Findings of WASSA 2023 Shared Task on Empathy, Emotion and Personality Detection in Conversation and Reactions to News Articles

Valentin Barriere[1], **João Sedoc**[2], **Shabnam Tafreshi**[3], **Salvatore Giorgi**[4]
[1]Centro Nacional de Inteligencia Artificial, Santiago, Chile
[2]New York University, New York, USA
[3]ARLIS, University of Maryland, College Park, USA
[4]University of Pennsylvania, Philadelphia, USA
valentin.barriere@cenia.cl,jsedoc@stern.nyu.edu
stafresh@umd.edu,sgiorgi@seas.upenn.edu

Abstract

This paper presents the results of the WASSA 2023 shared task on predicting empathy, emotion, and personality in conversations and reactions to news articles. Participating teams were given access to a new dataset from Omitaomu et al. (2022) comprising empathic and emotional reactions to news articles. The dataset included formal and informal text, self-report data, and third-party annotations. Specifically, the dataset contained news articles (where harm is done to a person, group, or other) and crowd-sourced essays written in reaction to the article. After reacting via essays, crowd workers engaged in conversations about the news articles. Finally, the crowd workers self-reported their empathic concern and distress, personality (using the Big Five), and multi-dimensional empathy (via the Interpersonal Reactivity Index). A third-party annotated both the conversational turns (for empathy, emotion polarity, and emotion intensity) and essays (for multi-label emotions). Thus, the dataset contained outcomes (self-reported or third-party annotated) at the turn level (within conversations) and the essay level. Participation was encouraged in five tracks: (i) predicting turn-level empathy, emotion polarity, and emotion intensity in conversations, (ii) predicting state empathy and distress scores, (iii) predicting emotion categories, (iv) predicting personality, and (v) predicting multi-dimensional trait empathy. In total, 21 teams participated in the shared task. We summarize the methods and resources used by the participating teams.

1 Introduction

Affect-related phenomena have been widely studied in the last two decades (Picard, 2000). They are crucial for social interactions between humans as they create a bond between the different social agents (Cassell, 2001), whether humans or machines. They are also essential to make machines understand the world and gain commonsense knowledge, which is essential when tackling complex human-related tasks. Studying the affective and social phenomena like opinions, emotions, empathy, distress, stances, persuasiveness (Buechel et al., 2018a; Barriere and Balahur, 2023; Park et al., 2014b) or speaker traits allows machine learning practitioners to dramatically improve the response from automated agents (Pelachaud et al., 2021; Zhao et al., 2016). Social skills like empathy are essential for human(-agent) communication (Parmar et al., 2022; Reis et al., 2017). Right now, it is helpful for as many applications such as an empathic agent (Rashkin et al., 2019; Zhong et al., 2020), a way to de-bias a corpus to train a language model (Lahnala et al., 2022b), or as a tool to help human to communicate or to find consensus (Pérez-Rosas et al., 2017; Sharma et al., 2023; Argyle et al., 2023). In general, empathic utterances can be emotional; therefore, examining emotion in text-based conversations may significantly impact predicting empathy. Moreover, according to (Lahnala et al., 2022a), many studies make an amalgam between empathy and emotion by poorly defining the former. Hence, studying emotion and empathy together can help to remove this bias, even though more psycho-linguistic work would be welcome.

This paper presents the WASSA 2023 Empathy Shared Task: Predicting Empathy, Emotion, and Personality in Conversations and Reaction to News Articles, which allows studying empathy and emotion in human interactions. Past WASSA shared tasks were also held on emotion, empathy, distress, or personality detection in text essays (Tafreshi et al., 2021; Barriere et al., 2022b). Thus, this year's task builds on past shared tasks, with data very similar to past years, plus a brand new type of data. We used a new dataset from (Omitaomu et al., 2022) containing reactions to news article data and annotations similar to (Buechel et al., 2018b) and (Tafreshi et al., 2021), including news articles that express harm to an entity (e.g., individual, group

of people, nature).

The news articles are accompanied by essays where authors express their empathy and distress in response to the content. Each essay is annotated for empathy and distress, and multi-label emotions. They are also enriched with additional information, such as the authors' personality traits, IRI, and demographic details, including age, gender, ethnicity, income, and education level. The new type of data introduced in this year's shared task consists in the subsequent conversations that the study participants had after writing their essays, which were annotated in perceived emotional polarity and intensity and perceived empathy. For more specific information, please refer to Section 3 in the paper.

Given this dataset as input, the shared task consists of five tracks (see Section 4 for each tracks' respective definitions of empathy and emotion):

1. Predicting Perceived Empathy and Emotion in Conversations (CONV): Teams develop models to predict several values linked to emotion and empathy for each speech turn in a textual conversation. The targets are third-party assessment of emotional polarity, emotional intensity, and empathy.
2. Predicting State Empathy (EMP): Teams develop models to predict, for each essay, empathy and distress scores quantified by Batson's empathic concern ("feeling for someone") and personal distress ("suffering with someone") (Batson et al., 1987) scales.[1]
3. Emotion Label Prediction (EMO): Teams develop models to predict, for each essay, a categorical emotion tag from the following Ekman's six basic emotions (sadness, joy, disgust, surprise, anger, or fear) (Ekman, 1971), as well as *hope* and *neutral* tag.
4. Personality Prediction (PER): Teams develop models to predict, for each essay, Big Five (OCEAN) personality traits (conscientiousness, openness, extraversion, agreeableness, emotional stability; John et al. 1999).
5. Predicting Multi-dimensional Trait Empathy (IRI): Teams develop models to predict, for each essay, multi-dimensional empathy (via the Interpersonal Reactivity Index; Davis, 1980): perspective taking, personal distress, fantasy, and empathic concern.

[1] *Distress* is a self-focused and negative affective state (*suffering with someone*) while *empathy* is a warm, tender, and compassionate state (*feeling for someone*).

2 Related Work

We provide related work for each track: affect-related phenomena in interactions (Section 2.1), emotion predictions (Section 2.2), empathy and distress (Section 2.3), and personality prediction (Section 2.4).

2.1 Affective Phenomena in Interactions

Affect-related phenomena in interactions is a field of study that comprises emotion recognition in conversations (McKeown et al., 2012; Ma et al., 2020; Firdaus et al., 2020; Ringeval et al., 2013), opinion analysis in interactions (Barriere et al., 2018, 2022a), first impressions assessment (Cafaro et al., 2017), or personality detection (Mairesse and Walker, 2006) among many others. The interest of these approaches is to use the interactional context in order to model the dynamics of the target phenomena within a conversation (Hazarika et al., 2018; Majumder et al., 2019; Poria et al., 2019b,a). Recent works are using speaker-dependent vectors (Majumder et al., 2019), graph neural networks to model the interactions (Ghosal et al., 2019), or dialog-aware attention mechanism (Shen et al., 2020).

