperform baseline method when all modalities are not being presented, while, slightly outperformed by the baseline when there are all three modalities. The multimodal model method in our proposed method which has the best performance overall was the gate layer model. The concat model and avg model has the similar result, in which concat model tend to perform slightly better but not in a significantly amount of improvement.

Comparing the results from regular train test split that randomly split the dialogue data, and the data using train test split considered the split by speakers, we found that the model slightly perform better when trained by the randomly split data. We speculate that the reason of the slight improvement came from the speaker's characteristic or behavior. The data using train test split considered the split by speakers will cause the model to have an unseen and unlearn behavior in the test dataset which make it harder to predict the results and cause a slight drop in the performance. However, the avg model tend to have the least difference in performance drop with the train test split method being changed. We also speculate that the method of averaging the output vectors from monomodal model cause the avg model to not recognize the origin of each output vector and being harder to be affected by the unseen characteristic of the data.

### 5.2.2 Results with Whisper

In this experiment, we test our model on ASR transcription from Whisper model. We also experimented on both the data with regular train test split that randomly split the dialogue data, and the data using train test split considered the split by speakers so that there will be no same speaker data across train and test data.

| Model | Training data | Test Data | 3-label | 7-label |
|---|---|---|---|---|
| | | | Accuracy | Accuracy |
| Concat | Human Transcription | L'+A | 49.2 | 31.7 |
| | | L'+A+V | 52.2 | 34.9 |
| | ASR | L'+A | 49.9 | 32.1 |
| | | L'+A+V | 52.9 | 35.5 |
| Avg | Human Transcription | L'+A | 50.1 | 32.0 |
| | | L'+A+V | 52.0 | 35.1 |
| | ASR | L'+A | 50.5 | 32.1 |
| | | L'+A+V | 53.1 | 35.8 |
| Gate Layer | Human Transcription | L'+A | **52.3** | **34.9** |
| | | L'+A+V | 54.4 | 35.2 |
| | ASR | L'+A | **52.5** | **34.6** |
| | | L'+A+V | 54.8 | 36.6 |
| Joint Transformer | Human Transcription | L'+A | 52.0 | 34.1 |
| | | L'+A+V | **60.3** | **41.0** |
| | ASR | L'+A | 52.2 | 34.2 |
| | | L'+A+V | **61.5** | **41.6** |

Table 6: The Whisper results evaluated on Hazumi dataset(online). Training data is split by speaker.

| Model | Training data | Test Data | 3-label | 7-label |
|---|---|---|---|---|
| | | | Accuracy | Accuracy |
| Concat | Human Transcription | L'+A | 49.5 | 31.7 |
| | | L'+A+V | 52.4 | 35.1 |
| | ASR | L'+A | 50.1 | 32.3 |
| | | L'+A+V | 52.8 | 35.3 |
| Avg | Human Transcription | L'+A | 50.3 | 32.2 |
| | | L'+A+V | 52.5 | 35.8 |
| | ASR | L'+A | 50.9 | 32.4 |
| | | L'+A+V | 53.3 | 36.3 |
| Gate Layer | Human Transcription | L'+A | **52.8** | **35.1** |
| | | L'+A+V | 55.0 | 35.7 |
| | ASR | L'+A | **53.1** | **35.3** |
| | | L'+A+V | 55.3 | 37.1 |
| Joint Transformer | Human Transcription | L'+A | 52.3 | 34.7 |
| | | L'+A+V | **61.2** | **41.7** |
| | ASR | L'+A | 52.9 | 34.8 |
| | | L'+A+V | **62.4** | **41.5** |

Table 7: The Whisper results evaluated on Hazumi dataset(online). Training data is randomly split.

The results also shows that our proposed method slightly outperform baseline method when all modalities are not being presented, while, slightly outperformed by the baseline when there are all three modalities. Similar to wave2vec result, the multimodal model method in our proposed method which has the best performance overall was the gate layer model. The concat model and avg model has the similar result, in which concat model slightly outperperform avg model but not in a significantly amount.

