Findings of WASSA 2024 Shared Task on Empathy and Personality Detection in Interactions

Salvatore Giorgi¹, João Sedoc², Valentin Barriere³, Shabnam Tafreshi⁴

¹National Institute on Drug Abuse, Baltimore, USA ²New York University, New York, USA

³Universidad de Chile – DCC | CENIA, Santiago, Chile

⁴inQbator AI at eviCore Healthcare, Evernorth Healthcare

sal.giorgi@gmail.com,jsedoc@stern.nyu.edu,

vbarriere@dcc.uchile.cl,stafresh@umd.edu

Abstract

This paper presents the results of the WASSA 2024 shared task on predicting empathy, emotion, and personality in conversations and reactions to news articles. Participating teams were given access to a new, unpublished extension of the WASSA 2023 shared task dataset. This task is both multi-level and multi-modal: data is available at the person, essay, dialog, and dialog-turn levels and includes formal (news articles) and informal text (essays and dialogs), self-report data (personality and distress), and third-party annotations (empathy and emotion). The shared task included a new focus on conversations between humans and LLM-based virtual agents which occur immediately after reading and reacting to the news articles. Participants were encouraged to explore the multilevel and multi-modal nature of this data. Participation was encouraged in four tracks: (i) predicting the perceived empathy at the dialog level, (ii) predicting turn-level empathy, emotion polarity, and emotion intensity in conversations, (iii) predicting state empathy and distress scores, and (iv) predicting personality. In total, 14 teams participated in the shared task. We summarize the methods and resources used by the participating teams.

1 Introduction

Empathy, emotions, and similar affective states are fundamental human relationships, informing complex social interactions and cognition (Cassell, 2001; Picard, 2000). These states can be consciously and unconsciously expressed through facial expressions, writing, speech, body language, and mimicry (Shoumy et al., 2020) and have effects on cooperation (Manson et al., 2013), romantic relationships (Ireland et al., 2011), and therapist ratings (Lord et al., 2015). In the context of humanagent interactions, these phenomena are essential to make machines understand the world and have humans actively and genuinely engage with them.

Dialog systems have become increasingly conversant, due to advances in generative artificial intelligence. As such, it has been suggested that these systems could be applied across a wide range of human facing applications, such as mental and behavioral healthcare and substance use recovery (Demszky et al., 2023; Stade et al., 2024b,a; Giorgi et al., 2024). Along with this, automatic agents are being designed with human-like traits such as empathy (Rashkin et al., 2019), emotion (Zhou and Wang, 2018; Huber et al., 2018), and personas (Roller et al., 2021; Cheng et al., 2023). There are also attempts to align such systems with human preferences, opinions, beliefs, and culture (Santurkar et al., 2023; Scherrer et al., 2024; Havaldar et al., 2023). Despite this active attention, implementations of such systems have had limited success and standardized data sets in which empathy, and related complex emotional states, can be modeled are in short supply Omitaomu et al. (2022).

While studying affect-related phenomena in the context of automatic dialog agents and human-bot interactions has become ubiquitous, Lahnala et al. (2022) has noted that concepts such as empathy are traditionally poorly defined. This shared task attempts to address these issues by presenting participants with multiple, psychologically grounded definitions of empathy allowing participants to study multiple forms of empathy, varying measurements of empathy (e.g., self-report and other-report), and their interactions. This includes trait empathy (empathy which is stable over time and systematically differs across people), state empathy (empathy experienced at a specific place in time), perceived empathy (how one's conversational partner views their empathy), and conversational turn empathy (the level of empathy expressed at each stage of a conversation).

This paper presents the WASSA 2024 Empathy Shared Task on Empathy and Personality Detection in Interactions, which allows studying empathy, personality, and perception in human-human and human-bot 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., 2022, 2023). 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., 2018a) 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 self-reported empathy and distress. 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. Similar to the WASSA 2023 shared task, we include subsequent conversations that the study participants had after writing their essays. Every turn in the conversation was third-party annotated for empathy, emotion, and emotional polarity. In this year's task, we introduced two new components, the first is that after each conversation, participants rated the empathy of their conversational partner. The second new component of the task is that we included conversations between people and a conversational AI system.