2.2 Emotion Prediction

Emotion classification is the task of predicting a single- or multi-label emotion classes (Ekman, 1971), or a value in the valence-arousal space, which has been widely studied in non-verbal language (Schuller et al., 2009; McKeown et al., 2012; Vinciarelli et al., 2008), or even in music (Soleymani et al., 2013). Emotion classification in text, more recently, has been studied thoroughly in terms of modeling, resources, and features as part of SemEval shared tasks for Affect computing and emotion classification (Strapparava and Mihalcea, 2007; Mohammad and Bravo-Marquez, 2017; Mohammad et al., 2018; Chatterjee et al., 2019; Sharma et al., 2020c).

Most emotion prediction models are learned in a supervised manner with feature engineering or continuous representation learned through pretrained language models (Peters et al., 2018; Devlin et al., 2018a), and now in an unsupervised way using emerging abilities of large language models (Choi et al., 2023; Brown et al., 2020). Acheampong et al. (2020); Murthy and Kumar (2021); Nandwani and Verma (2021); Acheampong et al. (2021); Ezzameli and Mahersia (2023) survey state-of-the-art

emotion detection techniques and resources and discuss open issues in this area.

2.3 Empathy and Distress

As seen in this shared task, empathy can have varying definitions: empathic utterances, state empathy (as measured via the Batson scale), and trait empathy (as measured via the IRI), among others. Thus, research on empathy in natural language processing often uses varying or even under-specified measures (Lahnala et al., 2022a). Prior work on modeling text-based empathy focused on the empathic concern, which is to share others' emotions in the conversations (Litvak et al., 2016; Fung et al., 2016). For instance, Xiao et al. (2015, 2016); Gibson et al. (2016) modeled empathy based on the ability of a therapist to adapt to the emotions of their clients; Zhou and Jurgens (2020) quantified empathy in condolences in social media using appraisal theory; Sharma et al. (2020b) developed a model based on fine-tuning contextualized language models to predict empathy specific to mental health in text-based platforms. Guda et al. (2021) additionally utilized demographic information (e.g., education, income, age) when fine-tuning contextualized language modeling for empathy and distress prediction. While empathy is vital for human(-agent) communication, some have argued that empathy is a poor guide for moral decision-making (Bloom, 2017). To this end, recent work has shown that language associated with empathy, when separated from compassion, is more self-focused and contains negative emotions (Yaden et al., 2023).

2.4 Personality Prediction

Vora et al. (2020) and Beck and Jackson (2022) survey and analyze personality prediction models, theories, and techniques. Ji et al. (2020) review such models specifically to detect suicidal behavior. Developing personality detection models range from feature engineering methods (Bharadwaj et al., 2018; Tadesse et al., 2018) to deep learning techniques (Yang et al., 2021; Ren et al., 2021; Lynn et al., 2020). Yang et al. (2021) developed a transformer-based model to predict users' personality based on Myers-Briggs Type Indicator (Myers et al., 1985, MBTI) personality trait theory given multiple posts of the user instead of predicting personality for a single post. Ren et al. (2021) utilized deep learning techniques to develop a multi-label personality prediction and sentiment analysis model based on MBTI and Big 5 datasets. Given the cost and time needed to collect personality survey responses, Vu et al. (2020) developed methods to predict out-of-sample survey questions. More recently, Large Language Models (such as GPT-3) have been used for zero-shot personality classification (Ganesan et al., 2023).

3 Data Collection and Annotation

The source of the data for the shared task is from Omitaomu et al. (2022). We extend this dataset with essay-level emotion annotations by the authors. Although the dataset is different from the data set of Buechel et al. (2018b) used in WASSA 2021 and 2022 shared task (Tafreshi et al., 2021; Barriere et al., 2022b), it can be considered an extension. Table Table 1 shows the train, development, and test splits. We first briefly present how the original dataset was collected and annotated in subsection 3.1. We discuss the additional emotion annotation in subsection 3.2.

	Train	Dev	Test
People	41	34	65
Conversations	386	114	50
Essays	792	208	100
Speech-Turns	9,176	2,000	1,425

Table 1: Corpus statistics detailing the number of annotations.

3.1 Initial Data Collection and Annotation

Here we provide a brief overview of the data collection process employed by Omitaomu et al. (2022). They recruited crowd workers from MTurk.com and utilized the Qualtrics survey platform and ParlAI for data collection. The data collection process began with an intake phase, during which crowd workers provided their demographic information and completed surveys for the Big Five (OCEAN) personality traits and the Interpersonal Reactivity Index (IRI). Next, pairs of crowd workers read news articles. Each pair read one article of the 100 articles. After reading the article, the crowd workers wrote an essay of 300 to 800 characters about the article they read and rated their empathy and distress levels using the Batson scale. Then, the pair of crowd workers engaged in online text conversation where they were instructed to talk about the article for a minimum of 10 turns per person in training and development sets and 15 turns per person in the test set.

After the conversations were collected, a new task was created to collect turn-level annotations for each conversation. The workers were asked to rate the empathy, emotional polarity, and emotional intensity of each turn. Three crowd workers annotated each turn and were given the context of the previous turns in the conversation.

3.2 External Emotion Annotation

We enriched the dataset by annotating the essays with multi-label emotion tags. We used the six Ekman's emotions to determine whether certain basic emotions are (Ekman, 1971) more correlated with empathy and distress. We added another emotion which is hope, as it is fairly present in our dataset and used in the GoEmotion dataset as a sub-emotion of Joy, and we wanted to separate them. With the neutral label, this gave us a total of 8 label tags. Three of the four coders annotated each essay using a maximum of two emotion tags (including neutral), yielding three to six tags for each essay. We used the LEAP protocol to reach a higher agreement between the annotators (Lee et al., 2023). We calculated the inter-annotator agreement using Krippendorff's α (Krippendorff, 2013) with the MASI distance (Passonneau, 2006) that has been proven helpful for multi-label annotations, and obtained 0.40 (0.44 with Jaccard distance). We computed the ground truth by labeling all the emotions with at least two tags among the three to six possible tags. The distribution of the train and development datasets are shown in Figure 1. The matrix of co-occurrences of the train/dev sets is shown in Figure 1. Disgust is positively correlated with two emotions: anger and surprise. The highest number of co-occurrences between two emotions is 36, which happens between disgust and anger. Neutral rarely happens with other emotions. Sadness is statistically more correlated to Fear and Hope.

