Thesis Proposal: Detecting Empathy Using Multimodal Language Model

Md Rakibul Hasan[1] **Md Zakir Hossain**[1] **Aneesh Krishna**[1]
Shafin Rahman[2] **Tom Gedeon**[1]
[1]Curtin University, Perth WA 6102, Australia
[2]North South University, Dhaka 1229, Bangladesh
{rakibul.hasan, zakir.hossain1, a.krishna}@curtin.edu.au
shafin.rahman@northsouth.edu, tom.gedeon@curtin.edu.au

Abstract

Empathy is crucial in numerous social interactions, including human-robot, patient-doctor, teacher-student, and customer-call centre conversations. Despite its importance, empathy detection in videos continues to be a challenging task because of the subjective nature of empathy and often remains under-explored. Existing studies have relied on scripted or semi-scripted interactions in text-, audio-, or video-only settings that fail to capture the complexities and nuances of real-life interactions. This PhD research aims to fill these gaps by developing a multimodal language model (MMLM) that detects empathy in audiovisual data. In addition to leveraging existing datasets, the proposed study involves collecting real-life interaction video and audio. This study will leverage optimisation techniques like neural architecture search to deliver an optimised small-scale MMLM. Successful implementation of this project has significant implications in enhancing the quality of social interactions as it enables real-time measurement of empathy and thus provides potential avenues for training for better empathy in interactions.

1 Introduction

The ability to understand and respond appropriately to the feelings, viewpoints, and beliefs of others is referred to as empathy (Decety and Jackson, 2004; Olderbak et al., 2014). Through engagement, this capacity can strengthen bonds among people and lessen stress and sadness. For instance, consider a situation where a family member falls ill, leading to personal distress. When sharing this sadness with a colleague, receiving genuine support can significantly mitigate unhappiness while enhancing the bond with that colleague. The significance of empathy is evident across a broad range of contexts, from socially assistive robots to human-to-human interactions (Hasan et al., 2023b).

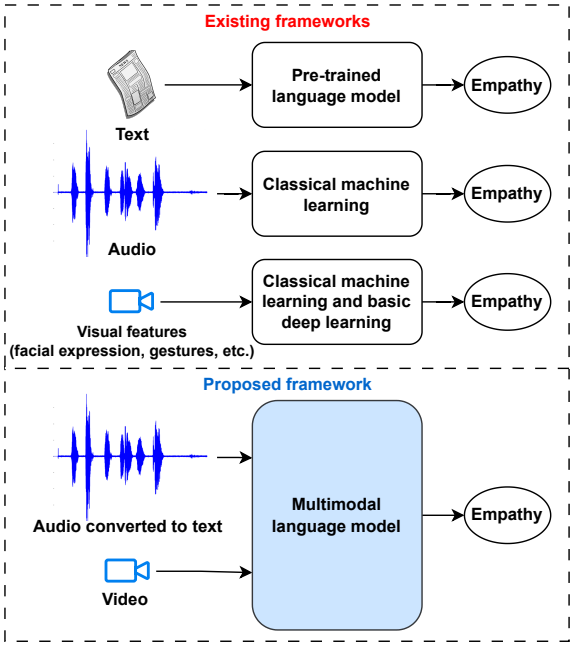

Figure 1: Comparison of our proposed approach with the literature. Existing works on text-based empathy detection (Barriere et al., 2022, 2023) leverage pre-trained language models. There exist a few works on audio- and video-based empathy detection (Alam et al., 2016; Mathur et al., 2021), where audio features or visual features are separately used mostly in classical machine learning algorithms. In contrast, we propose to leverage both audio and video information in a multimodal language model.

Assessment of empathy levels in real-life empathy-seeking interactions is crucial for determining the quality of such interactions (Bellet and Maloney, 1991). Empathy deficit often leads to conflicts and miscommunications, highlighting the importance of measuring empathy levels. Evaluating empathy levels can help answer questions such as 'To what extent does a teacher exhibit empathy towards students?', 'How empathic is a caregiver towards patients?' and 'Does an employer demonstrate empathy towards employees?', among others. Empathy detection allows for a compre-

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, pages 338–349

hensive evaluation of empathy levels in various contexts and can facilitate the development of effective strategies to promote empathy in social interactions.

The term empathy is itself subjective (Decety and Jackson, 2004), so annotated video datasets of empathy are also scarce (Hosseini and Caragea, 2021). To this front, we will collect and annotate audiovisual data of dyadic conversations. Our annotation protocol will consider the subjective experience of the participants involved in the dyadic conversation. Apart from our data, there are some public datasets (such as RealTalk (Geng et al., 2023)) where the data closely aligns with the scope of this research but not the annotation. To this end, we propose re-annotating some samples in terms of empathy and training the model using semi-supervised learning (Xu et al., 2021). Re-annotating these existing data can be challenging as we may no longer have the subjective experience of the study participants.

Empathy measurement is challenging for people (Lawrence et al., 2004), let alone automated systems. There are several works on empathy detection from textual contents (Barriere et al., 2022, 2023; Hasan et al., 2023b). Research in computational empathy from audio and audiovisual data is emerging with a few existing works (Alam et al., 2016; Barros et al., 2019). It is likely that a multimodal model with visual content – in addition to conventional textual content – can boost empathy detection performance because action and gesture play a crucial role in signalling the presence or absence of empathy.

