A Computational Framework to Identify Self-Aspects in Text

Anonymous ACL submission

Abstract

This Ph.D. proposal introduces a plan to develop a computational framework to identify Self-aspects in text. The Self is a multifaceted construct and it is reflected in language. While it is described across disciplines like cognitive science and phenomenology, it remains underexplored in natural language processing (NLP). Many of the aspects of the Self align with psychological and other well-researched phenomena (e.g., those related to mental health), highlighting the need for systematic NLP-based analysis. In line with this, we plan to introduce an ontology of Self-aspects and a gold-standard annotated dataset. Using this foundation, we will develop and evaluate conventional discriminative models, generative large language models, and embedding-based retrieval approaches against four main criteria: interpretability, ground-truth adherence, accuracy, and computational efficiency. Top-performing models will be applied in case studies in mental health and empirical phenomenology.

1 Introduction

The Self, superficially experienced as "the (perhaps sometimes elusive) feeling of being the particular person one is" (Siderits et al., 2013), is a complex phenomenon, amply discussed in philosophy and cognitive science (e.g., Zahavi, 2008). While there exist different views about the metaphysical nature of the Self (Siderits et al., 2013), in this work, we build on its phenomenological and behavioural manifestations. In everyday experience, the Self is characterised by multiple phenomenological and psychological aspects, including the experience of one's own body (Bermúdez, 2018) and a sense of agency (Gallagher, 2000), among others (Caporusso, 2022).

These Self-aspects are conceptually and empirically related to other well-established constructs—such as personality traits or experiential modes. For example, their relevance to contexts such as mental health research is supported in related work, which highlights the central role of Self-related processes in well-being and psychopathology, as well as in empirical phenomenology (i.e., the empirical investigation of experience, Aspers, 2009), where they are key to understanding altered states of consciousness (see Section 2).

Importantly, the specific ways in which Self-aspects are experienced by a person in a given moment are reflected in the language they use (e.g., see Pennebaker et al. (2003) and Section 2). The found correlations between textual features and Self-aspects can be further employed in downstream NLP tasks, for instance to detect psychological states (Caporusso et al., 2023; Du and Sun, 2022; Kolenik et al., 2024). However, the connections between textual features and many Self-aspects important for the identification of, e.g., mental health conditions and phenomenological states, are underexplored.

To address this shortcoming, we propose a computational framework capable of automatically detecting the presence and mode of Self-aspects in text. Existing tools such as LIWC (Linguistic Inquiry and Word Count; Boyd et al., 2022) and VADER (Valence Aware Dictionary and sEntiment Reasoner; Hutto and Gilbert, 2014) have shown that psychologically meaningful patterns can be computationally extracted from text using lexicons and interpretable features. Building on this tradition, our framework aims to go further: to detect nuanced, theoretically grounded aspects of Self-experience—such as agency, embodiment, or narrative coherence—through a combination of ontology design, annotated data, and a range of modelling approaches. The resulting method can be applied to tasks in domains such as mental health research and empirical phenomenology.

2 Related Work

2.1 Textual Features and Self-Aspects Correlations

This subsection surveys studies mapping text features to aspects of the Self.

Self Aspects Most research focuses on *I-talk*, i.e., the use of first-person pronouns as indicators of Self-focus (Pennebaker et al., 2003), which correlates with emotional pain, trauma, and depression (Tausczik and Pennebaker, 2010). Furthermore, pronoun usage hints at specific understandings of the Self vs others distinction (Na and Choi, 2009; Sharpless, 1985). The usage of active vs passive voice can shed light on the sense of agency of the author of a text (Simchon et al., 2023), while the Narrative Self (NS; i.e., "the narrative someone has of themselves, comprising their autobiographical memories and stories of who they are" Caporusso et al., 2024) is reflected in the structure and coherence of one's autobiographical accounts (Habermas and Köber, 2015; Holm et al., 2016; Jaeger et al., 2014; Waters and Fivush, 2015). In this context, Author profiling (AP) refers to the task of inferring personal characteristics of an author based on their writing, which has applications in, e.g., sociolinguistics and mental health analytics (Eke et al., 2019; Ouni et al., 2023b).

The correlation of text features with other aspects of the Self, such as the Minimal Self (MS; "the fact that experiences are presented to us in a fundamentally personal and subjective way" Caporusso et al., 2024), are less explored (Uno and Imaizumi, 2025).

Caporusso et al. (2024) investigated the LIWC categories associated with different aspects of the Self: MS, NS, Self as Agent (AS; "the experience of being an agent, i.e., in control, active"), Bodily Self (BS; "the experience of owning, controlling, and/or identifying with someone's own body (or parts of it)"), and Social Self (SS; "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"). Specifically, utilising a mixed approach to annotate the data, the authors classified text instances as presenting or not each of the mentioned self-aspects, and they analysed the obtained splits with LIWC.

Methods The methodological approaches utilised to detect correlations between textual features and Self-aspects can be broadly grouped into three main types:

- Approaches based on stylistic features such as punctuation, syntactic patterns, part-of-speech (POS) tags, sentence length, character/word n-grams, and structural features (e.g., number of paragraphs or capitalised words)—see Ouni et al. (2021); Vijayan and Govilkar (2019).

