

# Words of Warmth: Trust and Sociability Norms for over 26k English Words

Saif M. Mohammad

National Research Council Canada

saif.mohammad@nrc-cnrc.gc.ca

## Abstract

Social psychologists have shown that *Warmth* (*W*) and *Competence* (*C*) are the primary dimensions along which we assess other people and groups. These dimensions impact various aspects of our lives from social competence and emotion regulation to success in the work place and how we view the world. More recent work has started to explore how these dimensions develop, why they have developed, and what they constitute. Of particular note, is the finding that warmth has two distinct components: Trust (*T*) and Sociability (*S*). In this work, we introduce *Words of Warmth*, the first large-scale repository of manually derived word–warmth (as well as word–trust and word–sociability) associations for over 26k English words. We show that the associations are highly reliable. We use the lexicons to study the rate at which children acquire WCTS words with age. Finally, we show that the lexicon enables a wide variety of bias and stereotype research through case studies on various target entities. Words of Warmth is freely available at:

<http://saifmohammad.com/warmth.html>

## 1 Introduction

*Who goes there: friend or foe?*

This is a question human beings have asked from the earliest of times to the present day. A large body of social psychology research has shown that *warmth* (*W*) (friendliness, trustworthiness, and sociability) and *competence* (*C*) (ability, power, dominance, and assertiveness) are core dimensions of social cognition and stereotypes (Fiske et al., 2002; Bodenhausen et al., 2012; Fiske, 2018; Abele et al., 2016; Koch et al., 2024). That is, human beings quickly and subconsciously judge (assess) other people, groups of people, and even their own selves along the dimensions of warmth and competence—likely because of evolutionary pressures (MacDonald, 1992; Eisenbruch and Krasnow, 2022). Assessing *W* and *C* was central to early human survival

(e.g., to anticipate whether someone will help them build useful things or whether they might steal their resources).

The dimensions of *W* and *C* have been shown to have substantial implications on a wide variety of facets, including: interpersonal status (Swencionis et al., 2017), social class (Durante and Fiske, 2017), self-beliefs (Wojciszke et al., 2009), political perception (Fiske et al., 2014), child development (Roussos and Dunham, 2016), cultural analyses (Fiske and Durante, 2016), as well as professional and organizational outcomes, such as hiring, employee evaluation, and allocation of tasks and resources (Cuddy et al., 2011).

*W* and *C* are considered to be orthogonal and together they create four quadrants: high *W* and high *C*, low *W* and high *C*, low *W* and low *C*, high *W* and low *C*. The *Stereotype Content Model* (Fiske et al., 2002) argues that how we perceive others is influenced by whether they are considered to be members of the *ingroup* (the same country, political affiliation, language, etc.) or *outgroup* (a different country, political affiliation, language, etc.). Members of the *ingroup* are generally considered to be high *W* and high *C*, whereas members of the *outgroup* tend to be perceived consistent with the other quadrants. For example, it has been shown that the *stereotypical* view towards members of one’s own social class is that they are high *W* and *C*, whereas the poor and homeless are perceived as low *W* (cold) and low *C* (incompetent), the elderly are perceived as high *W* and low *C*, and accountants and business people are perceived as low *W* and high *C* (Fiske, 2018).

These perceptions and stereotypes (influenced by *ingroup* and *outgroup* memberships) evoke different emotions and behaviour. For example, a positive event associated with someone in our *ingroup* (considered warm and competent) evokes pride, whereas a positive event associated with someone in our *outgroup* (e.g., considered cold and compe-

tent) evokes envy. Thus determining W and C perceptions is tremendously valuable in understanding: why people act the way they do; what is driving the discourse in complex social interactions such as discussions about climate change; how different social groups (e.g., immigrants, disabled people, and elderly) are viewed by different groups; and whose view of the world is centered.

More recently, psychologists have shown that warmth should be modeled in terms of two separate dimensions: *Trust (T)* and *Sociability (S)* (Abele et al., 2016; Koch et al., 2024). T is the dimension of trust, morality, goodness, sincerity, and integrity, whereas S is the dimension of sociableness, friendliness, gregariousness, and conviviality. (As shorthand, we will refer to any set of dimensions by simply their letters: WTS for warmth, trust and sociability, WCTS for all four, etc.)

W and C are perceived through various modalities, including: facial expressions, body language, one's actions, what one says, how they say it (words used, tone, etc.). Language is of particular interest as it is a direct and vastly expressive medium. Often, work on W and C makes use of language in the form of responses to researcher questions in labs. However, a notable issue is that people can be reluctant to explicitly divulge their stereotypes towards certain target groups (Nosek et al., 2005; Maina et al., 2018; Hilton and Von Hippel, 1996). Thus, work with every-day utterances and social media data is attractive as a complement to traditional approaches. Further, the words one uses can often communicate W and C through associations (connotations), and can reveal perceptions and stereotypes (even if the speaker is not consciously aware of it).

Large manually compiled repositories of word-competence norms exist for English: e.g., the NRC VAD Lexicon ~ 20k words—used widely for sentiment analysis research. However, existing lexicons for W are much smaller: e.g., Nicolas et al. (2021) manually compiled a set of 341 words.

**Our Work.** We compiled sociability and trust association norms for over 26k English words. The lexicons were created by crowdsourcing and employing a slate of quality control measures. We show that the resulting association scores have high reliability (repeating the annotations leads to very similar scores and rankings). We created a third combined lexicon for warmth by taking the union of the entries for the trust and sociability lexicons.

Together, we refer to the set of three lexicons as the *Words of Warmth Lexicons*.

The three lexicons enable a wide variety of research and applications. Notably:

#### *In Psychology and Social Cognition*

- What kind of trust assessments do children develop first? And what kinds are developed later? (Trust can be of different kinds: care-based, character-based, consistency-based, etc.) Similarly for sociability.
- What are the mechanisms underpinning the development of WCST assessment capabilities in children? How does exposure to different conditions impact these capabilities?
- How different are the WCST capabilities of people in different cultures?
- What role do differences in language play in the development of WCST capabilities?

#### *In Computational Social Science, NLP*

- The lexicons can be used to study public discourse on topics of interest. For example, how are the levels of warmth, competence, trust, and sociability in online discussions about climate change or vaccines changing with time; how do these levels vary for different stakeholders?; what sub-aspects of climate change (or vaccines or any topic of interest) evoke the lowest amounts of warmth, competence, trust, and sociability? etc.
- How has the perceived WCST of a chosen target of interest (say government, banks, immigrants, etc.) changed over the last 100 years?

#### *In HCI and NLP*

- Understanding perceptions of WCST of people towards artificial agents.

#### *In Digital Humanities and NLP*

- What role do warmth, trust, sociability, and competence play in developing compelling characters and story arcs? How does this vary by genre and culture?

#### *In Commerce*

- Tracking warmth, trust, sociability, and competence towards one's product on social media. This can help understand product branding, tracking user satisfaction, and taking the appropriate remedial actions for product improvement and public-facing communications.
- Understanding how perceptions of warmth and competence of one's product impact customer behavior.

In the second half of this paper, we use the lexicons to explore:

1. At what rate do children acquire WCTS words? And how do these change with age? This sheds light on how social cognition develops and on the relative importance of the two dimensions. (5)
2. How do we use W and C words in social media, especially when mentioning various social groups? This sheds light on how our perceptions of W and C towards various social groups manifests in public discourse.

We make all of the lexicons and code freely available for research.<sup>1</sup>

## 2 Related Work

Despite the considerable importance of warmth and competence in social cognition and behaviour (as discussed in the Introduction), there is much we do not know about how these dimensions develop; how children assess W and C of those around them; and which dimension is of greater significance.

Some research argues that warmth is the primary component of valence, which in turn is evolutionarily central to the approach–avoid response, and so assessment of warmth emerges earlier than competence (Cuddy et al., 2007). This is the *primacy of valence* hypothesis. The view that, in children, competence emerges earlier than warmth, is known as the *primacy of competence hypothesis* (Roussos and Dunham, 2016). In support of this hypothesis are some studies that show that infants (even as young as 6 to 8 months) assess competence levels and show more trust in those they think are more competent (Koenig and Echols, 2003; Tummeltshammer et al., 2014). Finally, there are studies showing how warmth towards the child from the caregiver has tremendous positive benefits for the child, arguably again showing that warmth is more important for children than competence. For example, Altschul et al. (2016) show that spanking by the caregiver predicted increases in child aggression. In contrast, caregiver warmth (much more than spanking) predicted social competence.

Since words act as the principal carriers of meaning, and many words connote W and C, large lexicons of W and C associations can be powerful resources for understanding questions such as those discussed above. There exist many lexicons for competence (aka dominance), such as

the Warriner et al. (2013) and Mohammad (2018) for English; Moors et al. (2013) for Dutch, and Vö et al. (2009) for German. The largest among these is the NRC VAD lexicon (Mohammad, 2018, 2025): version 1 has entries for over 20,000 English words, and version 2 for over 44,000 unigrams and 10,000 bigrams (two-word sequences). However, manually compiled lexicons for word–warmth associations are much smaller. Most notably, Nicolas et al. (2021) manually compiled a set of 341 words. They also expanded this lexicon automatically using WordNet synonyms and word embeddings. However, even near synonyms and distributionally close term pairs can convey very different WST associations; for example, slip vs. fault and skinny vs. slender. In fact, Fraser et al. (2024) found that the automatically expanded lexicon was not effective in capturing W and C.

In response to the tremendous upsurge of social media content, generative AI, polarization, and rising misinformation, we have also seen growing interest in tackling stereotypes and bias research in NLP. Influential early work explored race and gender bias in word embeddings and automatic systems (Caliskan et al., 2017; Thelwall, 2018; Kiritchenko and Mohammad, 2018; Tan and Celis, 2019; Blodgett et al., 2020). A considerable amount of recent research explores bias and stereotypes in generative AI (Kotek et al., 2023; Zhou et al., 2024; Baines et al., 2024). Yet, a growing area of interest is work on exploring bias and stereotypes in large amounts of social media data from a computational social science perspective (Sánchez-Junquera et al., 2021; Ariza-Casabona et al., 2022; Bosco et al., 2023; Fraser et al., 2024; Schmeisser-Nieto et al., 2024).

