ACL 2026 SAN DIEGO JULY 2-7

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

Mutaz Ayesh, Saif M. Mohammad, **Nedjma Ousidhoum**

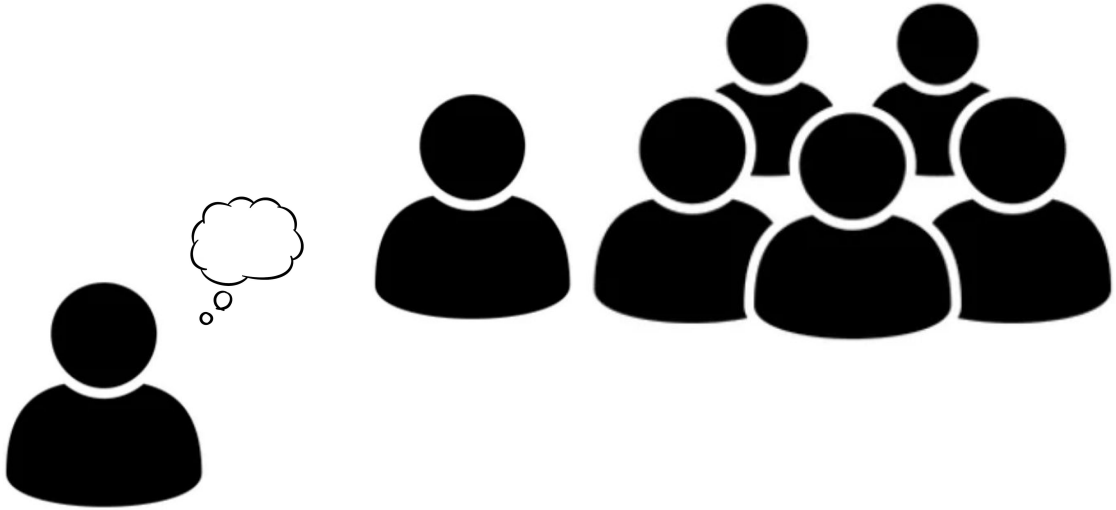
https://github.com/nedjmaou/W_C_Sent

Interpersonal Evaluation As a Concept

Social psychologists report that people primarily form impressions of others based Warmth and Competence