Large language models, such as GPT-4 and PaLM, have shown excellent results in performing various complex tasks. Recently, multimodal language models (MMLMs), by integrating other modalities, such as images, demonstrate promising proof of concept in complex audiovisual recognition tasks (Driess et al., 2023; Wu et al., 2023). MMLM, therefore, seems an appropriate and state-of-the-art approach for empathy detection from audiovisual data, and hence, we propose to leverage MMLM as the backbone of our prediction system.

However, training an MMLM requires huge computational budgets (e.g., multiple GPUs or TPUs for multiple days). With a limited computational budget, this research aims to experiment in two aspects: (1) using small-scale MMLMs and (2) prompt engineering with large-scale MMLMs. Firstly, small-scale MMLMs, such as MiniVLM (Wang et al., 2021) and SimVLM (Wang et al., 2022), or compressing comparatively large-scale MMLMs through knowledge distillation (Fang et al., 2021) will allow us to leverage MMLMs in our experimental setup. One of the key aims of this research is to find an optimised and small-scale MMLM suitable for empathy detection. To this end, we will employ techniques such as neural architecture search, lottery ticket hypothesis or knowledge distillation. Secondly, prompt engineering and few-shot learning (fine-tuning with few data samples) with large-scale MMLMs shall reduce the huge computational requirement of full training and is thus considered appropriate for our low-computation and low-data scenario.

1.1 Aims and research questions

The primary aim of this project is to develop a robust method for detecting empathy by leveraging a range of multimodal empathic cues associated with video and audio. This project starts with exploring existing text- and video-based datasets. We will collect real-life video conversations to address the inadequacy of existing datasets for empathy detection. This project endeavours to address the following key research questions:

RQ 1 Dataset

- **RQ 1.1** What methods can be employed to effectively collect and annotate audiovisual data for empathy detection?
- **RQ 1.2** How can we re-annotate existing audiovisual data with regard to empathy and leverage a semi-supervised learning technique?

RQ 2 Model development

- **RQ 2.1** How can an MMLM utilising video and audio be constructed to detect empathy?
- **RQ 2.2** How much can we optimise the initial MMLM by reducing computational requirements?
- **RQ 2.3** Through prompt engineering, how could we leverage large-scale MMLM with a limited computational budget?

2 Related work

2.1 Empathy from audiovisual data

The one-minute gradual empathy (OMG-Empathy) dataset, introduced by Barros et al. (2019), consists of semi-scripted storytelling videos between

a speaker and a responder, where the responder self-annotated their valence (as a continuous value from -1 to $+1$) of the interaction. It is worth noting that although this dataset is titled 'empathy', the output label space is actually a valence score. To explore the potential of this dataset, the OMG-Empathy 2019 prediction challenge[1] was organised. Participants of the challenge, such as Barbieri et al. (2019), employed a multimodal neural network, incorporating audio signals, transcripts, raw faces, facial landmarks, and full-body images to predict continuous valence scores. Similarly, Tan et al. (2019) utilised a multimodal long short-term memory (LSTM) network, whereas Hinduja et al. (2019) implemented a convolutional neural network, using hand-crafted and deep features. Moreover, Azari et al. (2019) employed both a support vector machine (SVM) and a neural network to predict valence scores (or arguably empathy) in the challenge. All these different approaches by diverse teams, however, could not outperform the baseline model consisting of VGG16, LSTM, and SVM models, which resulted in a maximum concordance correlation coefficient of 0.23 (Barros et al., 2019). These approaches indeed showcase the potential of utilising multimodal data and machine learning techniques to detect empathy from audiovisual data, but at the same time, there is potentially much room for improvement.

Zhu et al. (2023) introduced the MEDIC dataset consisting of psychotherapeutic counselling sessions and proposed baseline models to predict empathy. To combine video, audio, and text modalities, they experimented with the Tensor Fusion Network (Zadeh et al., 2017), the Sentimental Words Aware Fusion Network (SWAFN) (Chen and Li, 2020), and a simple concatenation model. The SWAFN model performed significantly better than the other two models, providing an accuracy of 86.4% and an F1 score of 86.3%.

The RealTalk dataset, introduced by Geng et al. (2023), includes dyadic conversations among various individuals. This dataset is not designed for empathy detection tasks, nor does it have empathy annotation. Nevertheless, such dyadic conversations could be leveraged for empathy research, provided that appropriate annotation is possible. Such a complex annotation task can succeed, especially because the dataset has socially appropriate and inappropriate scenarios, which could be considered as empathy and no empathy annotations.

Apart from human-human interaction, Mathur et al. (2021) and Spitale et al. (2022) investigated empathy between human and socially assistive robots. They conducted experiments involving a humanoid robot and human participants, in which the robot interacted with 46 students by telling different scripted stories. At the end of each interaction, participants rated their level of empathy through a survey with a 5-point Likert scale. The experiment resulted in a 6.9-hour video dataset and corresponding empathy labels. Conversations were labelled as either empathic or non-empathic based on a threshold (median value) empathy score calculated from the survey response. To predict empathy on this dataset, Mathur et al. (2021) used eight different machine learning and deep learning models that utilised various features such as eye gaze, facial action units and landmarks, head pose, and point distribution parameters of the face. Their best approach, which is an XGBoost model, achieved an accuracy of 69% and an area under the receiver operating characteristic curve (AUC) of 72%.