4 Shared Task

We set up all four tracks in CodaLab[2]. We describe each task separately in Section 4.1 and then describe dataset, resources, and evaluation metrics in Section 4.2. Note that the last four tracks are similar to the ones offered by WASSA 2022 shared task, even though this year it is possible to use the conversations to get more context.

[2] https://codalab.lisn.upsaclay.fr/competitions/11167

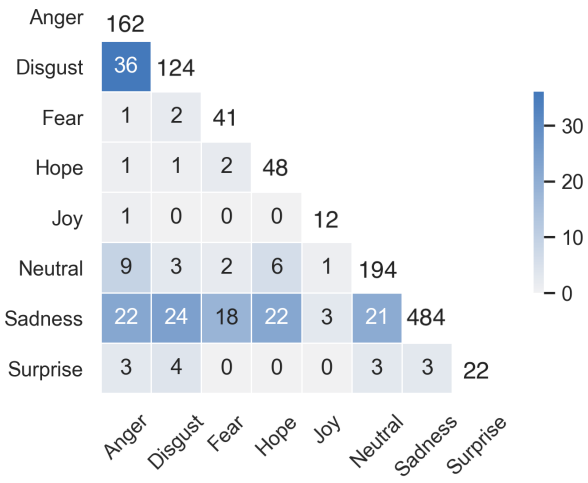

Figure 1: Co-occurrence matrix of the EMO labels on the train and dev sets

4.1 Tracks

Track 1 - Turn-level Empathy and Emotion in Conversations (CONV): The formulation of this task is to predict, for each conversational turn, the emotion polarity and intensity as well as the third party annotations of empathy. The targets are third-party assessment of emotional polarity (positive, negative, or neutral) and both emotional intensity and empathy coded on an ordinal scale from 1 to 5 with a not applicable option. This track is new to WASSA 2023.

Track 2 - State Empathy Prediction (EMP): The formulation of this task is to predict, for each essay, Batson's empathic concern ("feeling for someone") and personal distress ("suffering with someone") scores (Batson et al., 1987). Teams are expected to develop models that predict the empathy score for each essay (self-report data from the essay writer). Both empathy and distress scores are real values between 1 and 7. Empathy score is an average of 7-point scale ratings, representing each of the following states (warm, tender, sympathetic, softhearted, moved, compassionate); distress score is an average of 7-point scale ratings, representing each of the following states (worried, upset, troubled, perturbed, grieved, disturbed, alarmed, distressed). These are state measures: measures that vary within people across time. For optional use, we made personality, demographic information, and emotion labels available for each essay. This track was previously done in WASSA 2022 and 2021, but this year's task uses new data.

Track 3 - Emotion Label Prediction (EMO): The formulation of this task is to predict, for each

essay, one or more emotion labels from the following Ekman's six basic emotions (sadness, joy, disgust, surprise, anger, or fear) (Ekman, 1971), as well as *neutral* (like in (Barriere et al., 2022b)), and we also added *hope*.

The same set of metadata that we described above was also provided for each essay in this task. Participants optionally could use this information as features to predict emotion labels. The essay-level emotion labels are third party annotations. This task was also done in WASSA 2022 and 2021, but this year's task uses new data.

Track 4 - Personality Prediction (PER): To code personality information, the Big 5 personality traits were provided, also known as the OCEAN model (Gosling et al., 2003b). In the OCEAN model, the theory identifies five factors (openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism[3]). For each essay, the writer was asked to complete the Ten Item Personality Inventory (Gosling et al., 2003a). Thus, this is self-reported essay-level data. This task was previously done in WASSA 2022, but the data in this year's task (2023) is new.

Track 5 - Multi-dimensional Trait Empathy Prediction (IRI): We use the Interpersonal Reactivity Index (IRI), a measurement tool for the multidimensional assessment of empathy (Davis, 1980). The IRI consists of four subscales (Perspective Taking, Fantasy, Empathic Concern, and Personal Distress) where each subscale consists of 7 items, each on a 5-point Likert scale. The IRI is a trait-level empathy measure: a measure that is stable within people across time. Though a similar task was done in 2022, this self-reported essay-level data is new to this year's task.

Multi-task: We gave the participants a unique id for each conversation so that the participants could use multi-task learning methods to tackle all the tasks simultaneously. Moreover, speakers in the train, dev, and test datasets were given unique ids so that teams could use several of the participant's essays or conversations in order to improve the results. This was proven to help last year for the PER and IRI subtasks (Barriere et al., 2022b).

[3]For the shared task, neuroticism has been reverse coded as emotional stability

4.2 Setup

Dataset: Participants were provided the dataset described in Section 3. Participants were allowed to add the development set to the training set and submit systems trained on both. The test set was made available to the participants at the beginning of the evaluation period.

Resources and Systems Restrictions Participants were allowed to use any lexical resources (e.g., emotion or empathy dictionaries) of their choice, additional training data, or off-the-shelf emotion or empathy models. We did not put any restrictions on this shared task. We proposed several baseline models for this article, which are described in Section 4.3.

Systems Evaluation: The organizers published an evaluation script that calculates Pearson correlation for the predictions of the conversation, empathy, personality and IRI prediction tasks and precision, recall, and F1 measure for each emotion class as well as the micro and macro average for the emotion label prediction task. Pearson coefficient is the linear correlations between two variables, and it produces scores from -1 (perfectly inversely correlated) to 1 (perfectly correlated). A score of 0 indicates no correlation. The official competition metric for the empathy and emotion in conversation task (CONV) is the average of the three Pearson correlations. The official competition metric for the empathy prediction task (EMP) is the average of the two Pearson correlations. The official competition metric for the emotion evaluation is the macro F1-score, which is the harmonic mean between precision and recall. The official competition metric for the personality (resp. IRI prediction) task PER (resp. IRI) is the average of the Pearson correlations of the 5 (resp. 4) variables.

4.3 Baselines

CONV: Following Omitaomu et al. (2022), we fine-tuned a RoBERTa (base) pretrained language model (Liu et al., 2019a). The model was trained on the training set and used the development set for model validation. We trained one model for each of the turn-level label types. The training was for 50 epochs, and the model checkpoint with the best validation set Pearson correlation was kept.

EMP: Like the CONV models, we fine-tuned a RoBERTa (base) pretrained language model (Liu

et al., 2019a). For training, we used both the training data of the essays and the WASSA22 training data (Barriere et al., 2022b). We created separate models for empathy and distress, and used the same checkpoint and stopping criteria as the conv task models.