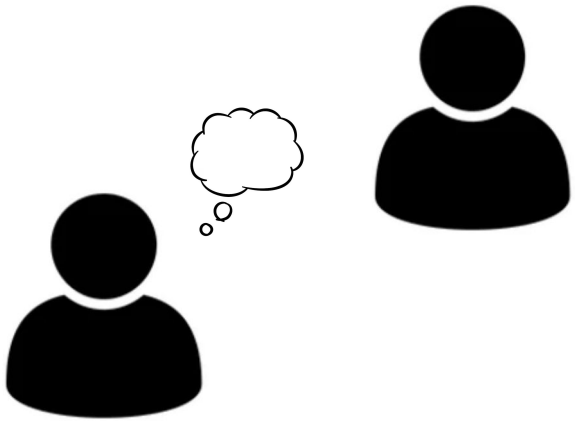
Warmth and Competence

Warmth and competence guide how individuals judge other people



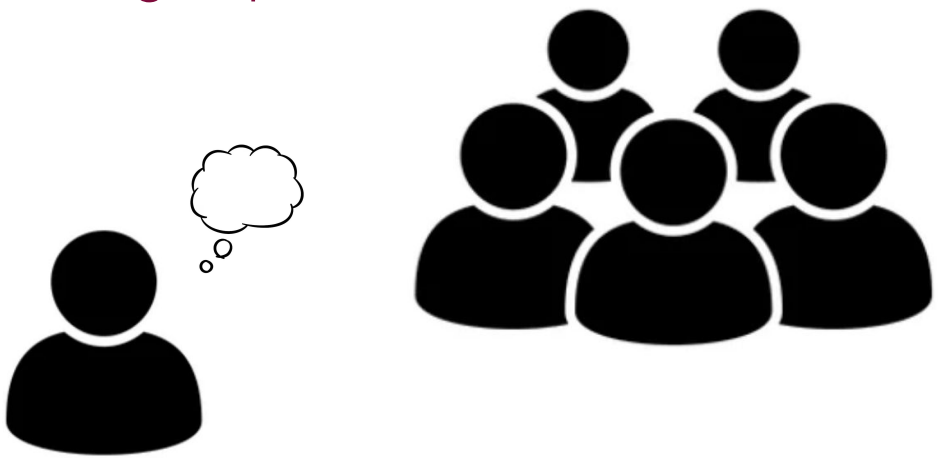
Warmth and Competence

Warmth and competence guide how individuals judge other people, i.e., individuals



Warmth and Competence

Warmth and competence guide how individuals judge other people, i.e., individuals and social groups



What do warmth and competence refer to?

Warmth

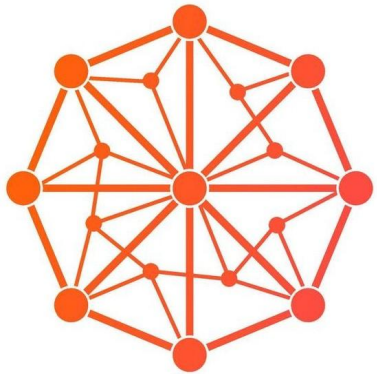
- Warmth refers to perceived friendliness and positive intent
- It is often divided into two dimensions

Warmth

- Warmth refers to perceived friendliness and positive intent
- It is often divided into two dimensions:
  - Trust (T): morality, sincerity, integrity, and benevolence

Warmth

- Warmth refers to perceived friendliness and positive intent
- It is often divided into two dimensions:
  - Trust (T): morality, sincerity, integrity, and benevolence
  - Sociability (S): friendliness, sociableness, gregariousness, and conviviality



Competence

- Competence (C) refers to perceived ability to effectively carry out intentions and achieve desired outcomes
  - Ability, skill, intelligence, power, dominance, assertiveness, influence

Why Warmth and Competence Matter

Warmth (Trust & Sociability) & Competence help us understand:

- social interactions

Why Warmth and Competence Matter

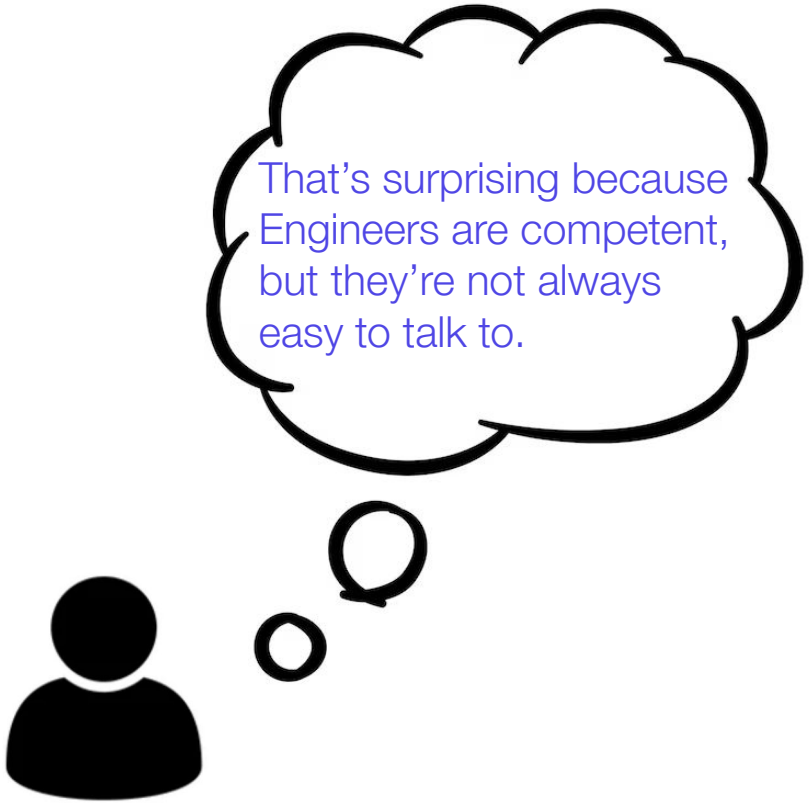
Warmth (Trust & Sociability) & Competence help us understand:

- social interactions
- emotional reactions

Why Warmth and Competence Matter

Warmth (Trust & Sociability) & Competence help us understand:

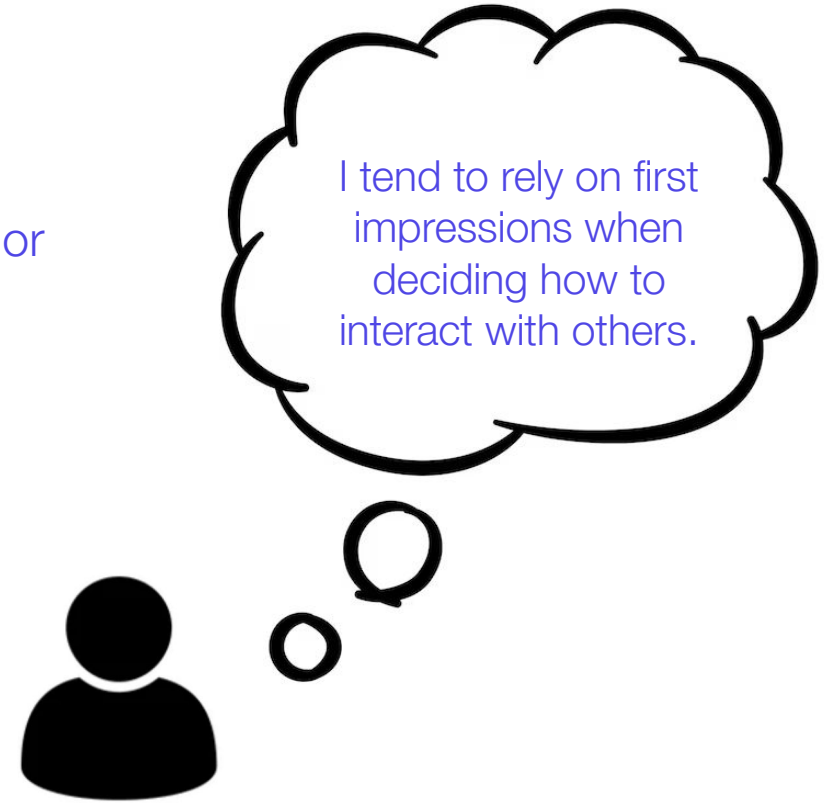
- social interactions
- emotional reactions
- stereotypes



Why Warmth and Competence Matter

Warmth (Trust & Sociability) & Competence help us understand:

- social interactions
- emotional reactions
- stereotypes
- broader human behavior



Warmth and Competence in NLP

- Warmth and competence remain relatively underexplored in NLP
- Most prior work relies on lexical-based approaches
  - They are simple and effective but limited for fine-grained ML tasks
  - They do not to capture contextual meaning

We introduce W&C-Sent, a sentence-level dataset designed to measure perceived warmth (trust and sociability) and competence toward targeted individuals and social groups

W&C-Sent Dataset – Individual Target Instance

Individual Target: Hillary 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) 

W&C-Sent Dataset – Group Target Instance

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 Dataset Construction – Data Selection

We used instances drawn from the SemEval-2016 Stance Dataset (90.8%) and the ABCDE Dataset (9.2%)

- @mention replaced by @user except for public figures, e.g., Hillary Clinton

W&C-Sent Dataset Construction – Data Selection

We used instances drawn from the SemEval-2016 Stance Dataset (90.8%) and the ABCDE Dataset (9.2%)

Instance from the SemEval-2016 Stance Dataset:

We need Obama out and @realDonaldTrump in the White House ASAP.

We annotate perceived social dimensions with Donald Trump as a target

W&C-Sent Dataset Construction – Data Selection

We re-purposed instances drawn from the SemEval-2016 Stance Dataset (90.8%) and the ABCDE Dataset (9.2%)

Instance from the SemEval-2016 Stance Dataset:

We need Obama out and @realDonaldTrump in the White House ASAP.