It is important to note that the participants' audio was not considered in this dataset, which might contribute towards better empathy prediction. Extracted visual features from this dataset are publicly available[2], but the audio conversation is unavailable. It is worth noting that they solely focused on predicting empathy in human-robot interaction as a binary classification, distinguishing between empathy and non-empathy. However, it did not account for predicting empathy levels as a continuous value or categorising empathy into more nuanced categories. These limitations raise the opportunity to incorporate a more comprehensive range of empathy levels in any human interaction. By doing so, we can obtain a more nuanced understanding of empathic responses and facilitate a more precise evaluation of empathic tendencies.

2.2 Empathy from audio data

There has been limited research that predicts empathy solely on audio conversations. Alam et al. (2016) investigated empathy prediction in human-to-human call-centre conversations using SVM and reported an unweighted average recall of 65.1%. Meanwhile, Gasteiger et al. (2022) explored the

[1] https://www2.informatik.uni-hamburg.de/wtm/omgchallenges/omg_empathy_description_19.html

[2] https://github.com/interaction-lab/empathy-modeling

empathy of computer-generated audio, where transcripts were subjected to sentiment analysis, and corresponding audio files were evaluated by a group of 89 human participants. The results revealed a 70% agreement between the sentiment analyser and human annotations. These studies demonstrate the promising potential of audio for predicting empathy.

2.3 Empathy from text data

The Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (WASSA) has organised a series of shared tasks[3] on empathy prediction on text data with some demographic information. The dataset consists of essays written in response to news articles that involve harm to individuals, organisations or nature. The recent 2023 version of this dataset further consists of written conversation (speech turn) between participants (Omitaomu et al., 2022; Barriere et al., 2023). The WASSA 2021, 2022, and 2023 shared tasks challenge to detect empathy levels as a continuous value. Several participants, such as Vasava et al. (2022); Chen et al. (2022); Qian et al. (2022); Del Arco et al. (2022); Lahnala et al. (2022); Ghosh et al. (2022); Hasan et al. (2024), have fine-tuned pre-trained language models (PLMs), such as RoBERTa (Liu et al., 2019), and BERT (Devlin et al., 2019), for these tasks. Overall, PLMs excelled in all WASSA empathy detection datasets (Tafreshi et al., 2021; Barriere et al., 2022, 2023), with a maximum Pearson correlation coefficient of 0.924 in predicting empathy in written essays (Hasan et al., 2024) and 0.708 in predicting empathy in speech turns (Lu et al., 2023).

Apart from WASSA competition, PLMs were also fine-tuned in predicting empathy in detecting empathy in medical students' essays about simulated patient-doctor interactions (Dey and Girju, 2022). They used various algorithms, including Naive Bayes, SVM, LSTM, and PLMs (BERT, RoBERTa) and found that PLMs are best suited for their empathy prediction setup, providing their best F1 score of 85%. These findings highlight the effectiveness of fine-tuning pre-trained models in text-based empathy prediction. Fine-tuning facilitates harnessing prior knowledge of PLMs, which helps enhance performance while minimising training time.

[3] https://wassa-workshop.github.io

3 Proposed methodology

3.1 Problem formulation

Our primary aim is to detect empathy in dyadic conversations. Denoting speaker as S, responder as R, video as v, audio as a, and other numerical data (e.g., response to a questionnaire) as n, the multimodal data can be represented as $X = \{x_S^v, x_R^v, x_S^a, x_R^a, x_S^n, x_R^n\}$. Using X, the task is to build a model $\mathcal{F}$ to detect empathy Y. Depending on the dataset, Y can be binary classes (empathy and non-empathy), multi-classes (multiple levels of empathy) or continuous degrees of empathy (regression problem).

3.2 Public datasets we will use

We plan to use four public datasets in our experiments (Table 1). Apart from these, we may utilise the human-robot interaction dataset proposed by Mathur et al. (2021), which includes visual features of the human participants. A model capable of predicting empathy even with unavailable modality (e.g., missing audio in the human-robot dataset) could probably lead to a more robust model that can be applied to more diverse circumstances.

3.2.1 NewsEmpathy

NewsEmpathy dataset, introduced by Buechel et al. (2018), consists of people's written essays in response to newspaper articles that are harmful to individuals, organisations or nature. To determine the annotation consistency, they calculated split-half reliability, which resulted in a 'very high' reliability value of 0.875. The dataset also consists of demographic information (age, sex, ethnicity, education, and income) of the study participants who wrote the essays. The annotation is done by the essay writers themselves on a continuous scale from 1 to 7. This dataset has undergone a series of improvements with additional data collection (Omitaomu et al., 2022; Barriere et al., 2023), which resulted in a total of 3,755 samples. Although it is not an audiovisual dataset, we plan to leverage it in our preliminary experiment with language models.

3.2.2 MEDIC

Zhu et al. (2023) introduced the MEDIC dataset to measure empathy in terms of three mechanisms: expression of experience, emotional reaction, and cognitive reaction. In each mechanism, the speech turns are annotated into three categories: no expression, weak expression, and strong expression. The

SL	Name	Data	# of samples	Annotation
1	NewsEmpathy	Written essay (text) in response to newspaper articles	3,755	Empathy
2	MEDIC	Counselling case videos	771 (total 11 hours)	Empathy
3	OMG-Empathy	Semi-scripted speaker-responder storytelling	80 (total 8 hours)	Valence
4	RealTalk	Unscripted conversations about diverse experiences	692 (total 115 hours)	Speaker presence
5	Ours	Unscripted dyadic conversations	*in progress*	Empathy

Table 1: Datasets to be leveraged in this project. As none of the public datasets completely aligns with the scope of this research, we plan to collect data, which is in progress.

dataset's annotation is a good match with the scope of the research; however, the data may not cover the complexities of real-life dyadic conversations envisioned in this project.