Given this dataset as input, the shared task consists of four tracks (see Section 4.1 for each tracks' respective definitions of empathy and emotion):

- 1. Empathy Prediction in Conversations (CONV-Dialog), which consists of predicting the perceived empathy at the dialog-level.
- 2. Empathy and Emotion Prediction in Conversations Turns (CONV-Turn), which consists in predicting the perceived empathy, emotion polarity, and emotion intensity at the speechturn-level in a conversation.
- 3. Empathy Prediction (EMP), which consists of predicting both the empathic concern and the personal distress at the essay-level.
- 4. Personality Prediction (PER), which consists of predicting the personality (openness, conscientiousness, extraversion, agreeableness, and emotional stability; OCEAN) of the essay writer, knowing all his/her essays, dialogs, and the news articles from which they reacted.

2 Related Work

2.1 Empathy and Distress

Several studies have attempted to both predict empathy from text (Litvak et al., 2016), model empathy in counseling (Xiao et al., 2015, 2016), and build empathetic conversational agents (Rashkin et al., 2019; Gao et al., 2021). Hosseini and Caragea (2021) looks at members of online health communities seek or provide empathy. Other studies have used appraisal theory to study empathy in Reddit conversations (Zhou and Jurgens, 2020; Yang and Jurgens, 2024). Computational methods have also been used to distinguish between good and bad types of empathy (Yaden et al., 2023; Abdul-Mageed et al., 2017), where "bad empathy" involves taking on the feelings of others, which can lead to empathetic burnout, while "good empathy" (also called compassion) is a prosocial motivator without emotional contagion. Given the extensive work on this subject, there are several in depth review articles (Shetty et al., 2024; Lahnala et al., 2022).

2.2 Personality Prediction

Predicting personality from text (including social media data) has been rigorously validated via psychometric tests (e.g., convergent validity, divergent validity, and test-retest reliability; Park et al., 2015). Dialog agents are also being designed with personalities (Liu et al., 2022). Similarly, large Language Models are being used for personality classification (Ganesan et al., 2023; Peters and Matz, 2024; Yang et al.) and for studying the personality of the model's themselves (Safdari et al., 2023; Salecha et al., 2024; Miotto et al., 2022). Similar to empathy, there are several survey papers on personality prediction models, theories, and techniques (Vora et al., 2020; Beck and Jackson, 2022).

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., 2022), it can be considered an extension. 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.

	Train	Dev	Test
People	75	83	34
Conversations	500	33	67
Essays	1,000	66	83
Speech-Turns	11,166	990	2,316

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 Par-IAI 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; Davis, 1983). 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 (Batson et al., 1987). 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.

4 Shared Task

We set up all four tracks in CodaLab¹. We describe each task separately in Section 4.1 and then describe dataset, resources, and evaluation metrics in Section 4.2. Tracks 2, 3, and 4 are similar to the ones offered by WASSA 2022 and 2023 shared tasks.

¹https://codalab.lisn.upsaclay.fr/ competitions/18810

4.1 Tracks

Track 1 - Empathy Prediction in Conversations (**CONV-Dialog**): The formulation of this task is to predict, for each conversation, the perceived empathy at the dialog-level. As described in Sec 3, immediately after each conversation participants were asked to rate the empathy of their conversational partner towards the patient of harm on a 1 to 7 ordinal scale. The participants were asked to predict the rated value of the partner. In the case of a human-bot conversation, there was only one rating since the conversational AI system was not tasked to do this rating. This track was newly introduced as part of this year's shared task.

Track 2 - Turn-level Empathy and Emotion in Conversations (CONV-Turn): 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 was introduced in WASSA 2023, but the data in this year's task (2024) is new.

Track 3 - 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 2023, 2022, and 2021, but this year's task uses new data.

Track 4 - Personality Prediction (PER): This task asked participants to predict personality scores for each essay. To code personality information, the Big 5 personality traits were provided, also known

Team	Perceived Empathy
Fraunhofer SIT	.193
ConText	.191
Chinchunmei	.172
EmpatheticFIG	.012
Baseline	.023

Table 2: Track 1 CONV-Dialog: Results of the teams participating in the EMP track (product moment correlations), order by descending effect size.

as the OCEAN model (Gosling et al., 2003b). In the OCEAN model, the theory identifies five factors (Openness to experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism²). Participants were asked to produce scores for each of the five factors. For each essay, the writer was asked to complete the Ten Item Personality Inventory (Gosling et al., 2003a), two items for each of the five factors. Thus, this is self-reported essay-level data. This task was previously done in WASSA 2022 and 2023, but the data in this year's task (2024) is new.