- Content-based approaches, relying on subject matter and vocabulary; features include term frequency-inverse document frequency (TF-IDF), topic models, and domain-specific keywords—see Ch and Cheema (2018); Ouni et al. (2023b)

- Hybrid approaches, where both stylistic and content-based features are analysed—see Fatima et al. (2017); Ouni et al. (2021, 2023b)

The use of LIWC or other lexicon-based techniques is the most common approach to investigate correlations between Self-aspects and textual features (Boyd and Schwartz, 2021; Pennebaker et al., 2003). More recently, however, NLP research has increasingly adopted machine learning (ML) methods—such as topic modelling and supervised classification—to analyse language patterns in a data-driven way (Eichstaedt et al., 2018; Ouni et al., 2021). Many studies used classical supervised learning methods, like support vector machines (SVMs; Chinea-Rios et al., 2022; HaCohen-Kerner, 2022; Vijayan and Govilkar, 2019), random forests (RFs; Fatima et al., 2017; Ouni et al., 2021), decision trees (Vijayan and Govilkar, 2019), and Naïve Bayes (NB; Mechti et al., 2020). Feature extraction in AP is critical: common strategies include Bag-of-Words (BoW) and TF-IDF (Ouni et al., 2023b), character and word n-grams (HaCohen-Kerner, 2022), POS and syntactic feature vectors (Mechti et al., 2020; Vijayan and Govilkar, 2019), word embeddings (Chinea-Rios et al., 2022; Fatima et al., 2017), semantic graphs and emotion tags (Ouni et al., 2023b). Furthermore, many studies employ qualitative approaches (Habermas and Köber, 2015; Waters and Fivush, 2015). However, deep learning (DL) models are increasingly employed as well, due to their capacity to automatically learn hierarchical feature representations from raw text and their superior performance on large-scale NLP tasks (Ouni et al., 2023a). Transformer-based models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) were adapted to AP tasks by fine-tuning on labelled AP datasets (Chinea-Rios et al., 2022). In recent work, LLMs have been explored for AP (see Huang et al., 2025). Huang et al. (2024) show that GPT-4 outperforms BERT-based models in zero-shot authorship attribution and verification, especially when guided by linguistic cues.

The type of text analysed varies widely, ranging from autobiographical essays (Adler, 2012; McAdams, 2001), stream-of-consciousness essays or narrative prompts (Pennebaker and Beall, 1986; Rude et al., 2004), transcripts of spoken conversations or interviews (Adler et al., 2008; Bamberg, 2008; Lysaker and Lysaker, 2002), diary entries and letters (Baumeister et al., 1994; Pennebaker and Francis, 1996), social media posts (Guntuku et al., 2019; Schwartz et al., 2013), to even published autobiographies or literature (Bruner, 2003; Freeman, 2009).

2.2 Downstream Applications

The correlations discussed in the previous subsection are often employed in downstream applications. For instance, Kolenik et al. (2024) utilised predefined sets of words and linguistic patterns that have been associated with specific psychological states, traits, or cognitive processes to train ML models that detect stress, anxiety, and depression. Similarly, Du and Sun (2022) leveraged linguistic features known to correlate with psychological states, like absolutist words and personal pronouns, to detect depression, anxiety, and suicidal ideation. In the context of the LT-EDI@RANLP 2023 shared task, first-person singular pronouns and time-related terms, recognised as indicative of depressive states (Ratcliffe, 2014), were employed to identify signs of depression in social media posts (Caporusso et al., 2023). Eichstaedt et al. (2018) utilised topic models to identify clusters of words that often appear together in Self-narratives, and supervised ML to predict an upcoming depression diagnosis from social media posts.

Outside of the context of NLP studies, works investigating, e.g., mental health issues or phenomenological states vastly address Self-aspects to identify the phenomenon of interest. For instance, an impacted sense of agency is registered in individuals with anxiety and depression, who experience a deficiency in estimating their control over positive outcomes (Mehta et al., 2023), while disturbances in interoception and Self-awareness were found to be correlated with anxiety and schizophrenia, among the others (Yang et al., 2024). Often, different Self-aspects correlate with disorders in a synergistic way, or there is an atypical disintegration of Self-aspects. For instance, Alzheimer's Disease (AD) and other conditions involving cognitive decline are associated with impaired Self-continuity, sense of personal history and future goals, capabilities of Self-reflection, and personal meaning (El Haj et al., 2015), resulting in a distorted narrative Self-identity. Along, and sometimes in support of, research in mental well-being, Self-aspects are relevant in the context of empirical phenomenology, among others. For example, a multitude of Self-aspects is examined in the investigation of experiences of dissolution (i.e., "experiential episodes during which the perceived boundaries between self and world (i.e., nonself) become fainter or less clear"; Caporusso, 2022; Nave et al., 2021), and bodily experience is investigated in the context of depersonalisation and derealisation disorders (Tanaka, 2018). In line with this, scales and symptom checklists have been developed to assess the presence and intensity of psychological or phenomenological states (Heering et al., 2016; Michal et al., 2014; Nour et al., 2016; Parnas et al., 2005; Sierra and Berrios, 2000).

2.3 Identified Gaps and Research Motivation

Disciplines like cognitive science, phenomenology, and psychology identify many different aspects of the Self, but NLP studies a) have dealt with only a few superficial ones and b) have only employed basic techniques. Indeed, while NLP started to employ the correlation between Self-aspects and textual features in various downstream tasks, the Self-aspects employed in, e.g., mental health research and empirical phenomenology are more varied and nuanced. For this reason, we believe that it would be helpful to identify further and more detailed connections between Self-aspects and textual features, and to develop a model to detect and analyse Self-aspects in text. This could be used by professionals of other disciplines, for instance to analyse patients' reports and transcripts of phenomenological interviews (e.g., see micro-phenomenology, Petitmengin et al., 2019).