3.2.3 OMG-Empathy

The OMG-Empathy dataset includes conversations of four speakers and ten responders (Barros et al., 2019). In each video, the speaker tells a semi-scripted story from a pool of eight stories, and the responder responds in a natural way. Following the session, the responder annotated the recorded video frame with a valence score from -1 to $+1$. Although the videos are a good match with the scope of this research, one can argue that the output annotations need to be re-considered, as the annotations are not based on empathy.

3.2.4 RealTalk

The RealTalk dataset comprises a wide variety of dyadic conversations among various individuals (Geng et al., 2023). The in-the-wild nature of this dataset makes it ideal to build a generalised AI model. However, this dataset does not have any empathy annotation.

3.3 Re-annotation of public datasets

Several datasets, such as MEDIC, adopt the Motivational Interviewing Treatment Integrity (MITI) code (Moyers et al., 2016) to annotate conversations in terms of empathy. MITI code is specifically designed to assess empathy of motivational interviewing-based treatment in healthcare and clinical sessions (Moyers et al., 2016). Although the conversations in OMG-Empathy and the RealTalk datasets may not be motivational interviewing, the technique of empathy annotation from the MITI code could still be useful in annotating samples from OMG-Empathy and RealTalk datasets. Further, we aim to leverage the existing annotations (e.g., valence in the OMG-Empathy dataset) as a guide while annotating for empathy. We aim to recruit and train multiple annotators, and we will calculate annotation consistency using standard interrater reliability assessment techniques. Subsequently, we aim to train a model using semi-supervised learning (Xu et al., 2021).

3.4 Open for collaboration

In addition to using public datasets, we welcome collaboration with scholars who wish to contribute their expertise and/or relevant private datasets in this domain. Additional compatible datasets would allow for more robust model development and validation.

3.5 Our dataset

As there are annotation mismatches with the OMG-Empathy and RealTalk datasets, we will collect and annotate new data. Details of our data collection experiment are discussed in the following subsections.

3.5.1 Study participants and their role

We will collect human-to-human dyadic conversations in empathy-seeking scenarios, where one person *(speaker)* talks about any concerning topics they face. Another person *(responder)* will interact with the speaker just like in a normal conversation. All participants (speaker and responder) will be free to use gestures (such as hand, head, or body)

Figure 2: A typical experimental setup for data collection with demo participants.

throughout the interaction. Our primary target participants are undergraduate students, postgraduate students, and staff at the host University. All participants must have normal vision and hearing abilities with necessary visual and hearing aids if required.

Speakers will be asked to reveal their emotions to talk about any concerning situations they have faced recently. They can choose to show any emotions in any situation. To help the participants decide on topics, the Geneva emotion wheel (Scherer, 2005) (Table 2) and some example topics (Table 3) will be made available to the participants before the data collection. A major portion of the example topics are prepared after brainstorming with a unit (course) coordinator regarding what sorts of situations are most common among our university students. We will advise speakers to choose a topic that they are comfortable with. If any responder finds the topic confronting, we will advise the speaker to choose another topic.

3.5.2 Equipment and data

To record the video and audio of the interactions between responders and the speaker, we will use an Insta360 ONE X2 camera[4], which has a built-in microphone and covers a 360-degree view. Participants will be seated on chairs. A typical setup is depicted in Figure 2.

We will maintain a logbook for each participant, consisting of participation ID (such as 01, 02, etc.), their seating spot (left, right), and their role (speaker or responder).

[4] https://store.insta360.com/product/one_x2

3.5.3 Questionnaire

At the end of the interaction, we will ask the participants (both speaker and responder) to fill in a questionnaire to collect demographic information and subjective ratings of the conversation regarding the degree of empathy. The questions will be hosted on the Qualtrics XM survey management system, which is a popular research survey management system used by other empathy research, such as (Gasteiger et al., 2022). All empathy assessment-related questions (other than demographic questions) are on an 11-point Likert scale (0 to 10, with 5 being the medium value) to provide many options to the participants, including a neutral opinion. The questionnaire includes the following four sections:

Participant information: This section includes eight questions, including the participant's demographic information and their role (speaker or responder). Depending on the role, the next set of questions is set to appear differently.

Speaker: This section includes 11 questions for speakers to reflect on their expressed emotions and assess their satisfaction with the conversation with the speakers. We designed four new questions, and the other seven are adapted from the consultation and relational empathy (CARE) scale proposed by Mercer et al. (2004). The CARE scale was designed to evaluate empathy in patient-doctor interactions, which resonates with our speaker-responder setup. In our experiment, speakers will share concerns with the other participants (responder), and the responder's empathy will be evaluated, which is similar to patients sharing their concerns with doctors and then evaluating the doctor's empathy, as in (Mercer et al., 2004).

Responder: This section includes 11 questions only for responders to assess their empathy towards the speakers. We developed two novel questions and adopted the remaining eight questions from empathy detection research by Mathur et al. (2021) and Shen (2010).

