
# WASSA 2022 Shared Task: Predicting Empathy, Emotion and Personality in Reaction to News Stories

Valentin Barriere, Shabnam Tafreshi, João Sedoc, Sawsan Alqahtani

05/26/22

# Motivation

- Nely: a social robot with **empathy**

# Motivation

- Nely: a social robot with **empathy**

- **Emotion** in disaster response

# Motivation

- Nely: a social robot with **empathy**

- **Emotion** in disaster response

- **Personality** detection and adaptation for a virtual agent

# Previous Task

**WASSA 2021 – Shared task on Predicting Empathy and Emotion 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

4 teams participated to EMP and 4 teams participated to EMO

# 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

# 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

# PER and IRI

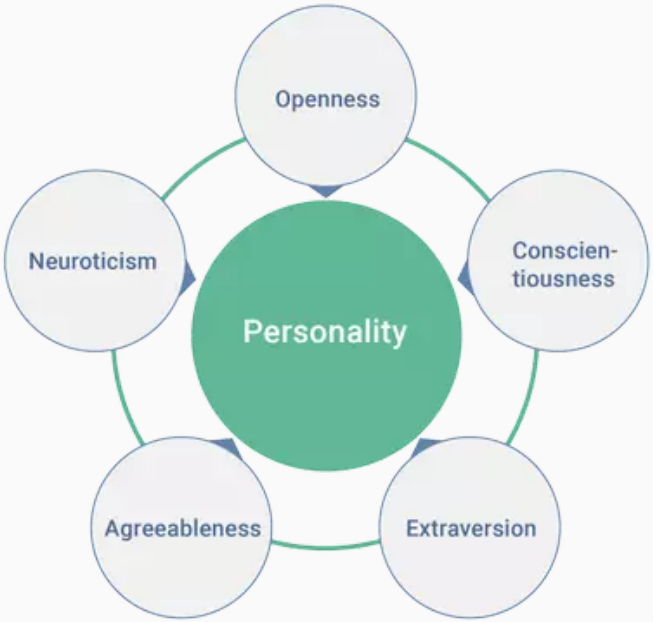



**Figure 1:** Big Five.

# PER and IRI

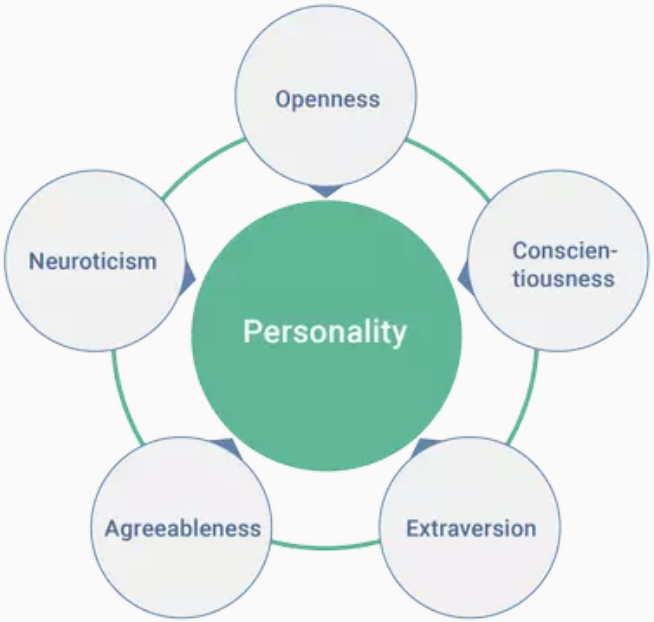



**Figure 1:** Big Five.

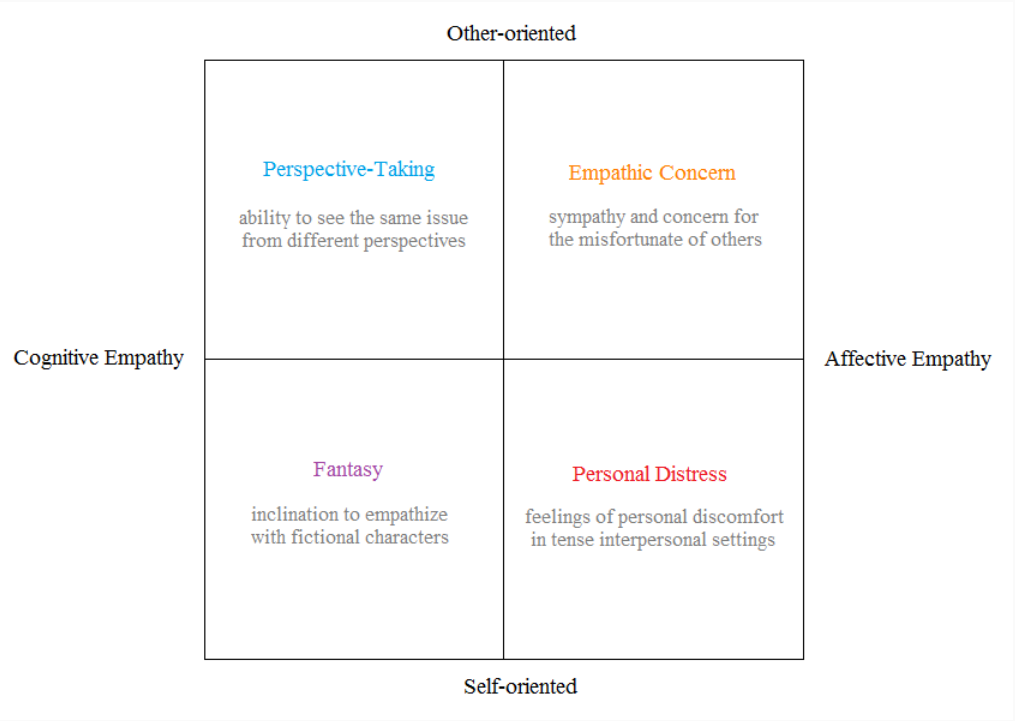
