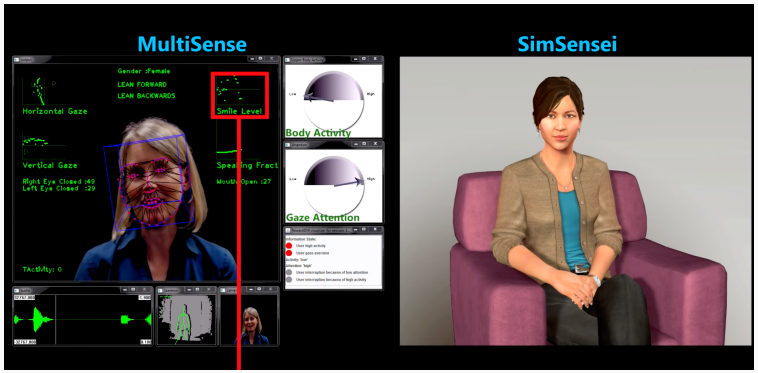
WASSA 2023 Shared Task: Empathy, Emotion and Personality Detection in Conversations and Reactions to News Articles

Valentin Barriere, João Sedoc, Shabnam Tafreshi, Salvatore Giorgi

07/14/23

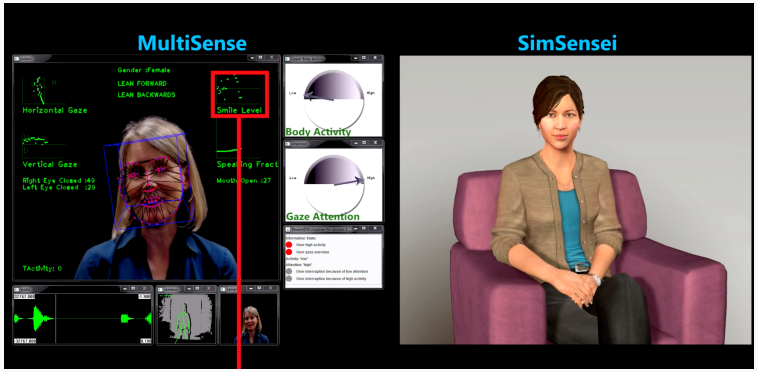
Motivation

- Social skills like **emotions** are essential for human(-agent) communication

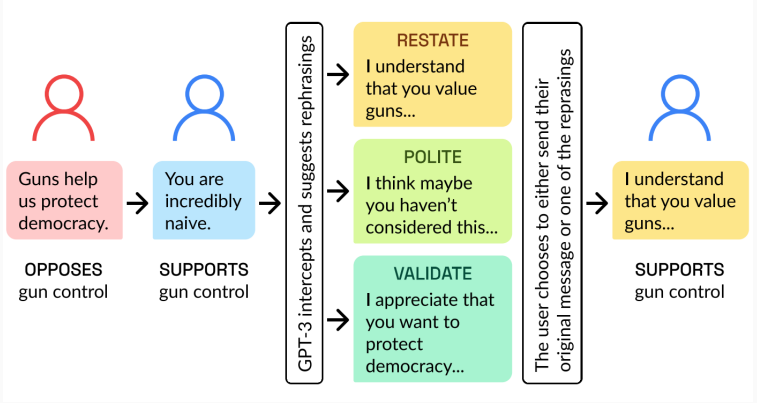


Motivation

- Social skills like **emotions** are essential for human(-agent) communication

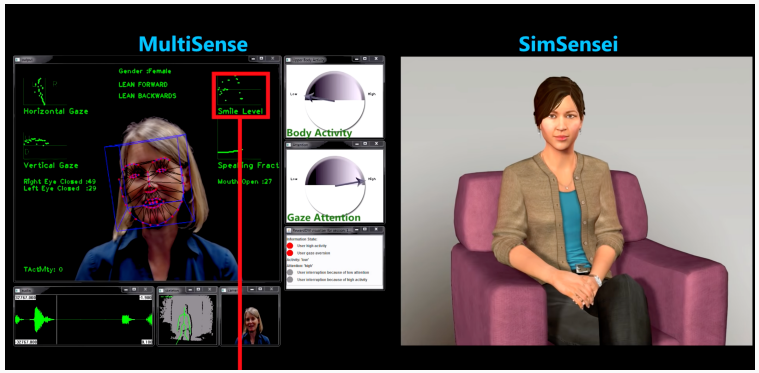


- **Empathy** can help human to better communicate or to find consensus

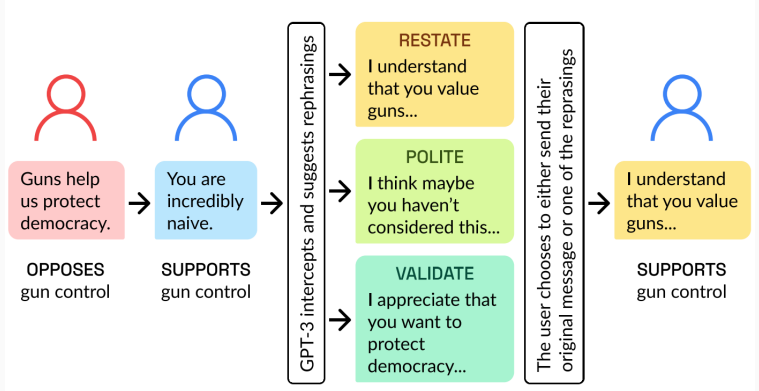


Motivation

- Social skills like **emotions** are essential for human(-agent) communication



- **Empathy** can help human to better communicate or to find consensus



- **Personality** analysis can be used to improve health and well-being

Predicting Empathy, Emotion and Personality in Reaction to News Stories

WASSA 2021 – Shared task on Predicting Empathy, Emotion and Personality in Reaction to News Stories