To this end, our proposed framework aligns in spirit with existing tools like LIWC and VADER. However, unlike these general-purpose approaches, our framework is specifically designed to capture a range of Self-aspects grounded in interdisciplinary theory. Moreover, while LIWC captures psychological correlates at a coarse granularity (e.g., affect, pronouns), we aim to represent structured components of Self-experience.

3 Research Proposal

This Ph.D. proposal seeks to explore the ways of developing a computational model to automatically detect Self-aspects in language. We plan to test the proposed approaches on different case studies from the fields of mental health and empirical phenomenology. Our Research Objectives (ROs) are as following:

- **RO1)** Detail an ontology of the Self aspects that would be relevant and sensible for a computational model to detect in text.
- **RO2)** Construct heterogeneous datasets with annotations relative to the identified Self-aspects.
- **RO3)** Define the desiderata of the computational model to detect Self-aspects in text and identify the approaches which would best fulfill them.
- **RO4)** Determine the evaluation approach and the applications for our computational model to detect Self-aspects in text.

We plan to produce the following outcomes: self ontology detailing and labelling instructions; heterogeneous annotated dataset; set of models to identify Self-aspects in text.

4 Self Ontology (RO1)

We aim to develop a comprehensive ontology of Self-aspects which are a) relevant to possible applications and b) detectable in text data. Each Self-aspect (e.g., bodily Self) is characterised by different elements (e.g., body ownership, body awareness), each of which is specified in different modes (e.g., body ownership: weak). Some of the Self-aspects investigated are identified through previous studies which developed similar lists or ontologies (e.g., Caporusso, 2022; Nave et al., 2021). The ontology, still a work-in-progress, is built collaboratively by adopting both bottom-up and a top-down approaches. That is to say, we utilise literature detailing the elements and modes of various Self-aspects (e.g., Moore, 2016; Serino et al., 2013) along with studies from disciplines like psychology and neuroscience detailing the Self-aspects relevant to the construct of interest (e.g., Petkova et al., 2011). Furthermore, we will be meeting with experts from fields which could benefit from our final model (e.g., mental health professionals and empirical phenomenologists) to better identify the specific Self-aspects, elements, and modes which could be relevant for their work. While analysing literature and consulting with experts, we will be exploring textual data itself. For each Self-aspect, element, and mode, we will provide a definition, both a positive and a negative example from textual data, and notes to guide the identification and/or distinction. Constructing the Self ontology presents various challenges, most of all regarding how the different components relate with each other. For example, most of the aspects and elements, if not all, appear to not be mutually exclusive, and there are aspects (e.g., sense of agency) that could apply to other aspects (e.g., sense of agency over bodily Self).

5 Datasets (RO2)

The datasets (aiming for at least 10; see Section 8), which will be annotated with the labels developed (see Section 4), need to vary in type as it is desired for the model to be able to analyse Self-aspects across different kinds of data. We plan to utilise transcripts from phenomenological interviews, clinical tasks, and structured or unstructured interviews. These will include employ already existing datasets and construct new ones. Importantly, all data collection—whether previously conducted or ongoing—is carried out within the scope of pre-approved research projects. Part of the phenomenological interviews data has already been collected (six subjects), and clinical interviews are being conducted in the context of an existing larger project. We aim to utilise datasets from different languages, in order to create a multilingual model. The annotated datasets will serve as training and testing data, as well as ground truth. The length of the text chunk considered as a labelling instance is determined case by case, based on what is sufficient to meaningfully express the presence of a specific Self-aspect or mode. In general, this can range from a single sentence to a short paragraph, depending on the complexity of the expression.

5.1 Annotation

Multiple annotators (e.g., three, possibly the same researchers compiling the Self ontology and the annotation guidelines) will independently annotate the datasets or part of them. Inter-annotator agreement will be calculated to assess consistency and reliability of the annotations. The first author, who will take part in and lead the annotation, has experience in conducting qualitative analysis and annotation of textual data, including mostly phenomenological interviews, but also, e.g., social media posts, with a focus on the Self. In the first phase of the annotation process, the annotators will meet and discuss their decisions, so to come to a similar understanding of the guidelines. This can bring to further adjustments of the guidelines themselves. In the case that it proves too expensive to manually label the entire dataset, we will adopt large language models (LLMs) for automatic annotation of the remaining instances—following an approach similar to Caporusso et al. (2024). Specifically, LLMs fine-tuned for instruction following (Brown et al., 2020) will be evaluated against a manually annotated subset to ensure quality. Importantly, LLM-based annotations will be used to augment training data for conventional discriminative models, embedding-based retrieval approaches, and, in principle, for fine-tuning LLMs—provided such synthetic data is excluded from evaluation (see Section 7). LLMs themselves will be evaluated separately, using only the manually labelled portion of the data to avoid circularity. This ensures a clean separation between training supervision and model evaluation.

6 Desiderata (RO3a)

Here, we discuss our desiderata for the models.