The WST lexicons we created are useful in studying the core dimensions of social cognition, how they develop in children, how they impact our traits, and how they shape our views and stereotypes.

## 3 Obtaining Human Ratings for Trust, Sociability, and Warmth

We describe the main steps below.

**1. Term Selection.** We wanted to include a large set of common English words. Further, we wanted to especially include terms with emotion associations (as opposed to lots of emotionally neutral terms). Thus, we chose the NRC VAD Lexicon (Mohammad, 2018) as the source of terms. Version 2 includes ~44k unigrams annotated for

<sup>1</sup><http://saifmohammad.com/warmth.html>

valence, arousal, and dominance. The valence (or sentiment) scores go from -1 (maximum negativity) to +1 (maximum positivity). Scores between -0.33 and +0.33 correspond to neutral valence. After manual examination of the valence scores, we chose to exclude terms with a valence score between -0.2 and +0.2 (keeping all of the non-neutral terms as well as some neutral terms). This resulted in a set of 26,188 unigrams.

**2. Trust and Sociability Questionnaires.** The questionnaires used to annotate the data were developed after several rounds of pilot annotations. Detailed directions, including notes directing respondents to consider predominant word sense (in case the word is ambiguous) and example questions (with suitable responses) were provided. (See Appendix.) The primary instruction and the questions presented to annotators are shown below.

Consider trustworthiness to be a broad category that includes: *trustworthy, honesty, fairness, dependability, reliability, morality, virtuousness, sincerity, honorableness, etc.*

Consider untrustworthiness to be a category that includes: *unfairness, dishonesty, untrustworthiness, dubiousness, immorality, sinfulness, insincerity, dishonorableness, etc.*

Q1. <term> is often associated with feeling:

- |   |                              |
|---|------------------------------|
| 3: very trustworthy                                       | -1: slightly untrustworthy   |
| 2: moderately trustworthy                                 | -2: moderately untrustworthy |
| 1: slightly trustworthy                                   | -3: very untrustworthy       |
| 0: not associated with being trustworthy or untrustworthy |                              |

Consider social warmth to be a broad category that includes: *warmness, sociableness, generosity, helpfulness, tolerance, understanding, thoughtfulness, etc.*

Consider social coldness to be a broad category that includes: *coldness, antisocialness, ungenerosity, unhelpfulness, intolerance, indifferent, thoughtlessness, etc.*

Q1. <term> is often associated with feeling:

- |   |                           |
|---|---------------------------|
| 3: very sociable                                    | -1: slightly unsociable   |
| 2: moderately sociable                              | -2: moderately unsociable |
| 1: slightly sociable                                | -3: very unsociable       |
| 0: not associated with being sociable or unsociable |                           |

**3. Quality Control Measures.** About 2% of the data was annotated beforehand by the authors and interspersed with the rest. These questions are referred to as *gold* (aka *control*) questions. Half of the gold questions were used to provide immediate feedback to the annotator (in the form of a pop-up on the screen) in case they mark them incorrectly. We refer to these as *popup gold*. This helps prevent the situation where one annotates a large number of instances without realizing that they are doing so incorrectly. It is possible, that some annotators share answers to gold questions with each other (despite this being against the terms of annotation). Thus, the other half of the gold questions were

also separately used to track how well an annotator was doing the task, but for these gold questions no popup was displayed in case of errors. We refer to these as *no-popup gold*.

**4. Crowdsourcing.** We setup the annotation tasks on the crowdsourcing platform, *Mechanical Turk*. In the task settings, we specified that we needed annotations from nine people for each word. (Since we got some additional funding later, three more annotations per word were obtained for trust.) We obtained annotations from native speakers of English residing around the world. Annotators were free to provide responses to as many terms as they wished. The annotation task was approved by the National Research Council Canada’s Institutional Review Board.

**Demographics:** Hundreds of annotators participated in each of the annotation tasks. About 69% of the respondents live in USA. The rest were from India, United Kingdom, and Canada. The average age of the respondents was 39.2 years. Among those that provided a response to the gender question: about 48% entered female, 52% said male, and no one marked themselves as nonbinary (or provided any other response).

**5. Filtering.** If an annotator’s accuracy on the gold questions (popup or non-popup) fell below 80%, then they were refused further annotation, and all of their annotations were discarded (despite being paid for). See Table 1 for summary statistics.

**6. Aggregation.** Every response was mapped to an integer from -3 (very untrustworthy/unsociable) to 3 (very trustworthy/sociable). The final score for each term is simply the average score it received from each of the annotators. We also created a categorical version of the sociability (S) lexicon by labeling all words that got an average S score  $\geq 2.5$  as *high S*,  $\geq 1.5$  and  $< 2.5$  as *moderate S*,  $\geq 0.5$  and  $< 1.5$  as *slight S*,  $> -0.5$  and  $< 0.5$  as *neither sociable nor unsociability*, and so on. The categorical version of the trust (T) lexicon was created similarly.

Social cognition research (Abele et al. 2016, Koch et al., 2024) points to how we perceive a person (or group) to be warm because we associate them with kindness, honesty, gregariousness, thoughtfulness, or some other quality associated with warmth. It is not required that one is both kind and gregarious or both honest and gregarious, etc. Thus, we created a warmth (W) lexicon by taking



Dataset	#words	Annotators	#Annotations	MAI	SHR ( $\rho$ )	SHR ( $r$ )
sociability	26,123	US, India, UK, Canada	205,475	7.9	0.965	0.969
trust	26,185	US, India, UK, Canada	299,365	11.4	0.943	0.957
warmth	26,085	US, India, UK, Canada	229,580	8.8	0.965	0.974

Table 1: A summary of the Words of Warmth annotations. MAI = mean annotations per word. SHR, measured through both Spearman rank and Pearson’s correlations (last two columns), indicate high reliability.

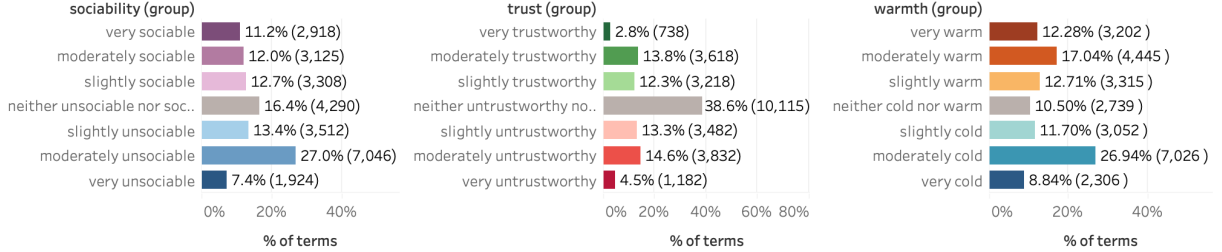


Figure 1: Distribution of terms in Words of Warmth: percentage and number of terms associated with each class.

the union/or’ing of the entries for T and S.<sup>2</sup>

For a given word  $x$ , if the absolute value of the T score for  $x$  is greater than the absolute value of the S score for  $x$ , then the W score for  $x$  is taken to be the T score. If the absolute value of the S score for  $x$  is greater than the absolute value of the T score for  $x$ , then the W score for  $x$  is taken to be the S score. If the two scores are the same, then the same score is taken as the W score. Thus, for example, for the word *uplifting* with an S score of 3 and a T score 0.67, the W score is 3; *birdbrain* with an S score of -1.71 and a T score of -2.62 gets a W score of -2.62. Figure 12 in the Appendix shows a scatter plot of words on the T–S space, colour coded as per their W score.

We refer to the list of words along with their scores and categorical labels for WST as the *Words of Warmth Lexicons*. (Table 2 in the Appendix shows example entries.) Since warmth analyses are often done along with competence (aka dominance) analyses, we also include in the lexicon the competence scores for the terms (taken from the NRC VAD Lexicon v2 (Mohammad, 2025)). Thus we also refer to this suite of lexicons as the *Warmth–Competence Lexicons*, or the *WCST Lexicons*.

Figure 1 shows the distribution of the different classes. As expected, most entries for trust are associated with neither trustworthiness nor untrustworthiness (38.6%), but it is worth noting that 28.9% of the words are associated with trustworthiness (to some degree) and 32.4% of the words are associated with untrustworthiness (to some degree). The pattern is different for sociability, wherein, a

large number of inanimate objects are seen as moderately unsociable (the most frequent category). In the warmth lexicon, 10.5% of the entries are marked as neither cold nor warm, whereas 42% have some association with warmth and 47.5% have some association with coldness. (Figure 13 in the Appendix shows a further break down of the percentage of terms in each of the warmth classes into the percentage of entries obtained from the trust lexicon and the percentage of entries obtained from the sociability lexicon.)

## 4 Reliability of the Annotations

A useful measure of quality is the reproducibility of the end result—repeated independent manual annotations from multiple respondents should result in similar scores. To assess this reproducibility, we calculate average *split-half reliability* (SHR) over 1000 trials.<sup>3</sup> All annotations for an item are randomly split into two halves. Two separate sets of scores are aggregated, just as described in Section 3 (bullet 6), from the two halves. Then we determine how close the two sets of scores are (using a metric of correlation). This is repeated 1000 times and the correlations are averaged. The last two columns in Table 1 show the results (split half-reliabilities). Spearman rank and Pearson correlation scores of around 0.95 indicate very high reliability of the real-valued scores obtained from the annotations.<sup>4</sup>

<sup>3</sup>SHR is a common way to determine reliability of responses to generate scores on an ordinal scale (Weir, 2005).

<sup>4</sup>For reference, if the annotations were random, then repeat annotations would have led to an SHR of 0. Perfectly consistent repeated annotations lead to an SHR of 1. Also, similar past work on word–anxiety associations had SHR scores in the 0.8s (Mohammad, 2024).

<sup>2</sup>That said, we also make the individual S and T scores available. So one can easily take the mean or some other function if that is more suitable for their particular application.