- Subtask I: **EMP** – Predicting the Empathy and Distress one essay optionally using demographic and personality as features
- Subtask II: **EMO** – Predicting the Emotion of one essay optionally using demographic and personality as features

Predicting Empathy, Emotion and Personality in Reaction to News Stories

WASSA 2022 – Shared task on Predicting Empathy, Emotion and Personality in Reaction to News Stories

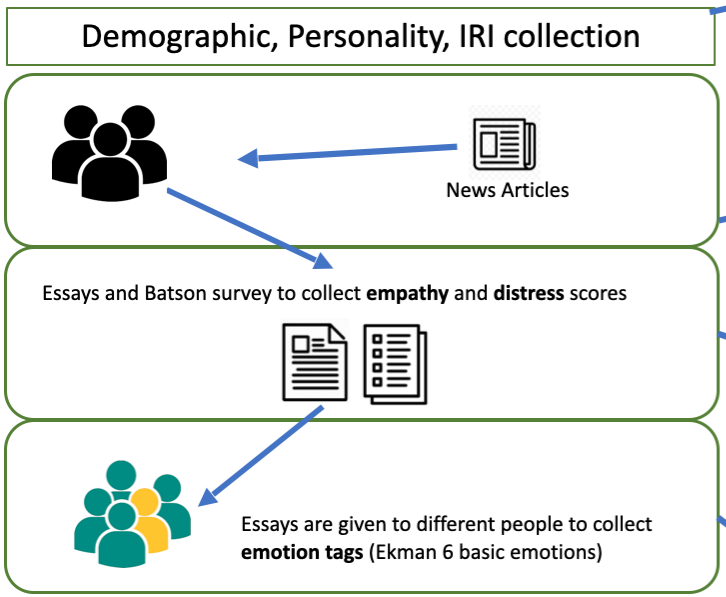
- Subtask I: **EMP** – Predicting the Empathy and Distress one essay optionally using demographic ~~and personality~~ as features
- Subtask II: **EMO** – Predicting the Emotion of one essay optionally using demographic ~~and personality~~ as features
- Subtask III: **PER** – Predicting the Big Five of one writer using its essays and optionally using demographic, and the articles as features
- Subtask IV: **IRI** – Predicting the Interpersonal Reactivity Index of one writer using its essays and optionally using demographic, and the articles as features