Interpretability, which in the context of ML refers to the extent to which a human can understand the internal mechanism of a model leading from input to output (Lipton, 2018; Molnar, 2020) is to be differentiated from explainability, which often involves post-hoc approximations of a model's behaviour (Molnar, 2020). This distinction is particularly crucial for our task for three main reasons. First, the target applications of our framework include implementations in sensitive domains like healthcare. Indeed, in such cases, the use of interpretable ML models is preferable to post-hoc explanations for black-box models, as the latter may be incomplete or misleading and do not ensure transparency, trust, and ethical decision-making (Ahmad et al., 2018; Amann et al., 2020; Bohlen et al., 2024; Chaddad et al., 2023; Doshi-Velez and Kim, 2017; Ennab and Mcheick, 2024; Lipton, 2018; Lu et al., 2023; Rudin, 2019; Tjoa and Guan, 2020). Some examples of this are studies by Gao et al. (2023); Wang et al. (2023). Second, generic explainability approaches are often insufficient in NLP due to the inherent ambiguity, subjectivity, and domain sensitivity of language data, necessitating explanations that align with the linguistic and reasoning norms of specific application areas (Mohammadi et al., 2025). Some examples are studies by Saha et al. (2022, 2023); Wang et al. (2023). Third, interpretability is desirable because it enables traceability—the ability to identify the specific passage or linguistic marker that led to a given classification. This is particularly important in applications such as studies based on the analysis of empirical phenomenological interviews, where it is necessary to provide illustrative examples for each identified experiential category (e.g., a specific mode of a Self-aspect).

Ground-Truth Basis requires that model outputs be directly derived from verified, annotated data, rather than inferred through non-transparent or heuristic reasoning (Goodfellow et al., 2016). Once again, this principle is especially critical in sensitive domains where decisions must be accountable and ethically sound (Mittelstadt, 2019; Varshney and Alemzadeh, 2017), and in NLP, where the inherent ambiguity and subjectivity of language complicate evaluation (Hovy and Prabhumoye, 2021). In many NLP tasks (e.g., Evkoski and Pollak, 2023) a degree of approximation is often tolerated in favour of pragmatic utility, and models are evaluated based on what is useful or convincing to downstream consumers. By contrast, in our work, it is strongly desirable that model predictions remain traceable to the actual input provided by us. This grounding is not only central to scientific rigour, but also to ensuring justifiability and trust in use cases such as clinical assessments and the analysis of phenomenological interviews, where outputs may influence human understanding of complex experiences.

Importantly, ground-truth basis is complementary to interpretability. While interpretability focuses on making the model's decision process understandable, ground-truth basis ensures that its outputs are substantively anchored in verified data rather than emergent patterns from opaque pretraining. Together, these two properties are essential for making computational predictions trustworthy and usable by stakeholders such as clinicians and phenomenologists.

As expected, achieving high classification **accuracy** remains a central objective, and considering all the other desiderata, a model with a lower **computational cost** is to be preferred. Additionally, given the sensitivity of the data, we prioritise tools that guarantee full control over processing and prevent third-party access.

7 Proposed Approaches (RO3b)

In this subsection, we refer to literature in order to compare the various proposed approaches with regard to each of our desiderata. The proposed approaches are: conventional discriminative models, including traditional AI and neural networks (NNs); generative LLMs, fine-tuned or with few-shot learning; and embedding-based retrieval approaches.

As the NLP landscape—particularly in relation to LLMs, interpretability, and domain-specific adaptation—continues to evolve rapidly, the methodological choices outlined below are intended as a flexible, revisable framework rather than a rigid pipeline. We anticipate that developments over the course of the Ph.D. will inform and potentially shift our implementation strategies, especially in response to emerging technologies and best practices in ethical, explainable NLP. In line with this flexible and modular approach, we also propose the investigation of a mixture-of-experts (MoE) architecture.

To train our models, we plan to employ both learned textual features—such as embeddings or TF-IDF representations—and predefined features derived from both previous studies (e.g., Pennebaker et al., 2003) and further investigations based on Caporusso et al. (2024)'s framework. This hybrid feature strategy supports both data-driven learning and interpretability through grounded linguistic markers.

Preliminary experiments are described in the Appendix A.

7.1 Conventional Discriminative Models

Conventional discriminative models include both traditional ML methods (Bishop and Nasrabadi, 2006) and NNs (LeCun et al., 2015). Examples include SVMs (Cristianini and Shawe-Taylor, 2000), Logistic Regression (LR), decision trees, and feed-forward or recurrent NNs (RNNs) (Goodfellow et al., 2016) trained for classification purposes. They are often employed in the context of supervised learning, where the model learns from labelled data (Murphy, 2012).

Conventional discriminative models represent a good approach to our goal, assuming the availability of high-quality annotated datasets. Once trained, such models can directly classify a given text instance into predefined categories—such as BS, NS, or AS—and further specify the mode for each element (e.g., bodily ownership: present; agency over the body: partial). **Interpretability** in this approach depends largely on the choice of model: while rule-based models like decision trees or LR are inherently transparent, NNs are less interpretable and often require post-hoc explanation methods. Regarding **ground-truth** alignment, conventional discriminative models are optimal, since their outputs are entirely dependent on the patterns found in the labelled examples. When sufficient and representative training data is available, these models can be very **accurate**. Furthermore, they can be highly efficient **computationally**.

7.2 Generative LLMs

Generative LLMs (e.g., GPT; Radford et al., 2018) are designed to produce new outputs—in the case of language models, in the form of text—by learning the underlying distribution of the training data (Bengio et al., 2003; Radford et al., 2018).

Although flexible, they come with a few challenges. For example, even in the case that their output looks plausible, it might be incorrect. This is referred to as *hallucination*, and it is due to the fact that these models generate responses solely based on learned statistical patterns (Zhang et al., 2022). Additionally, they reflect biases present in their training data and lack transparent mechanisms for interpreting or verifying their outputs (Bolukbasi et al., 2016).

Ideally, generative LLMs will be applied to our task either through prompt-based few-shot learning or via fine-tuning on labelled datasets (Wei et al., 2022; Wolf et al., 2020), which generally improves accuracy and control over outputs (Howard and Ruder, 2018).