Comparing the results from regular train test split that randomly split the dialogue data, and the data using train test split considered the split by speakers, we also found the similar occurence to wav2vec experiment that the model slightly perform better when trained by the randomly split data. We speculate that the reason of the slight improvement are the same which came from the speaker's characteristic or behavior. The data using train test split considered the split by speakers will cause the model to experience an unseen test dataset which obstruct the prediction of the results and hinder the performance.

In comparison to wav2vec result, we could see that the overall accuracy from Whisper model are higher compared to the wav2vec result. However, With all three modalities being presented, the baseline has a sudden improvement in the accuracy which we could speculate the the quality of the text generated from Whisper is better than wav2vec which make the baseline perform at its maximum potential that it was designed for, while, our proposed method tends to perform better in more noisy environment.

## 5.3 Analysis

In this section, we will further analyze the results and evaluate more into our speculation resulted from the results in experiments mention above in Section 5.2

| ASR Model | CER |
|---|---|
| wav2vec | 39.7 |
| Whisper | 16.1 |

Table 8:  CER results from ASR transcription

| Transcription | Text |
|---|---|
| Human | Youtube とかライブ配信とか見てたらけっこう時間経ちますもんね |
| wav2vec | 遊中部と辛を林心とかみてたら結構時間立ちまそもね |
| Whisper | YouTube とかライブ配信とか見てたら結構時間経ちますもんね |
| Human | もう恒例行事みたいな感じで行ってます |
| wav2vec | も高礼行心みたいな漢字で行てます |
| Whisper | もう恒例行事みたいな感じで言ってます |
| Human | ベトナムではホイアンというところで灯篭流しをみました |
| wav2vec | ベト並ではほやんといと所で登ろ流子を見ました |
| Whisper | ベトナムではウィアンというところで道路を流しを見ました |

Table 9:  Examples of ASR transcription

### 5.3.1 Word error from ASR

We calculated the character error rate (CER) [34] of the transcription from both wav2vec and Whisper data. CER is a measure of accuracy used to evaluate the performance of a speech recognition system. It is a measure of the rate of errors made by the system in recognizing characters (letters and numbers) in a given utterance. CER is calculated by measuring the number of errors that the system makes in recognizing characters in an utterance and dividing it by the total number of characters in the utterance. The results are expressed as a percentage. CER is one of the most commonly used measures of accuracy in speech recognition systems and is used to compare different systems and algorithms.

CER results shown in Table 8 confirmed that the transcription from Whisper are significantly better than wav2ve. Futhermore, with an example transcription shown in Table 9, we could see the error happened in lots of wav2vec transcription, while an error from Whisper rarely occurs unless the text is containing specific words, names of places, or rare vocabulary.

### 5.3.2 Results from an offline Hazumi dataset

In this experiment, we test our concat model on the offline Hazumi dataset. We experimented on only the data with regular train test split that randomly split the dialogue data. The offline dataset are consisted of 3 versions: Hazumi1712, Hazumi1902, and Hazumi1911. In these version, the volunteers and agent will have a conversation with the duration of approximately 15 minutes. Each conversation and interaction will be annotated by multiple annotators through a offline questionaire. There was also a recording of Kinect and Body movement sensor. The number of Experiments and dialogues is shown in Table10. The results in Table 11 shows that the model has similar results in accuracy compared to the concat model trained with online dataset.

| Dataset | Number of Experiments | Number of Dialogues |
|---|---|---|
| Hazumi1712 | 29 | 2422 |
| Hazumi1902 | 30 | 2514 |
| Hazumi1911 | 30 | 2859 |
| Total | 89 | 7795 |

Table 10: The number of dialogues in each Hazumi dataset (offline)

| Test Data | 3-label | 7-label |
|---|---|---|
| | Accuracy | Accuracy |
| L | 45.4 | 25.4 |
| L+A | 50.1 | 32.1 |
| L+A+V | 55.4 | 38.7 |

Table 11: The results evaluated on Hazumi dataset(offline) with concat model.