SemEval-2016 Labels

- Original Target Donald Trump
- Original Label

  Stance: In favor

W&C-Sent Dataset Construction – Data Selection

We re-purposed instances drawn from the SemEval-2016 Stance Dataset (90.8%) and the ABCDE Dataset (9.2%)

Instance from the SemEval-2016 Stance Dataset:

We need Obama out and @realDonaldTrump in the White House ASAP.

SemEval-2016 Labels

- Original Target Donald Trump
- Original Label

  Stance: In favor

W&C-Sent Labels

- Extracted Target Barack Obama
- Social Dimensions
  - Trust: –2 (Moderate Distrust)
  - Sociability: –1 (Slight Unsociability)
  - Competence: -2 (Moderate Incompetence)

W&C-Sent Dataset Construction – Data Selection

We re-purposed instances drawn from the SemEval-2016 Stance Dataset (90.8%) and the ABCDE Dataset (9.2%)

Instance from the SemEval-2016 Stance Dataset:

We need Obama out and @realDonaldTrump in the White House ASAP.

SemEval-2016 Labels

- Original Target Donald Trump
- Original Label

  Stance: In favor

W&C-Sent Labels

- Extracted Target Barack Obama
- Social Dimensions
  - Trust: –2 (Moderate Distrust)
  - Sociability: –1 (Slight Unsociability)
  - Competence: -2 (Moderate Incompetence)

We do the same for posts about climate change, feminism, abortion, atheism

W&C-Sent Dataset Construction – Annotation

- To reduce cognitive load, trust, sociability, and competence were annotated separately
- Guidelines were adapted from Words of Warmth (Mohammad, 2025)
- Each post-target pair was rated on a 7-point ordinal scale (from –3 to +3):
  - –3 = very low trust/sociability/competence
  - 0 = neutral, not applicable, or no information
  - +3 = very high trust/sociability/competence

W&C-Sent Dataset Construction – Annotation

- To reduce cognitive load, trust, sociability, and competence were annotated separately
- Guidelines were adapted from Words of Warmth (Mohammad, 2025)
- Each post-target pair was rated on a 7-point ordinal scale (-3 to +3)

We need Obama out and @realDonaldTrump in the White House ASAP.

- Trust: –2 (Moderate Distrust) 
- Sociability: –1 (Slight Unsociability) 
- Competence: -2 (Moderate Incompetence) 

W&C-Sent Dataset Construction – Quality Control

- The data instances were labeled using Prolific
- Annotator selection criteria:
  - Must be fluent English speakers
  - Must be based in English-speaking countries
  - Must have an approval rate above 99% on Prolific
  - Paid $16–22 per hour
- To ensure data quality and reduce bot or inattentive responses, annotators had to complete attention checks
    i. Selecting a designated response following specific instructions
    ii. Selecting the right gold-standard annotations

W&C-Sent Dataset Construction – Annotation Quality

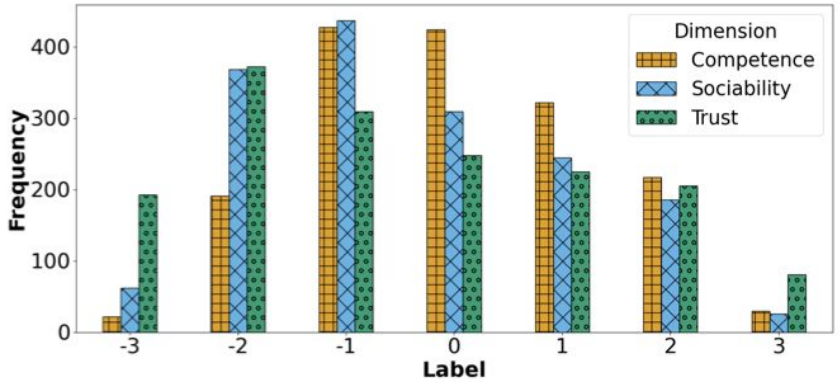
- Split-half reliability (SHR) was used to assess consistency:
  - Trust: 0.76
  - Sociability: 0.68
  - Competence: 0.56
- Krippendorff's α (IAA):
  - Trust: 0.60
  - Sociability: 0.50
  - Competence: 0.30
- When collapsing scores into coarse categories (low/neutral/high), agreement improves:
  - 0.63 (for both trust and sociability)
  - 0.52 (competence)