EMO: We created two baselines for the EMO subtask. As a first baseline, we fine-tuned a pretrained base RoBERTa transformer model (Liu et al., 2019b) over the GoEmotions dataset (Demszky et al., 2020) in a multi-label way. This lead to a macro-averaged F1 score of 0.64 on the GoEmotions test set with emotions grouped using Ekman's taxonomy, in line with the original article. We applied this model directly to the WASSA test set. This model is called $\text{Baseline}_{\text{FT}}$. As a second baseline, we fine-tuned once again this model with the essays from the training set in a multi-label way. This second model is called $\text{Baseline}_{\text{FT}}$. The pretrained models that we used were made available online using the `transformers` library (Wolf et al., 2019). We used the Adam algorithm (Kingma and Ba, 2014) with early stopping for the optimization of the training loss, using a learning rate of 10^{-5}. We followed the official partitions is both cases.

PER: We used a Big 5 personality model developed by Park et al. (2014a). This model was trained on Facebook status updates and questionnaire-based self-reported Big Five personality traits from 66,732 people. This model used ngrams and topics extracted from the Facebook status updates in an ℓ_2 penalized Ridge regression and resulted in an out-of-sample accuracy (Pearson r) of 0.43 (Openness), 0.37 (Conscientiousness), 0.42 (Extraversion), 0.35 (Agreeableness), and 0.35 (Neuroticism). This model was then applied to each essay in the test set for the shared task, producing Big 5 estimates for each.

IRI: We use the Empathic Concern model built by Giorgi et al. (2023) and train additional models for the three remaining dimensions of the IRI: Fantasy, Perspective Taking, and Personal Distress. These models were built over existing data sets where 2,805 consenting participants shared their lifetime Facebook status updates and responded to the IRI questionnaire (Abdul-Mageed et al., 2017). For each participant, we extract RoBERTa embeddings (averaging word embeddings across sentences, sentence embeddings averaged across Facebook status updates, and status embeddings averaged within participants). We used the second to last layer of the roberta-large model, producing a 1,024-dimensional vector for each participant. Using a penalized Ridge regression (with a ℓ_2 regularization strength of 10^5; tuned during nested cross-validation) in a 10-fold cross-validation resulted in a prediction accuracy (Pearson r) of 0.276 (Empathic Concern), 0.294 (Fantasy), 0.116 (Perspective Taking), and 0.291 (Personal Distress). This model was then applied to each essay in the test set for the shared task after extracting RoBERTA embeddings from each essay.

5 Results and Discussion

5.1 Empathy Prediction (CONV)

Table 2 shows the results of the track on Emotion Polarity (Emo Pol), Emotion Intensity (Emo Int) and Observed Empathy (Emp). All are regression tasks and were evaluated using Pearson correlation. The participants are ranked using the average of all three metrics. Nine teams submitted results to this track. The best system is *HIT-SCIR* obtaining the best results overall (averaged $r = .758$) but also for all the targets: an emotion polarity ($r = .852$), emotion intensity ($r = .714$) and perceived empathy ($r = .708$).

Team	Emo Pol	Emo Int	Emp	Avg
HIT-SCIR	**0.852**	**0.714**	**0.708**	**0.758**
YNU-HPCC	0.824	0.693	0.674	0.730
Hawk	0.809	0.701	0.665	0.725
NCUEE-NLP	0.803	0.698	0.669	0.724
warrior1127	0.770	0.701	0.660	0.710
CAISA	0.783	0.686	0.652	0.707
Curtin OCAI	0.750	0.683	0.573	0.669
sushantkarki	0.778	-0.030	-0.023	0.242
Cordyceps	-0.005	0.039	0.018	0.017
Baseline	0.781	0.692	0.660	0.711

Table 2: Results of the teams participating in the CONV track (Pearson correlations).

5.2 Empathy Prediction (EMP)

Table 3 shows the main results of the track on empathy (Emp) and distress (Dis) prediction. 9 teams submitted results and the best scoring system is *NCUEE-NLP* team (averaged r = .418). They also obtain the best separate scores for empathy and distress with respective r of .415 and .421.

Comparison with previous results: In (Buechel et al., 2018b), the best-performing system obtained r=.404 for empathy and r=.444 for distress. These

Team	Emp	Dis	Avg
NCUEE-NLP	**0.415**	**0.421**	**0.418**
CAISA	0.348	0.420	0.384
PICT-CLRL	0.358	0.334	0.346
zex	0.293	0.391	0.342
HIT-SCIR	0.329	0.354	0.342
YNU-HPCC	0.331	0.245	0.288
Curtin OCAI	0.187	0.344	0.266
Hawk	0.270	0.207	0.238
Cordyceps	-0.020	0.096	0.038
Baseline	**0.536**	**0.575**	**0.555**

Table 3: Results of the teams participating in the EMP track (Pearson correlations).

results were achieved only on the training set using ten-fold cross-validation experiments, which is not comparable to the results in this shared task. In WASSA 2021 and 2022 (Tafreshi et al., 2021; Barriere et al., 2022b), that had the largest training sets, the best scoring systems reached an averaged *r* of .545 and .540. These past scores are in line with the one of the baseline that we proposed, which was trained also using the past years' datasets and gives far better performances than the systems of this year's participants (average *r*=.555 compared to *r*=.418).

5.3 Emotion Recognition (EMO)

Table 4 presents the results for 13 teams for emotion prediction models. The best-performing system in terms of Macro F1 (70.1%) as well as micro-Jaccard (60.1%) is *Adityapatkar* which is significantly higher than the remaining emotion prediction models. To get more insight, we also provide a breakdown of the results by emotion class in Table 7. Fear was easily predicted by the majority of the participant's systems, as per the neutral and sadness classes that are the most present in the dataset. The results are very heterogenous among the participants in the breakdown for all emotion labels. The emotion model submitted by team *LingJing* outperforms the other models on Disgust, while team *andeldiko* performs best on Anger.

5.4 Personality and Interpersonal Reactivity Prediction (PER/IRI)

The results of the tracks on personality and IRI predictions are presented in Table 5. Five and six teams submitted results to respectively the PER and IRI subtasks. The best scoring system for both tasks is the one of *YNU-HPCC*. For the PER task,

Team	P	R	F1	Jac
Adityapatkar	**0.810**	0.677	**0.701**	**0.600**
Bias Busters	0.630	**0.731**	0.647	0.538
HIT-SCIR	0.721	0.631	0.644	0.562
zex	0.699	0.637	0.643	0.562
lazyboy.blk	0.776	0.601	0.613	0.554
Converge	0.596	0.560	0.565	0.539
amsqr	0.752	0.479	0.533	0.507
surajtc	0.463	0.668	0.522	0.451
YNU-HPCC	0.575	0.502	0.514	0.542
VISU	0.257	0.301	0.272	0.421
Cordyceps	0.191	0.236	0.202	0.241
Sidshank	0.295	0.211	0.150	0.287
mimmu3302	0.092	0.200	0.126	0.271
$\text{Baseline}_{\text{FT}}$	0.631	0.645	0.632	0.551
$\text{Baseline}_{\text{EXT}}$	0.860	0.539	0.602	0.522

Table 4: Results of the teams participating in the EMO track (macro-averaged precision (P), recall (R), F1-score (F1) and micro-Jaccard (Jac)).

like last year, participants obtained negative correlations in all the tasks. There were four submitted systems with negative correlations for Extraversion and three with negative correlations for Conscientiousness prediction. For the IRI task, the best results on the different dimensions were distributed over the different teams and the baseline: team *CAISA* obtained the best *r* for the *perspective taking*, team *Xuao* for the *personal distress*, team *Hawk* for the *fantasy* and our baseline for the *empathic concern*.