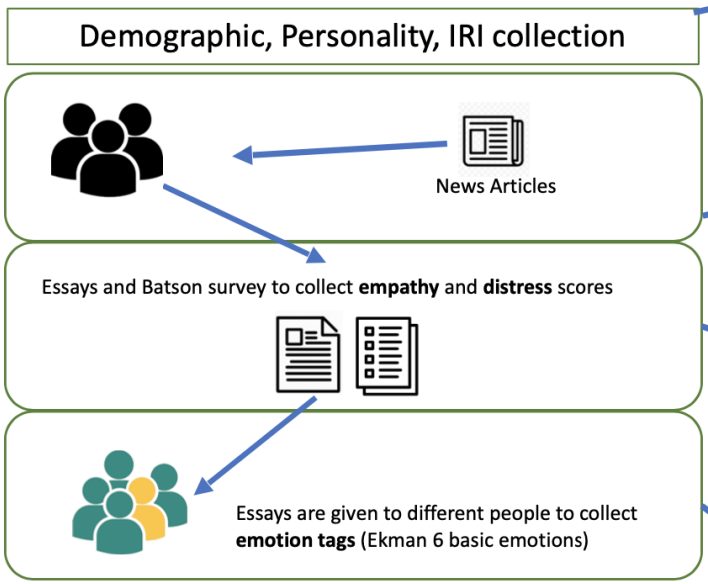



**Figure 2:** IRI .

# 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 5 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 [Buechel et al., EMNLP 2018].

Different group of annotators read the essays and used Ekman 6 basic emotion labels + *no-emotion* and provided **emotion tags** per essay [Tafreshi et al., WASSA 2021].

# Dataset Statistics

| | joy | sadness | disgust | fear | anger | surprise | no-emo |
|---|---|---|---|---|---|---|---|
| **Train** | 82 | 647 | 149 | 194 | 349 | 164 | 275 |
| **Dev** | 14 | 98 | 12 | 31 | 76 | 14 | 25 |
| **Test** | 33 | 177 | 28 | 70 | 122 | 40 | 55 |
| **Total** | 129 | **922** | 189 | 295 | **547** | 218 | 355 |

**Table 1:** Distribution of emotion labels in the datasets.

**Dataset Split**

| Type | Train | Dev | Test | Total |
|---|---|---|---|---|
| Essay | 1860 | 270 | 525 | 2655 |
| Writer | 372 | 54 | 105 | 531 |

**Table 2:** Train, dev and test set splits.

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.

