A Phenomenologically-Inspired Computational Analysis of Self-Categories in Text

Jaya Caporusso[1,2], Boshko Koloski[1,2], Maša Rebernik[3], Senja Pollak[1], Matthew Purver[1,4]

[1]Jožef Stefan Institute, Ljubljana, Slovenia – jaya.caporusso96@gmail.com, {boshko.koloski, senja.pollak}@ijs.si

[2]Jožef Stefan International Postgraduate School, Jamova cesta 39, 1000 Ljubljana, Slovenia

[3]Institute for Developmental and Strategic Analysis, Ljubljana, Slovenia

[4]Queen Mary University of London, UK

Abstract

The self is a pervasive aspect of human experience, influencing crucial areas like mental health and manifesting in the texts we produce. Previous research indicates a significant correlation between the use of self-related expressions—terms and linguistic structures individuals use to refer to themselves, such as first-person pronouns—and various personal attributes, including personality traits, mental states, and psychological disorders. These findings enable the construction of simple yet explainable and effective representations, which can be later utilised for downstream tasks like classification, clustering, and segmentation. We present an approach to investigate the self in text data in a more detailed manner, expanding its understanding by adopting aspects of the self as defined by cognitive science and phenomenology. We employ the large language model GPT3.5 to classify text as to whether it presents these self-aspects, and we analyse the obtained splits with LIWC-22. This exploratory study aims to bridge the gap between the knowledge about using self-references in text, Natural Language Processing techniques and applications, and the phenomenological understanding(s) of the self, opening new venues in all three directions.

Keywords: self, statistical analysis, phenomenology, classification, large language models

1. Introduction

Every secret of a writer's soul, every experience of his life; every quality of his mind is written large in his works.

--- Virginia Woolf, Orlando

The *sense* of self is understood as "the (perhaps sometimes elusive) feeling of being the particular person one is" (Siderits et al., 2011). However, what the self per se is, or whether it even exists, has been a topic of discussion in the fields of philosophy and cognitive science for a long time, and different definitions have been developed (see Siderits et al., 2011). These different understandings of the self (some of which are further addressed in Section 3.1) are not necessarily mutually exclusive and are often seen as different aspects of what constitutes the self. Previous studies have shown that in daily life all of these different self-elements present themselves together, in a coherent way, while in specific experiences, such as dissolution experiences ("experiential episodes during which the perceived boundaries between self and world (i.e., non self) become fainter or less clear [and] a sense of unity with the world or elements of it [is felt]"; Caporusso, 2022), different parts of the self gradually lose intensity and clarity and/or disappear, leaving only the so-called *minimal self* present, or, arguably, as last to go (Caporusso, 2022). As these aspects interact with critical subjects of research interest such as mental health (e.g., Parnas and Henriksen, 2014), it is particularly

important to foster an interdisciplinary investigation of the self. This can be done through the investigation of language.

Indeed, how we shape our language reflects not only the situation we are in but also who we are. This is true when we choose the topics to address, but, most importantly, it is true when we choose the style in which to cover that content. For example, researchers found differences in our linguistic style correlated with age, gender, personality type, mood, mental health, and even physical health state. Specifically, the stylistic differences can be found in the use of words with positive vs negative valence, the tense of verbs, and the use of intensifiers (for an overview, see Pennebaker et al., 2003). Interestingly, the self-related expression (i.e., how we talk about or address ourselves) that appears often in this kind of analysis is the use of first-person singular pronouns (i.e., *I* and *me*). We believe that the definition of self that such analysis can capture is limiting. Therefore, we consider definitions of self provided by cognitive science and phenomenology, and we develop a framework to investigate which textual aspects (specifically, the categories provided by LIWC-22, Boyd et al., 2022) correspond to the presence (or absence) of each of the selected self-aspects. We do so by employing a dataset of Reddit posts on various topics, manual annotation, annotation by a GPT model, and LIWC-22.

In Section 2 we present related work, in Section 3 our method, and in Section 4 our results.

2. Related work

Following, we present studies that address the correlation between self-expressions—in particular, *I-talk*—and the traits and states of the individual producing the analysed language. These studies often utilise the Linguistic Inquiry and Word Count (LIWC), a text analysis software developed to analyse linguistic and psychosocial constructs connected to various textual aspects (Boyd et al., 2022). LIWC has been applied to one of the most widely studied theories of personality, the Big Five (Goldberg, 1990), which comprises a set of five traits. Various studies have found them to be associated with linguistic features: e.g., neuroticism with first-person singular pronouns (e.g., Yarkoni, 2010). Furthermore, multiple studies have found that the use of the first-person singular is associated with depression. For instance, it was found that college-aged individuals experiencing depression tend to use more first-person singular pronouns when writing about their college experiences (Rude et al., 2004), and during natural speech captured over several days (Mehl and Pennebaker, 2003), compared to non-depressed individuals. This has been employed in classification tasks (e.g., Caporusso et al., 2023).

3. Method

Our study includes three main parts: a) manual annotation of data, b) classification with GPT3.5, and c) statistical text analysis.