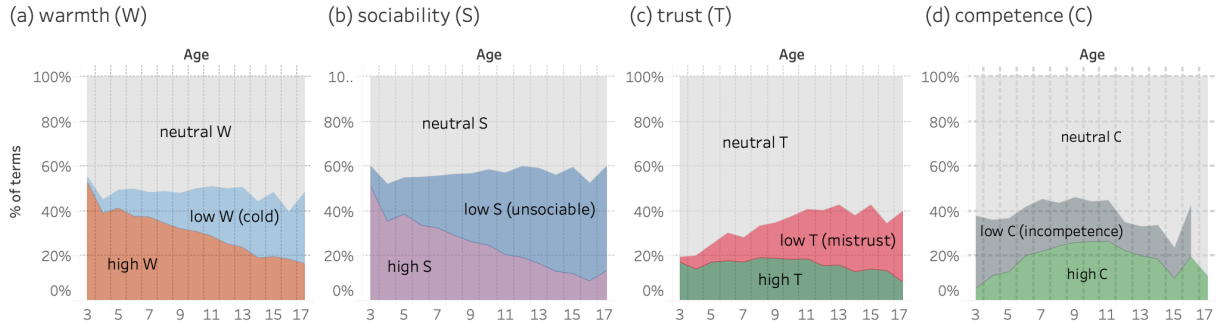


Figure 2: Stream charts of the percentages of high-, low-, and neutral WCTS words acquired in ages 3 to 17. (The three percentages for each age sum up to 100%.)

## 5 At what rate do children acquire words associated with warmth, competence, trust, and sociability?

As discussed in the Related Work, there is much we do not know about how warmth and competence develop and which dimension is of greater significance. To shed some light on this, we use the WCTS lexicons in combination with an age of acquisition dataset (Kuperman et al., 2012) to examine when children acquire WCTS-associated words. The age of acquisition dataset includes the age at which  $\sim 30K$  English words are commonly acquired by children.

For this set of experiments, every dimension is split into three regions: low, neutral, and high. For WST: scores between  $-1.5$  and  $1.5$  are considered neutral; scores  $\leq -1.5$  are considered low (low T or untrustworthy, low S or unsociable, low W or cold); and scores  $\geq 1.5$  are considered high.<sup>5</sup> We will refer to the set of words from the high- and low-regions as polar words. (In other words, for a given dimension, the set of polar words includes all words except the neutral words.)

Figures 2 a through d show stream graphs of the percentages of high, low, and neutral words acquired each year by children. (For every x-axis value, the three percentages sum up to 100%.) Observe that from age 3 onward, the percentage of high-W words decreases steadily with age, whereas, the percentage of low-W words increases with age. Overall, the number of polar warmth words acquired with age stays steady at about 50%. The pattern for competence (d) is markedly different. The rate at which high-C words are acquired increases gradually, peaking at about 10 years of age, and then decreasing again. The rate of acqui-

sition of low-C words is highest in the early years and decreases steadily with age. The rate of acquisition of polar words has a slight inverted U pattern (peaking at 10 years). Overall, the percent of polar W words acquired at each of the ages is higher than the percent of polar C words.

The rate of acquisition of high- and low-S words (shown in (b)) is similar to that of high- and low-W words. The rate for high-T words starts off high and stays steady till about 10 to 11 years, after which there is a slight but steady decline. The rate of acquisition of low-T words is very small at age 3, but it increases steadily with age.

*Discussion:* Overall, we see clear trends in the acquisitions of words for each of the dimensions. The higher percentages for polar (non-neutral) warmth words vs. polar competence words is consistent with the primacy of valence hypothesis (as opposed to the primacy of competence hypothesis). The markedly higher percentages for polar sociability words as opposed to polar trust words in early years, is consistent with the notion that the dimension of sociability is more important than the dimension of trust (and morality) in the early years. Among the polar words, it is interesting that the early years are marked with a greater percentage of high-WST words, as well as low-C words. This is consistent with the notion that children receive more warmth earlier in life than later. The higher percentage of low-C words is consistent with the fact that children are more heavily dependent on caregivers in early years than in later years.

These findings have implications in developmental psychology and evolutionary linguistics. They are also relevant to understanding how children develop these key dimensions of social cognition and their role in shaping traits such as social competence and emotion regulation (Roussos and Dunham, 2016; Wojciszke et al., 2009).

<sup>5</sup>For C scores (taken from the NRC VAD lexicon): scores between  $-0.33$  and  $0.33$ : neutral; between  $-1$  and  $-0.33$ : low C; between  $0.33$  and  $1$ : high C.

## 6 Case Studies of W and C Stereotypes

Our stereotypes about people often manifest in language. The WCTS lexicons (with WTS scores from our newly created lexicon and competence scores from the NRC VAD Lexicon v2) can be used in combination with large amounts of text to shed light on human stereotypes. We make use of two methods which provide different windows into human stereotype towards various targets commonly studied in stereotype research (Morabito et al., 2024): Direct Target Lookup of target terms in the WCTS lexicon (**Direct WCTS**) and Examining WCTS of terms co-occurring with the target terms in text (**Co-terms WCTS**). For our Co-terms WCTS experiments we use a corpus of lemmatized and lower-cased American and Canadian geo-located posts on X (formerly Twitter) from 2015 to 2021 (Vishnubhotla and Mohammad, 2022; Wahle et al., 2025).<sup>6</sup>

**Direct WCTS:** One can directly look up WCTS scores of target terms in the lexicons. For example, Figure 3 (a) shows the W and C scores of terms referring to various social groups. The shading marks the quadrants: high W and C (white), high W and low C (yellow), low W and high C (green), and low W and C (blue). Note that the average W and C scores of all 2,086 terms in the lexicons are 0.002 and 0.001, respectively (very close to 0). Therefore, since the term *worker* has a positive score for both W and C, it means that people perceive *worker* as being more warm and more competent than the average term in the lexicon. Note how the concept of *god* is perceived as highly warm and highly competent; *disabled* as very low C; *criminal* as very low W; and people outside of the socially preferred weight class as low W and C. Some of the terms in our original list such as *lgbtq*, *muslim*, and *jew* do not occur in the lexicon. These words can be analyzed through the WCTS of their co-occurring terms described ahead.

**Co-terms WCTS:** We obtain co-occurrence based WCTS scores by examining the lexicon entries of terms co-occurring with the target terms. Steps:

1. Manually identify minimally ambiguous terms commonly used to refer to the target. Collect posts that include mentions of the target term(s). E.g., using the terms *nurse* and *nurses* to collect posts about nursing professionals.

2. Calculate co-term WCTS scores for each target term. Following (Teodorescu and Mohammad, 2023; Turney, 2002) we calculate the percentage of high-W words in the target corpus minus the percentage of low-W words in the target corpus.<sup>7</sup> CTS scores are determined similarly.

The WCTS scores obtained using co-terms give an indication of the extent to which we use high- and low-WCTS terms in utterances that include the target term. Higher scores indicate more high-WCTS words and fewer low-WCTS words. Some important points should be noted regarding how to further interpret these scores:

1. The co-term scores need not correlate with the target scores. This can happen for a number of reasons, including: how we respond when directly asked about a target may differ from our true feelings; mentions of the target may be in a restricted context not representing the full set of contexts in which the target is talked about; etc.
2. Different groups of people may use different terms to refer to the same target entity. For example, people who use the term *lgbtq* tend to view the group more positively than people who do not (e.g., those who use identity terms dispreferred by the group).
3. Even though the co-term scales have the same range as the direct scores (−1 to 1), the two metrics are not directly comparable. For one, target scores have a normal distribution around 0, whereas the co-term scores have normal distributions at an offset. For example, the average W and C scores of all 3.1 million tweets in our tweets corpus are 0.5001 and 0.1370, respectively.<sup>8</sup> Thus, in the analyses below we examine relative positions of targets w.r.t. the average on the W–C plots. We also consider quadrants in the W–C space with respect to the W and C averages (and not w.r.t. 0, 0).
4. We checked for stability of the WCTS scores for a given target entity, by looking at how much the scores vary for each of the years from 2015 to 2021, and also by examining scores for morphological variants of the target term. The closeness

<sup>6</sup>The TUSC tweets (Vishnubhotla and Mohammad, 2022) with WCTS features is now part of the ABCDE dataset for Computational Affective Science (Wahle et al., 2025).

<sup>7</sup>Teodorescu and Mohammad (2023) and Turney (2002) show that this formula accurately captures the degree of emotions in various corpora for valence, arousal, anger, sadness, etc. Other similar formulae may also be used. Our goal was to use a simple and interpretable approach that has been shown to work well for aggregate-level analysis.

<sup>8</sup>This is because in a sentence we often use many warmth words, whereas even one or two coldness words are sufficient to convey a strong overall coldness tone.

of these scores indicate stability of the WCTS scores. For example, in Fig 14 in the Appendix we plot the values for the pronouns for every year; and in Fig 5 we plot the values for both *america* and *american*, as well as *canada* and *canadian*. Due to limited space and for clarity we omit other plots showing scores for different years and morphological variants.

5. We use the W-C plots as the primary mechanism to showcase the kind of analyses the WCTS lexicons enable, but note that similar analysis can be done with T-C, S-C, etc. For example, one notable aspect we found in our analysis was that while many of the social groups considered had T and S scores that were close to each other, there exist terms such as *homosexual* whose T score was quite different (in this case, much lower) than their S score. This helps us understand and track how the discourse about gay people is still polarizing and a section of society uses low-trust (morality) terms when talking about this group (mirroring the known negative and harmful stereotypes against them).

We show some case studies below. (The Appendix has a supplemental case study of professions.)

**1. Social Groups.** Figure 3 (b) shows the co-terms based W and C scores of various social groups. Some notable observations include:

- *muslim*, *jew*, and *immigrant* get low-W scores (consistent with known negative stereotypes towards them in US and Canada).
- *elderly* and *underweight* get low-C and high-W scores; whereas *overweight* gets an even lower C score. *obese* gets low-W and low-C scores.
- *god* gets high direct W and C scores (Figure 3 (a)), but the discourse around *god* on X is such that the term gets lower co-terms-based C score than many other terms (b).