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 in WASSA 2022 for the PER and IRI subtasks (Barriere et al., 2022).

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.

Team	Emotion Intensity	Emotion Polarity	Empathy	Avg
ConText	.622	.679	.577	.626
Chinchunmei	.607	.680	.582	.623
EmpatheticFIG	.601	.671	.559	.610
Last-min-submission-team	.589	.663	.534	.595
hyy3	.581	.644	.544	.590
Empathify	.584	.638	.541	.588
empaths	.473	.422	.534	.477
Fraunhofer SIT	.032	018	.034	007
Zhenmei	043	020	027	030
Baseline	.417	.646	.694	.586

Table 3: Track 2 CONV-Turn: Results of the teams participating in the CONV-Turn track (product moment correlations), order by average descending effect size.

Systems Evaluation: The organizers published an evaluation script that calculates product moment correlations for the predictions of the four tasks. The product moment correlation coefficient is the linear correlation 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 in conversations task (CONV-Dialog) is the product moment correlation for perceived empathy. The official competition metric for the empathy and emotion in conversation task (CONV-Turn) is the average of the three product moment correlations (emotion intensity, emotion polarity, and empathy). The official competition metric for the state empathy prediction task (EMP) is the average of the two product moment correlations (empathy and distress). The official competition metric for the personality task (PER) is the average of the product moment correlations of the five factors.

4.3 Baselines

CONV-Dialog: Similar to Omitaomu et al. (2022), we fine-tuned a RoBERTa (base) pretrained language model (Liu et al., 2019). The model was trained on the training set and used the development set for model validation. We trained one model for all dialog turns of the person being assessed. The model was trained using regression since this was on a 7 point scale. The training was for 30 epochs, and the model checkpoint with the best validation set product moment correlation was kept.

CONV-Turn: Following Omitaomu et al. (2022), we fine-tuned a RoBERTa (base) pretrained language model (Liu et al., 2019). The model was trained on the training set and used the development set for model validation. We trained one

²For the shared task, neuroticism has been reverse coded as emotional stability

Team	Empathy	Distress	Avg.
RU	.523	.383	.453
Chinchunmei	.474	.311	.393
Fraunhofer SIT	.375	.395	.385
1024m	.361	.327	.344
ConText	.390	.252	.321
Empathify	.290	.217	.253
Daisy	.345	.082	.213
Baseline	.629	.477	.553

Table 4: Track 3 EMP: Results of the teams participating in the EMP track (product moment correlations).

model for each of the turn-level label types. The training was for 30 epochs, and the model check-point with the best validation set product moment correlation was kept.

EMP: Like the CONV models, we fine-tuned a RoBERTa (base) pretrained language model (Liu et al., 2019). For training, we used both the training data of the essays and the WASSA22 and WASSA23 training data (Barriere et al., 2023, 2022). We created separate models for empathy and distress, and used the same checkpoint and stopping criteria as the CONV task models.

PER: Similar to the 2023 shared task, we used a Big 5 personality model developed by Park et al. (2015). This model was trained on Facebook status updates from 66,732 people who self-reported questionnaire-based Big Five personality traits. This model used ngrams and topics extracted from the Facebook status updates in an ℓ_2 penalized Ridge regression. This model was then applied to all text generated by each person in the test set (i.e., essays and conversations), producing Big 5 estimates for each.

5 Results and Discussion

A total of 14 teams participated in this year's shared task, with 4, 9, 7, and 3 teams across the four tracks, respectively. The results for each task are summarized below.

5.1 Empathy Prediction (CONV-Dialog)

Table 2 shows the results for Track 1. Here participants were asked to predict (via a regression task) perceived empathy, as rated by conversational partners. A total of four teams participated in this track, and were evaluated via product moment correlation. The system with the highest test set correlation was Fraunhofer SIT (r = .193), though it should be noted that no team had a statistically significant correlation (p < .05).