3.1. Dataset

For our study, we need a dataset of text data that is not focused on a specific topic or context. We use a dataset by Völske (2017) with the purpose of automatic summarisation and constructed by scraping various subreddits.[1] For our study, we sample the first 1,000 rows of the dataset. On average, a document is 1120 characters long, consisting of 235.12 words

[1] https://huggingface.co/datasets/webis/tldr-17

situated in 11 sentences. Figure 1 displays the top-10 subreddits present in the dataset so-obtained.

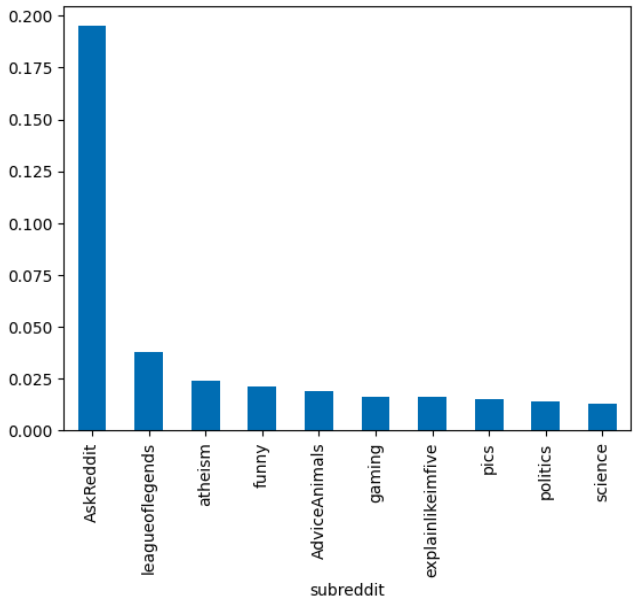


Figure 1: Top-10 subreddits present in the dataset.

3.2. Manual annotation

We are interested in classifying text instances based on five self-categories: *Minimal Self* (MS), *Narrative Self* (NS), *Self as Agent* (AS), *Bodily Self* (BS), and *Social Self* (SS). In our framework, each of the self-categories can either be present or not present in a specific text instance. We first construct the instructions for annotation, where the five self-categories of interest are clearly defined and accompanied by positive and negative examples. This is done collaboratively by two of the authors, who independently annotate more than 25 text instances and then discuss the choices made, coming to an agreement. Following, for each category we provide its definition and one example for each condition.

3.2.1. Minimal self.

Also referred to as core self, it refers to the aspects of mineness or for-me-ness of experience—that is, the fact that experiences are presented to us in a fundamentally personal and subjective way.

Example of Reddit post with MS present: "*I think it should be fixed on either UTC standard or UTC+1 year around, with the current zone offsets. Moving timescales add a lot of complexity to the implementation of timekeeping systems and have [dubious value](I think seasonal shifting time made sense in the pre-electric past, when timekeeping was more flexible and artificial light was inefficient and often dangerous. (...)*"

Example of Reddit post with MS not present: "*This picture doesn't follow too well, as defining characteristics of major characters are left out in both sections.*"

3.2.2. Narrative self.

The narrative someone has of themselves, comprising their autobiographical memories and stories of who they are. In the words of Alasdair Macintyre (1985), the narrative self represents "the unity of a narrative which links birth to life to death as narrative beginning to

middle to end".

Example of Reddit post with NS present: "*About two months ago I had a very vivid and detailed dream about Extraterrestrials invading earth. I believe there was also some kind of plane crash involved - perhaps having to do with the ETs disabling all Earth-en electronics? Anyway, in the dream as the spaceships were descending, I felt this overwhelming terror and realization that my life would never be the same, and that I would very likely die soon. Woke up really scared from that one.*"

Example of Reddit post with NS not present: "*If the number of sides of any circle => 4, then yes.*"

3.2.3. Self as agent.

The experience of being an agent, i.e., in control, active.

Example of Reddit post with AS present: *"So you're saying "try it, I might not mind losing access to directions that follow my only available mode of transportation (public)"? This isn't a it might be ok but some people don't like it issue like Siri not listening to you well . This is removing an entire function that I use all the time. It's not worth it and I won't be upgrading. (...)*"

Example of Reddit post with AS not present: "*Didn't they lose 6 games in a row? Just because you're close for some of the games doesn't mean that you're not a lot weaker than that team. I love Arkansas razorbacks football. 2 years (i think) we lost to Alabama and LSU by a field goal that we missed from less than 40 yards. This year we lost to Alabam 52-0. Our team isnt' young and we we're ranked 8. Our quarterback did get injured against UL monroe, but that doesn't make up 49 points.*"

3.2.4. Bodily self.

The experience of owning, controlling, and/or identifying with someone's own body (or parts of it)—that is to say, the "distinctive ways" in which we are aware of our own body (Bermudez, 2018).