**2. Genders.** Figure 4 (a) and (b) show the direct and co-terms based W and C scores of various gender groups. Observe that:

- When asked for W and C assessment directly, people consider all these terms as high W, but perceive substantial variations in their competence. *father* and *mother* are seen as high C whereas *grandmother* is seen as low C.
- In contrast, the co-term plots show that our language has marked differences for these terms not just for C but also for W. We use the most high W and fewer low W words when mentioning *grandfather* (even more than *daughter*, and

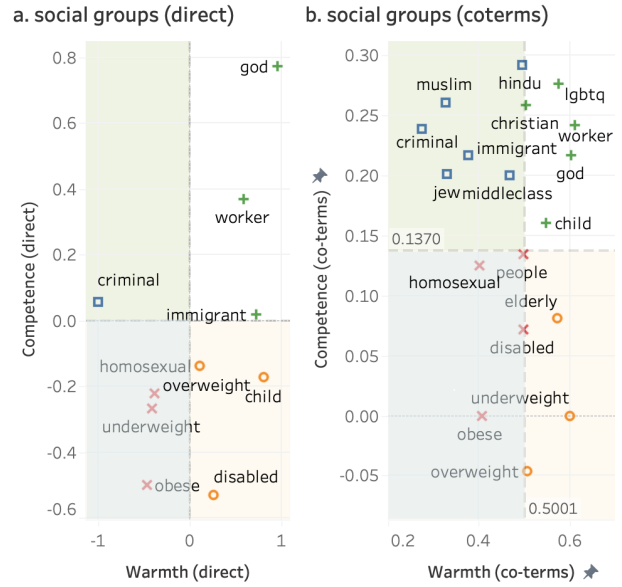


Figure 3: W-C plots for social groups.

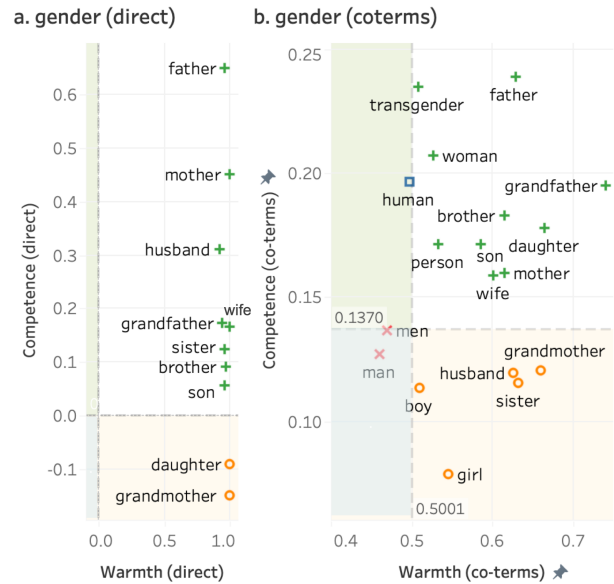


Figure 4: W-C plots for various gender terms.

*grandmother*) compared to *son*, *wife*, and *husband*. The co-term plot also shows certain additional related terms for comparison such as *person* and *human*.

- Importantly, we see clear gender stereotypes reflected in these scores with males getting higher C scores and females getting higher W scores (with the exception of *grandfather*).

**3. Ingroup-Outgroup Impact.** Since we know which tweets were posted in Canada and which in the USA, we can use that information to examine ingroup and outgroup behavior. Specifically, we examine the W and C words used by Canadians and Americans when they refer to each other. Figure 5 shows the co-terms based W and C scores



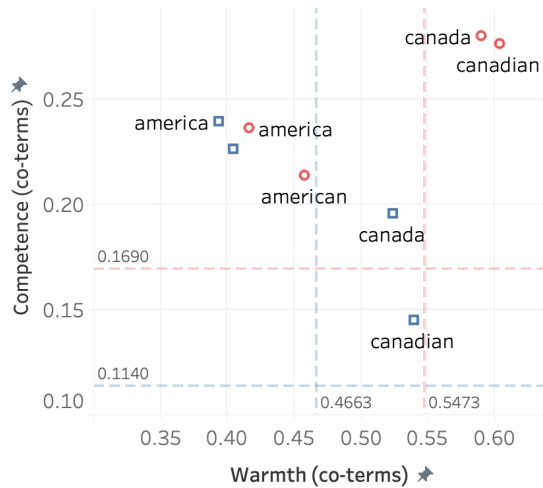


Figure 5: Co-terms W-C plot for tweets by Canadians (red) and Americans (blue) mentioning each other.

for: *america*, *american*, *canada*, and *canadian* obtained from the tweets by each group (Americans and Canadians). Note the blue dashed lines that indicate the average W and C scores of all posts by Americans and red dashed lines that indicate the averages for Canadians. We observe that posts by Canadians in general have higher W and C scores than posts by Americans. Further, the scores provide some evidence that Canadians view themselves as more competent and much warmer than their neighbours (consistent with ingroup and outgroup stereotypes). In contrast, while Americans view themselves as more competent than Canadians, they too perceive Canadians as warmer (suggesting that the *Canadians are nicer* stereotype overrides the outgroup stereotype in this case).

**4. Pronouns.** Analyzing posts with pronouns gives us interesting insights into how we speak about ourselves (1st person) vs. a person we are directly engaging with (2nd person) vs. a third person. How do the W and C scores differ for these different types of posts? Figure 14 in the Appendix shows the plots.

- Firstly, we note that different pronouns fall in markedly separate areas of the W-C plots (can also be seen from the yearly plot shown in (b)).
- Secondly, first-person-singular mentions (*I* and *me*) are associated with low C, whereas and second person pronoun mentions (*you*) are associated with high C. This is consistent with findings by Kacewicz et al. (2014) and (Pennebaker, 2011), who show that *I* is often overused by people in low-power/status relations, whereas *you* and *we* are often used by people in high-

power/status relationships. Additionally, we speculate this is because social media interactions often involve people who may not personally know each other, and so as a matter of being socially congenial to the addressee we prioritize competence over warmth and tend to use more high-C and fewer low-C words.

- Mentions of *we* are associated with high W. This is consistent with findings that show that *we* commonly occurs in more positive contexts (Sendén et al., 2014; Pennebaker, 2011).

With the lexicons made freely available, we hope they will spur further detailed exploration into the targets discussed above as well as numerous others.

## 7 Conclusion

We created Words of Warmth — a large lexicon of word-warmth, word-trust, and word-sociability association scores from over 700,000 responses from hundreds of respondents. We showed that the scores are highly reliable ( $>0.94$  SHR). We used the lexicon to study the rate at which WCTS words are acquired with age. Finally, we presented case studies on how the lexicons can be used to track stereotypes.

We make the lexicons freely available to foster further research, notably on understanding: how human beings develop their inner stereotype model; how early childhood experiences can impact our stereotype model, social competence, emotion regulation, and personality traits; and tracing stereotype and bias of a source population of interest towards a target entity of interest. Other areas of future work, include: using the lexicons to study text produced by generative AI (the degree of stereotypes it reflects and under what conditions); creating warmth lexicons for more languages and cultures to enable cross-cultural comparisons; developing automatic systems to assess systematic and consistent trends in WCTS biases on various social media channels (such as various subreddits); and analyzing one’s own writing (over a period of time) to understand how we assess ourselves in terms of WCTS (and its implications on mental health). One can even use the lexicons to assess perceptions of complex social issues such as how we should deal with climate change and immigration. Thus, we see wide applicability of Words of Warmth in psychology, computational affective science, NLP, public health, digital humanities, political science, and social science research.

## 8 Limitations

The large body of social psychology work on the dimensions of social cognition and stereotype are based on human responses. Traditionally, these studies only included responses from people in the western world. This can lead to over generalizations. However, more recently there has been growing work that confirms the importance of warmth and competence across cultures (Fiske and Durante, 2016; Grigoryev et al., 2019) and showing the evolutionary basis of these dimensions (MacDonald, 1992; Cuddy et al., 2007; Eisenbruch and Krasnow, 2022). Nonetheless, it is entirely possible, that in some cultures W and C are not the primary dimensions of social cognition.

This work develops WTS lexicons for English, based on responses primarily from the US, Canada, UK, and India. Thus it is important to contextualize any conclusions as those applying to English speakers, and that too mainly US speakers. Just as the social psychology work, true global conclusions can only be drawn from many such works on many languages and cultures. We see this work as a first step that paves the way for more work in various other languages and cultures.

See discussions of limitations in how the lexicons can be used and interpreted in the Ethics Statement below (§9).

## 9 Ethics and Data Statement

The crowd-sourced task presented in this paper was approved by our Institutional Research Ethics Board. Our annotation process stored no information about annotator identity and as such there is no privacy risk to them. The individual words selected did not pose any risks beyond the risks of occasionally reading text on the internet. The annotators were free to do as many word annotations as they wished. The instructions included a brief description of the purpose of the task (Figures 6 and 8).

WCTS assessments are complex, nuanced, and often instantaneous mental judgments. Additionally, each individual may use language to convey these assessments slightly differently. We discuss below notable ethical considerations when computationally analyzing WCTS through language.

Importantly, Words of Warmth should not be used as a standalone tool for detecting stereotypes and bias in individual utterances. At minimum, it must be used in combination with various other

sources of information, large amounts of texts, and appropriate contextualization (the same text may mean different things in different contexts). See considerations below, which also apply broadly to any lexical dataset of association norms (many of these build on similar issues for emotions, discussed in Mohammad (2023, 2022)):

1. *Coverage*: We sampled a large number of English words from other lexical sources (which themselves sample from many sources). Yet, the words included do not cover all domains, genres, and people of different locations, socio-economic strata, etc. equally. It likely includes more of the vocabulary common in the United States with socio-economic and educational backgrounds that allow for technology access.
2. *Word Senses and Dominant Sense Priors*: Words when used in different senses and contexts may be associated with different degrees of WCTS associations. The entries in Words of Warmth are indicative of the associations with the predominant senses of the words. This is usually not problematic because most words have a highly dominant main sense (which occurs much more frequently than the other senses). In specialized domains, some terms might have a different dominant sense than in general usage. Entries in the lexicon for such terms should be appropriately updated or removed. Further, any conclusions using the lexicon should be made based on relative change of associations using a large number of textual tokens. For example, if there is a marked increase in coldness words from one period to the next, where each period has thousands of word tokens, then the impact of word sense ambiguity is minimal, and it is likely that some broader phenomenon is causing the marked increase in coldness words. (See last two bullets.)
3. *Not Immutable*: The WCTS scores do not indicate an inherent unchangeable attribute. The associations can change with time (e.g., the decrease in coldness and immorality associated with *inter-race relationships* over the last 100 years), but the lexicon entries are largely fixed. They pertain to the time they are created. However, they can be updated with time.
4. *Socio-Cultural Biases*: The annotations for WCTS capture various human biases. These biases may be systematically different for different socio-cultural groups. Our data was annotated by mostly US, Canadian, UK, and Indian English

speakers, but even within these countries there are many diverse socio-cultural groups. Notably, crowd annotators on Amazon Mechanical Turk do not reflect populations at large. In the US for example, they tend to skew towards male, white, and younger people. However, compared to studies that involve just a handful of annotators, crowd annotations benefit from drawing on hundreds and thousands of annotators (such as this work).