Responder – in general empathy: To assess the responders' empathy according to an established empathy measurement scale, this section includes the full questionnaire from the Toronto Empathy Questionnaire (Spreng et al., 2009), which includes 17 questions.

Anger	Sadness	Shame	Disappointment	Fear	Disgust	Hate
Regret	Guilt	Pride	Joy	Pleasure	Contentment	Love
Admiration	Relief	Compassion	Amusement	Interest	Contempt	Other

Table 2: List of emotions speaker can choose to express (Scherer, 2005).

SL	Topics
1	Facing challenges in academic work (exams or other assessments) or finding it difficult to keep up with studies
2	Feelings of homesickness or missing family/friends/pets back home
3	Being bullied by peers or classmates
4	Dealing with financial difficulties, such as not being able to afford food, housing, or other basic needs
5	Accomplishing a personal goal, such as winning a competition or achieving a high grade on a test
6	Overcoming a personal challenge, such as overcoming a fear or learning a new skill
7	Choice of the project in a course
8	In a group project, conflict among group members about what direction should the project go in
9	Deciding who will do what role based on a skillset in a group project
10	Deciding who will do what work and how long in a group project
11	Contributions of team members in a group project
12	Dealing with a physical illness/injury of self or relative
13	Recent bad experiences in travel
14	Experiencing difficulties or challenges related to family, such as divorce
15	Experiencing difficulties in relationships with friends, romantic partners, or family members
16	Feeling lonely or disconnected from others
17	Struggling with cultural or personal identity, such as feeling marginalised or discriminated against
18	Having trouble adapting to a new school or community
19	Coping with the loss of a loved one, such as a pet or family member
20	Struggling with substance abuse or addiction
21	Facing harassment or discrimination based on factors such as race, gender, or sexual orientation
22	Excited with graduation from school/college/university
23	Starting a new romantic relationship or friendship
24	Celebrating a special occasion, such as a birthday
25	Travelling or experiencing new adventures, such as travelling to a new country or trying a new hobby
26	Struggling with mental health issues such as anxiety, depression or stress
27	Cannot find preferred accommodation
28	Other

Table 3: List of sample topics the speaker can choose to discuss with the responder.

3.6 Multimodal language model

The primary aim is to detect the responder's empathy towards the speaker's perspective. As depicted in Figure 3, conversation transcripts and visual data will be fed into a multimodal language model (MMLM) for final empathy detection. The distinct patterns of hand and facial gestures observed in the top and bottom interactions exhibited in Figure 3 indicate a higher level of empathy than in the middle. Consequently, integrating facial expressions and hand gestures should significantly enhance the accuracy of empathy detection and thus be considered in this project.

3.6.1 Audiovisual data pre-processing

We will use Insta360 Studio video editing software[5] to extract the listener and responder video frames.

[5] https://www.insta360.com/download/insta360-onex2

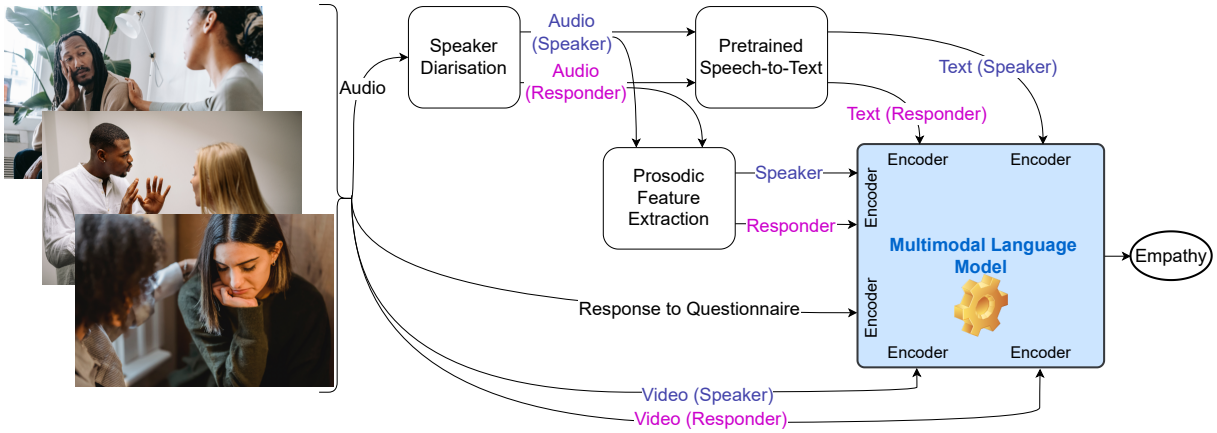

Figure 3: Our proposed framework of empathy detection. To process audiovisual data, we first separate the audio and video of the speaker and the responder. Following speech-to-text conversion and prosodic feature extraction, we will leverage a multimodal language model to detect empathy.

The first step in audio processing is to diarise the audio, which means separating the speech of the speaker and the responder from any background noise. To extract the audio from the video, we will use FFMPEG[6] as done in (Azari et al., 2019). Following this, three speech-to-text converters – OpenAI Whisper[7], Google speech-to-text[8], and Watson speech-to-text[9] – will be used to convert the audio conversation into text. Agreement among the three converters will be checked for each word, and any disagreement will be settled through manual intervention. The converted textual information will be used as features for empathy detection. In addition, prosodic features, such as pitch and loudness, will be extracted from audio and leveraged in the empathy detection pipeline.