### 5.3.3 Results from a human transcription

In this experiment, we test our model on an original human trancription. We only experimented on the data with regular train test split that randomly split the dialogue data.

Without all three modalities, the results shows that our proposed method slightly outperform baseline method. However, our gate layer model was slightly outperformed by the baseline when there are all three modalities. Similar to prior result, the gate layer model has the best performance overall in all three of our proposed method . The concat model and avg model also has the similar results like the experiment with both wav2vec and Whisper.

From the speculation we mentioned in Whisper experiment that we could see that the overall accuracy from Whisper model are higher compared to the wav2vec result. We could further confirm that those improvement are similar to a correlation

| Model | Test Data | 3-label | 7-label |
|---|---|---|---|
| | | Accuracy | Accuracy |
| Concat | L | 50.2 | 33.2 |
| | L+A | 51.2 | 33.7 |
| | L+A+V | 53.1 | 36.1 |
| Avg | L | 51.0 | 33.1 |
| | L+A | 51.7 | 33.4 |
| | L+A+V | 52.8 | 35.7 |
| Gate Layer | L | 52.1 | 34.0 |
| | L+A | 53.3 | 35.3 |
| | L+A+V | 54.6 | 36.7 |
| Joint Transformer | L | 49.3 | 30.2 |
| | L+A | 52.5 | 34.4 |
| | L+A+V | 61.2 | 41.4 |

Table 12: The results evaluated on Hazumi dataset(online) with Human transcription

in human transcription result. We could speculate the the quality of the text generated from Whisper is more similar to human transcription which make the model behave similar to the model trained with human transcription.

# Chapter 6

## Conclusion

Research in the field of sentiment analysis has traditionally focused on textual data, but with the growing popularity of online communication, there is an ever-increasing amount of multimodal data available, which has the potential to provide valuable insights into the sentiment of a text. Current sentiment analysis models are limited to a specific domain or context, and thus are not able to accurately interpret sentiment when the text is unavailable, such as in a video or image.

In this paper, we proposed a robust multimodal sentiment classification method that can effectively use multi-modal data in various noisy environments, and showed that our method can retain the model performance when data quality of one or more modalities is low. The proposed method in this study addresses the problem of current multimodal sentiment analysis models by using a gate network and modal independent classifiers to mitigate the effect of data noise and improve the performance of the model.

The result showed that our proposed method slightly outperform baseline method when all modalities are not being presented. Although, being outperformed by the baseline when there are all three modalities. The result confirmed that our proposed method help mitigate the loss in performance when noises or an absence is presented in multimodal data. It is worth noting that this proposed method can also be applied to other areas of natural language processing (NLP) such as language translation, text summarization and dialogue generation.

For future study, the proposed method could be implemented in various applications that involve multimodal NLP or sentiment analysis, and can also be extended to other areas of NLP such as language translation, text summarization, and dialogue generation. With the usage of simple classification and already-prepared feature extraction method, the model can be further improved with a specifically design monomodal model for each specific task and a specific type of feature extraction method for some kind of multimodal data that isn't mentioned in this study.

# Acknowledgements

I would like to begin by thanking my advisor, Associate Professor Naoki Yoshinaga for his insight and guidance throughout the course of this project. I owe a tremendous debt of gratitude for his helpful comments. His knowledge in NLP have been a great help to me in understanding the NLP concepts and the research. Without him, I would have been lost. His enthusiasm and knowledge inspired me to delve further into the subject and discover the answers I was looking for. His suggestion and guidance helped me got through this NLP project, even with my trouble with Japanese language, the research topic, or even the academic conference. The research would not be completed without all those suggestion, comments, and supports.

I also would like to thank Professor Masashi Toyoda and Associate Professor Kazuo Goda for their kind guidance during the conferences and their valuable comments and advice for improvement during laboratory meeting and the thesis defense practices.