Example of Reddit post with BS present: *"All but one of my nails were in the ballpark of 1 1/8" - 1 1/2" long when my ring finger nail broke to the quick on Monday! It was the second break on the same hand in about a month, so I finally had to get compulsive and make all the nails the same length! (...)*"

Example of Reddit post with BS not present: "*Art is about the hardest thing to categorize in terms of good and bad. To consider one work or artist as dominate over another comes down to personal opinion. Sure some things maybe blatantly better than other works, but it ultimately lies with the individual. (...)*"

3.2.5. Social self.

The self as it is shaped and/or perceived when in an interaction or relationship of sorts with other people or entities to whom we attribute qualities of an inner life.

Example of Reddit post with SS present: *"I don't know what kind of relationships you've been in, but I've never suffered or lost friends because of a relationship. My girlfriends have been nice, friendly people that my friends loved. If we had a problem, we'd discuss it.*"

Example of Reddit post with SS not present: "*This picture doesn't follow too well, as defining characteristics of major characters are left out in both sections.*"

3.2. Classification with LLMs

Large language models (LLMs) like GPT-3.5 are pre-trained to model the language on a big collection of documents consisting of trillions of tokens via causal language modeling (where part of a document is written and the model is asked to complete it). This strategy represents a self-supervised learning methodology that enables learning from a vast range of language resources via minimal human intervention. However, one should note that models pre-trained

on vast and largely uncontrolled data are prone to epistemological difficulties, like inherent model biases (see Beck et al., 2024). To tune LLMs for instruction following, an instruction-tuning task is introduced where models are prompted with instructions and are asked to complete the document by providing the correct answer. This enables models to be efficient instruction followers and to learn from in-context examples (see Brown et al., 2020).

In order to assess the in-context learning performance of LLMs for automatic annotation of larger data collection, we evaluate their performance on the batch of examples annotated by experts. Following, we leverage the in-context learning capabilities of auto-regressive GPT models (in our case GPT3.5-turbo-instruct), coupled with prompt engineering, to perform one-shot, two-way classification of documents (where each document is assigned one label at a time as present / not present) within our corpora (see Brown et al., 2020). We structure the prompt in three distinct sections:

- *Personalisation*: the model is given a personalisation (Koloski et al., 2024; Beck et al., 2024), with which we condition the classification as we guide the model to provide answers from the perspective of the given persona. We employ two distinct personalisations for annotations to enrich our automatic annotation process: an expert in *sociology* and an expert in *phenomenology*.
- *Definition*: we provide clear definitions for each label as defined by the expert annotators, along with detailed annotation instructions to ensure precise classification.
- *Exemplars*: we include a single positive and negative example for each label, accompanied by expert explanations.

We first check the models' performance compared to the experts' annotations, by prompting the model for each of the five categories. The two models show the following agreement rates with the annotators: *MS*: 87.2%; *NS*: 92.4%; *SA:* 86.0%; *BS*: 93.0%; *SS*: 92.9%. We then prompt the two models for each category on the data from Section 3.1. We only consider data points where both models agree on the classifications to build the datasets for further analysis. We further split the documents in groups, where each group represents a combination of category and label (e.g. present AS, not-present AS). Table 1 shows the statistics of each dataset.

Table 1: Datasets statistics

	not present			**present**		
self-category	*avg. len.*	*avg. words*	*avg sent.*	*avg. len.*	*avg. words*	*avg sent.*
agent	1213.17	256.72	12.65	886.48	186.92	8.88
bodily	1144.38	240.55	11.63	869.34	189.72	9.80
minimal	1319.25	280.16	13.86	1016.34	214.60	10.32
narrative	935.36	193.47	9.10	1154.85	246.26	12.05
social	815.44	169.14	8.12	1239.53	263.37	12.86

3.1. LIWC-22 analysis

Due to the exploratory nature of our study, we opt to include all the LIWC-22 categories and subcategories in our analysis. For each of these, LIWC provides scores relative to the text length.

3.2. Statistical text analysis

After obtaining the scores from LIWC-22, we check for normality for each LIWC category in each dataset. When both datasets of the same pair (e.g., MS present and MS not present) present a normal distribution, we use the paired t-test to check whether the two datasets are statistically different with regard to that LIWC category; otherwise, we use the Wilcoxon signed-rank test. While the latter could always be used, a t-test is preferable when a normal distribution is present, due to its higher statistical efficiency.

4. Results

In this section, we report on a selection of LIWC categories that are significantly ($p < 0.05$) more present in each dataset. For each of them, we provide the median difference[2] between the present and not-present condition (D*med*) and the pooled standard deviation (SD) in the footnotes. A description of the LIWC categories is provided by Boyd et al. (2022).

4.1. Minimal Self[3]

Between the categories significantly more present when the MS is **present** compared to when it is not, these are the ten with the highest median differences (see Figure 2): *Authentic*, *Cognitive processes*, *Linguistic*, *Affect* (D*med*: 1.03, SD: 3.06), *"I"*, *Cognition*, *Positive tone*, *Insight*, *Emotion*, and *Social behaviour*. Furthermore, interestingly also the following categories show statistical significance in this condition: *Negative tone*, *Adjectives*, *Certitude*, *Positive emotion*. Between the categories significantly more present when the MS is **not present** compared to when it is, these are the ten with the lowest median differences: *Word count*[4], *All punctuation*, *Other punctuation*, *Numbers*, *Negative Emotion*, *Anger*, *Technology*, *Leisure*, *Money*, *Fatigue*, and *Assent*.