5.2 Turn-level Empathy and Emotion Prediction (CONV-Turn)

Table 3 shows the results of Track 2. Participants were asked to predict third-party assessments emotion intensity, emotion polarity, and empathy for each turn in the dialogs. Teams were evaluated via the average product moment correlation across the three measures. Team *ConText* had the highest average correlation (r = .626), as well as the highest correlation for emotion intensity (r = .622). Team *Chinchunmei* had the highest correlations for emotion polarity (r = .680) and empathy (r = .582), and also had the second highest average correlation (r = .623). The RoBERTa-base baseline outperforms on Empathy, but underperformed across all other dimensions. All top correlations were statistically significant (p < .05).

5.3 State Empathy and Distress Predictions (EMP)

Table 4 shows the results for Track 3. A total of 7 teams participated in this track. Teams were ranked via the average product moment correlation across both empathy and distress. Team *RU* had the highest ranked system with an average correlation of r = .453. This team also had the highest empathy correlation (r = .523). Team *Fraunhofer SIT* had the highest distress correlation (r = .395). Our baseline RoBERTa-base model outperformed all other models likely due to the inclusion of the data from the prior years. All top correlations were statistically significant (p < .05).

5.4 Personality Predictions (PER)

Table 5 shows the results for Track 4. Three teams participated in this task. Team *amsqr* attained the highest average correction (r = .300). This team's results for the Agreeableness (r = .540) and Emotional Stability (r = .757) factors were the only significant correlations in the results (p < .05). Only one team that participated in this task submitted a system description paper. Thus, it is unclear how *amsqr* achieved their results.

5.5 Comparison with previous results

Table 6 shows the results of each track across previous shared tasks from 2021 (Tafreshi et al.,

Team	Openness	Conscientiousness	Extraversion	Agreeableness	Emotional Stability	Avg.
amsqr	.170	.303	272	.540	.757	.300
NU	103	.102	085	.154	.279	.069
1024m	.059	032	128	015	009	025
Baseline	.042	.207	.300	.127	012	.133

Table 5: Track 4 PER: Results of the teams participating in the PER track (product moment correlations).

2021), 2022 (Barriere et al., 2022), and 2023 (Barriere et al., 2023). Note that the numbers reported for (Buechel et al., 2018b) are the average empathy and distress scores for their best system (r = .404for empathy and r = .444 for distress). This system used 10-fold cross validation, rather than dedicated train, development, and test sets, and thus, these results are not comparable to those in this shared task. In WASSA 2023, which shared three out of four tracks, results for Track 2 (CONV-Turn) were higher in magnitude, with eight out of ten systems (including baseline) scoring above (highest average r = .758) the best team in this year's shared task (ConText). For Track 3 (EMP), this year's best scoring team (RU) outperformed all of the nine teams that participated in this track in 2023 (highest average r = .418). Finally, for Track 4 (PER) the highest performing team this year (amsqr) outperformed last year's teams (highest average r = .252).

6 Overview of Submitted Systems

Below we summarize the algorithms and resources used by the teams.

6.1 Machine Learning Architectures

The machine learning architectures used by the participating teams are summarized in Table 7. Similar to last year's shared task, most teams relied on BERT (Devlin et al., 2019) and related variants: RoBERTa (Liu et al., 2019), DeBERTa (He

	CONV Dialog	CONV Turn	EMP	PER
Buechel et al. (2018b)	-	-	.242	-
WASSA 2021	-	-	.545	-
WASSA 2022	-	-	.540	.230
WASSA 2023	-	.758	.418	.252
This year (2024)	.193	.626	.453	.300

Table 6: Comparison of best performing scores across each track for previous years of the shared task. Reported average product moment correlation. et al.), SieBERT (Hartmann et al., 2023), and similar variants finetuned for sentiment (Barbieri et al., 2020). These models were used outof-the-box, and in custom architectures, such as those used to create history-dependent embedding. New this year are systems based on more modern large language models, such as Llama (Touvron et al., 2023) and GPT variants (e.g., ChatGPT and GPT-40; Brown et al., 2020). Two systems utilized psychologically-grounded or theory-based features, where features were chosen based on their known relationships to emotion and empathy. Finally, other transformer-based models were used, though less often: T5 (Raffel et al., 2020) and Longformer (Beltagy et al., 2020).