6 Overview of Submitted Systems

A total of 21 teams participated in the shared tasks, with 9, 9, 13, 5, and 6 teams participating for the five tracks, respectively. In this section, we provide a summary of the machine learning algorithms and resources that were used by the teams.

6.1 Machine Learning Architectures and Resources

Architectures: The majority of the proposed systems are based on neural networks architectures and transformers (Vaswani et al., 2017). Approaches include classical models like BERT (Devlin et al., 2018b), RoBERTa (Liu et al., 2019b), DeBERTaV2 and DeBERTaV3 (He et al., 2020, 2021), ELECTRA (Clark et al., 2020), ALBERT (Lan et al., 2020), DistilBERT (Sanh et al., 2019), SimCSE (Gao et al., 2021), and also transformers for long texts like BigBird and LongFormer (Zaheer et al., 2020; Beltagy et al., 2020). Some of them trained their model using Parameter Efficient Fine Tuning methods like Low-Rank Adapters

Team	Consc.	Open.	Extr.	Agree.	Stab.	**PER**	Persp.	Distr.	Fant.	Emp.	**IRI**
YNU-HPCC	0.289	**0.372**	-0.130	**0.410**	**0.317**	**0.252**	0.102	0.256	0.033	**0.226**	**0.154**
Xuhao	∅	∅	∅	∅	∅	∅	0.132	**0.366**	0.036	0.076	0.153
CAISA	**0.323**	0.327	-0.197	0.290	0.256	0.200	**0.158**	-0.188	-0.056	0.180	0.024
Curtin OCAI	0.186	0.152	0.014	-0.038	0.183	0.099	-0.092	0.193	-0.014	-0.114	-0.007
Cordyceps	-0.059	-0.187	**0.160**	0.101	-0.010	0.001	0.004	0.191	-0.018	0.089	0.067
Hawk	-0.082	0.066	-0.109	-0.119	-0.114	-0.072	-0.013	-0.020	**0.138**	-0.153	-0.012
Baselines	-0.131	-0.037	-0.134	0.195	0.081	-0.005	0.107	-0.046	0.063	**0.340**	0.116

Table 5: Results of the teams participating in the PER/IRI tracks (Pearson correlations).

(Hu et al., 2021) or AdapterHub (Pfeiffer et al., 2020). Only two of the submitted systems used an interaction-aware model. The first one is based on Kim and Vossen (2021), which is able to learn intra- and inter-speaker states and context to predict the emotion of a current speaker. The second one is a RoBERTa transformer using a context window containing past and future utterances. One team used Large Language Models with GPT3 (Brown et al., 2020) and GPT4 (OpenAI, 2023) with in-context learning but also by fine-tuning them. One team used bidirectional LSTM (Hochreiter and Schmidhuber, 1997) with classical text embeddings like Glove (Pennington et al., 2014), Fastext (Bojanowski et al., 2016) and Flair (Akbik et al., 2018). Finally, one team proposed to use a Multinomial Naive Bayes.

Two systems proposed integrating the writer's metadata using in-context learning, one by rewriting the sentences with natural language templates and another by prompting the table.

Resources: Two teams used RoBERTa transformers that were already fine-tuned for sentiment and emotion tasks before fine-tuning them on the data. These models were trained on nearly 58M tweets and fine-tuned for sentiment analysis and emotion recognition using the TweetEval benchmark (Barbieri et al., 2020). One team used the Epitome empathy dataset of (Sharma et al., 2020a) composed of support-seeker and responder posts on Reddit in order to pre-train the weights of their adapter layers. Finally, one team used an interaction-aware model trained on emotion recognition in conversations (Kim and Vossen, 2021).

Others: Three teams used data augmentation to create new examples: two by paraphrasing the under-represented classes using a T5 (Raffel et al., 2019), and one by generating examples of the under-represented classes with a GPT-4. One team used a FLAN-T5 model (Chung et al., 2022) to summarize the long articles in order to reduce the number of tokens used as input to their classifier. Four teams used ensemble methods, which are classics for coding competitions.

ML Alg.	# of team	CONV	EMP	EMO	PER/IRI
BERT-like	11	✓	✓	✓	✓
Ensemble	4	✓	✓	✓	
Data-Aug.	3	✓	✓	✓	✓
Adapters	2	✓	✓	✓	✓
LLM	1			✓	
biLSTM	1			✓	
Naive Bayes	1			✓	

Table 6: Algorithms used by the different teams. We listed all the techniques that teams reported in their system description papers.

7 Conclusion

In this paper, we presented the shared task on empathy, emotion, and personality detection in essays and conversations in reactions to news articles, to which 21 teams participated and 12 submitted a paper. Like last year, neural models are the major parts of the submissions, especially transformer models. The systems obtaining the best results for the five subtasks relied on BERT, RoBERTa, and DeBERTa models. Nobody used task-related features extracted from lexicons, as was the case in the previous editions. External data still helps improve the results, like leveraging Emotion, Sentiment, and Empathy external datasets. Nevertheless, more is needed to make the systems competitive enough to beat fine-tuned bigger models like the biggest DeBERTa (1.3B) used by the winning teams of the CONV, PER, and IRI subtasks. Likewise, using a finely crafted model for interactions cannot compete with a model 10 times its size, using a simple window to integrate context. Finally, some participants used features from a track to give more context, but no approach has considered using multi-task learning between the tracks, even though it was possible to do it. Surprisingly, no teams used the identifier of the speaker to integrate their conversations in order to get more context to find the empathy or emotion of the essay.

Limitations

The test dataset size makes it difficult to draw meaningful conclusions for the Tracks 2 to 5. Similarly, the text data associated with this task (i.e., reaction essays) may make it difficult to infer person-level traits using preexisting models, which may be trained on other domains of text (e.g., social media data). This could explain the negative correlations with extraversion and conscientiousness in Table 5 for the baseline model. Finally, annotating text for emotions and perceived empathy are difficult, subjective tasks. Often statements in the essays are ambiguous and could be interpreted in various ways, especially considering the fact that these are written essay and void of speech cues and the body language of the speaker. Thus, the third-party annotators' own reactions to the news articles could influence how the reaction essays are perceived (e.g., interpreting reactions to the Syrian civil war may depend on the political beliefs of the annotator). Therefore, cultural and social biases may be present in the third party annotations.