Predicting Empathy, Emotion and Personality in Reaction to News Stories

WASSA 2023 – Shared task on Predicting Empathy, Emotion and Personality in Reaction to News Stories

- Subtask I: **EMP** – Predicting the Empathy and Distress one essay optionally using demographic ~~and personality~~ as features
- Subtask II: **EMO** – Predicting the Emotion of one essay optionally using demographic ~~and personality~~ as features
- Subtask III: **PER** – Predicting the Big Five of one writer using its essays and optionally using demographic, and the articles as features
- Subtask IV: **IRI** – Predicting the Interpersonal Reactivity Index of one writer using its essays and optionally using demographic, and the articles as features
- Subtask V: **CONV** – Predicting speech-turn-level Empathy, Emotion polarity, and Emotion intensity in conversations

Annotation and collections of the data and labels



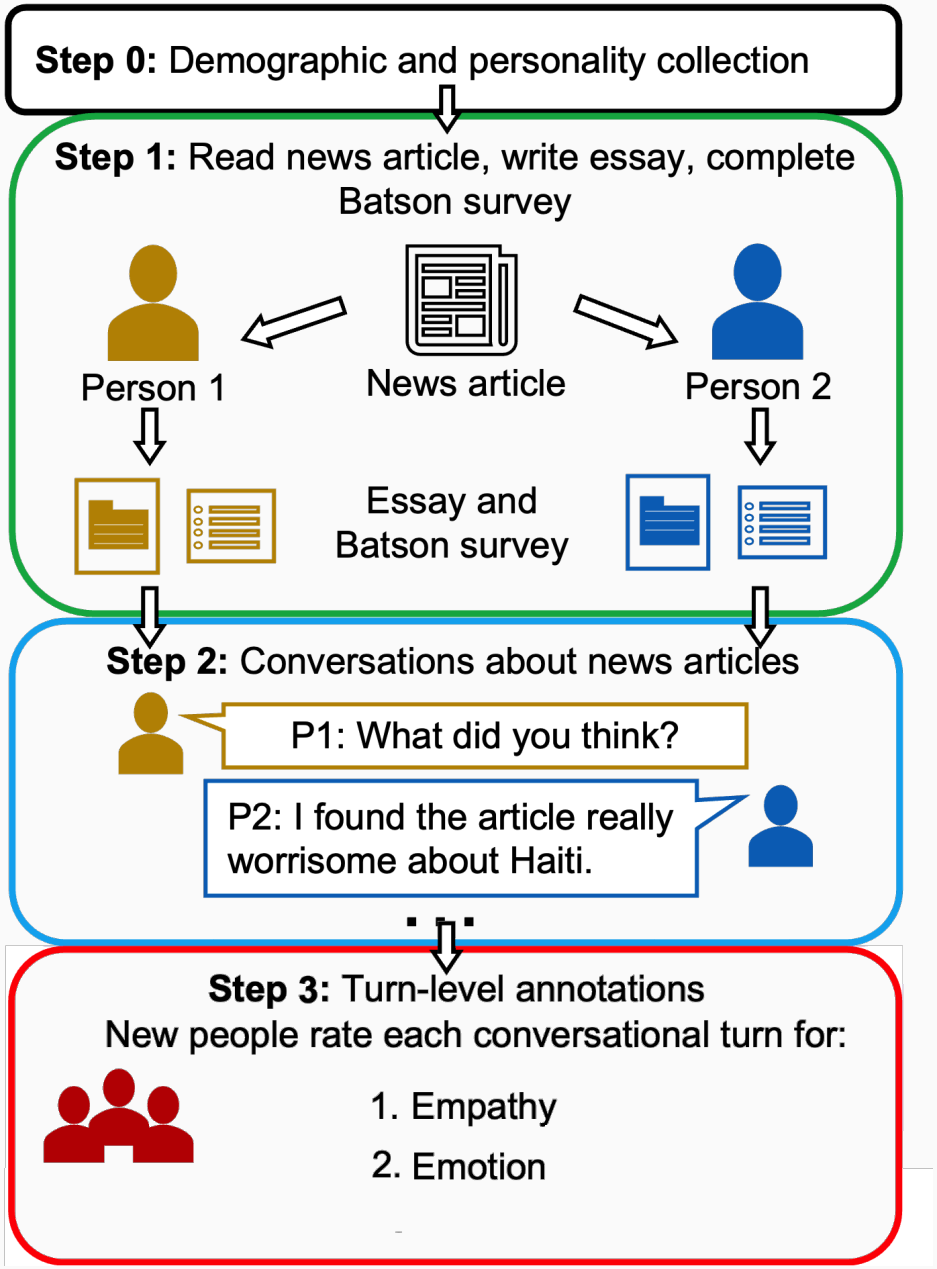
Collecting annotators **demographic, personality** (Big Five model) and **Interpersonal Reactivity Index** (IRI)