My appreciation also extends to those who provided me with invaluable feedback and advice throughout the duration of this project. I am especially thankful for the willingness of the seniors, colleagues, and juniors I consulted who gave me their valuable time and expertise. I am also thankful to the people who have taken the time to review and critique my work. Their constructive comments and feedback helped me to make this project better.

I am also thankful for all the laboratory staff members who help me through all the documents for both university and outside academic presentation despite the trouble with my difference in document compared to Japanese students. I would not have such an easy time dealing with all documents without their help.

I am grateful staff at my university who provided me with support as I am one of the international students. I thank them for their assistance.

I would also like to thank the University of Tokyo who provided me with the opportunity to pursue this project. It has been a great learning experience for me.

I am also grateful to my family and friends for their unwavering support. I am especially thankful to my parents for their encouragement and understanding that made it possible for me to embark on this journey. Without their love and support, this research would not have been possible.

Once again, I am deeply grateful to everyone who has helped and supported me for this project. I could not have done this without them.

# References

[1] Dang, N. C., Moreno-GarcÃŋa, M. N. & De la Prieta, F. Sentiment analysis based on deep learning: A comparative study. *Electronics* **9** (2020). URL https://www.mdpi.com/2079-9292/9/3/483.

[2] Ngiam, J. *et al.* Multimodal deep learning. In Getoor, L. & Scheffer, T. (eds.) *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, 689–696 (Omnipress, 2011). URL https://icml.cc/2011/papers/399_icmlpaper.pdf.

[3] Fendji, J. L. K. E., Tala, D. C. M., Yenke, B. O. & Atemkeng, M. Automatic speech recognition and limited vocabulary: A survey (2021). URL https://arxiv.org/abs/2108.10254.

[4] Schneider, S., Baevski, A., Collobert, R. & Auli, M. wav2vec: Unsupervised pre-training for speech recognition (2019). URL https://arxiv.org/abs/1904.05862.

[5] Komatani, K. & Okada, S. Multimodal human-agent dialogue corpus with annotations at utterance and dialogue levels. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 1–8 (2021).

[6] Mcculloch, W. S. & Pitts, W. H. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics* **5**, 115–133 (1943).

[7] Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning Representations by Back-propagating Errors. *Nature* **323**, 533–536 (1986). URL http://www.nature.com/articles/323533a0.

[8] Popescu, M.-C., Balas, V. E., Perescu-Popescu, L. & Mastorakis, N. Multilayer perceptron and neural networks. *WSEAS Trans. Cir. and Sys.* **8**, 579â588 (2009).

[9] Vaswani, A. *et al.* Attention is all you need. In Guyon, I. *et al.* (eds.) *Ad-*

*vances in Neural Information Processing Systems*, vol. 30 (Curran Associates, Inc., 2017). URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[10] Caruana, R. Multitask learning - machine learning. URL https://link.springer.com/article/10.1023/A:1007379606734.

[11] Crawshaw, M. Multi-task learning with deep neural networks: A survey (2020). URL https://arxiv.org/abs/2009.09796.

[12] Shazeer, N. *et al.* Outrageously large neural networks: The sparsely-gated mixture-of-experts layer (2017). URL https://arxiv.org/abs/1701.06538.

[13] Dang, N. C., Moreno-GarcÃŋa, M. N. & De la Prieta, F. Sentiment analysis based on deep learning: A comparative study. *Electronics* **9** (2020). URL https://www.mdpi.com/2079-9292/9/3/483.

[14] Guo, X., Yu, W. & Wang, X. An overview on fine-grained text sentiment analysis: Survey and challenges. *Journal of Physics: Conference Series* **1757**, 012038 (2021). URL https://dx.doi.org/10.1088/1742-6596/1757/1/012038.

[15] Emotion recognition and detection methods: A comprehensive survey (2020).