6.2 Resources

Two teams used a RoBERTa model finetuned on the GoEmotions data set (Demszky et al., 2020), which contains Reddit posts annotated for 27 emotions. One team analyzed their model using a SHAP (SHapley Additive exPlanations) analysis (Lundberg and Lee, 2017) in tandem with the GoEmotions-based emotions classifier, which highlighted which emotions were most associated with empathy and distress. Several teams used large language models for data augmentation. Llama and GPT-related models were used for paraphrasing and for predicting psychological indicators (e.g., perspective-taking, sympathy, and compassion) used downstream to predict turn-level empathy. Similarly, mT5 (Xue et al., 2021) was used to extract figurative language (metaphor, idiom, and hyperbole) labels for conversational turns. While not an external resource, one team used a built a knowledge graph from self reported demographics (age, gender, income, and education) and empathy (trait level; IRI). Finally, LIWC (Boyd et al., 2022) was used to understand what types of turnlevel text were most associated emotional intensity, emotional polarity, and empathy.

Alg.	# of teams	CONV Dialog	CONV Turn	EMP	PER
BERT-like	9	\checkmark	\checkmark	\checkmark	\checkmark
GPT-like	3	-	\checkmark	\checkmark	-
Llama-like	3	\checkmark	\checkmark	\checkmark	-
Theory based	2	\checkmark	\checkmark	-	-

Table 7: Algorithms used by the different teams. We listed all the techniques that teams reported in their system description papers. Note that not all participating teams submitted system papers.

7 Conclusion

In this paper, we presented the shared task on empathy and personality detection in essays and conversations in reactions to news articles, to which 14 teams participated and 12 submitted a paper. While this year saw an increase in large language models (i.e., GPT and Llama variants), though most teams still relied on BERT and RoBERTa variants. A small number of teams used task related resources, such as models finetuned on sentiment and emotion datasets. No external data sets were used, though generative AI systems were used for data augmentation. Two teams used empathy and emotion related features, such as figurative language and related psychologically-grounded indicators. One team used historical conversation context, when predicting turn-level labels. A single team used speaker self-report data, such as demographics (age, gender, race), socioeconomics (income and education), and trait-level empathy (IRI) alongside text-based features. Finally, no teams used used multi-task learning, leveraged the multi-modal or multi-level nature of the data, or made a distinction between bot and human data.

Limitations

This shared task and the associated data are limited in several ways. First, despite the multi-modal aspects of the data, the number of words collect per person is limited (by both essay and conversation length). Thus, it may be more difficult to predict trait level measures (e.g., personality) than state or turn-level measures (e.g., emotion and empathy). This could explain the low performance in Track 4 PER. We note that past work on detecting personality from social media has used a minimum of 1000 words for accurate predictions (Lynn et al., 2020). Similarly, personality models are typically trained on larger personality questionnaires, such as the 20- or 100-item personality inventory (Park et al., 2015). The 10-item version could lead to noisier personality estimates which are more difficult to predict from text. Second, the person- and essaylevel tasks both have small sample sizes, which could explain why none of the correlations in Track 1 were significant (i.e., the task was under powered). Finally, we only consider English language news articles and English conversations from crowd workers in the U.S., thus limiting the study of empathy to these cultures and languages/dialects.

Ethics Statement

There are several ethical concerns one should consider when predicting affective measures. There is mounting evidence that emotion recognition systems are being use as part of mass surveillance systems by governments and private entities worldwide (Barkane, 2022). These systems are used in high stakes settings, such as law enforcement, and are known to discriminate and violate rights to privacy (Kieslich and Lünich, 2024). While there are many prosocial applications to embedding machines with emotions and empathy, such as mental health related chatbots, there are also several similar concerns. Such empathetic or emotional systems could increase trust with their end users through anthropromorphisms (Abercrombie et al., 2023), which could have several nefarious use cases (for example, spreading misinformation or political ads which elicit empathetic responses). Similarly, overly empathetic or agreeable systems could ignore or agree with unsafe or toxic input (Kim 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 Eichstaedt, and Lyle Ungar. 2017. Recognizing pathogenic empathy in social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, pages 448–451.
- Gavin Abercrombie, Amanda Cercas Curry, Tanvi Dinkar, Verena Rieser, and Zeerak Talat. 2023. Mirages. on anthropomorphism in dialogue systems. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4776–4790.

- Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association* for Computational Linguistics: EMNLP 2020, pages 1644–1650.
- Irena Barkane. 2022. Questioning the eu proposal for an artificial intelligence act: The need for prohibitions and a stricter approach to biometric surveillance 1. *Information Polity*, 27(2):147–162.
- 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.
- 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 megaanalysis 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. *arXiv* preprint arXiv:2004.05150.
- Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. 2022. The development and psychometric properties of liwc-22. *Austin, TX: University of Texas at Austin,* 10.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901.
- Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and João Sedoc. 2018a. 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.
- 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, Brussels, Belgium. Association for Computational Linguistics.

- Justine Cassell. 2001. Embodied conversational agents: Representation and intelligence in user interfaces. *AI Magazine*, 22(4):67–83.
- Myra Cheng, Tiziano Piccardi, and Diyi Yang. 2023. CoMPosT: Characterizing and evaluating caricature in LLM simulations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10853–10875, Singapore. Association for Computational Linguistics.
- Mark H Davis. 1983. Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of personality and social psychology*, 44(1):113.
- Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. arXiv preprint arXiv:2005.00547.
- Dorottya Demszky, Diyi Yang, David S Yeager, Christopher J Bryan, Margarett Clapper, Susannah Chandhok, Johannes C Eichstaedt, Cameron Hecht, Jeremy Jamieson, Meghann Johnson, et al. 2023. Using large language models in psychology. *Nature Reviews Psychology*, 2(11):688–701.
- 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.
- 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.
- Jun Gao, Yuhan Liu, Haolin Deng, Wei Wang, Yu Cao, Jiachen Du, and Ruifeng Xu. 2021. Improving empathetic response generation by recognizing emotion cause in conversations. In *Findings of the association for computational linguistics: EMNLP 2021*, pages 807–819.
- Salvatore Giorgi, Kelsey Isman, Tingting Liu, Zachary Fried, Brenda Curtis, et al. 2024. Evaluating generative ai responses to real-world drug-related questions. *Psychiatry Research*, page 116058.
- Samuel D Gosling, Peter J Rentfrow, and William B Swann Jr. 2003a. A very brief measure of the bigfive 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 bigfive personality domains. *Journal of Research in Personality*, 37:504–528.
- Jochen Hartmann, Mark Heitmann, Christian Siebert, and Christina Schamp. 2023. More than a feeling: Accuracy and application of sentiment analysis. *International Journal of Research in Marketing*, 40(1):75–87.
- Shreya Havaldar, Bhumika Singhal, Sunny Rai, Langchen Liu, Sharath Chandra Guntuku, and Lyle Ungar. 2023. Multilingual language models are not multicultural: A case study in emotion. In Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 202–214.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.
- Mahshid Hosseini and Cornelia Caragea. 2021. It takes two to empathize: One to seek and one to provide. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 13018–13026.
- Bernd Huber, Daniel McDuff, Chris Brockett, Michel Galley, and Bill Dolan. 2018. Emotional dialogue generation using image-grounded language models. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, page 277. ACM.
- Molly E. Ireland, Richard B. Slatcher, Paul W. Eastwick, Lauren E. Scissors, Eli J. Finkel, and James W. Pennebaker. 2011. Language style matching predicts relationship initiation and stability. *Psychological Science*, 22(1):39–44. PMID: 21149854.
- Kimon Kieslich and Marco Lünich. 2024. Regulating ai-based remote biometric identification. investigating the public demand for bans, audits, and public database registrations. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 173–185.
- Hyunwoo Kim, Youngjae Yu, Liwei Jiang, Ximing Lu, Daniel Khashabi, Gunhee Kim, Yejin Choi, and Maarten Sap. 2022. ProsocialDialog: A prosocial backbone for conversational agents. In *Proceedings* of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 4005–4029, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Allison Lahnala, Charles Welch, David Jurgens, and Lucie Flek. 2022. A critical reflection and forward perspective on empathy and natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2139–2158.
- Marina Litvak, Jahna Otterbacher, Chee Siang Ang, and David Atkins. 2016. Social and linguistic behavior and its correlation to trait empathy. Association for Computational Linguistics.