Collection of **100 empathic news articles** where there is harm to a person, group, or other, and randomly selected articles are given to each annotator.

Each annotator reads these articles and **write essays** to share his/her **empathy towards** the third person/ group/other from the articles. In addition, the **empathy** and **distress** is collected based on Batson survey [Omitaomu et al., 2022].

Different group of annotators read the essays and used Ekman 6 basic emotion labels + *hope* + *no-emotion* and provided **emotion tags** per essay

Annotation and collections of the data and labels



PER and IRI

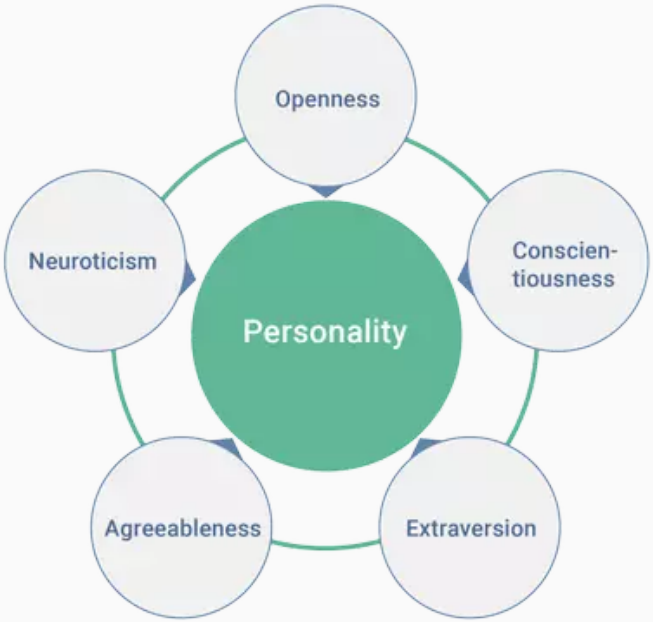


Figure 1: Big Five.

PER and IRI

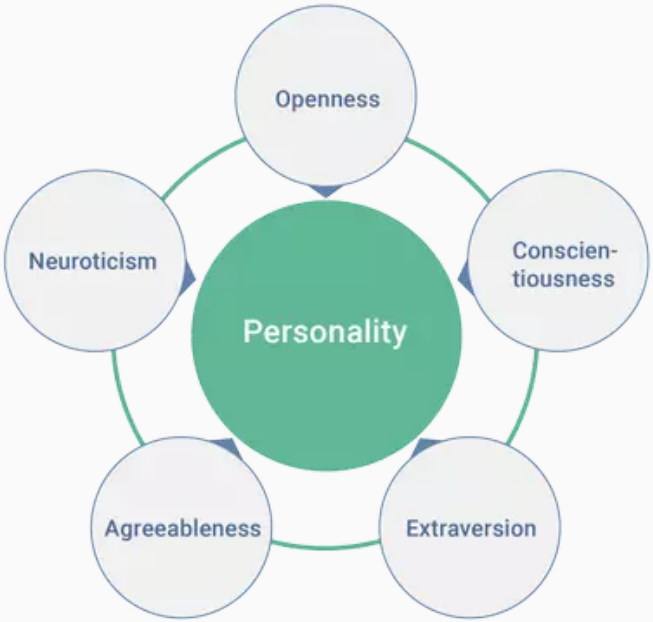


Figure 1: Big Five.