5. *Inappropriate Biases*: Our biases impact how we view the world, and some of the biases of an individual may be inappropriate. For example, one may have race or gender-related biases that may percolate subtly into one's notions of WCTS associated with words. Our dataset curation was careful to avoid words from problematic sources. We also ask people annotate terms based on what most English speakers think (as opposed to what they themselves think). This helps to some extent, but the lexicon may still capture some historical WTS associations with certain identity groups. This can be useful for some socio-cultural studies; but we also caution that WCTS associations with identity groups be carefully contextualized to avoid false conclusions.
6. *Perceptions (not "right" or "correct" labels)*: Our goal here was to identify common perceptions of WTS association. These are not meant to be "correct" or "right" answers, but rather what the majority of the annotators believe based on their intuitions of the English language.
7. *Avoid Essentialism*: When using the lexicon alone, it is more appropriate to make claims about WCTS word usage rather than the WCTS of the speakers. For example, '*the use of trust words in the context of the target group grew by 20%*' rather than '*trust in the target group grew by 20%*'. In certain contexts, and with additional information, the inferences from word usage can be used to make broader claims.
8. *Avoid Overclaiming*: Inferences drawn from larger amounts of text are often more reliable than those drawn from small amounts of text. For example, '*the use of warmth words grew by 20%*' is informative when determined from hundreds, thousands, tens of thousands, or more instances. Do not draw inferences about a single sentence or utterance from the WCTS associations of its constituent words.

9. *Embrace Comparative Analyses*: Comparative analyses can be much more useful than stand-alone analyses. Often, WCTS word counts and percentages on their own are not very useful. For example, '*the use of warmth words grew by 20% when compared to [data from last year, data from a different person, etc.]*' is more useful than saying '*on average, 5 warmth words were used in every 100 words*'.

We recommend careful reflection of ethical considerations relevant for the specific context of deployment when using Words of Warmth.

## Acknowledgments

Thanks to Susan Fiske, Tara Small, Nedjma Oousidhoum, Jan Philip Wahle, and Kathleen Fraser for helpful discussions. Thanks to Jan Philip Wahle for the early access to the ABCDE dataset.

## References

- Andrea E Abele, Nicole Hauke, Kim Peters, Eva Louvet, Aleksandra Szymkow, and Yanping Duan. 2016. Facets of the fundamental content dimensions: Agency with competence and assertiveness—communion with warmth and morality. *Frontiers in psychology*, 7:1810.
- Inna Altschul, Shawna J Lee, and Elizabeth T Gershoff. 2016. Hugs, not hits: Warmth and spanking as predictors of child social competence. *Journal of Marriage and Family*, 78(3):695–714.
- Alejandro Ariza-Casabona, Wolfgang S Schmeisser-Nieto, Montserrat Nofre, Mariona Taulé, Enrique Amigó, Berta Chulvi, and Paolo Rosso. 2022. Overview of detests at iberlef 2022: Detection and classification of racial stereotypes in spanish. *Procesamiento del lenguaje natural*, 69:217–228.
- Alexander Baines, Lidia Gruia, Gail Collyer-Hoar, and Elisa Rubegni. 2024. Playgrounds and prejudices: Exploring biases in generative ai for children. In *Proceedings of the 23rd Annual ACM Interaction Design and Children Conference*, pages 839–843.
- Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in nlp. *arXiv preprint arXiv:2005.14050*.
- Galen V Bodenhausen, Sonia K Kang, and Destiny Peery. 2012. Social categorization and the perception of social groups. *The Sage handbook of social cognition*, pages 311–329.
- Cristina Bosco, Viviana Patti, Simona Frenda, Alessandra Teresa Cignarella, Marinella Paciello, and Francesca D'Errico. 2023. Detecting racial stereotypes: An italian social media corpus where psychology meets nlp. *Information Processing & Management*, 60(1):103118.

- Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.
- Amy JC Cuddy, Susan T Fiske, and Peter Glick. 2007. The bias map: behaviors from intergroup affect and stereotypes. *Journal of personality and social psychology*, 92(4):631.
- Amy JC Cuddy, Peter Glick, and Anna Beninger. 2011. The dynamics of warmth and competence judgments, and their outcomes in organizations. *Research in organizational behavior*, 31:73–98.
- Federica Durante and Susan T Fiske. 2017. How social-class stereotypes maintain inequality. *Current opinion in psychology*, 18:43–48.
- Adar B Eisenbruch and Max M Krasnow. 2022. Why warmth matters more than competence: A new evolutionary approach. *Perspectives on Psychological Science*, 17(6):1604–1623.
- Susan Fiske, Amy Cuddy, Peter Glick, and Jun Xu. 2002. A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, 82:878–902.
- Susan T Fiske. 2018. Stereotype content: Warmth and competence endure. *Current directions in psychological science*, 27(2):67–73.
- Susan T Fiske and Federica Durante. 2016. Stereotype content across cultures. *Handbook of advances in culture and psychology*, 6:209–258.
- Susan T Fiske, Federica Durante, et al. 2014. Never trust a politician? collective distrust, relational accountability, and voter response. *Power, politics, and paranoia: Why people are suspicious of their leaders*, pages 91–105.
- Kathleen Fraser, Svetlana Kiritchenko, and Isar Nejadgholi. 2024. [How does stereotype content differ across data sources?](#) In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (\*SEM 2024)*, pages 18–34, Mexico City, Mexico. Association for Computational Linguistics.
- Dmitry Grigoryev, Susan T Fiske, and Anastasia Batkhina. 2019. Mapping ethnic stereotypes and their antecedents in russia: The stereotype content model. *Frontiers in psychology*, 10:1643.
- James L Hilton and William Von Hippel. 1996. Stereotypes. *Annual review of psychology*, 47(1):237–271.
- Ewa Kacwicz, James W Pennebaker, Matthew Davis, Moongee Jeon, and Arthur C Graesser. 2014. Pronoun use reflects standings in social hierarchies. *Journal of Language and Social Psychology*, 33(2):125–143.
- Svetlana Kiritchenko and Saif Mohammad. 2018. [Examining gender and race bias in two hundred sentiment analysis systems](#). In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana. Association for Computational Linguistics.
- Alex Koch, Austin Smith, Susan T Fiske, Andrea E Abele, Naomi Ellemers, and Vincent Yzerbyt. 2024. Validating a brief measure of four facets of social evaluation. *Behavior Research Methods*, 56(8):8521–8539.
- Melissa A Koenig and Catharine H Echols. 2003. Infants’ understanding of false labeling events: The referential roles of words and the speakers who use them. *Cognition*, 87(3):179–208.
- Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of the ACM collective intelligence conference*, pages 12–24.
- Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 english words. *Behavior research methods*, 44:978–990.
- Kevin MacDonald. 1992. Warmth as a developmental construct: An evolutionary analysis. *Child development*, 63(4):753–773.
- Ivy W Maina, Tanisha D Belton, Sara Ginzberg, Ajit Singh, and Tiffani J Johnson. 2018. A decade of studying implicit racial/ethnic bias in healthcare providers using the implicit association test. *Social science & medicine*, 199:219–229.
- Saif Mohammad. 2018. [Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. Association for Computational Linguistics.
- Saif M. Mohammad. 2022. Ethics sheet for automatic emotion recognition and sentiment analysis. *Computational Linguistics*, 48(2):239–278.
- Saif M. Mohammad. 2023. Best practices in the creation and use of emotion lexicons. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, Dubrovnik, Croatia. Association for Computational Linguistics.
- Saif M. Mohammad. 2024. [WorryWords: Norms of anxiety association for over 44k English words](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16261–16278, Miami, Florida, USA. Association for Computational Linguistics.



