The Supervised Semantic Differential (SSD): learning semantic directions aligned with psychological scales

Plisiecki, H., Lenartowicz, P., Pokropek, A., Flakus, M.

Motivation

Psychological text data are often too small for end-to-end fine-tuning, yet contain semantically rich latent constructs (e.g., trust, identity, ideology) that are not lexically explicit.

Traditional dictionary or feature-based approaches (e.g., LIWC) impose predefined semantic categories, while large LMs can encode these dimensions implicitly - but their directions are entangled and opaque.

Our objective is to develop a transparent and data-efficient method that:

- Learns a single interpretable vector direction in embedding space that best explains variance in a target psychological variable,
- Works with frozen pre-trained embeddings (no fine-tuning required),
- Produces interpretable semantic axes analogous to classical affective dimensions (Osgood's evaluation–potency–activity).

Method

Step 1: Base embedding Space

Use 300-d Word2Vec (NKJP + Wikipedia), L2-normalize all word vectors and apply ABTT (m=1) at the model level to improve isotropy.

Step 2: Document representation

For essay i, average SIF-weighted context vectors around seed occurrences (±3 tokens), then L2-normalize: $\mathbf{d}_i = \text{normalize}\left(\frac{1}{|O_i|}\sum_{o\in O_i}\frac{\sum_{t\in N(o)}\alpha_t\,\mathbf{x}_t}{\sum_{t\in N(o)}\alpha_t}\right)$, with SIF weights: $\alpha_t = \frac{a}{a+p(t)}$.

Step 3: Dimensionality reduction

Apply PCA to the document matrix to limit multidimensionality for OLS.

Step 4: Regression & back-mapping

Fit OLS in PCA space ($y = Zw$), back-project to embedding space and unit-normalize the gradient:

$$\widehat{\beta} = \text{normalize}\left((\,\mathbf{C}\,\mathbf{w}\,)\odot\frac{1}{\boldsymbol{\sigma}}\right)$$

where C are PCA loadings and $\boldsymbol{\sigma}$ are pre-PCA feature scales.

Step 5: Interpretation

Rank base model's vocabulary by cosine to $\widehat{\beta}$ (and $-\widehat{\beta}$), excluding high-frequency noise, numerals, and proper names, to read the positive / negative semantic poles.

To aid interpretation beyond single words, top-100 neighbors were grouped into clusters using k-means ($k=4$), each each represented by its centroid's alignment with the gradient and internal coherence.

Finally, for each cluster, we identified sentences in which semantically related words appeared, extracting short context snippets (the sentence itself and its immediate neighbors) to illustrate how these semantic directions manifest in actual text. (not shown here)

Example Datasets

Example Open Ended Question Format:

"For the next 5 minutes, please write about everything that comes to your mind when you think about Poland and the Polish people, as well as other people and countries that surround us. Please reflect on what your feelings and impressions are when you turn your attention to this topic.Write down these thoughts as they come to you and follow them wherever your mind naturally takes you."

Dataset 1. Collective Narcissism
Scale: Collective Narcissism (0 to 30)
Open-ended question: above
Size: 1320 responses
Mean length: 30 words
SEED: country, state, homeland, nation (country), nation (people)

Dataset 2. Climate Change
Scale: Readiness to counteract Climate change (1-5 Likert)
Open-ended question: similar to above, but about climate change
Size: 665 responses
Mean length: 49 words
SEED: change, climate (noun), relating to climate (adj)

Results

Collective Narcissism

The gradient explained a modest but reliable share of variance, R² = 0.071, F = 3.47, p < .001, r = .267 (N = 926 kept; 394 dropped). The slope magnitude was ||β|| = 2.10 SD per +1.0 cosine, equivalent to +0.89 points per +0.10 cosine; the IQR effect was 1.50 points.

Examples of sentences most aligned with the beta:
"Poland is a wonderful nation with wonderful people, which is why I wouldn't want to mix cultures by letting in all immigrants."
"We are a nation of wonderful people and great values."

Examples of sentences least aligned with the beta:
"Poland is definitely a country closed off to differences such as sexual orientations other than heterosexual, which I consider a sign of being backward."
"I have a neutral opinion about other people and countries, but I also see that other governments can make bad decisions for their citizens."

Positive Clusters

size	Centroid cos beta	coherence	top	Interpretation
27	0.55	0.41	long-term, social activist, association, foundation	Community & Legacy
27	0.51	0.48	enormous, huge, powerful, gigantic	Grandeur & Prosperity
28	0.43	0.53	glory, bless, beloved, benefactor	Sacralized Patriotism
18	0.39	0.6	to want, to decide, to intend, to resolve	Agency & Determination

Negative Clusters

size	Centroid cos beta	coherence	top	Intepretation
18	-0.54	0.67	to articulate, to specify, to clarify, to distinguish	Analytical Elaboration
25	-0.58	0.62	generalization, interpretation, to evaluate, formalism	Philosophical Reflection
24	-0.6	0.61	semantic, to blur, vague, ambiguous, to generalize	Ambiguity & Contrast
33	-0.62	0.59	formulation, phrasing, definition, connotation	Meta-Linguistic Commentary

Climate Change

The gradient predicted scores with R² = 0.095, F = 2.85, p < .001, r = .308 (N = 565 kept; 90 dropped). The slope was ||β|| = 2.61 SD per +1.0 cosine, ≈ +0.28 points per +0.10 cosine; the IQR effect was 0.46 points.

Examples of sentences most aligned with the beta:
"Poland is a wonderful nation with wonderful people, which is why I wouldn't want to mix cultures by letting in all immigrants."
"We are a nation of wonderful people and great values."

Examples of sentences least aligned with the beta:
"Poland is definitely a country closed off to differences such as sexual orientations other than heterosexual, which I consider a sign of being backward."
"I have a neutral opinion about other people and countries, but I also see that other governments can make bad decisions for their citizens."

Positive Clusters

size	Centroid cos beta	coherence	top	Interpretation
46	0.49	0.57	fatigue, heat, stress, pain, nervousness	Somatic Discomfort
18	0.49	0.56	difficulty, inconvenience, hardship, frugality	Coping & Practical Constraints
19	0.49	0.54	yard, courtyard, balcony, back	Everyday Physical Surroundings
17	0.47	0.59	motivation, satisfaction, anxiety, appetite	Emotional Regulation & Self-Perception

Negative Clusters

size	Centroid cos beta	coherence	top	Intepretation
7	-0.42	0.74	to recreate, to reconstruct, to rebuild	Reconstruction & Revision
30	-0.51	0.59	to characterize, to describe, to present	Scientific Description & Definition
22	-0.52	0.58	descriptive, semantic, abstract, multidimensional	Methodological Formalization
41	-0.58	0.54	contemporary, timeless, archetype, postmodern	Temporal & Cultural Framing