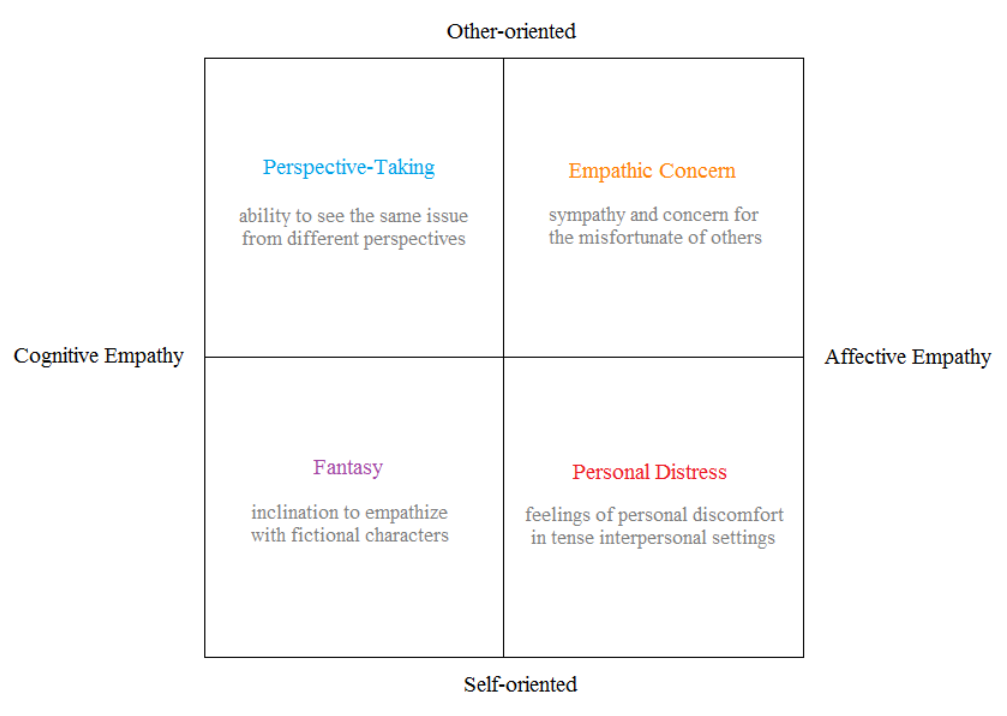


Figure 2: IRI .

Dataset Statistics

	Train	Dev	Test
People	41	34	65
Conversations	396	104	50
Essays	792	208	100
Speech-Turns	9,176	2,000	1,425

Table 1: Corpus statistics.

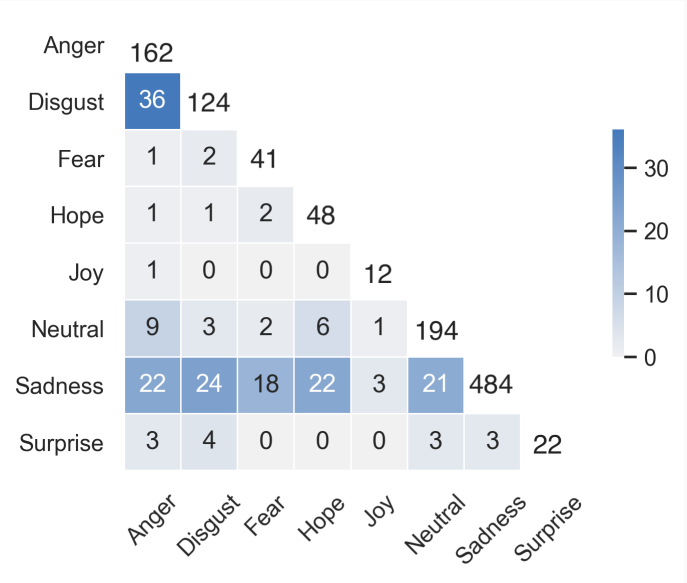


Figure 3: Co-occurrence matrix of the EMO labels on the train and dev sets

For the CONV task, the prediction is at the speech-turn level. For the EMP and EMO tasks, the prediction is at the essay-level while for the PER and IRI task the prediction is at the writer-level.