3.6.2 Features and output labels from questionnaire

We will use the following features from the questionnaire:

a. Speaker's demographic information
b. Responder's demographic information
c. Speaker's response to the questionnaire on revealed emotion
d. Responder's general empathy

We will aggregate the answers to the responder's questions and calculate a single empathy score (ES) for each interaction between one speaker and one responder:

$$ES = \sum_{i=1}^{N} q_i \quad (1)$$

where N is the number of empathy assessment-related questions ($N = 9$) answered by the responder, and q is the value of the Likert scale (0 to 10). This empathy score will be used as the continuous ground truth empathy score.

This project envisages modelling empathy prediction both as a regression (continuous empathy score) and a classification (empathy levels) problem. Motivated by a recent study on text-based empathy detection (Montiel-Vázquez et al., 2022), we will annotate each interaction into five categories from a third-person perspective: (1) not empathic at all, (2) a little empathic, (3) somewhat empathic, (4) empathic, and (5) very much empathic.

3.6.3 Model development

The model primarily takes in audio and video data from both the speaker and the responder and uses it to infer the level of empathy that the responder feels towards the speaker. The visual and text data go through independent encoders to obtain encoded representations for each modality per speaker. We will leverage a video vision transformer, ViViT, (Arnab et al., 2021) to encode video sequences. Depending on the language model backbone (e.g., BERT, RoBERTa), we will leverage the corresponding tokeniser to encode text transcripts. These encoded representations likely capture various cues: visual cues like facial expressions and hand gestures and linguistic cues from the transcripts. The representations are then fused together with numerical information from the questionnaire using

[6] http://www.ffmpeg.org/
[7] https://openai.com/research/whisper
[8] https://cloud.google.com/speech-to-text/
[9] https://www.ibm.com/cloud/watson-speech-to-text

methods like early or late fusion. The fused multimodal context representation is input to an empathy detection model, which detects if the responder empathises towards the speaker.

3.6.4 Optimisation

One of the key aims of this project is to find out an optimised small-scale MMLM. To this front, we aspire to leverage techniques such as neural architecture search (NAS), lottery ticket hypothesis, and knowledge distillation.

NAS involves a systematic exploration of various neural network alternatives using automated testing to identify the best-performing architecture (Elsken et al., 2019). The lottery ticket hypothesis suggests that large networks may contain smaller efficient subnetworks, and pruning techniques can be used to find these lottery ticket subnetworks. This can lead to smaller, faster, and more efficient models without significantly reducing performance. Finally, the knowledge distillation technique compresses and optimises a large teacher model into a smaller student model while retaining most of its capabilities.

3.6.5 Model evaluation

We will evaluate the performance of the final model using cross-validation and compare it with baseline models. To compare with existing studies, we will use specific data from our dataset, such as using audiovisual data to compare with Mathur et al. (2021); Barbieri et al. (2019); Tan et al. (2019), audio data to compare with Alam et al. (2016); Gasteiger et al. (2022), and text data with Vasava et al. (2022); Chen et al. (2022); Qian et al. (2022); Del Arco et al. (2022); Lahnala et al. (2022); Ghosh et al. (2022); Barriere et al. (2022, 2023); Hasan et al. (2023a). As for the evaluation metrics, we will adopt established metrics corresponding to each public dataset so that we can compare our results with the literature. For our collected dataset, we will provide results in multiple established evaluation metrics: (1) Pearson correlation coefficients, Spearman's correlation coefficient, and concordance correlation coefficient for continuous degree of empathy prediction, and (2) accuracy, precision, recall, F1 score, and AUC score for empathy level prediction.

4 Preliminary experiments

We have experimented with the NewsEmpathy datasets (Omitaomu et al., 2022; Barriere et al., 2023), where we experimented with fine-tuning three PLMs (ALBERT, DistillBERT, and BERT). The dataset has numerical demographic information, which enhanced empathy detection because of the subjective nature of empathy. To this end, we constructed meaningful sentences from the numeric demographic information, which in fact, boosted empathy detection performance. As for data-centric improvement, we also leveraged T5-based PLMs for text summarising and rephrasing (Hasan et al., 2023a).

As a follow-up study, we leveraged GPT-3.5 LLM to mitigate annotation noise in crowdsource datasets. Crowdsourcing is a faster and cheaper way to collect data and annotation in computational social science research, such as empathy. However, crowdsourcing involves many different people who may undertake such jobs only for financial benefit, and thus it becomes difficult to maintain the quality of collected data and annotation (Sheehan, 2018). To this end, we proposed re-annotating noisy and misleading annotations using GPT-3.5 LLM and mixing these new annotations with human-provided good annotations. Apart from this, we also leveraged GPT-3.5 in converting numerical demographic information into meaningful sentences and data augmentation through paraphrasing (Hasan et al., 2024).

5 Conclusion

The ability to detect and understand empathy is central to improving social interaction. This PhD research proposes a multimodal framework by modelling empathy based on video and speech transcripts in an integrated manner. We start with public datasets and further collect data tailor-made for empathy detection, as most currently available corpora do not fully match the target objectives. With the recent success of multimodal language models (MMLMs), this project aims to use cross-modality dynamics and joint representations of multimodal audiovisual data in an MMLM. To accommodate MMLM in a resource-constrained environment, this research aims to adopt optimisation techniques, such as neural architecture search, lottery ticket hypothesis, and knowledge distillation. The proposed MMLM can be used to detect empathy in various settings, such as in education, healthcare, and businesses.