# 4 shared tasks on Codalab

**Evaluation**

- 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/834`

# Overall submissions

- X registered
  - 10 teams participated to EMP
  - 13 teams participated to EMO
  - 2 teams participated to PER
  - 3 teams participated to IRI
- Models:
  - PTLM: BERT, RoBERTa, LongFormer, DeBERTA, GPT-3,...
  - SVM
- Features and Ressources:
  - Dataset: GoEmotions, EPITOME, CARER, XED
  - Embeddings: Emotion-enriched word embeddings
  - Lexicons: NRC
- Others:
  - Data-Augmentation: Random, Balanced, Punctuation Substitution, Back-translation..
  - Prompt-based method to integrate meta-data
  - Ensemble Methods: Boosting, Bagging
  - Zero-shot models, fine-tuning, adapting
  - Binary classification, Multi-label classification, Single or Multi-task

# Empathy and Distress results – Evaluation Phase

| Team | Emp | Dis | Avg |
|---|---|---|---|
| IUCL-1 | 0.537 | **0.543** | **0.540** |
| SINAI | **0.541** | 0.519 | 0.530 |
| IUCL-2 | 0.512 | **0.543** | 0.527 |
| SURREY-CTS-NLP | 0.501 | 0.498 | 0.499 |
| LingJing | 0.508 | 0.489 | 0.499 |
| PHG | 0.470 | 0.506 | 0.488 |
| IITP-AINLPML | 0.479 | 0.488 | 0.483 |
| mantis | -0.028 | -0.064 | -0.048 |
| PVG (WASSA2021) | 0.517 | 0.574 | 0.545 |

**Table 3:** Results of the teams participating in the EMP track.

Best score is from last year: note that participants were allowed to use the personality and IRI information as features.

# Empathy and Distress results – Post-Evaluation Phase

| Team | Emp | Dis | Avg | Rank Eval |
|---|---|---|---|---|
| IUCL-1 | 0.537 | 0.543 | **0.540** | 1 |
| SINAI | **0.541** | 0.519 | 0.530 | 2 |
| IUCL-2 | 0.512 | **0.547** | 0.529 | 3 |
| CAISA | 0.524 | 0.521 | 0.523 | ∅ |
| SURREY-CTS-NLP | 0.504 | 0.530 | 0.517 | 4 |
| LingJing | 0.508 | 0.489 | 0.499 | 5 |
| PHG | 0.470 | 0.506 | 0.488 | 6 |
| IITP-AINLPML | 0.479 | 0.488 | 0.483 | 7 |
| mantis | 0.484 | 0.453 | 0.468 | 8 |
| phuonglh | 0.196 | 0.183 | 0.190 | 9 |
| PVG (WASSA2021) | 0.517 | 0.574 | **0.545** | 0 |

**Table 4:** Results of the teams participating in the EMP track.

Best score is from last year: note that participants were allowed to use the personality and IRI information as features.

# Emotions results – Evaluation Phase

| Team | P | R | F1 | Acc |
|---|---|---|---|---|
| LingJing | **0.740** | **0.679** | **0.698** | **0.754** |
| himanshu.1007 | 0.594 | 0.584 | 0.585 | 0.661 |
| IUCL-2 | 0.572 | 0.552 | 0.557 | 0.638 |
| SURREY-CTS-NLP | 0.101 | 0.100 | 0.101 | 0.101 |
| SINAI | 0.589 | 0.535 | 0.553 | 0.636 |
| mantis | 0.142 | 0.142 | 0.142 | 0.202 |
| blueyellow | 0.571 | 0.531 | 0.544 | 0.623 |
| IUCL-1 | 0.564 | 0.539 | 0.544 | 0.611 |
| shantpat | 0.537 | 0.527 | 0.527 | 0.606 |
| PHG | 0.557 | 0.529 | 0.531 | 0.611 |
| IITP-AINLPML | 0.527 | 0.585 | 0.524 | 0.585 |
| PVG AI Club | 0.497 | 0.464 | 0.473 | 0.571 |
| IITK (WASSA2021) | 0.57 | 0.55 | 0.55 | 0.62 |