While LLMs offer great flexibility and generalisation capabilities, they are not **interpretable**. Although post-hoc explanation methods like LIME (Local Interpretable Model-agnostic Explanations; Alvarez-Melis and Jaakkola, 2018; Ribeiro et al.,

2016) or SHAP (SHapley Additive exPlanations; Jin et al., 2020; Lundberg and Lee, 2017) can provide some superficial insight, they do not guarantee true transparency or fidelity to the model's internal reasoning. Furthermore, LLMs are not grounded in **ground-truth** data. Even when fine-tuned, it remains unclear whether these models' predictions are derived from the data used for fine-tuning or the huge corpora used for pre-training. Furthermore, their outputs can change even from subtle shifts in prompt wording. This affects the consistency and reliability of the model. **Accuracy** is often high in LLMs, but it depends on prompt design and the complexity of the task. Inconsistent results could result from similar inputs, particularly when the classification schema is fine-grained, such as distinguishing between modes of Self-experience. Finally, generative LLMs are **computationally** expensive.

7.3 Embedding-Based Retrieval

Embedding-based retrieval is a type of retrieval-based approach which involves mapping the input into a shared vector space using models such as BERT (Devlin et al., 2019) or Sentence-BERT (Reimers and Gurevych, 2019). The vector representations of the inputs are compared to the already existing vector space, i.e., the knowledge base (Karpukhin et al., 2020). The initial vector space can be fine-tuned to task specific data, enhancing the model performance, and the semantic similarity between the reference and the input texts can be measured via cosine similarity or other distance metrics (Cer et al., 2018; Xiong et al., 2020).

For our purpose, embedding-based retrieval is especially useful in the case that a well-curated repository of annotated examples is available. The model can retrieve similar past instances that have already been labelled, allowing it to infer the classification of the new instance by analogy. While the embedding process itself is not inherently **interpretable**, the example-based reasoning enabled by retrieval models provides a form of implicit transparency: it is possible to inspect the retrieved examples and their labels to understand the basis of the model's recommendation. This makes the approach more explainable than generative LLMs, although not as transparent as rule-based classifiers. In terms of **ground-truth** alignment, embedding-based retrieval performs strongly. The model's decisions are anchored in annotated, verified data, and it does not generate new content but rather identifies the closest match among existing cases. In RAG-style architectures (retrieval-augmented generation; Lewis et al., 2020), this grounding helps reduce—but does not eliminate—the risk of hallucination during generation. **Accuracy** depends heavily on the quality and diversity of the dataset: if the database covers a broad range of expressions for different Self-aspects and modes, the model can achieve high classification performance. **Computationally**, this approach is efficient. Embeddings can be pre-computed, and retrieval operations (e.g., cosine similarity search) are lightweight.

7.4 Mixture of Experts

We also plan to explore a MoE architecture based on the work by Swamy et al. (2025), who proposed an interpretable MoE model designed for human-centric applications. In such architectures, different sub-networks—i.e., *experts*—are selectively activated depending on the input, enabling instance-specific reasoning and the possibility of **interpretability** where needed. This design offers a compelling balance between flexibility and transparency: it allows the integration of both interpretable models and black-box models within a unified framework. For our purposes, this means we can assign interpretable models to Self-aspect categories where explanation is critical (e.g., clinical applications), while using more complex models for noisier or less constrained categories.

The modular nature of MoE architectures also aligns well with our Self-aspect ontology. Since each expert can be specialised to a distinct subset of Self-aspects or linguistic patterns, this structure supports both conceptual clarity and efficient scalability (**computational cost**). Moreover, because only a few experts are activated per instance, the resulting predictions can offer local insight into the decision process, particularly when interpretable experts are selected. Importantly, expert modules trained on annotated data can maintain clear ties to their training supervision, preserving **ground-truth basis** at the module level. We believe this architecture is a promising direction to address the trade-off between **accuracy** and interpretability across the wide range of Self-related phenomena we aim to model.

8 Evaluation (RO4)

8.1 Intrinsic Evaluation

To evaluate and compare the effectiveness of different classification methods for identifying Self-aspects and their elements and modes in text, the approach proposed by Demšar (2006) to compare the performance of multiple classifiers across multiple datasets will be adopted. To use this method, a minimum of five different datasets is necessary, although it is recommended to employ at least 10. In the context of this Ph.D., a diverse range of models will be used to perform the classification (see Section 7). Despite their varied architectures and learning paradigms, they all can be evaluated in a comparable way. That is to say, by producing predictions over shared, annotated datasets and assessing them using standard performance metrics such as accuracy, F1-score, or macro-averaged precision and recall. By using Demšar (2006)'s framework, the evaluation will not only focus on raw performance, but also support robust conclusions about the relative strengths of each approach in the context of supervised Self-aspect classification. This is essential for making informed methodological choices, particularly when weighing the benefits of interpretable and ground-truth-aligned models against those of more flexible, data-driven generative LLMs. For the purposes of evaluation, we adopt an instance-based setup, treating each labelled unit (e.g., sentence or utterance) as a classification instance. Future work may explore span-based evaluation to capture finer-grained textual markers of Self-aspect expression.

8.2 Extrinsic Evaluation

We also plan to evaluate our framework by how useful it proves to be in downstream tasks. As it is likely that different trade-offs of desiderable features are best for different applications, we do not aim to propose one singular model, but a collection of models. They will ideally be implemented in a user-friendly software that will allow the selection of the desired model, along with information and suggestions regarding each of them. Additionally, similarly to LIWC (Boyd et al., 2022), the user will be able to select which Self-aspects to analyse, and to which degree of granularity. It will be possible to determine at which level should the analysis be conducted, e.g., at the sentence, paragraph, or document level.