- Yifan Liu, Wei Wei, Jiayi Liu, Xianling Mao, Rui Fang, and Dangyang Chen. 2022. Improving personality consistency in conversation by persona extending. In Proceedings of the 31st ACM International Conference on Information & Knowledge Management, pages 1350–1359.
- 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.
- Sarah Peregrine Lord, Elisa Sheng, Zac E Imel, John Baer, and David C Atkins. 2015. More than reflections: Empathy in motivational interviewing includes language style synchrony between therapist and client. *Behavior therapy*, 46(3):296–303.
- Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.
- 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.
- Joseph H Manson, Gregory A Bryant, Matthew M Gervais, and Michelle A Kline. 2013. Convergence of speech rate in conversation predicts cooperation. *Evolution and Human Behavior*, 34(6):419–426.
- Marilù Miotto, Nicola Rossberg, and Bennett Kleinberg. 2022. Who is GPT-3? an exploration of personality, values and demographics. In Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS), pages 218–227, Abu Dhabi, UAE. Association for Computational Linguistics.
- 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.
- Gregory Park, H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Michal Kosinski, David J Stillwell, Lyle H Ungar, and Martin EP Seligman. 2015. Automatic personality assessment through social media language. *Journal of personality and social psychology*, 108(6):934.
- Heinrich Peters and Sandra C Matz. 2024. Large language models can infer psychological dispositions of social media users. *PNAS nexus*, 3(6).
- Rosalind W Picard. 2000. *Affective computing*. MIT press.

- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.
- Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic opendomain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381.
- Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, et al. 2021. Recipes for building an open-domain chatbot. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 300–325.
- Mustafa Safdari, Greg Serapio-García, Clément Crepy, Stephen Fitz, Peter Romero, Luning Sun, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models. *arXiv preprint arXiv:2307.00184*.
- Aadesh Salecha, Molly E. Ireland, Shashanka Subrahmanya, João Sedoc, Lyle H. Ungar, and Johannes C. Eichstaedt. 2024. Large language models show human-like social desirability biases in survey responses. *Preprint*, arXiv:2405.06058.
- Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR.
- Nino Scherrer, Claudia Shi, Amir Feder, and David Blei. 2024. Evaluating the moral beliefs encoded in llms. *Advances in Neural Information Processing Systems*, 36.
- Vishal Anand Shetty, Shauna Durbin, Meghan S Weyrich, Airín Denise Martínez, Jing Qian, and David L Chin. 2024. A scoping review of empathy recognition in text using natural language processing. *Journal of the American Medical Informatics Association*, 31(3):762–775.
- Nusrat J Shoumy, Li-Minn Ang, Kah Phooi Seng, DM Motiur Rahaman, and Tanveer Zia. 2020. Multimodal big data affective analytics: A comprehensive survey using text, audio, visual and physiological signals. *Journal of Network and Computer Applications*, 149:102447.
- Elizabeth C Stade, johannes C Eichstaedt, Jane P Kim, and Shannon W Stirman. 2024a. Readiness for ai deployment and implementation (readi): A proposed framework for the evaluation of ai-mental health applications.

- Elizabeth C Stade, Shannon Wiltsey Stirman, Lyle H Ungar, Cody L Boland, H Andrew Schwartz, David B Yaden, João Sedoc, Robert J DeRubeis, Robb Willer, and Johannes C Eichstaedt. 2024b. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *NPJ Mental Health Research*, 3(1):12.
- 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.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.
- 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.
- 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.
- Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings* of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 483–498.
- 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*.
- Jiamin Yang and David Jurgens. 2024. Modeling empathetic alignment in conversation. In *Proceedings* of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 3127–3148.
- Tao Yang, Tianyuan Shi, Fanqi Wan, Xiaojun Quan, Qifan Wang, Bingzhe Wu, and Jiaxiang Wu. Psycot: Psychological questionnaire as powerful chain-ofthought for personality detection. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

- 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), pages 609– 626.
- Xianda Zhou and William Yang Wang. 2018. MojiTalk: Generating emotional responses at scale. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1128–1137, Melbourne, Australia. Association for Computational Linguistics.