**Table 5:** Results of the teams participating in the EMO track.

# Emotions results – Post-Evaluation Phase

| Team | P | R | F1 | Acc | Rank Eval |
|---|---|---|---|---|---|
| LingJing | **0.740** | **0.679** | **0.698** | **0.754** | 1 |
| CAISA | 0.625 | 0.592 | 0.604 | 0.669 | ∅ |
| himanshu.1007 | 0.594 | 0.584 | 0.585 | 0.661 | 2 |
| IUCL-2 | 0.599 | 0.555 | 0.572 | 0.646 | 3 |
| SURREY-CTS-NLP | 0.595 | 0.559 | 0.571 | 0.646 | 12 |
| SINAI | 0.589 | 0.535 | 0.553 | 0.636 | 4 |
| mantis | 0.594 | 0.528 | 0.548 | 0.632 | 11 |
| blueyellow | 0.571 | 0.531 | 0.544 | 0.623 | 5 |
| IUCL-1 | 0.564 | 0.539 | 0.544 | 0.611 | 6 |
| shantpat | 0.552 | 0.532 | 0.534 | 0.623 | 8 |
| PHG | 0.557 | 0.529 | 0.531 | 0.611 | 7 |
| IITP-AINLPML | 0.527 | 0.585 | 0.524 | 0.585 | 9 |
| PVG AI Club | 0.473 | 0.467 | 0.464 | 0.560 | 10 |
| IITK (WASSA2021) | 0.57 | 0.55 | 0.55 | 0.62 | 4.5 |

**Table 6:** Results of the teams participating in the EMO track.

# Personality and IRI results

| **Team** | Consc. | Open. | Extr. | Agree. | Stab. | **PER** |
|---|---|---|---|---|---|---|
| LingJing | .165 | .337 | .098 | .246 | .305 | **.230** |
| IITP | .134 | .092 | .102 | -.176 | .086 | .047 |
| SINAI | .145 | -.215 | .087 | .030 | -.047 | .000 |
| Aggreg (Org.) | .207 | .506 | .123 | .310 | .383 | **.306** |
| $WD(Train\|\|Test)$ | .12 | .20 | .14 | .29 | .17 | |

**Table 7:** Results of the teams participating in the PER tracks.

| **Team** | Persp. | Distr. | Fant. | Emp. | **IRI** |
|---|---|---|---|---|---|
| LingJing | .139 | .245 | .377 | .257 | **.255** |
| IITP | .039 | .004 | .011 | .252 | .076 |
| Aggreg (Org.) | .166 | .29 | .495 | .374 | **.331** |

**Table 8:** Results of the teams participating in the IRI tracks.

Best system do not tag at the writer level but at the essay level. Better results obtained when aggregating the predictions at the writer level.

# Personality and IRI results

| **Team** | Consc. | Open. | Extr. | Agree. | Stab. | **PER** |
|---|---|---|---|---|---|---|
| LingJing | .165 | .337 | .098 | .246 | .305 | **.230** |
| IITP | .134 | .092 | .102 | -.176 | .086 | .047 |
| SINAI | .145 | -.215 | .087 | .030 | -.047 | .000 |
| Aggreg (Org.) | .207 | .506 | .123 | .310 | .383 | **.306** |
| $WD(Train\|\|Test)$ | .12 | .20 | .14 | .29 | .17 | |

**Table 7:** Results of the teams participating in the PER tracks.

| **Team** | Persp. | Distr. | Fant. | Emp. | **IRI** |
|---|---|---|---|---|---|
| LingJing | .139 | .245 | .377 | .257 | **.255** |
| IITP | .039 | .004 | .011 | .252 | .076 |
| Aggreg (Org.) | .166 | .29 | .495 | .374 | **.331** |

**Table 8:** Results of the teams participating in the IRI tracks.

Best system do not tag at the writer level but at the essay level. Better results obtained when aggregating the predictions at the writer level.

# Best approaches

## EMP: IUCL-1

- RoBERTA(-*large*) models
- single-task models for empathy and distress
- Not using the demographic metadata (results were worst)

## EMO: LingJing

- DeBERTa(-*v2-xxl*) models pre-trained over GoEmotions and XED
- Child-tuning method
- Bagging algorithm

## PER/IRI: LingJing

- DeBERTa(-*v3-large*) model
- Prompt-based method to integrate the demographic metadata
- Data-augmentation using punctuation insertion
- Ensemble model

# Summary

- **Prompt-based method** to integrate the demographic metadata inside transformers
- **Data-augmentation** helped for the EMO/PER/IRI tasks
- **Ensemble** methods are always helping a lot
- **Lexicons** are almost forgotten
- Sometimes **simple** is better (EMP best result)
- Sometimes **not** (EMO/PER/IRI, bigger models with complex methods)
- Nobody used the **news articles** to add interactional context !

Special mention: SINAI team using zero-shot models trained over MNLI

# Acknowledgements

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

# Questions?

# References i