4.2. Narrative Self[5]

Between the categories significantly more present when the NS is **present** compared to when it is not, these are the ten with the highest D*med*s (see Figure 3): *Word Count* (we excluded

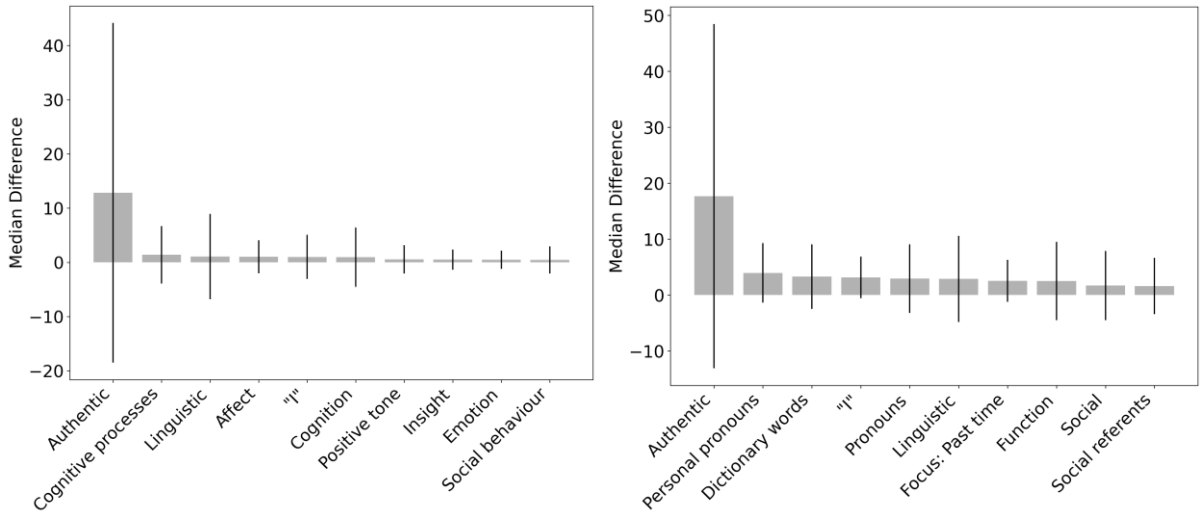
[2]The median difference tends to be closer to +1 when the variable of interest is present and to -1 when it is not.

[3]**Minimal self present.** *Authentic*: D*med*: 12.83, SD: 31.34; *Cognitive processes*: D*med*: 1.40, SD: 5.30; *Linguistic dimensions*: D*med*: 1.06, SD: 7.87; *Affect*: D*med*: 1.03, SD: 3.06; *"I"*: D*med*: 1.00, SD: 4.07; *Cognition*: D*med*: 0.95, SD: 5.48; *Positive tone*: D*med*: 0.54, SD: 2.60; *Insight*: D*med*: 0.49, SD: 1.85; *Emotion*: D*med*: 0.47, SD: 1.68; *Social behaviour*: D*med*: 0.44, SD: 2.49; *Negative tone*: D*med*: 0.41, SD: 1.71; *Adjectives*: D*med*: 0.29, SD: 4.41; *Certitude*: D*med*: 0.22, SD: 1.68; *Positive emotion*: D*med*: 0.2, SD: 1.29. **Minimal self not present.** *Word count*: D*med*: -47.00, SD: 205.54; *All punctuation*: D*med*: -1.84, SD: 10.46; *Other punctuation*: Dmed: -1.10, SD: 4.25; *Numbers*: D*med*: -0.15, SD: 2.76; *Negative Emotion*: D*med*: 0.00, SD: 0.95; *Anger*: D*med*: 0.00, SD: 0.60; *Technology*: D*med*: 0.00, SD: 1.19; *Leisure*: D*med*: 0.00, SD: 1.30; *Money*: D*med*: 0.00, SD: 2.06; *Fatigue*: D*med*: 0.00, SD: 0.19; and *Assent*: D*med*: 0.00, SD: 0.57.

[4]Although LIWC scores are relative to the text length, this finding highlights systematic variations in the total number of words across the compared conditions.