- Saif M. Mohammad. 2025. [NRC VAD Lexicon v2: Norms for Valence, Arousal, and Dominance for over 55k English Terms](#). *arXiv preprint arXiv:2503.23547*.
- Agnes Moors, Jan De Houwer, Dirk Hermans, Sabine Wanmaker, Kevin Van Schie, Anne-Laura Van Harmelen, Maarten De Schryver, Jeffrey De Winne, and Marc Brysbaert. 2013. Norms of valence, arousal, dominance, and age of acquisition for 4,300 dutch words. *Behavior research methods*, 45(1):169–177.
- Robert Morabito, Sangmitra Madhusudan, Tyler McDonald, and Ali Emami. 2024. Stop! benchmarking large language models with sensitivity testing on offensive progressions. *arXiv preprint arXiv:2409.13843*.
- Gandalf Nicolas, Xuechunzi Bai, and Susan T Fiske. 2021. Comprehensive stereotype content dictionaries using a semi-automated method. *European Journal of Social Psychology*, 51(1):178–196.
- Brian A Nosek, Anthony G Greenwald, and Mahzarin R Banaji. 2005. Understanding and using the implicit association test: II. method variables and construct validity. *Personality and Social Psychology Bulletin*, 31(2):166–180.
- James W Pennebaker. 2011. The secret life of pronouns. *New Scientist*, 211(2828):42–45.
- Gina Roussos and Yarrow Dunham. 2016. [The development of stereotype content: The use of warmth and competence in assessing social groups](#). *Journal of Experimental Child Psychology*, 141:133–144.
- Javier Sánchez-Junquera, Berta Chulvi, Paolo Rosso, and Simone Paolo Ponzetto. 2021. How do you speak about immigrants? taxonomy and stereoisimmigrants dataset for identifying stereotypes about immigrants. *Applied Sciences*, 11(8):3610.
- Wolfgang S Schmeisser-Nieto, Alessandra Teresa Cignarella, Tom Bourgeade, Simona Frenda, Alejandro Ariza-Casabona, Mario Laurent, Paolo Giovanni Cicirelli, Andrea Marra, Giuseppe Corbelli, Farah Benamara, et al. 2024. Stereohoax: a multilingual corpus of racial hoaxes and social media reactions annotated for stereotypes. *Language Resources and Evaluation*, pages 1–39.
- Marie Gustafsson Sendén, Torun Lindholm, and Sverker Sikström. 2014. Biases in news media as reflected by personal pronouns in evaluative contexts. *Social Psychology*.
- Jillian K Swencionis, Cydney H Dupree, and Susan T Fiske. 2017. Warmth-competence tradeoffs in impression management across race and social-class divides. *Journal of Social Issues*, 73(1):175–191.
- Yi Chern Tan and L Elisa Celis. 2019. Assessing social and intersectional biases in contextualized word representations. *Advances in neural information processing systems*, 32.
- Daniela Teodorescu and Saif Mohammad. 2023. [Evaluating emotion arcs across languages: Bridging the global divide in sentiment analysis](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4124–4137, Singapore. Association for Computational Linguistics.
- Mike Thelwall. 2018. Gender bias in sentiment analysis. *Online Information Review*, 42(1):45–57.
- Kristen Swan Tummeltshammer, Rachel Wu, David M Sobel, and Natasha Z Kirkham. 2014. Infants track the reliability of potential informants. *Psychological science*, 25(9):1730–1738.
- Peter Turney. 2002. [Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 417–424, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Krishnapriya Vishnubhotla and Saif M. Mohammad. 2022. [Tweet Emotion Dynamics: Emotion word usage in tweets from US and Canada](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4162–4176, Marseille, France. European Language Resources Association.
- Melissa LH Võ, Markus Conrad, Lars Kuchinke, Karolina Urton, Markus J Hofmann, and Arthur M Jacobs. 2009. The berlin affective word list reloaded (bawl-r). *Behavior research methods*, 41(2):534–538.
- Jan Philip Wahle, Krishnapriya Vishnubhotla, Bela Gipp, and Saif M. Mohammad. 2025. Affect, body, cognition, demographics, and emotion: The abcd of text features for computational affective science. *arXiv*.
- Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, 45(4):1191–1207.
- Joseph P Weir. 2005. Quantifying test-retest reliability using the intraclass correlation coefficient and the sem. *The Journal of Strength & Conditioning Research*, 19(1):231–240.
- Bogdan Wojciszke, Andrea E Abele, and Wiesław Baryla. 2009. Two dimensions of interpersonal attitudes: Liking depends on communion, respect depends on agency. *European Journal of Social Psychology*, 39(6):973–990.
- Mi Zhou, Vibhanshu Abhishek, Timothy Derdenger, Jaymo Kim, and Kannan Srinivasan. 2024. Bias in Generative AI. *arXiv preprint arXiv:2403.02726*.

## A APPENDIX

### A.1 FAQ

Q1. When should one use the warmth score and when should one use trust and sociability scores?

Ans. Decades of social cognition research has converged on two primary dimensions: competence and warmth. Thus, for many applications it is useful to examine the warmth and competence dimensions (using corresponding lexicon entries). More nuanced analysis is enabled by splitting the warmth dimension into sub-categories. This is particularly appropriate when trust and sociability are expected to diverge: e.g., modern-day politicians are often seen as untrustworthy, yet sociable. One may also use a specific sub-dimension (trust or sociability) if that is the focus of the work: e.g., if one is interested in exploring trust in AI-generated text towards targets of interest, then the trust lexicon can be used.

Q2. What is the purpose of the popup feedback during the annotation process?

Ans. Annotation can be a tedious process. So it is unfortunate when one misunderstands some directions and spends time producing a large number of poor annotations. The popup feedback is there to let annotators know (as they are annotating) when they get certain instances wrong so that they can assess whether they have misunderstood something. This way they get immediate feedback. Secondly, it helps with quality control—people tend to refrain from clicking randomly when they know these checks exist.

### A.2 AMT Questionnaires

Screenshots of the trust and sociability detailed instructions, sample question, and examples presented to the annotators are shown in Figures 6 through 11. Participants were informed that they may work on as many instances as they wish.

### A.3 Distribution of Words of Warmth

Words of Warmth is made freely available on the project website as a compressed file. Terms of use will require that users not re-distribute the file and not post any form of the lexicon on the web. This is to prevent the resource being included in the data scrape fed to a large language model. See full list of terms of use at the project home page. Table 2 shows entries for a random sample of words from Words of Warmth.

Word	Sociability	Trust	Warmth
consoler	3.00	2.00	3.00
cohesiveness	3.00	2.18	3.00
wedding	2.88	2.22	2.88
blessed	2.83	2.27	2.83
conversant	2.75	0.89	2.75
folk	2.67	1.30	2.67
luckiest	2.57	0.27	2.57
ethicist	1.00	2.50	2.50
epidemiologist	-0.71	2.36	2.36
neuropsychologist	-0.62	2.27	2.27
sumptuously	2.14	0.55	2.14
dauntless	2.00	1.73	2.00
grief	2.00	0.20	2.00
sundeck	1.88	0.27	1.88
schoolbook	1.14	1.80	1.80
teetotal	0.00	1.73	1.73
equalization	0.50	1.64	1.64
irresistibility	1.57	0.55	1.57
teenage	1.50	0.00	1.50
bikini	1.38	0.00	1.38
navigation	1.25	0.50	1.25
cardamom	1.12	0.25	1.12
fertileness	1.00	0.00	1.00
gainful	0.86	0.77	0.86
climax	0.75	0.08	0.75
collectable	0.57	0.62	0.62
posthaste	0.00	0.50	0.50
enamelware	0.17	0.42	0.42
metaphoric	0.33	0.18	0.33
directionality	0.22	0.00	0.22
switchover	0.12	0.10	0.12
appendix	0.00	0.00	0
minuscule	-0.14	0.12	-0.14
bobber	0.12	-0.42	-0.42
miniaturization	-0.62	0.25	-0.62
misrecognition	-0.75	-0.80	-0.80
impel	-1.00	0.17	-1.00
unselect	-1.12	-0.67	-1.12
dodgers	-0.89	-1.20	-1.20
nonplussed	-1.29	0.08	-1.29
stifled	-1.38	-0.44	-1.38
impractical	-1.50	-0.55	-1.50
imperceptivity	-1.56	-0.56	-1.56
varicella	-1.62	-0.20	-1.62
prattler	-1.67	-0.92	-1.67
gentrify	-1.75	-1.50	-1.75
smoking	-1.75	-1.45	-1.75
rant	-1.86	-1.55	-1.86
defoliate	-1.88	-0.64	-1.88
notoriously	-1.89	-1.55	-1.89
debilitating	-2.00	-0.60	-2.00
paralyze	-2.00	-0.67	-2.00
pettiness	-2.00	-1.89	-2.00
rift	-2.00	-1.09	-2.00
bacteria	-2.11	-0.36	-2.11
detractor	-2.14	-1.75	-2.14
egocentrism	-2.25	-2.00	-2.25
illegitimate	-2.00	-2.30	-2.30
curse	-2.43	-1.75	-2.43
slovenliness	-2.38	-2.55	-2.55
inbreed	-2.67	-1.82	-2.67
horrible	-2.62	-2.78	-2.78
denigration	-2.88	-2.44	-2.88
stalker	-3.00	-2.67	-3.00
narcism	-2.43	-3.00	-3.00

Table 2: Randomly sampled terms and their anxiety-association score from Words of Warmth.

## Instructions

Summary

**Detailed Instructions**

Examples

### Introduction:

1. Attempt these questions only if you are fluent in English.
2. Your responses are confidential.

### Task:

Words can be associated with different degrees of trustworthiness or untrustworthiness. While there is some variation from person to person, there is also a fair amount of consensus. For example, most people will agree that the term:

- *mother* is often associated with perceptions of being **very trustworthy**
- *approved* is often associated with perceptions of being **moderately trustworthy**
- *nod* is often associated with perceptions of being **slightly trustworthy**
- *table* is often **not associated** with perceptions of being trustworthy or untrustworthy
- *unsure* is often associated with perceptions of being **slightly untrustworthy**
- *bullies* is often associated with perceptions of being **moderately untrustworthy**
- *fraudster* is often associated with perceptions of being **very untrustworthy**

In this multiple choice task, you will be given common English terms and you have to select the options that best describe the degree of trustworthiness or untrustworthiness associated with them.

Consider **trustworthiness** to be a broad category that includes:

- trustworthiness, honesty, fairness, dependability, reliability, morality, virtuousness, sincerity, honorableness, uprightness, equity, etc.

Consider **untrustworthiness** to be a broad category that includes:

- unfairness, dishonesty, untrustworthiness, dubiousness, immorality, sinfulness, insincerity, dishonorableness, crookedness, iniquity, etc.

This task is not about sentiment. For example, something can be positive and not relevant to trustworthiness (such as cheerful, sunny, and generous); and similarly, something can be negative and not relevant to trustworthiness (such as earthquake, miserly, and sad).

This task is not about intelligence or capability. For example, wily (which means skilled at gaining an advantage deceitfully) is associated with high capability but a fair amount of untrustworthiness.

Give answers that capture what most English speakers would agree.

If you do not know the meaning of a word or are unsure, you can look it up in a dictionary (e.g., the [Merriam Webster](#)) or on the internet.

### Purpose of the task:

Your responses will be used in a research study to better understand how trustworthiness and untrustworthiness manifest in language.