Ethics Statement

The main ethical concern is the possibility of misuse of the data and models for manipulation of others. For example, models could be used to produce political ads which elicit empathetic responses which further influence voting or donations. Models could be used to deploy malicious bots on social media platforms (Giorgi et al., 2021), design public health messages (which could be especially problematic around sensitive topics such as vaccines), or spread misinformation (Himelein-Wachowiak et al., 2021). More information are available in the original dataset article (Omitaomu et al., 2022).

Acknowledgements

V.B. has been funded by the grant National Center for Artificial Intelligence CENIA FB210017, Basal ANID.

References

Muhammad Abdul-Mageed, Anneke Buffone, Hao Peng, Salvatore Giorgi, Johannes C Eichstaedt, and Lyle H Ungar. 2017. Recognizing pathogenic empathy in social media. In *ICWSM*, pages 448–451.

Francisca Adoma Acheampong, Henry Nunoo-Mensah, and Wenyu Chen. 2021. Transformer models for text-based emotion detection: a review of bert-based approaches. *Artificial Intelligence Review*, 54(8):5789–5829.

Francisca Adoma Acheampong, Chen Wenyu, and Henry Nunoo-Mensah. 2020. Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports*, 2(7):e12189.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual String Embeddings for Sequence Labeling. In *Coling*.

Lisa P. Argyle, Ethan Busby, Joshua Gubler, Chris Bail, Thomas Howe, Christopher Rytting, and David Wingate. 2023. AI Chat Assistants can Improve Conversations about Divisive Topics. *ArXiv*.

Francesco Barbieri, Jose Camacho-collados, and Leonardo Neves Luis Espinosa-anke. 2020. TWEETEVAL : Unified Benchmark and Comparative Evaluation for Tweet Classification. pages 1644–1650.

Valentin Barriere and Alexandra Balahur. 2023. Multilingual Multi-target Stance Recognition in Online Public Consultations. *accepted to MDPI Mathematics*.

Valentin Barriere, Chloe Clavel, and Slim Essid. 2018. Attitude Classification in Adjacency Pairs of a Human-Agent Interaction with Hidden Conditional Random Fields. In *ICASSP*.

Valentin Barriere, Slim Essid, and Chloé Clavel. 2022a. Opinions in interactions : New annotations of the SEMAINE database. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7049–7055, Marseille, France. European Language Resources Association.

Valentin Barriere, Shabnam Tafreshi, João Sedoc, and Sawsan Alqahtani. 2022b. 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.

C Daniel Batson, Jim Fultz, and Patricia A Schoenrade. 1987. Distress and empathy: Two qualitatively distinct vicarious emotions with different motivational consequences. *Journal of personality*, 55(1):19–39.

Emorie D Beck and Joshua J Jackson. 2022. A mega-analysis of personality prediction: Robustness and boundary conditions. *Journal of Personality and Social Psychology*, 122(3):523.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer.

Srilakshmi Bharadwaj, Srinidhi Sridhar, Rahul Choudhary, and Ramamoorthy Srinath. 2018. Persona traits identification based on myers-briggs type indicator (mbti)-a text classification approach. In *2018 international conference on advances in computing,*

communications and informatics (ICACCI), pages 1076–1082. IEEE.

Paul Bloom. 2017. *Against empathy: The case for rational compassion*. Random House.

Piotr Bojanowski, Edouard Grave, Armand Joulin, Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Tomas Mikolov, Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, Samuel J. Gershman, and Jürgen Schmidhuber. 2016. Bag of Tricks for Efficient Text Classification. *arXiv:1604.00289v1[cs.AI]*, pages 1–55.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners.

Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and João Sedoc. 2018a. Modeling empathy and distress in reaction to news stories. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, (2017):4758–4765.

Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and João Sedoc. 2018b. 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.

Angelo Cafaro, Johannes Wagner, Tobias Baur, Soumia Dermouche, Mercedes Torres Torres, Catherine Pelachaud, Elisabeth André, and Michel Valstar. 2017. The NoXi database: multimodal recordings of mediated novice-expert interactions. In *ICMI*.

Justine Cassell. 2001. Embodied conversational agents: Representation and intelligence in user interfaces. *AI Magazine*, 22(4):67–83.

Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal. 2019. SemEval-2019 task 3: EmoContext contextual emotion detection in text. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 39–48, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Minje Choi, Jiaxin Pei, Sagar Kumar, Chang Shu, and David Jurgens. 2023. Do LLMs Understand Social Knowledge? Evaluating the Sociability of Large Language Models with SocKET Benchmark. (1).

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H Chi, Jeff Dean, Jacob Devlin, Adam Robert, Denny Zhou, Quoc V Le, and Jason Wei. 2022. Scaling Instruction-Finetuned Language Models.

Kevin Clark, Minh-Thang Luong, Quoc Le, and Christopher D. Manning. 2020. Pre-Training Transformers as Energy-Based Cloze Models. pages 285–294.

Mark H Davis. 1980. *Interpersonal Reactivity Index*. Edwin Mellen Press.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A Dataset of Fine-Grained Emotions. pages 4040–4054.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018a. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jacob Devlin, Ming-wei Chang, Kenton Lee, and Kristina Toutanova. 2018b. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Paul Ekman. 1971. Universals and cultural differences in facial expressions of emotion. In *Nebraska symposium on motivation*. University of Nebraska Press.

K Ezzameli and H Mahersia. 2023. Emotion recognition from unimodal to multimodal analysis: A review. *Information Fusion*, page 101847.

Mauajama Firdaus, Hardik Chauhan, Asif Ekbal, and Pushpak Bhattacharyya. 2020. MEISD : A Multimodal Multi-Label Emotion , Intensity and Sentiment Dialogue Dataset for Emotion Recognition and Sentiment Analysis in Conversations. In *COLING*, pages 4441–4453.

Pascale Fung, Dario Bertero, Yan Wan, Anik Dey, Ricky Ho Yin Chan, Farhad Bin Siddique, Yang Yang, Chien-Sheng Wu, and Ruixi Lin. 2016. Towards empathetic human-robot interactions. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 173–193. Springer.

Adithya V Ganesan, Yash Kumar Lal, August Håkan Nilsson, and H. Andrew Schwartz. 2023. Systematic evaluation of gpt-3 for zero-shot personality estimation. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*, pages 6894–6910.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. *arXiv*, 2.