[5]**Narrative self present.** *Word Count*: D*med*: 62.00, SD: 167.63; *Authentic*: D*med*: 17.67, SD: 30.77; *Personal pronouns*: D*med*: 3.95, SD: 5.32; *Dictionary words*: D*med*: 3.28, SD: 5.77; *"I"*: D*med*: 3.17, SD: 3.72; *Pronouns*: D*med*: 2.93, SD: 6.15; *Linguistic dimensions*: D*med*: 2.88, SD: 7.71; *Focus: Past time*: D*med*: 2.53, SD: 3.77; *Function words*: D*med*: 2.51, SD: 7.02; *Social*: D*med*: 1.69, SD: 6.20; *Social referents*: D*med*: 1.62, SD: 5.04; *Affect*: D*med*: 1.34, SD: 3.16; *Physical:* D*med*: 1.27, SD: 2.33; *Affiliation*: D*med*: 0.95, SD: 2.24; *Emotion*: D*med*: 0.86, SD: 1.72; *Time*: D*med*: 0.86, SD: 2.96; *Negative tone*: D*med*: 0.73, SD: 1.78; *Drives*: D*med*: 0.70, SD: 3.34; *Positive tone*: D*med*: 0.54, SD: 2.52; *Adverbs*: D*med*: 0.54, SD: 3.20; *Conjunctions*: D*med*: 0.46, SD: 2.92; *Positive emotion*: D*med*: 0.43, SD: 1.32; *"She/He"*: D*med*: 0.37, SD: 2.82; and *Negative emotion*: D*med*: 0.33, SD: 1.00. **Narrative self not present**. *Analytic*: D*med*: -13.18, SD: 26.79; *Clout*: D*med*: -9.02, SD: 31.31; *Big words*: D*med*: -2.00, SD: 6.33; *Cognitive processes*: D*med*: -1.76, SD: 5.58; *Cognition*: D*med*: -1.64, SD: 5.79; *Focus: Present time*; D*med*: -1.20, SD: 3.35; *Other pronouns:* D*med*: -1.01, SD: 4.23; *"You"*: D*med*: -0.94, SD: 3.32; *All punctuation:* D*med*: -0.82, SD: 10.08; *Differentiation*: D*med*: -0.66, SD: 2.66; *Tentative*: D*med*: -0.65, SD: 2.41; *"We"*: D*med*: 0.00, SD: 1.65; *Anxiety:* D*med*: 0.00, SD: 0.33; *Anger:* D*med*: 0.00, SD: 0.54; *Sadness*: D*med*: 0.00, SD: 0.28; *Swear words*: D*med*: 0.00, SD: 0.86; *Moralisation:* D*med*: 0.00, SD: 0.70; *Culture*: D*med*: 0.0, SD: 1.71; *Health*: D*med*: 0.00, SD: 1.26; *Mental health*: D*med*: 0.00, SD: 0.32; *Auditory*: D*med*: 0.00, SD: 0.67; *Feeling*: D*med*: 0.00, SD: 0.69; and *Prosocial behaviour*: D*med*: 0.13, SD: 0.86.

this category from Figure 1 due to the different proportions with the other categories), *Authentic*, *Personal pronouns*, *Dictionary words*, *"I"*, *Pronouns*, *Linguistic dimensions*, *Focus: Past time*, *Function words*, and *Social*. Furthermore, interestingly also the following categories show statistical significance in this condition: *Social referents* (we added this category to Figure 1 as well), *Affect*, *Physical*, *Affiliation*, *Emotion*, *Time*, *Negative tone*, *Drives*, *Positive tone*, *Adverbs, Conjunctions*, *Positive emotion*, *"She/He"*, and *Negative emotion*. Between the categories significantly more present when the NS is **not present** compared to when it is, these are the ten with the lowest D*med*s: *Analytic*, *Clout*, *Big words*, *Cognitive processes*, *Cognition*, *Focus: Present time*, *Other pronouns*, *"You"*, *All punctuation*, *Differentiation*, and *Tentative*. Furthermore, interestingly also the following categories show statistical significance in this condition: *"We"*, *Anxiety*, *Anger*, *Sadness*, *Swear words*, *Moralisation*, *Culture*, *Health*, *Mental health*, *Auditory*, *Feeling*, and *Prosocial behaviour*.



Figures 2-3: Dmeds and Pooled SDs of LIWC Categories for Minimal Self (2) and Narrative Self (3): Present Condition.

4.3. Self as an Agent [6]

Between the categories significantly more present when the AS is **present** compared to when it is not, these are the ten with the highest D*med*s (see Figure 4): *Personal*, *Pronouns*, *"I"*, *Affect*, *Dictionary words*, *Emotion*, *Function words*, *Positive tone*, *Positive emotion*, and *Certitude*. Between the categories significantly more present when the AS is **not present** compared to when it is, these are the ten with the lowest D*med*s: *Word Count*, *Other pronouns*, *Numbers*, *Negative emotions*, *Friend*, *Sexual*, *Feeling*.

