Annotating Dimensions of Social Perception in Text: A Sentence-Level Dataset of Warmth and Competence

Mutaz Ayesh, Saif M. Mohammad, Nedjma Ousidhoum
Contact: OusidhoumN@Cardiff.ac.uk

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

National Research Council Canada
Conseil national de recherches Canada

Warmth and Competence (W and C)

Primary dimensions along which people form impressions on individuals and social groups

- Warmth (W) Sociability (S)+Trust (T): morality, goodness, sincerity, integrity
- Competence (C) ability, power, dominance, assertiveness

Why a sentence-based dataset for W & C?

- Existing NLP work on W & C relies primarily on word-level lexicons
- While useful, lexical approaches do not fully capture contextual meaning
- A need for sentence-level modeling to better capture social perception in context

We introduce W&C-Sent, a sentence-level dataset for measuring warmth (trust, sociability) and competence toward individuals and social groups

Individual Target: Hilary Clinton
"Would you wanna be in a long-term relationship with someone who hides her emails and lies to your face? Then #DontVote"

- Trust: –3 (High Distrust)
- Sociability: –2 (Moderate Unsociability)
- Competence: +2 (Moderate Competence)

Target Group: Religious People
"Could all those who believe in a god please leave. The meeting will now continue for the grown ups only."

- Trust: 0 (Neutral Trust)
- Sociability: –3 (High Unsociability)
- Competence: -3 High Incompetence)

W&C-Sent Construction

- 1,633 English sentence–target pairs annotated for social perception
- Instances from: SemEval 2016 stance (90%) and ABCDE dataset (10%)
- Covers 7 targets (3 individuals, 4 groups)
- Each instance annotated by 4–7 annotators
- 7-point ordinal scale: –3 (very low), 0 (neutral), +3 (very high)
- Includes manually derived fine-grained association scores for all pairs

Dimension	Low (%)	Neutral (%)	High (%)
Trust	57.9	6.4	35.7
Sociability	62.0	5.2	32.8
Competence	47.3	9.9	42.7

Target	Count	%
Hillary Clinton	617	37.7
Donald Trump	409	25.0
Women	313	19.0
Barack Obama	113	7.0
Religious People	105	6.5
Climate Activists	40	2.5
Atheists	36	2.3
Total	1,633	100

Quality Control

- Includes several attention checks, and reliability assessment procedures
- SHR = 0.76 for trust, 0.68 for sociability, and 0.56 for competence

Experimental Setup

- Dataset split: 60/20/20 train/dev/test, with the test set used for evaluation
- **Task:**
  Predict trust, sociability, and competence scores (-3 to +3) from sentence–target pairs

Models and Baselines

- **Setup:** One classification model per dimension
- **LLMs:**
  a. Prompted with definitions and sentence-level evaluation instructions
  b. Tested in zero-shot (ZS) and few-shot (FS) settings

Model	F1	Acc	±1 Acc
Majority Baseline	0.08	0.22	0.58
LR	0.30	0.30	0.65
Fine-tuned BERTweet	0.31	0.35	0.83
Gemma3 ZS	0.34	0.36	0.78
Gemma3 FS	0.33	0.38	0.79
Qwen3 ZS	0.30	0.32	0.78
Qwen3 FS	0.26	0.30	0.78
GPT-4o ZS	0.42	0.43	0.91
GPT-4o FS	0.40	0.39	0.84
GPT-4o-mini ZS	0.26	0.27	0.60
GPT-4o-mini FS	0.17	0.20	0.54
GPT-5.2 ZS	0.38	0.41	0.87
GPT-5.2 FS	0.39	0.40	0.89

1. Fine-grained Trust Classification

Model	F1	Acc	±1 Acc
Majority Baseline	0.11	0.26	0.67
LR	0.26	0.31	0.70
Fine-tuned BERTweet	0.34	0.46	0.88
Gemma3 ZS	0.27	0.34	0.82
Gemma3 FS	0.23	0.31	0.76
Qwen3 ZS	0.30	0.35	0.81
Qwen3 FS	0.24	0.31	0.82
GPT-4o ZS	0.40	0.44	0.92
GPT-4o FS	0.35	0.38	0.87
GPT-4o-mini ZS	0.14	0.13	0.52
GPT-4o-mini FS	0.20	0.24	0.59
GPT-5.2 ZS	0.30	0.31	0.83
GPT-5.2 FS	0.37	0.40	0.88

2. Fine-grained Sociability Classification

Model	F1	Acc	±1 Acc
Majority Baseline	0.09	0.24	0.60
LR	0.16	0.19	0.62
Fine-tuned BERTweet	0.24	0.36	0.86
Gemma3 ZS	0.19	0.22	0.58
Gemma3 FS	0.27	0.35	0.70
Qwen3 ZS	0.18	0.22	0.60
Qwen3 FS	0.21	0.25	0.67
GPT-4o ZS	0.31	0.34	0.80
GPT-4o FS	0.24	0.24	0.64
GPT-4o-mini ZS	0.11	0.09	0.37
GPT-4o-mini FS	0.20	0.22	0.60
GPT-5.2 ZS	0.25	0.28	0.73
GPT-5.2 FS	0.24	0.27	0.71

3. Fine-grained Competence Classification

Findings

- Competence is the most difficult dimension to predict
- LLMs struggle to reliably capture subtle social perception cues
- Zero-shot prompting often outperforms few-shot prompting
- Performance improves substantially when labels are coarsened
- High-quality task-specific training can outperform LLMs

Ethical Considerations

- Associations and stereotypes; not inherent properties
- Consider coverage, domain, ambiguity,
- socio-cultural effects, etc.
- Ethics Sheet for Emotion Recognition (Mohammad, 2022, CL)
- Best Practices in the Use of Emotion Lexicons (Mohammad, 2020)

Check out the paper and data here:https://github.com/nedjmaou/W_C_Sent