### Quality Control:

- Responses that are not in accordance with the instructions will not be paid for.
- Some questions have pre-determined correct answers. If you mark these questions incorrectly, we will often give you immediate feedback in a pop-up box. An occasional misanswer is okay. In addition, for some questions, we record the misanswers, but do not show a popup. If the **rate of misanswering is high (e.g., >20%), then \*\*all\*\* of one's HITs may be rejected.**
- If you see that you are getting quite a few of the gold questions wrong (e.g. more than 2 in every 10 HITs), then do not accept more HITs.
- If you disagree with the answer for a gold HIT, include the correct response in the Feedback textbox. Note that missing an occasional gold question will not lead to the rejection of your responses.
- This quality control measure promotes fairness for those who do the task responsibly.

### Notes:

- If a term has more than one meaning, consider the most common meaning.
- A rule of thumb is that a term associated with more trustworthiness tends to often occur in sentences that convey trustworthiness, whereas a term associated with more untrustworthiness tends to often occur in sentences that convey untrustworthiness.
- Try not to overthink the answer. **Let your instinct guide you.**

Figure 6: Trust Questionnaire: Detailed instructions.

## Summary Instructions

This task is about words and their association with trustworthiness/untrustworthiness.

Consider **trustworthiness** to be a broad category that includes:

- trustworthiness, honesty, fairness, dependability, reliability, morality, virtuousness, sincerity, honorableness, uprightness, equity, etc.

Consider **untrustworthiness** to be a broad category that includes:

- unfairness, dishonesty, untrustworthiness, dubiousness, immorality, sinfulness, insincerity, dishonorableness, crookedness, iniquity, etc.

This task is not about sentiment. For example, something can be positive and not relevant to trustworthiness (such as cheerful, sunny, and generous); and similarly, something can be negative and not relevant to trustworthiness (such as earthquake, miserly, and sad).

This task is not about intelligence or capability. For example, wily (which means skilled at gaining an advantage deceitfully) is associated with high capability but a fair amount of untrustworthiness.

If you do not know the meaning of a word or are unsure, you can look it up in a dictionary (e.g., the [Merriam Webster](#)) or on the internet.

A rule of thumb is that a term associated with more trustworthiness tends to often occur in sentences that convey trustworthiness, whereas a term associated with more untrustworthiness tends to often occur in sentences that convey untrustworthiness.

### Quality Control

- Some questions have pre-determined correct answers. If you mark these questions incorrectly, we will often give you immediate feedback in a pop-up box. An occasional misanswer is okay. In addition, for some questions, we record the misanswers, but do not show a popup. If the **rate of misanswering is high (e.g., >20%)**, then **\*\*all\*\* of one's HITs may be rejected**.

### Demographics

Provide your age, country, and gender in the first HIT (or at least one HIT) that you do. You can leave the text boxes blank in subsequent HITs. This information will be used to study trends of age, location, etc. with emotion.

Your Age (in years):   
Your Country (where you live):   
Gender (male, female, nonbinary, etc.):

Select the options that **\*\*most English speakers\*\*** will agree with.

Q1. *unaffectionate* is often associated with perceptions of being:

- ☐ 3: very trustworthy
- ☐ 2: moderately trustworthy
- ☐ 1: slightly trustworthy
- ☐ 0: not associated with perceptions of being trustworthy or untrustworthy
- ☐ -1: slightly untrustworthy
- ☐ -2: moderately untrustworthy
- ☐ -3: very untrustworthy

Feedback (optional):

Figure 7: Trust Questionnaire: Sample question.



### Introduction:

1. Attempt these questions only if you are fluent in English.
2. Your responses are confidential.

### Task:

Words can be associated with different degrees of social warmth or social coldness. While there is some variation from person to person, there is also a fair amount of consensus. For example, most people will agree that the term:

- *motherhood* is often associated with perceptions of being **very warm and sociable**
- *cooperate* is often associated with perceptions of being **moderately warm and sociable**
- *acquaintance* is often associated with perceptions of being **slightly warm and sociable**
- *clever* is often **not associated** with perceptions of being warm and sociable or cold and unsociable
- *irrelevant* is often associated with perceptions of being **slightly cold and unsociable**
- *argument* is often associated with perceptions of being **moderately cold and unsociable**
- *psychopath* is often associated with perceptions of being **very cold and unsociable**

In this multiple choice task, you will be given common English terms and you have to select the options that best describe the degree of social warmth or social coldness associated with them.

Consider **social warmth** to be a broad category that includes:

- warmth, sociableness, generosity, helpfulness, tolerance, understanding, thoughtfulness, etc.

Consider **social coldness** to be a broad category that includes:

- coldness, antisocialness, ungenerosity, unhelpfulness, intolerance, indifferent, thoughtlessness, etc.

This task is not about physical warmth/coldness (physical temperature, etc.).

This task is not about cleverness or competence. (For example, one can view someone as clever and cold; and someone else as clever and warm.)

This task is not about sentiment. (For example, something can be positive and cold (such as automated search engine systems) or positive and warm (such as friendship).)

Give answers that capture what most English speakers would agree.

If you do not know the meaning of a word or are unsure, you can look it up in a dictionary (e.g., the [Merriam Webster](#)) or on the internet.

### Purpose of the task:

Your responses will be used in a research study to better understand how social warmth and social coldness manifest in language.

### Quality Control:

- Responses that are not in accordance with the instructions will not be paid for.
- Some questions have pre-determined correct answers. If you mark these questions incorrectly, we will often give you immediate feedback in a pop-up box. An occasional misanswer is okay. In addition, for some questions, we record the misanswers, but do not show a popup. If the **rate of misanswering is high (e.g., >20%), then *\*\*all\*\** of one's HITs may be rejected.**
- If you see that you are getting quite a few of the gold questions wrong (e.g. more than 2 in every 10 HITs), then do not accept more HITs.
- If you disagree with the answer for a gold HIT, include the correct response in the Feedback textbox. Note that missing an occasional gold question will not lead to the rejection of your responses.
- This quality control measure promotes fairness for those who do the task responsibly.

### Notes:

- If a term has more than one meaning, consider the most common meaning.
- A rule of thumb is that a term associated with more social warmth tends to often occur in sentences that convey social warmth, whereas a term associated with more social coldness tends to often occur in sentences that convey social coldness.
- Try not to overthink the answer. **Let your instinct guide you.**

Figure 8: Sociability Questionnaire: Detailed instructions.

## Summary Instructions

This task is about words and their association with social warmth/coldness.  
Consider **social warmth** to be a broad category that includes:

- warmth, sociableness, generosity, helpfulness, tolerance, understanding, thoughtfulness, etc.

Consider **social coldness** to be a broad category that includes:

- coldness, antisocialness, ungenerosity, unhelpfulness, intolerance, indifferent, thoughtlessness, etc.

This task is not about physical warmth/coldness (physical temperature, etc.).

This task is not about cleverness or competence. (For example, one can view someone as clever and cold; and someone else as clever and warm.)

This task is not about sentiment. (For example, something can be positive and cold (such as automated search engine systems) or positive and warm (such as friendship).)

If you do not know the meaning of a word or are unsure, you can look it up in a dictionary (e.g., the [Merriam Webster](#)) or on the internet.

A rule of thumb is that a term associated with more social warmth tends to often occur in sentences that convey MH1>, whereas a term associated with more social coldness tends to often occur in sentences that convey social coldness.

### Quality Control

- Some questions have pre-determined correct answers. If you mark these questions incorrectly, we will often give you immediate feedback in a pop-up box. An occasional misanswer is okay. In addition, for some questions, we record the misanswers, but do not show a popup. If the **rate of misanswering is high (e.g., >20%), then \*\*all\*\* of one's HITs may be rejected.**

### Demographics

Provide your age, country, and gender in the first HIT (or at least one HIT) that you do. You can leave the text boxes blank in subsequent HITs. This information will be used to study trends of age, location, etc. with emotion.

Your Age (in years):   
Your Country (where you live):   
Gender (male, female, nonbinary, etc.):

Select the options that **\*\*most English speakers\*\*** will agree with.

Q1. *capitalist* is often associated with perceptions of being:

- ☐ 3: very warm and sociable
- ☐ 2: moderately warm and sociable
- ☐ 1: slightly warm and sociable
- ☐ 0: not associated with perceptions of being warm and sociable or cold and unsociable
- ☐ -1: slightly cold and unsociable
- ☐ -2: moderately cold and unsociable
- ☐ -3: very cold and unsociable

Feedback (optional):

Figure 9: Sociability Questionnaire: Sample question.

Very trustworthy:

- justice, sincere, noble, respected, tolerant, parent, doctor, referee, inspire, devoted

Moderately trustworthy:

- good plan, approving, commendation, head coach, alarm clock,

Slightly trustworthy:

- good attendance, neighbor, junior employee, quote in an article,

Not associated with trustworthiness or untrustworthiness:

- table, envelope, cheerful, sunny, generous, earthquake, miserly, sad, utencil, tree, paint, garage,

Slightly untrustworthy:

- lapse, breaking a minor rule, naughty, mouth off, slight exaggeration

Moderately untrustworthy:

- bullied, favouritism, misconduct, mistreat, used car salesperson, suspicious, flimsy, misrepresent

Very untrustworthy:

- murder, corrupt, biased, xenophobic, fraud, vice, bigotted, cheat

Figure 10: Trust Questionnaire: Examples.

Very warm and sociable:

- helpful, loved, mother, nurse, family, adore

Moderately warm and sociable:

- beach, greeting card, neighbor, treatment, cooperate, gathering, encourage, volunteering, invite, popular

Slightly warm and sociable:

- nod, dinner, town, acquaintance, work meetinig, follow, university

Not associated with social warmth or social coldness:

- clever, table, envelope, utencil, tree, calendar, paint, garage

Slightly cold and unsociable:

- irrelevant, computer, automated, absent, paper work, order, night, shadow

Moderately cold and unsociable:

- moody, irritable, argument, fight, ignore, unpopular, vain

Very cold and unsociable:

- psychopath, murder, indifferent, loner, jail, hated

Figure 11: Sociability Questionnaire: Examples.