We plan to conduct at least two case studies in which we will apply one or more of our developed models to different tasks.

In the context of an ongoing project on NLP approaches to cognitive decline, we plan to analyse comparable texts produced by clinical vs non-clinical population by using one or more of our proposed models. In particular, this will serve to test hypothesis on the differences in Self-aspects, but also, potentially, to identify features that could be used to detect cognitive decline.

In the context of the larger attempt to develop a computational framework to support the analysis of phenomenological interviews, one or more of our developed models will be adopted to support the analysis of the phenomenology of the Self, fundamental to most, if not all, experiences. This could help highlight how the Self is experienced differently across an episode (e.g., a dissolution experience; Caporusso, 2022), or how it is experienced by different populations, e.g., affected or not by derealisation.

9 Conclusion

We presented a proposal to design a computational model capable of detecting Self-aspects in text, grounded in a structured ontology and supported by diverse, annotated datasets curated by us. Our approach bridges conceptual insights from fields such as psychology and phenomenology with empirical techniques in NLP, enabling interpretable and application-oriented analysis of Self in language. Rather than relying on a single architecture, we propose and evaluate a range of computational models—rule-based, embedding-based, and generative LLMs—each assessed in light of desiderata such as interpretability, ground-truth basis, accuracy, and computational cost. By aligning technical development with ethical considerations and application-specific constraints, we aim to contribute not only a functional model, but also a thoughtful framework for the computational study of the Self.

10 Limitations

Our work presents various limitations. The Self-aspects specified in our ontology may be insufficient or suboptimal for the range of tasks we intend to address. Additionally, although our datasets are heterogeneous, this may still be insufficient for generalisability—particularly across cultural contexts where expressions of Self may vary significantly. The heterogeneity of the datasets, along with the flexible granularity of labelling units, may also introduce inconsistencies. Furthermore, many of the computational approaches we propose require substantial resources, including large volumes of annotated data. Moreover, there is a risk of overfitting to the specific theoretical assumptions embedded in our ontology, particularly if it privileges certain conceptions of the Self over others, potentially narrowing the interpretive scope of our models. Reconciling the need for interpretability and ground-truth adherence with high classification performance remains a central challenge in our methodological design.

11 Ethical Considerations

As this study relies on the use of existing datasets or datasets collected within the scope of other projects, the ethical considerations pertaining to each dataset are governed by the terms under which the data have been or will be collected. For datasets obtained through restricted access, we will comply with all necessary data use agreements and institutional requirements. We are committed to ensuring the anonymisation of all textual data prior to model training. Since our datasets and LLMs may reflect cultural or demographic biases, we acknowledge the risk of reproducing or amplifying such biases in our outputs. We emphasise that the computational models developed in this research are intended to function as support tools rather than as standalone decision-makers.

References

Jonathan M Adler. 2012. Living into the story: agency and coherence in a longitudinal study of narrative identity development and mental health over the course of psychotherapy. *Journal of personality and social psychology*, 102(2):367.

Jonathan M Adler, Lauren M Skalina, and Dan P McAdams. 2008. The narrative reconstruction of psychotherapy and psychological health. *Psychotherapy research*, 18(6):719–734.

Muhammad Aurangzeb Ahmad, Carly Eckert, and Ankur Teredesai. 2018. Interpretable machine learning in healthcare. In *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*, pages 559–560.

David Alvarez-Melis and Tommi S Jaakkola. 2018. On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*.

Julia Amann, Alessandro Blasimme, Effy Vayena, Dietmar Frey, Vince I Madai, and Precise4Q Consortium. 2020. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC medical informatics and decision making*, 20:1–9.

Patrik Aspers. 2009. Empirical phenomenology: A qualitative research approach (the cologne seminars). *Indo-pacific journal of phenomenology*, 9(2).

Michael Bamberg. 2008. Considering counter narratives. In *Considering counter-narratives: Narrating, resisting, making sense*, pages 351–371. John Benjamins Publishing Company.

Roy F Baumeister, Arlene M Stillwell, and Todd F Heatherton. 1994. Guilt: an interpersonal approach. *Psychological bulletin*, 115(2):243.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

José Luis Bermúdez. 2018. *The bodily self: Selected essays*. MIT Press.

Christopher M Bishop and Nasser M Nasrabadi. 2006. *Pattern recognition and machine learning*, volume 4. Springer.

Lasse Bohlen, Julian Rosenberger, Patrick Zschech, and Mathias Kraus. 2024. Leveraging interpretable machine learning in intensive care. *Annals of Operations Research*, pages 1–40.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. 2022. The development and psychometric properties of liwc-22. *Austin, TX: University of Texas at Austin*, 10:1–47.

Ryan L Boyd and H Andrew Schwartz. 2021. Natural language analysis and the psychology of verbal behavior: The past, present, and future states of the field. *Journal of Language and Social Psychology*, 40(1):21–41.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jerome Seymour Bruner. 2003. *Making stories: Law, literature, life*. Harvard University Press.

Jaya Caporusso. 2022. Dissolution experiences and the experience of the self: an empirical phenomenological investigation (unpublished master's thesis). university of vienna. *Advisor: Assist. Prof. Dr. Maja Smrdu*.

Jaya Caporusso, Boshko Koloski, Maša Rebernik, Senja Pollak, and Matthew Purver. 2024. A phenomenologically-inspired computational analysis of self-categories in text. In *Proceedings of JADT 2024*.