James Gibson, Dogan Can, Bo Xiao, Zac E Imel, David C Atkins, Panayiotis Georgiou, and Shrikanth Narayanan. 2016. A deep learning approach to modeling empathy in addiction counseling. *Commitment*, 111:21.

Salvatore Giorgi, Shreya Havaldar, Farhan Ahmed, Zuhaib Akhtar, Shalaka Vaidya, Gary Pan, Lyle H Ungar, H Andrew Schwartz, and Joao Sedoc. 2023. Human-centered metrics for dialog system evaluation. *arXiv preprint arXiv:2305.14757*.

Salvatore Giorgi, Lyle Ungar, and H. Andrew Schwartz. 2021. Characterizing social spambots by their human traits. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5148–5158, Online. Association for Computational Linguistics.

Samuel D Gosling, Peter J Rentfrow, and William B Swann Jr. 2003a. A very brief measure of the big-five personality domains. *Journal of Research in personality*, 37(6):504–528.

Samuel D Gosling, Peter J Rentfrow, and Williams B Swann Jr. 2003b. A very brief measure of the big-five personality domains. *Journal of Research in Personality*, 37:504–528.

Bhanu Prakash Reddy Guda, Aparna Garimella, and Niyati Chhaya. 2021. Empathbert: A bert-based framework for demographic-aware empathy prediction. *arXiv preprint arXiv:2102.00272*.

Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. 2018. ICON : Interactive Conversational Memory Network for Multimodal Emotion Detection. In *EMNLP*, pages 2594–2604.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. pages 1–17.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. *arXiv*, pages 1–21.

McKenzie Himelein-Wachowiak, Salvatore Giorgi, Amanda Devoto, Muhammad Rahman, Lyle Ungar, H Andrew Schwartz, David H Epstein, Lorenzo Leggio, and Brenda Curtis. 2021. Bots and misinformation spread on social media: Implications for covid-19. *Journal of medical Internet research*, 23(5):e26933.

Sepp Hochreiter and Jurgen Schmidhuber. 1997. LONG SHORT-TERM MEMORY. *Neural Computation*, 9(8):1735–1780.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Shaoxiong Ji, Shirui Pan, Xue Li, Erik Cambria, Guodong Long, and Zi Huang. 2020. Suicidal ideation detection: A review of machine learning methods and applications. *IEEE Transactions on Computational Social Systems*, 8(1):214–226.

Oliver P John, Sanjay Srivastava, et al. 1999. *The Big-Five trait taxonomy: History, measurement, and theoretical perspectives*, volume 2. University of California Berkeley.

Taewoon Kim and Piek Vossen. 2021. EmoBERTa: Speaker-Aware Emotion Recognition in Conversation with RoBERTa.

Diederik Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations*, pages 1–13.

Klaus Krippendorff. 2013. Content Analysis: An Introduction to Its Methodology. In *Content Analysis: An Introduction to Its Methodology*.

Allison Lahnala, Charles Welch, David Jurgens, and Lucie Flek. 2022a. A Critical Reflection and Forward Perspective on Empathy and Natural Language Processing. *Findings of the Association for Computational Linguistics: EMNLP 2022*, (3):2139–2158.

Allison Lahnala, Charles Welch, Béla Neuendorf, and Lucie Flek. 2022b. Mitigating toxic degeneration with empathetic data: Exploring the relationship between toxicity and empathy. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4926–4938, Seattle, United States. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Kevin Gimpel, Sebastian Goodman, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A LITE BERT FOR SELF-SUPERVISED LEARNING OF LANGUAGE REPRESENTATIONS. In *ICLR*.

Seunggun Lee, Alexandra DeLucia, Nikita Nangia, Praneeth S. Ganedi, Ryan Min Guan, Rubing Li, Britney A. Ngaw, Aditya Singhal, Shalaka Vaidya, Zijun Yuan, Lining Zhang, and João Sedoc. 2023. Common law annotations: Investigating the stability of dialog system output annotations. In *Findings of the Association for Computational Linguistics: ACL 2023*, Toronto, Canada. Association for Computational Linguistics.

Marina Litvak, Jahna Otterbacher, Chee Siang Ang, and David Atkins. 2016. Social and linguistic behavior and its correlation to trait empathy. In *Proceedings*

of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES), pages 128–137.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. RoBERTa: A Robustly Optimized BERT Pretraining Approach. (1).

Veronica Lynn, Niranjan Balasubramanian, and H. Andrew Schwartz. 2020. Hierarchical modeling for user personality prediction: The role of message-level attention. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5306–5316, Online. Association for Computational Linguistics.

Yukun Ma, Khanh Linh Nguyen, Frank Z. Xing, and Erik Cambria. 2020. A survey on empathetic dialogue systems. *Information Fusion*, 64(May):50–70.

François Mairesse and Marilyn a. Walker. 2006. Automatic recognition of personality in conversation. *Proceedings of the Human Language Technology Conference of the NAACL*, (June):85–88.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. DialogueRNN: An Attentive RNN for Emotion Detection in Conversations. In *AAAI*.

Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schröder. 2012. The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3(1):5–17.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 1–17.

Saif M Mohammad and Felipe Bravo-Marquez. 2017. Wassa-2017 shared task on emotion intensity. *arXiv preprint arXiv:1708.03700*.

Ashritha R Murthy and KM Anil Kumar. 2021. A review of different approaches for detecting emotion from text. In *IOP Conference Series: Materials Science and Engineering*, volume 1110, page 012009. IOP Publishing.

Isabel Briggs Myers, Mary H McCaulley, and Robert Most. 1985. *Manual, a guide to the development and use of the Myers-Briggs type indicator*. consulting psychologists press.

Pansy Nandwani and Rupali Verma. 2021. A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining*, 11(1):1–19.

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.

OpenAI. 2023. GPT-4 Technical Report. 4:1–100.

Gregory Park, H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L Kern, Michal Kosinski, David J. Stillwell, Lyle H. Ungar, and Martin E.P. Seligman. 2014a. Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*.

Sunghyun Park, Han Suk Shim, Moitreya Chatterjee, Kenji Sagae, and Louis-Philippe Morency. 2014b. Computational Analysis of Persuasiveness in Social Multimedia: A Novel Dataset and Multimodal Prediction Approach. *Proceedings of the 16th International Conference on Multimodal Interaction - ICMI '14*, pages 50–57.

Dhaval Parmar, Stefan Olafsson, Dina Utami, Prasanth Murali, and Timothy Bickmore. 2022. Designing empathic virtual agents: manipulating animation, voice, rendering, and empathy to create persuasive agents. *Autonomous Agents and Multi-Agent Systems*, 36(1):17.

Rebecca Passonneau. 2006. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006*, pages 831–836.