4.4. Bodily Self [7]

[6]**Self as an agent present.** *Personal pronouns:* D*med*: 2.17, SD: 5.32; *Pronouns*: D*med*: 2.10, SD: 6.10; *"I"*: D*med*: 1.89, SD: 4.24; *Affect*: D*med*: 0.89, SD: 3.23; *Dictionary words:* D*med*: 0.79, SD: 5.50; *Emotion*: D*med*: 0.50, SD: 1.68; *Function words*: D*med*: 0.48, SD: 6.80; *Positive tone*: D*med*: 0.48, SD: 2.68; *Positive emotion*: D*med*: 0.38, SD: 1.19; and *Certitude* (D*med*: 0.30, SD: 1.43. **Self as an agent not present.** *Word count*: D*med*: -70.50, SD: 167.23; *Other pronouns*: D*med*: -0.41, SD: 3.63; *Numbers*: D*med*: -0.21, SD: 2.82; *Negative emotions:* D*med*: 0.00, SD: 1.07; *Friend*: D*med*: 0.00, SD: 0.64; *Sexual;* D*med*: 0.00, SD: 0.76; and *Feeling*: D*med*: 0.00, SD: 0.85.

[7]**Bodily self present.** *Authentic*: D*med*: 16.23, SD: 30.57; *"I":* D*med*: 5.59, SD: 3.81; *Personal pronouns:* D*med*: 4.43, SD: 4.44; *Focus: Past time*: D*med*: 3.15, SD: 4.02; *Pronouns*: D*med*: 2.79, SD: 4.94; *Dictionary words*: D*med*: 2.66, SD: 5.31; *Linguistic dimensions*: D*med*: 2.61, SD: 6.62; *Physical:* D*med*: 2.52, SD: 2.82; *Function words*: D*med*: 2.20, SD: 6.25; *Time*: D*med*: 0.91, SD: 2.54; *Emotion*: D*med*: 0.73, SD: 1.67; *Health*: D*med*: 0.35, SD: 1.38; *"She/He"*: D*med*: 0.28, SD: 3.05; and

Between the categories significantly more present when the BS is **present** compared to when it is not, these are the ten with the highest D*med*s (see Figure 5): *Authentic*, *"I"*, *Personal pronouns*, *Focus: Past time*, *Pronouns*, *Dictionary words*, *Linguistic dimensions*, *Physical*, *Function words*, and *Time*. Furthermore, interestingly also the following categories show statistical significance in this condition: *Emotion*, *Health, "She/He"*, and *Negative emotion*. Between the categories significantly more present when the BS is **not present** compared to when it is, these are the ten with the lowest D*med*s: *Word count*, *Clout*, *Focus: Present time*, *Big words*, *Cognitive processes*, *Auxiliary verbs*, *Differentiation*, *"You"*, *Cognition*, and *Power*.

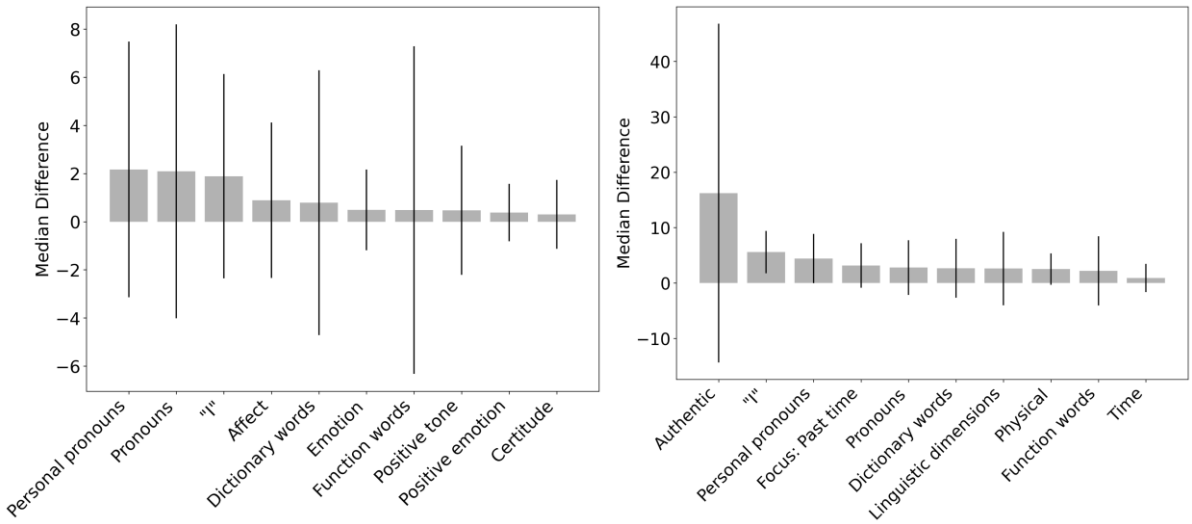


Figure 4-5: Dmeds and Pooled SDs of LIWC Categories for Self as an Agent (4) and Bodily Self (5): Present.