Jaya Caporusso, Thi Hong Hanh Tran, and Senja Pollak. 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*, pages 172–178, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Muhammad Waqas Anjum Ch and Waqas Arshad Cheema. 2018. A study of content based methods for author profiling in multiple genres. *International Journal of Scientific Engineering Research*, 9(9):322–327.

Ahmad Chaddad, Jihao Peng, Jian Xu, and Ahmed Bouridane. 2023. Survey of explainable ai techniques in healthcare. *Sensors*, 23(2):634.

Mara Chinea-Rios, Thomas Müller, Gretel Liz De la Peña Sarracén, Francisco Rangel, and Marc Franco-Salvador. 2022. Zero and few-shot learning for author profiling. In *International Conference on Applications of Natural Language to Information Systems*, pages 333–344. Springer.

Nello Cristianini and John Shawe-Taylor. 2000. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press.

Janez Demšar. 2006. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Xiaowei Du and Yunmei Sun. 2022. Linguistic features and psychological states: A machine-learning based approach. *Frontiers in psychology*, 13:955850.

Johannes C Eichstaedt, Robert J Smith, Raina M Merchant, Lyle H Ungar, Patrick Crutchley, Daniel Preoţiuc-Pietro, David A Asch, and H Andrew Schwartz. 2018. Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, 115(44):11203–11208.

Christopher Ifeanyi Eke, Azah Anir Norman, Liyana Shuib, and Henry Friday Nweke. 2019. A survey of user profiling: State-of-the-art, challenges, and solutions. *IEEE Access*, 7:144907–144924.

Mohamad El Haj, Pascal Antoine, Jean Louis Nandrino, and Dimitrios Kapogiannis. 2015. Autobiographical memory decline in alzheimer's disease, a theoretical and clinical overview. *Ageing research reviews*, 23:183–192.

Mohammad Ennab and Hamid Mcheick. 2024. Enhancing interpretability and accuracy of ai models in healthcare: a comprehensive review on challenges and future directions. *Frontiers in Robotics and AI*, 11:1444763.

Bojan Evkoski and Senja Pollak. 2023. Xai in computational linguistics: Understanding political leanings in the slovenian parliament. *arXiv preprint arXiv:2305.04631*.

Mehwish Fatima, Komal Hasan, Saba Anwar, and Rao Muhammad Adeel Nawab. 2017. Multilingual author profiling on facebook. *Information Processing & Management*, 53(4):886–904.

Mark Freeman. 2009. *Hindsight: The promise and peril of looking backward*. Oxford University Press.

Shaun Gallagher. 2000. Philosophical conceptions of the self: implications for cognitive science. *Trends in cognitive sciences*, 4(1):14–21.

Xiaoquan Gao, Sabriya Alam, Pengyi Shi, Franklin Dexter, and Nan Kong. 2023. Interpretable machine learning models for hospital readmission prediction: a two-step extracted regression tree approach. *BMC medical informatics and decision making*, 23(1):104.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. 2016. *Deep learning*, volume 1. MIT press Cambridge.

Sharath Chandra Guntuku, Rachelle Schneider, Arthur Pelullo, Jami Young, Vivien Wong, Lyle Ungar, Daniel Polsky, Kevin G Volpp, and Raina Merchant. 2019. Studying expressions of loneliness in individuals using twitter: an observational study. *BMJ open*, 9(11):e030355.

Tilmann Habermas and Christin Köber. 2015. Autobiographical reasoning in life narratives buffers the effect of biographical disruptions on the sense of self-continuity. *Memory*, 23(5):664–674.

Yaakov HaCohen-Kerner. 2022. Survey on profiling age and gender of text authors. *Expert Systems with Applications*, 199:117140.

Henriëtte Dorothée Heering, Saskia Goedhart, Richard Bruggeman, Wiepke Cahn, Lieuwe de Haan, René S Kahn, Carin J Meijer, Inez Myin-Germeys, Jim van Os, and Durk Wiersma. 2016. Disturbed experience of self: psychometric analysis of the self-experience lifetime frequency scale (self). *Psychopathology*, 49(2):69–76.

Tine Holm, Dorthe Kirkegaard Thomsen, and Vibeke Bliksted. 2016. Life story chapters and narrative self-continuity in patients with schizophrenia. *Consciousness and cognition*, 45:60–74.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and linguistics compass*, 15(8):e12432.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Baixiang Huang, Canyu Chen, and Kai Shu. 2024. Can large language models identify authorship? *arXiv preprint arXiv:2403.08213*.

Baixiang Huang, Canyu Chen, and Kai Shu. 2025. Authorship attribution in the era of llms: Problems, methodologies, and challenges. *ACM SIGKDD Explorations Newsletter*, 26(2):21–43.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

Jeff Jaeger, Katie M Lindblom, Kelly Parker-Guilbert, and Lori A Zoellner. 2014. Trauma narratives: It's what you say, not how you say it. *Psychological Trauma: Theory, Research, Practice, and Policy*, 6(5):473.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8018–8025.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *EMNLP (1)*, pages 6769–6781.

Tine Kolenik, Günter Schiepek, and Matjaž Gams. 2024. Computational psychotherapy system for mental health prediction and behavior change with a conversational agent. *Neuropsychiatric Disease and Treatment*, pages 2465–2498.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature*, 521(7553):436–444.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

X Alice Li and Devi Parikh. 2019. Lemotif: An affective visual journal using deep neural networks. *arXiv preprint arXiv:1903.07766*.