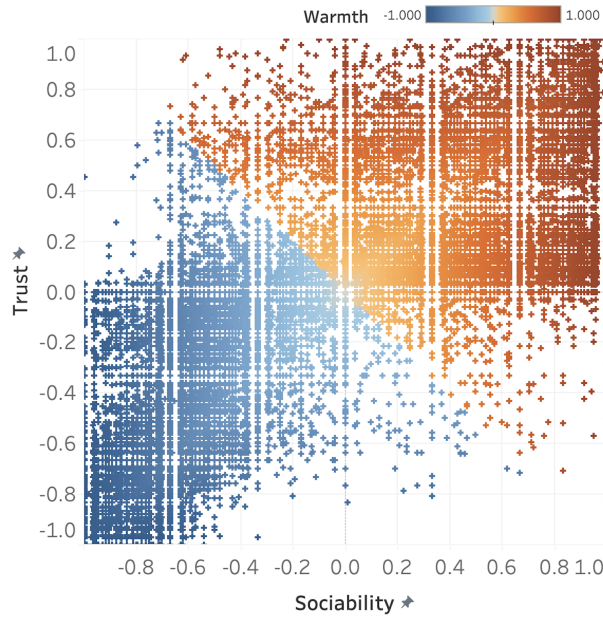


Figure 12: Scatterplot of words on the trust-sociability space. Individual points are colored as per their W score.

#### A.4 Case Study: Professions

Figure 15 shows W-C plots for various professions. The direct lexicon lookup of the targets (Fig 15 (a)) shown that people perceive engineers, doctors, and teachers to be high competence, whereas nurses and teachers are considered very warm. In contrast, being jobless is perceived as cold and incompetent. The coterms plot (b) shows that mentions of CEO get an even higher competence score than engineer and teacher, and in fact the mentions of doctor get a lower competence score than nurse.<sup>9</sup> This indicates that even though doctors are considered as competent, their mentions in social media are more in contexts where one is expressing a lack of competence/power/situational control (e.g., not having access to a doctor).

#### A.5 Supplementary Figures and Tables

Figure 12 shows a scatter plot of words on the T-S space, colour coded as per their W score (described in Section 5). Figure 13 shows a break down of the percentage of terms in each of the warmth classes into the percentage of entries obtained from the trust lexicon and the percentage of entries obtained from the sociability lexicon.

Figure 14 (a) shows the W-C plot for various pronouns. Figure 14 (b) is the same plot except the pronouns are plotted separately for every year. (These plots were described in Section 7.)

<sup>9</sup>The term CEO is not in the WC lexicons.

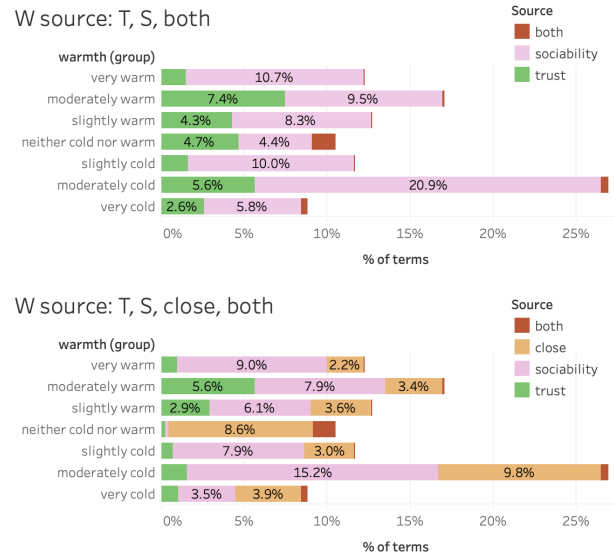


Figure 13: Stacked bar charts showing a break down of the percentage of terms in each of the warmth classes into the percentage of entries obtained from the trust lexicon and the percentage of entries obtained from the sociability lexicon. (a) shows a break down into 3 classes: percentage of terms for which the W score is the same as the T score and different from the S score ( $\text{abs}(\text{T score}) > \text{abs}(\text{S score})$ ), percentage of terms for which the W score is the same as the S score and different from the T score ( $\text{abs}(\text{S score}) > \text{abs}(\text{T score})$ ), and percentage of terms for which the W score is the same as both the T and S scores ( $\text{abs}(\text{T score}) = \text{abs}(\text{S score})$ ). (b) is similar to (a) except that a fourth class, *close*, is added to show the cases where the S and T scores are close to each other ( $\text{T score} - \text{S score} < 0.5$ ).

#### A.6 Computational Resources and Carbon Footprint

A nice advantage of using simple lexicon-based approaches is the low carbon footprint and computational resources required. All of the experiments described in the paper were conducted on a regular personal laptop.



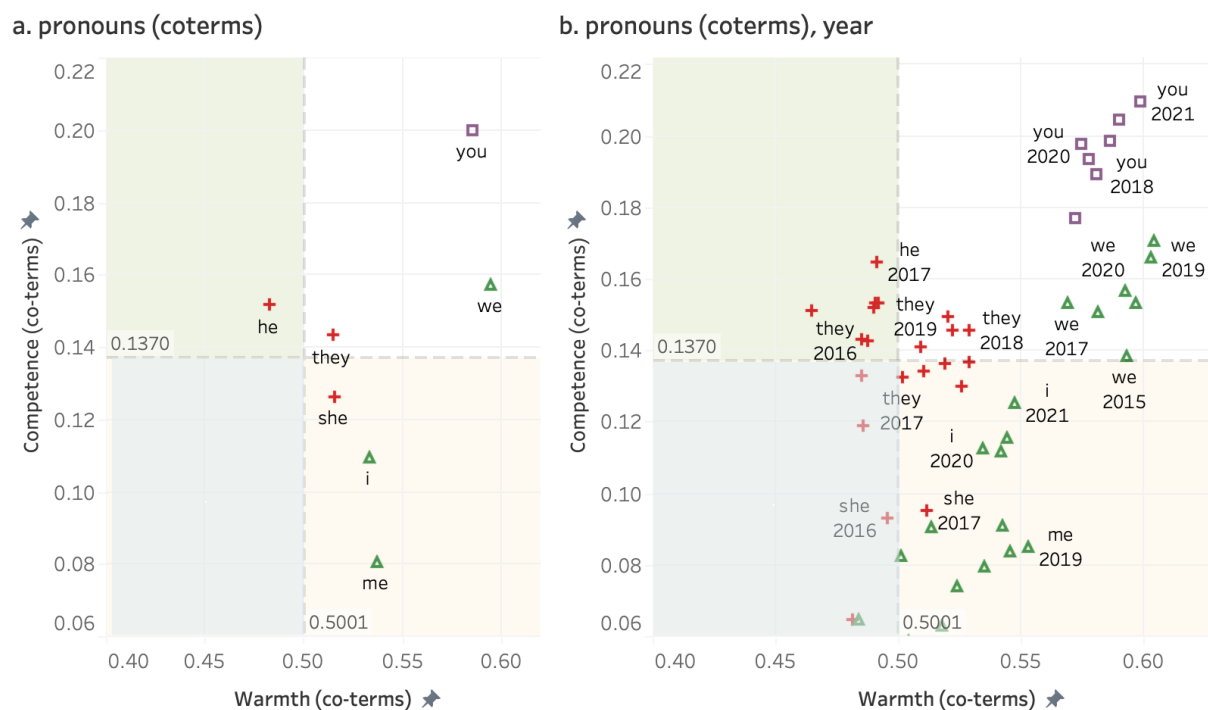


Figure 14: Direct and co-term W-C plots for various pronouns.

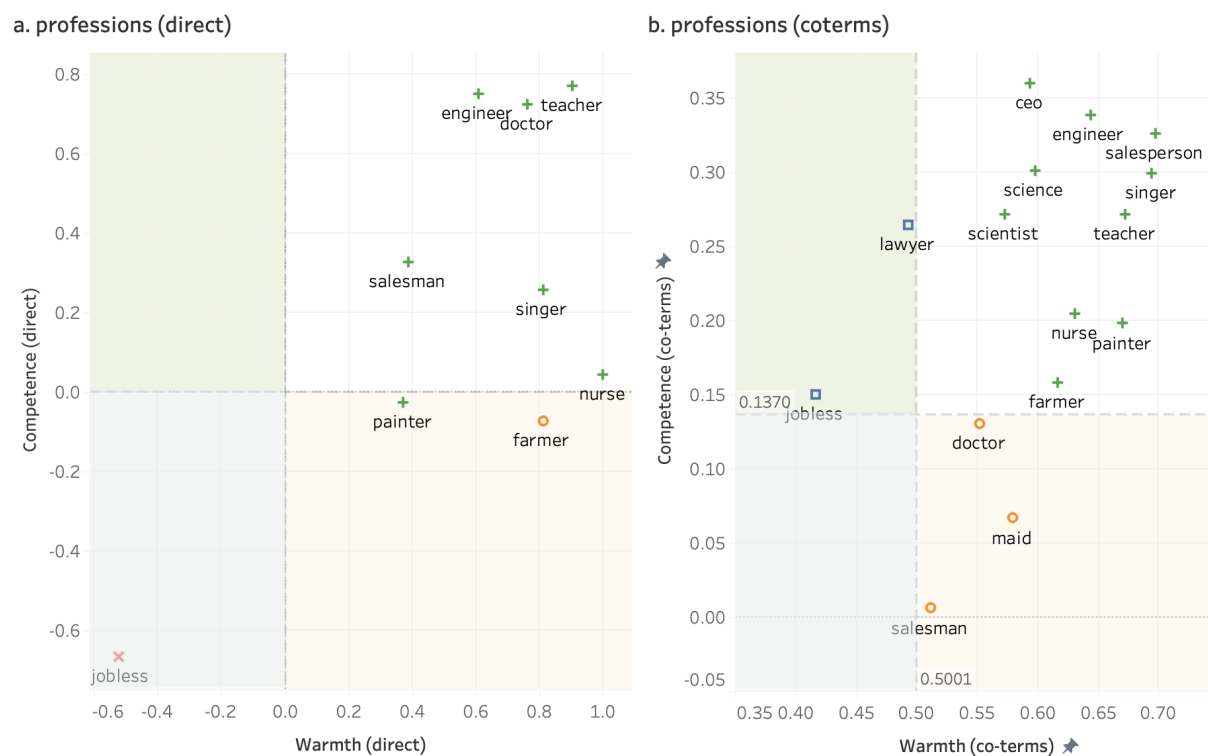


Figure 15: Direct and co-term W-C plots for various professions.