4.5. Social Self [8]

Between the categories significantly more present when the SS is **present** compared to when it is not, these are the ten with the highest D*med*s (see Figure 6): *Word count*, *Authentic*, *Social*, *Dictionary words*, *Personal pronouns*, *Linguistic*, *Social referents*, *Pronouns*, *Function words*, and *"I"*. Furthermore, interestingly also the following categories show statistical significance in this condition: *Focus: Past time*, *Affect*, *Drives*, *Emotion*, *Physical*, *Negative tone*, *Positive tone*, *Verbs*, *"They"*, *Time*, *Insight*, *Positive emotion*, *"She/He"*,

Negative emotion: D*med*: 0.12, SD: 1.04. and. **Bodily self not present**. *Word count*: D*med*: -70.00, SD: 158.09; *Clout:* D*med*: -15.37, SD: 33.31; *Focus: Present time*: D*med*d: -3.13, SD: 3.13; *Big words*: D*med*: -2.04, SD: 5.72; *Cognitive processes*: Dmed: -1.78, SD: 4.45; *Auxiliary verbs*: D*med*: -1.71, SD: 3.30; *Differentiation*: D*med*: -1.42, SD: 2.31; *"You":* D*med*: -1.20, SD: 2.64; *Cognition*: D*med*: -1.18, SD: 4.70; and *Power*: D*med*: -0.99, SD: 1.54.

[8]**Social self present.** *Word count*: D*med*: 69.00, SD: 154.64; *Authentic*: D*med*: 11.37, SD: 31.53; *Social*: D*med*: 4.60, SD: 5.68; *Dictionary words:* D*med*: 4.12, SD: 5.66; *Personal pronouns*: D*med*: 3.64, SD: 5.17; *Linguistic*: D*med*: 3.29, SD: 7.52; *Social referents*: D*med*: 3.15, SD: 4.61; *Pronouns*: D*med*: 2.97, SD: 5.96; *Function words:* D*med*: 2.61, SD: 6.93; and *"I"*: D*med*: 2.47, SD: 4.16; *Focus: Past time:* D*med*: 1.57, SD: 4.02; *Affect*: D*med*: 1.43, SD: 3.19; *Drives*: D*med*: 1.25, SD: 3.20; *Emotion:* D*med*: 1.11, SD: 1.75; *Physical*: D*med*: 1.1, SD: 2.37; *Negative tone:* D*med*: 0.89, SD: 1.69; *Positive tone*: D*med*: 0.74, SD: 2.67; *Verbs*: D*med*: 0.73, SD: 5.37; *"They"*: D*med*: 0.71, SD: 1.42; *Time*: D*med*: 0.62, SD: 3.02; *Insight*: D*med*: 0.61, SD: 1.83; *Positive emotion*: D*med*: 0.46, SD: 1.36; *"She/He"*: D*med*: 0.43, SD: 2.48; *Negative emotion*: D*med*: 0.37, SD: 1.00; and *Conjunctions:* D*med*: 0.21, SD: 3.04. **Social self not present.** *Analytic*: D*med*: -16.17, SD: 26.43; *Articles*: D*med*: -1.09, SD: 3.49; *Numbers:* D*med*: -0.34, SD: 3.03; *Feeling*: D*med*: 0.00, SD: 0.79; *Curiosity*: D*med*: 0.00, SD: 0.69; *Sexual*: D*med*: 0.00, SD: 0.76; *Substances*: D*med*: 0.00, SD: 0.32; *Mental:* D*med*: 0.00, SD: 0.31; *Illness*: D*med*: 0.00, SD: 0.74; *Health*: D*med*: 0.00, SD: 1.28; *Technology*: D*med*: 0.00, SD: 1.35; *"We"*: D*med*: 0.00, SD: 1.56; *Anxiety*: D*med*: 0.00, SD: 0.35; *Anger*: D*med*: 0.00, SD: 0.54; *Sadness:* D*med*: 0.00, SD: 0.26; *Swear words*: D*med*: 0.00, SD: 0.83; and *Moralisation*: D*med*: 0.00, SD: 0.71.

Negative emotion, and *Conjunctions*. Between the categories significantly more present when the SS is **not present** compared to when it is, these are the ten with the lowest D*med*s: *Analytic*, *Articles*, *Numbers*, *Feeling*, *Curiosity*, *Sexual*, *Substances*, *Mental*, *Illness*, and *Health*. Furthermore, interestingly also the following categories show statistical significance in this condition: *Technology*, *"We"*, *Anxiety*, *Anger*, *Sadness*, *Swear words*, and *Moralisation*.

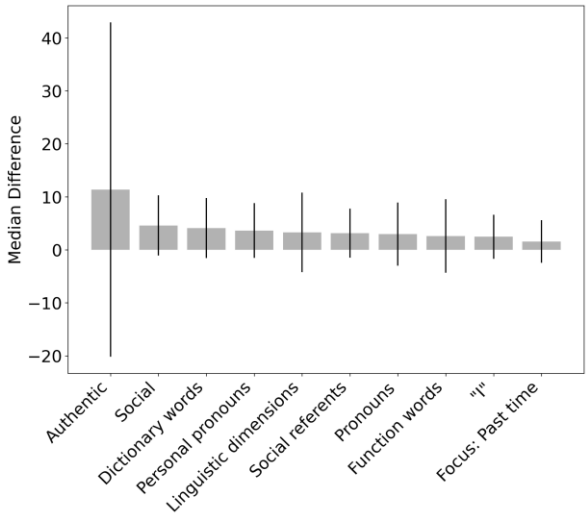


Figure 6: Dmeds and Pooled SDs of LIWC Categories for Social Self (6): Present.