Ethics statement

The project involves collecting video and audio from human participants. Therefore, to conduct the data collection experiments, necessary ethics approval will be sought from the Human Research Ethics Committee of the host university. Before each data collection session, we will brief the participant and provide an information form and consent form for signing, including how the data will be stored and utilised, the purpose of the experiment and the data collection methods. We will collect signed consent forms from all participants.

References

Firoj Alam, Morena Danieli, and Giuseppe Riccardi. 2016. Can we detect speakers' empathy?: A real-life case study. In *2016 7th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, pages 000059–000064. IEEE.

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. 2021. ViViT: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846.

Bita Azari, Zhitian Zhang, and Angelica Lim. 2019. Towards an emocog model for multimodal empathy prediction. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–4. IEEE.

Francesco Barbieri, Eric Guizzo, Federico Lucchesi, Giovanni Maffei, Fermín Moscoso del Prado Martín, and Tillman Weyde. 2019. Towards a multimodal time-based empathy prediction system. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–5.

Valentin Barriere, João Sedoc, Shabnam Tafreshi, and Salvatore Giorgi. 2023. Findings of WASSA 2023 shared task on empathy, emotion and personality detection in conversation and reactions to news articles. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 511–525, Toronto, Canada. Association for Computational Linguistics.

Valentin Barriere, Shabnam Tafreshi, João Sedoc, and Sawsan Alqahtani. 2022. WASSA 2022 shared task: Predicting empathy, emotion and personality in reaction to news stories. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 214–227, Dublin, Ireland. Association for Computational Linguistics.

Pablo Barros, Nikhil Churamani, Angelica Lim, and Stefan Wermter. 2019. The OMG-empathy dataset: Evaluating the impact of affective behavior in storytelling. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–7.

Paul S. Bellet and Michael J. Maloney. 1991. The importance of empathy as an interviewing skill in medicine. *JAMA*, 266(13):1831–1832.

Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and João Sedoc. 2018. Modeling empathy and distress in reaction to news stories. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4758–4765, Brussels, Belgium. Association for Computational Linguistics.

Minping Chen and Xia Li. 2020. SWAFN: Sentimental words aware fusion network for multimodal sentiment analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1067–1077, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yue Chen, Yingnan Ju, and Sandra Kübler. 2022. Iucl at wassa 2022 shared task: A text-only approach to empathy and emotion detection. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*.

Jean Decety and Philip L Jackson. 2004. The functional architecture of human empathy. *Behavioral and cognitive neuroscience reviews*, 3(2):71–100.

Flor Miriam Del Arco, Jaime Collado-Montañez, L Alfonso Ureña, and María-Teresa Martín-Valdivia. 2022. Empathy and distress prediction using transformer multi-output regression and emotion analysis with an ensemble of supervised and zero-shot learning models. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 239–244.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Priyanka Dey and Roxana Girju. 2022. Enriching deep learning with frame semantics for empathy classification in medical narrative essays. In *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, pages 207–217.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence.

2023. PaLM-e: An embodied multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 8469–8488. PMLR.

Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. 2019. Neural architecture search: A survey. *The Journal of Machine Learning Research*, 20(1):1997–2017.

Zhiyuan Fang, Jianfeng Wang, Xiaowei Hu, Lijuan Wang, Yezhou Yang, and Zicheng Liu. 2021. Compressing visual-linguistic model via knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1428–1438.

Norina Gasteiger, JongYoon Lim, Mehdi Hellou, Bruce A MacDonald, and Ho Seok Ahn. 2022. Moving away from robotic interactions: Evaluation of empathy, emotion and sentiment expressed and detected by computer systems. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1365–1370. IEEE.

Scott Geng, Revant Teotia, Purva Tendulkar, Sachit Menon, and Carl Vondrick. 2023. Affective faces for goal-driven dyadic communication. *arXiv preprint arXiv:2301.10939*.

Soumitra Ghosh, Dhirendra Maurya, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Team iitp-ainlpml at wassa 2022: Empathy detection, emotion classification and personality detection. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 255–260.

Md Rakibul Hasan, Md Zakir Hossain, Tom Gedeon, and Shafin Rahman. 2024. LLM-GEm: Large language model-guided prediction of people's empathy levels towards newspaper article. In *Findings of the Association for Computational Linguistics: EACL 2024*, St. Julians, Malta. Association for Computational Linguistics.

Md Rakibul Hasan, Md Zakir Hossain, Tom Gedeon, Susannah Soon, and Shafin Rahman. 2023a. Curtin OCAI at WASSA 2023 empathy, emotion and personality shared task: Demographic-aware prediction using multiple transformers. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 536–541, Toronto, Canada. Association for Computational Linguistics.

Md Rakibul Hasan, Md Zakir Hossain, Shreya Ghosh, Susannah Soon, and Tom Gedeon. 2023b. Empathy detection using machine learning on text, audiovisual, audio or physiological signals. *arXiv preprint arXiv:2311.00721*.

Saurabh Hinduja, Md Taufeeq Uddin, Sk Rahatul Jannat, Astha Sharma, and Shaun Canavan. 2019. Fusion of hand-crafted and deep features for empathy prediction. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–4. IEEE.