Catherine Pelachaud, Carlos Busso, and Dirk Heylen. 2021. Multimodal Behavior Modeling for Socially Interactive Agents. *The Handbook on Socially Interactive Agents*, 1:259–310.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543.

Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, and Lawrence An. 2017. Understanding and predicting empathic behavior in counseling therapy. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1426–1435, Vancouver, Canada. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulic, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. AdapterHub : A Framework for Adapting Transformers. *arXiv preprint arXiv:2007.07779*.

Rosalind W Picard. 2000. *Affective computing*. MIT press.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019a. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In *ACL*, pages 527–536.

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019b. Emotion Recognition in Conversation: Research Challenges, Datasets, and Recent Advances. pages 1–10.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. pages 1–53.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Harry T Reis, Edward P Lemay Jr, and Catrin Finkenauer. 2017. Toward understanding understanding: The importance of feeling understood in relationships. *Social and Personality Psychology Compass*, 11(3):e12308.

Zhancheng Ren, Qiang Shen, Xiaolei Diao, and Hao Xu. 2021. A sentiment-aware deep learning approach for personality detection from text. *Information Processing & Management*, 58(3):102532.

Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. 2013. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2013*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv*, pages 2–6.

Björn Schuller, Stefan Steidl, and Anton Batliner. 2009. The INTERSPEECH 2009 emotion challenge. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 312–315.

Ashish Sharma, Inna W. Lin, Adam S. Miner, David C. Atkins, and Tim Althoff. 2023. Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence*, 5(1):46–57.

Ashish Sharma, Adam S. Miner, David C. Atkins, and Tim Althoff. 2020a. A computational approach to understanding empathy expressed in text-based mental health support WARNING: This paper contains content related to suicide and self-harm. *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pages 5263–5276.

Ashish Sharma, Adam S Miner, David C Atkins, and Tim Althoff. 2020b. A computational approach to understanding empathy expressed in text-based mental health support. *arXiv preprint arXiv:2009.08441*.

Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas PYKL, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Bjorn Gamback. 2020c. Semeval-2020 task 8: Memotion analysis–the visuolingual metaphor! *arXiv preprint arXiv:2008.03781*.

Weizhou Shen, Junqing Chen, Xiaojun Quan, and Zhixian Xie. 2020. DialogXL: All-in-One XLNet for Multi-Party Conversation Emotion Recognition.

M. Soleymani, M. N. Caro, E. M. Schmidt, and Y. H. Yang. 2013. The MediaEval 2013 brave new task: Emotion in music. In *Working Notes Proceedings of the MediaEval 2015 Workshop*, volume 1043.

Carlo Strapparava and Rada Mihalcea. 2007. Semeval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74.

Michael M Tadesse, Hongfei Lin, Bo Xu, and Liang Yang. 2018. Personality predictions based on user behavior on the facebook social media platform. *IEEE Access*, 6:61959–61969.

Shabnam Tafreshi, Orphee 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, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *arXiv:1706.03762 [cs]*.

Alessandro Vinciarelli, Maja Pantic, Hervé Bourlard, and Alex Pentland. 2008. Social signals, their function, and automatic analysis. *Proceedings of the 10th international conference on Multimodal interfaces - IMCI '08*, page 61.

Hetal Vora, Mamta Bhamare, and Dr K Ashok Kumar. 2020. Personality prediction from social media text: An overview. *Int. J. Eng. Res*, 9(05):352–357.

Huy Vu, Suhaib Abdurahman, Sudeep Bhatia, and Lyle Ungar. 2020. Predicting responses to psychological questionnaires from participants' social media posts and question text embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1512–1524.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing.

Bo Xiao, Chewei Huang, Zac E Imel, David C Atkins, Panayiotis Georgiou, and Shrikanth S Narayanan. 2016. A technology prototype system for rating therapist empathy from audio recordings in addiction counseling. *PeerJ Computer Science*, 2:e59.

Bo Xiao, Zac E Imel, Panayiotis G Georgiou, David C Atkins, and Shrikanth S Narayanan. 2015. " rate my therapist": automated detection of empathy in drug and alcohol counseling via speech and language processing. *PloS one*, 10(12):e0143055.

David B Yaden, Salvatore Giorgi, Matthew Jordan, Anneke Buffone, Johannes C Eichstaedt, H Andrew Schwartz, Lyle Ungar, and Paul Bloom. 2023. Characterizing empathy and compassion using computational linguistic analysis. *Emotion*.

Feifan Yang, Xiaojun Quan, Yunyi Yang, and Jianxing Yu. 2021. Multi-document transformer for personality detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14221–14229.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Others. 2020. Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33:17283–17297.

Ran Zhao, Tanmay Sinha, Alan W Black, and Justine Cassell. 2016. Socially-aware virtual agents: Automatically assessing dyadic rapport from temporal patterns of behavior. In *Proceedings of Intelligent Virtual Agents (IVA 2016).*, volume 10011 LNAI.

Peixiang Zhong, Chen Zhang, Hao Wang, Yong Liu, and Chunyan Miao. 2020. Towards persona-based empathetic conversational models. *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pages 6556–6566.

Naitian Zhou and David Jurgens. 2020. Condolence and empathy in online communities. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

A Emotion-level scores

The scores of the participants' systems by emotion label are visible in Table 7. The classes *hope* and *surprise* are absent from the test set.

	Anger			Disgust			Fear			Neutral			Sadness		
Team	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Adityapatkar	86	25	39	67	40	**50**	100	100	**100**	68	89	77	85	85	**85**
Bias Busters	83	42	**56**	22	40	28	83	100	91	53	94	68	73	89	80
HIT-SCIR	88	28	44	25	20	22	100	100	**100**	67	86	76	80	80	80
zex	78	28	42	25	20	22	100	100	**100**	68	89	77	79	80	80
lazyboy.blk	100	12	22	33	20	25	100	100	**100**	67	92	**78**	88	76	81
Converge	50	25	33	0	0	0	100	80	89	69	86	77	79	89	84
amsqr	100	21	34	33	20	25	100	40	56	64	81	72	78	78	78
surajtc	50	28	37	6	20	9	56	100	71	51	100	68	68	85	76
YNU-HPCC	62	21	31	0	0	0	75	60	67	67	83	74	83	87	**85**
VISU	0	0	0	0	0	0	0	0	0	63	53	57	65	98	78
Cordyceps	14	4	6	0	0	0	0	0	0	37	53	43	44	61	51
Sidshank	0	0	0	0	0	0	0	0	0	100	6	11	47	100	64
mimmu3302	0	0	0	0	0	0	0	0	0	0	0	0	46	100	63
Baseline	56	54	55	14	20	17	100	80	89	67	83	74	78	85	81

Table 7: Emotion-level participants performances