4.6. General trends

Authenticity is a key factor in differentiating between the presence or absence of self-categories, ranking in the top two for the highest D*med* across four of the five categories evaluated (MS, NS, BS, and SS). The Reddit posts including self-categories also appear to include more function words (NS, AS, SS), pronouns and personal pronouns (in the case of NS, AS, BS, and SS), and first-person personal pronouns (in all the cases). They furthermore show to have both a positive (MS, NS, AS, and SS) and negative tone (MS, NS, and SS), emotion (MS, NS, AS, BS, and SS), and positive (MS, NS, AS, and SS) and negative emotion (NS, BS, and SS), while specific cases of negative emotion, such as anxiety (not-NS and not-SS), sadness (not-MS, not-NS, not-BS, and not-SS), and anger (not-MS, not-NS, and not-SS), are more often found in the posts with the absence of a certain self-category. Where a focus on past time is often present in correlation with self-categories (NS, BS, and SS), a focus on present time is more often present when the self-categories are not present (NS and BS). The AS shows to be related to certitude words, which were found by Fast and Horvitz (2016) to be present in Reddit posts high on dogmatism. It is interesting to note that the posts containing AS are not correlated with the clout and power LIWC categories.

5. Conclusions

Based on definitions from cognitive science and phenomenology, we built a multi-classifier of five different self-categories: *Minimal Self, Narrative Self, Self as Agent, Bodily Self,* and *Social Self.* We constructed five pairs of datasets, each pair including Reddit posts either presenting or not one of the self-categories. We then employed LIWC-22 to analyse them. Further work is needed to further explore the connection between self-categories and LIWC categories, generalising it on other kinds of data, and to inquire the possibility of directly using LIWC categories as textual indicators of a specific self-category. We hope that our

exploratory study can bring significant benefits since it fosters the development of interpretable and transparent models, essential for sensitive domains, such as clinical settings.

Acknowledgments

We acknowledge the financial support from the Slovenian Research Innovation Agency (ARIS) core research program Knowledge Technologies (P2-0103). The Young Researcher Grant (PR-12394) supported the work of BK. JC wishes to thank Dr. Tine Kolenik for participating in the brainstorming process.

Bibliography

Beck T., Schuff H., Lauscher A. and Gurevych, I. (2024). Sensitivity, Performance, Robustness: Deconstructing the Effect of Sociodemographic Prompting. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, 1, 2589-2615.

Bermúdez J. L. (2018). *The bodily self: Selected essays*. MIT Press.

Boyd R.L., Ashokkumar A., Seraj S. and Pennebaker J.W. (2022). The development and psychometric properties of LIWC-22. *Austin, TX: University of Texas at Austin*, 1-47. https://www.liwc.app/

Brown T., Mann B., Ryder N., Subbiah M., Kaplan J.D., Dhariwal P., Neelakantan A., Shyam P., Sastry G., Askell A. and Agarwal S. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.

Caporusso J. (2022). Dissolution experiences and the experience of the self: An empirical phenomenological investigation. [Master's thesis, University of Vienna]. https://10.25365/thesis.71694

Caporusso J., Tran T.H.H. and Pollak S. (2023). IJS@LT-EDI : Ensemble Approaches to Detect Signs of Depression from Social Media Text. In *Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion*, 172–178, Varna, Bulgaria.

Fast E. and Horvitz E. (2016). Identifying dogmatism in social media: Signals and models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 690-699.

Goldberg L. R. (1990). An alternative 'description of personality': The Big-Five factor structure. Journal of Personality and Social Psychology, 59, 1216–29.

Koloski B., Lavrač N., Cestnik B., Pollak S., Škrlj B. and Kastrin, A. (2023). AHAM: Adapt, Help, Ask, Model--Harvesting LLMs for literature mining. *arXiv preprint arXiv:2312.15784*.

MacIntyre A. (1985). After virtue: A study in moral theory (2nd edition). Duckworth, London.

Mehl M.R. and Pennebaker, J.W. (2003). The sounds of social life: a psychometric analysis of students' daily social environments and natural conversations. *Journal of personality and social psychology*, *84*(4), 857.

Parnas J., & Henriksen, M. G. (2014). Disordered self in the schizophrenia spectrum: a clinical and research perspective. Harvard review of psychiatry, 22(5), 251-265.

Pennebaker J. W., Mehl M. R. and Niederhoffer K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, *54*(1), 547–577.

Rude S., Gortner E.M. and Pennebaker J. (2004). Language use of depressed and depression-vulnerable college students. *Cognition & Emotion, 18*(8), 1121–1133.

Siderits M., Thompson E. and Zahavi, D. (Eds.). (2011). *Self, no self?: Perspectives from analytical, phenomenological, and Indian traditions*. Oxford University Press.

Völske M., Potthast M., Syed S., & Stein, B. (2017). Tl; dr: Mining reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, 59–63.

Yarkoni, T. (2010). Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. Journal of Research in Personality, 44, 363–73.