Mahshid Hosseini and Cornelia Caragea. 2021. Distilling knowledge for empathy detection. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3713–3724.

Allison Lahnala, Charles Welch, and Lucie Flek. 2022. Caisa at wassa 2022: Adapter-tuning for empathy prediction. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 280–285.

Emma J Lawrence, Philip Shaw, Dawn Baker, Simon Baron-Cohen, and Anthony S David. 2004. Measuring empathy: reliability and validity of the empathy quotient. *Psychological medicine*, 34(5):911–920.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Xin Lu, Zhuojun Li, Yanpeng Tong, Yanyan Zhao, and Bing Qin. 2023. HIT-SCIR at WASSA 2023: Empathy and emotion analysis at the utterance-level and the essay-level. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 574–580, Toronto, Canada. Association for Computational Linguistics.

Leena Mathur, Micol Spitale, Hao Xi, Jieyun Li, and Maja J Matarić. 2021. Modeling user empathy elicited by a robot storyteller. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8.

Stewart W Mercer, Margaret Maxwell, David Heaney, and Graham Cm Watt. 2004. The consultation and relational empathy (care) measure: development and preliminary validation and reliability of an empathy-based consultation process measure. *Family practice*, 21(6):699–705.

Edwin Carlos Montiel-Vázquez, Jorge Adolfo Ramírez Uresti, and Octavio Loyola-González. 2022. An explainable artificial intelligence approach for detecting empathy in textual communication. *Applied Sciences*, 12(19):9407.

Theresa B. Moyers, Lauren N. Rowell, Jennifer K. Manuel, Denise Ernst, and Jon M. Houck. 2016. The motivational interviewing treatment integrity code (miti 4): Rationale, preliminary reliability and validity. *Journal of Substance Abuse Treatment*, 65:36–42.

Sally Olderbak, Claudia Sassenrath, Johannes Keller, and Oliver Wilhelm. 2014. An emotion-differentiated perspective on empathy with the emotion specific empathy questionnaire. *Frontiers in Psychology*, 5.

Damilola Omitaomu, Shabnam Tafreshi, Tingting Liu, Sven Buechel, Chris Callison-Burch, Johannes Eichstaedt, Lyle Ungar, and João Sedoc. 2022. Empathic conversations: A multi-level dataset of contextualized conversations. *arXiv preprint arXiv:2205.12698*.

Shenbin Qian, Constantin Orašan, Diptesh Kanojia, Hadeel Saadany, and Félix Do Carmo. 2022. Surrey-cts-nlp at wassa2022: An experiment of discourse and sentiment analysis for the prediction of empathy, distress and emotion. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 271–275.

Klaus R Scherer. 2005. What are emotions? and how can they be measured? *Social science information*, 44(4):695–729.

Kim Bartel Sheehan. 2018. Crowdsourcing research: Data collection with amazon's mechanical turk. *Communication Monographs*, 85(1):140–156.

Lijiang Shen. 2010. On a scale of state empathy during message processing. *Western Journal of Communication*, 74(5):504–524.

Micol Spitale, Sarah Okamoto, Mahima Gupta, HAO Xi, and Maja J Matarić. 2022. Socially assistive robots as storytellers that elicit empathy. *ACM Transactions on Human-Robot Interaction (THRI)*, 11(4):1–29.

R Nathan Spreng, Margaret C McKinnon, Raymond A Mar, and Brian Levine. 2009. The toronto empathy questionnaire: Scale development and initial validation of a factor-analytic solution to multiple empathy measures. *Journal of personality assessment*, 91(1):62–71.

Shabnam Tafreshi, Orphée De Clercq, Valentin Barriere, Sven Buechel, João Sedoc, and Alexandra Balahur. 2021. Wassa 2021 shared task: Predicting empathy and emotion in reaction to news stories. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 92–104. Association for Computational Linguistics.

Zhi-Xuan Tan, Arushi Goel, Thanh-Son Nguyen, and Desmond C Ong. 2019. A multimodal LSTM for predicting listener empathic responses over time. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–4. IEEE.

Himil Vasava, Pramegh Uikey, Gaurav Wasnik, and Raksha Sharma. 2022. Transformer-based architecture for empathy prediction and emotion classification. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 261–264.

Jianfeng Wang, Xiaowei Hu, Pengchuan Zhang, Xiujun Li, Lijuan Wang, Lei Zhang, Jianfeng Gao, and Zicheng Liu. 2021. MiniVLM: A smaller and faster vision-language model. *arXiv preprint arXiv:2012.06946*.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2022. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*.

Wenhao Wu, Zhun Sun, and Wanli Ouyang. 2023. Revisiting classifier: Transferring vision-language models for video recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37(3), pages 2847–2855.

Yi Xu, Jiandong Ding, Lu Zhang, and Shuigeng Zhou. 2021. DP-SSL: Towards robust semi-supervised learning with a few labeled samples. In *Advances in Neural Information Processing Systems*, volume 34, pages 15895–15907. Curran Associates, Inc.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114, Copenhagen, Denmark. Association for Computational Linguistics.

Zhouan Zhu, Chenguang Li, Jicai Pan, Xin Li, Yufei Xiao, Yanan Chang, Feiyi Zheng, and Shangfei Wang. 2023. MEDIC: A multimodal empathy dataset in counseling. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6054–6062.