W&C-Sent Dataset

7 targets in total (3 individuals and 4 groups)

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

W&C-Sent Dataset Distribution



- Neutral is the most frequent overall, particularly for competence
- Negative scores are more prevalent than positive ones likely due to data source which includes controversial topics (e.g., elections and reproductive rights)
- Positive labels are generally less frequent, with +3 being the rarest category
- +3 and -3 are the least common labels in general

Detecting Social Dimensions in Text with W&C-Sent

- We predict trust, sociability, and competence scores (-3 to +3)
- Setup
  - Split 60/20/20 train/dev/test, with the test set used for evaluation
  - One classification model per dimension
  - LLMs (closed and open models)
    - prompted with definitions and sentence-level evaluation instructions
    - Tested in zero-shot (ZS) and few-shot (FS) settings

Experimental Results – Trust Classification

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
Qwen2.5 ZS	0.25	0.27	0.72
Qwen2.5 FS	0.35	0.35	0.76
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

Experimental Results – Sociability 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
Qwen2.5 ZS	0.20	0.20	0.69
Qwen2.5 FS	0.25	0.25	0.67
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

Experimental Results – Competence 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

Findings

- Competence is the most difficult dimension to predict
- LLMs struggle to reliably capture subtle social perception cues
- Performance improves substantially when labels are coarsened
- Fine-tuning a PTLM can outperform LLMs
- Zero-shot prompting often outperforms few-shot prompting, i.e., potential calibration bias
    - In the 102 cases, we notice that
        - FS prompting amplifies surface cues such as ALL-CAPS and exclamation marks
        - There was a strong bias toward neutral predictions (i.e., by default)
        - Misinterpretation of sarcasm and figurative language as literal meaning leading to neutral or overly penalised labels for posts with sarcastic content

Analysis – High Subjectivity in Social Perception Annotation

- Exact fine-grained label agreement is very low (3%)
- Polarity-level agreement is higher (26.9%)
  - mostly on negative instances
- Agreement patterns differ by dimension
  - Higher for trust toward individuals and sociability toward groups.
  - This likely reflects differences in how dimensions are interpreted
    - individual: personal/moral vs. group: relational/social)
- Meaningful cross-dimension disagreement appears in 159 sentence–target pairs with opposite polarity signs.
- A common pattern is negative trust/sociability but positive competence.
  - Annotators consistently preserve competence judgments even when trust and sociability are negative, following annotation guidelines.

Analysis – Stance and Sentiment Vs. Social Perceptions

- We analyse 854 W&C-Sent instances (52.3%) sharing targets with the SemEval-2016 Stance dataset
- Stance strongly conditions social perception in both direction and intensity of judgment
- Sentiment shows a similar pattern, with slightly higher correlations for trust and sociability

Analysis – Are the Three Dimensions interrelated?

- The three dimensions are strongly interrelated showing notable links in human perception
  - Trust correlates most strongly with sociability ($\rho = 0.79$)
  - Trust and competence are moderately correlated ($\rho = 0.67$),
  - Sociability and competence are also moderately correlated ($\rho = 0.68$)

Conclusion

- W&C-Sent includes over 1,600 instances covering seven targets
- Each instance is annotated by fluent English speakers for trust (T), sociability (S), and competence (C)
- Experiments show that models struggle to capture subtle social cues
    - The dataset can serve as a benchmark for evaluating models' ability to capture nuanced social dimensions
- We release W&C-Sent publicly, including individual-level labels
- Applications include text analytics, bias and stereotype evaluation, and related NLP and CSS tasks

Thank you!