5 shared tasks on Codalab

Evaluation

- CONV: Pearson Correlation (average mean over empathy-level and emotion polarity and emotion intensity)
- EMP: Pearson Correlation (average mean over empathy and distress)
- EMO: f1-score, precision, recall, and accuracy (macro and overall the 6 emotions + neutral)
- PER: Pearson Correlation (average mean over openness to experience, conscientiousness, extraversion, agreeableness and neuroticism)
- IRI: Pearson Correlation (average mean over perspective taking, fantasy, empathic concern and personal distress)

Codalab competition:
`https://codalab.lisn.upsaclay.fr/competitions/11167`

Overall submissions

- 79 participants registered
  - 7 teams participated to CONV
  - 8 teams participated to EMP
  - 10 teams participated to EMO
  - 5 teams participated to PER
  - 6 teams participated to IRI
- Models: PTLM (BERT, RoBERTa, BigBird, DeBERTA, ELECTRA, AlBERT, DistilBERT, ...), LLM (GPT-3, chatGPT, GPT4), LSTM, Interaction-aware model
- Ressources: GoEmotions, EPITOME, Sentiment and Emotion datasets of tweets or conversations
- Others:
  - Data-Augmentation: Paraphrasing or Generating underrepresented classes, Back-translation
  - Demographic meta-data integrated as embeddings or using prompt-based methods as sentence or as a table
  - Ensemble methods (models or adapters)
  - In-Context-Learning, fine-tuning, adapting

Empathy and Emotion in Conversations results – Evaluation Phase

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).

Empathy and Distress results – Evaluation Phase

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).

The baseline was trained also using the 2022 data.

Emotions results – Evaluation Phase

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)).

Personality and IRI results

Team	Consc.	Open.	Extr.	Agree.	Stab.	**PER**
YNU-HPCC	0.289	**0.372**	-0.130	**0.410**	**0.317**	**0.252**
CAISA	**0.323**	0.327	-0.197	0.290	0.256	0.200
Curtin OCAI	0.186	0.152	0.014	-0.038	0.183	0.099
Cordyceps	-0.059	-0.187	**0.160**	0.101	-0.010	0.001
Hawk	-0.082	0.066	-0.109	-0.119	-0.114	-0.072
Baselines	-0.131	-0.037	-0.134	0.195	0.081	-0.005

Table 5: Results of the **PER** tracks (Pearson correlations).

Team	Persp.	Distr.	Fant.	Emp.	**IRI**
YNU-HPCC	0.102	0.256	0.033	**0.226**	**0.154**
Xuhao	0.132	**0.366**	0.036	0.076	0.153
CAISA	**0.158**	-0.188	-0.056	0.180	0.024
Curtin OCAI	-0.092	0.193	-0.014	-0.114	-0.007
Cordyceps	0.004	0.191	-0.018	0.089	0.067
Hawk	-0.013	-0.020	**0.138**	-0.153	-0.012
Baselines	0.107	-0.046	0.063	**0.340**	0.116

Table 6: Results of the **IRI** tracks (Pearson correlations).

Best approaches

CONV: HIT-SCIR

- Ensemble of DeBERTA(-*xl* and -*xxl*) models
- Context window to integrate the interactions

EMP: NCUEE-NLP

- Ensemble of RoBERTA(-*base*) models
- PT on Sentiment in tweets and Emotion in Conversations

EMO: Adityapatkar

- Ensemble of BERT and RoBERTA(-*base*) models
- Learned a probability threshold

PER/IRI: YNU-HPCC

- DeBERTA(-*xxl*) with an adapter
- Data-augmentation using back-translation

Summary

- Demographic were used this year to split the dataset in order to test models robustness
- **Data-augmentation** still common
- **Ensemble** methods are always helping a lot
- GPT-4 was beaten
- But bigger fine-tunable models lead
- Nobody tried to model the all 5 tasks at the same time !

Acknowledgements

We want to thank all the participants of this shared-task!

Questions?

References i