[16] Bebis, G. & Georgiopoulos, M. Feed-forward neural networks. *IEEE Potentials* **13**, 27–31 (1994).

[17] Schmidt, R. M. Recurrent neural networks (rnns): A gentle introduction and overview. *CoRR* **abs/1912.05911** (2019). URL http://arxiv.org/abs/1912.05911. 1912.05911.

[18] Shenoy, A. & Sardana, A. Multilogue-net: A context-aware RNN for multi-modal emotion detection and sentiment analysis in conversation. In *Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML)*, 19–28 (Association for Computational Linguistics, Seattle, USA, 2020). URL https://aclanthology.org/2020.challengehml-1.3.

[19] Rabiner, L. & Juang, B. An introduction to hidden markov models. *IEEE ASSP Magazine* **3**, 4–16 (1986).

[20] Ghezzi, A., Gabelloni, D., Martini, A. & Natalicchio, A. Crowdsourcing: A review and suggestions for future research. *Wiley-Blackwell: International Journal of Management Reviews* (2018).

[21] Sutskever, I., Vinyals, O. & Le, Q. V. Sequence to sequence learning with neural networks (2014). URL https://arxiv.org/abs/1409.3215.

[22] Baevski, A., Zhou, H., Mohamed, A. & Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. URL https://arxiv.org/abs/2006.11477.

[23] Radford, A. *et al.* Robust speech recognition via large-scale weak supervision (2022). URL https://arxiv.org/abs/2212.04356.

[24] Pennington, J., Socher, R. & Manning, C. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543 (Association for Computational Linguistics, Doha, Qatar, 2014). URL https://aclanthology.org/D14-1162.

[25] Eyben, F., Wöllmer, M. & Schuller, B. Opensmile: The munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, 1459 – 1462 (Association for Computing Machinery, New York, NY, USA, 2010). URL https://doi.org/10.1145/1873951.1874246.

[26] Bozkurt, E., Erzin, E., Erdem, ÃĞ. E. & Erdem, A. T. Interspeech 2009 emotion recognition challenge evaluation. In *2010 IEEE 18th Signal Processing and Communications Applications Conference*, 216–219 (2010).

[27] BaltruÅąaitis, T., Robinson, P. & Morency, L.-P. Openface: An open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1–10 (2016).

[28] Minaee, S. *et al.* Deep learning based text classification: A comprehensive review (2020). URL https://arxiv.org/abs/2004.03705.

[29] Agarap, A. F. Deep learning using rectified linear units (relu) (2018). URL https://arxiv.org/abs/1803.08375.

[30] Liu, W., Wen, Y., Yu, Z. & Yang, M. Large-margin softmax loss for convolutional neural networks (2016). URL https://arxiv.org/abs/1612.02295.

[31] Hajdinjak, M. & Mihelic, F. Wizard of oz experiments. In *The IEEE Region 8 EUROCON 2003. Computer as a Tool.*, vol. 2, 112–116 vol.2 (2003).

[32] Zhang, Z. Microsoft kinect sensor and its effect. *IEEE MultiMedia* **19**, 4–10 (2012).

[33] Auer, E. *et al.* ELAN as flexible annotation framework for sound and image processing detectors. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)* (European Language Resources Association (ELRA), Valletta, Malta, 2010). URL http://www.lrec-conf.org/proceedings/lrec2010/pdf/234_Paper.pdf.

[34] Morris, A., Maier, V. & Green, P. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. (2004).

# Publications

### 査読なし国内会議

- Tiyajamorn Nattapong, 吉永直樹, Multimodal Sentiment Classification using Modal-independent Classifiers, NLP 若手の会 (YANS) 第 17 回シンポジウム, 2022.

### 査読あり国際会議 (修士論文のテーマ以外行った発表)

- Tiyajamorn Nattapong, Tomoyuki Kajiwara, Yuki Arase, and Makoto Onizuka. Language-agnostic Representation from Multilingual Sentence Encoders for Cross-lingual Similarity Estimation. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2021.