Zachary C Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Sheng-Chieh Lu, Christine L Swisher, Caroline Chung, David Jaffray, and Chris Sidey-Gibbons. 2023. On the importance of interpretable machine learning predictions to inform clinical decision making in oncology. *Frontiers in oncology*, 13:1129380.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Paul Henry Lysaker and John Timothy Lysaker. 2002. Narrative structure in psychosis: Schizophrenia and disruptions in the dialogical self. *Theory & Psychology*, 12(2):207–220.

Dan P McAdams. 2001. The psychology of life stories. *Review of general psychology*, 5(2):100–122.

Seifeddine Mechti, Nabil Khoufi, and Lamia Hadrich Belguith. 2020. Improving native language identification model with syntactic features: Case of arabic. In *Intelligent Systems Design and Applications: 18th International Conference on Intelligent Systems Design and Applications (ISDA 2018) held in Vellore, India, December 6-8, 2018, Volume 2*, pages 202–211. Springer.

Marishka M Mehta, Soojung Na, Xiaosi Gu, James W Murrough, and Laurel S Morris. 2023. Reward-related self-agency is disturbed in depression and anxiety. *PloS one*, 18(3):e0282727.

Matthias Michal, Bettina Reuchlein, Julia Adler, Iris Reiner, Manfred E Beutel, Claus Vögele, Hartmut Schächinger, and Andre Schulz. 2014. Striking discrepancy of anomalous body experiences with normal interoceptive accuracy in depersonalization-derealization disorder. *PloS one*, 9(2):e89823.

Brent Mittelstadt. 2019. Principles alone cannot guarantee ethical ai. *Nature machine intelligence*, 1(11):501–507.

Hadi Mohammadi, Ayoub Bagheri, Anastasia Giachanou, and Daniel L Oberski. 2025. Explainability in practice: A survey of explainable nlp across various domains. *arXiv preprint arXiv:2502.00837*.

Christoph Molnar. 2020. *Interpretable machine learning*. Lulu. com.

James W Moore. 2016. What is the sense of agency and why does it matter? *Frontiers in psychology*, 7:1272.

Kevin P Murphy. 2012. *Machine learning: a probabilistic perspective*. MIT press.

Jinkyung Na and Incheol Choi. 2009. Culture and first-person pronouns. *Personality and Social Psychology Bulletin*, 35(11):1492–1499.

Ohad Nave, Fynn-Mathis Trautwein, Yochai Ataria, Yair Dor-Ziderman, Yoav Schweitzer, Stephen Fulder, and Aviva Berkovich-Ohana. 2021. Self-boundary dissolution in meditation: A phenomenological investigation. *Brain Sciences*, 11(6):819.

Matthew M Nour, Lisa Evans, David Nutt, and Robin L Carhart-Harris. 2016. Ego-dissolution and psychedelics: validation of the ego-dissolution inventory (edi). *Frontiers in human neuroscience*, 10:190474.

Sarra Ouni, Fethi Fkih, and Mohamed Nazih Omri. 2021. Toward a new approach to author profiling based on the extraction of statistical features. *Social Network Analysis and Mining*, 11(1):59.

Sarra Ouni, Fethi Fkih, and Mohamed Nazih Omri. 2023a. Novel semantic and statistic features-based author profiling approach. *Journal of Ambient Intelligence and Humanized Computing*, 14(9):12807–12823.

Sarra Ouni, Fethi Fkih, and Mohamed Nazih Omri. 2023b. A survey of machine learning-based author profiling from texts analysis in social networks. *Multimedia Tools and Applications*, 82(24):36653–36686.

Josef Parnas, Paul Møller, Tilo Kircher, Jørgen Thalbitzer, Lennart Jansson, Peter Handest, and Dan Zahavi. 2005. Ease: examination of anomalous self-experience. *Psychopathology*, 38(5):236.

James W Pennebaker and Sandra K Beall. 1986. Confronting a traumatic event: toward an understanding of inhibition and disease. *Journal of abnormal psychology*, 95(3):274.

James W Pennebaker and Martha E Francis. 1996. Cognitive, emotional, and language processes in disclosure. *Cognition & emotion*, 10(6):601–626.

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

Claire Petitmengin, Anne Remillieux, and Camila Valenzuela-Moguillansky. 2019. Discovering the structures of lived experience: Towards a micro-phenomenological analysis method. *Phenomenology and the Cognitive Sciences*, 18(4):691–730.

Valeria I Petkova, Malin Björnsdotter, Giovanni Gentile, Tomas Jonsson, Tie-Qiang Li, and H Henrik Ehrsson. 2011. From part-to whole-body ownership in the multisensory brain. *Current Biology*, 21(13):1118–1122.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Matthew Ratcliffe. 2014. *Experiences of depression: A study in phenomenology*. OUP Oxford.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you? explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Stephanie Rude, Eva-Maria Gortner, and James Pennebaker. 2004. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, 18(8):1121–1133.

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215.

Rupsa Saha, Ole-Christoffer Granmo, and Morten Goodwin. 2023. Using tsetlin machine to discover interpretable rules in natural language processing applications. *Expert Systems*, 40(4):e12873.

Rupsa Saha, Ole-Christoffer Granmo, Vladimir I Zadorozhny, and Morten Goodwin. 2022. A relational tsetlin machine with applications to natural language understanding. *Journal of Intelligent Information Systems*, pages 1–28.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791.

Andrea Serino, Adrian Alsmith, Marcello Costantini, Alisa Mandrigin, Ana Tajadura-Jimenez, and Christophe Lopez. 2013. Bodily ownership and self-location: components of bodily self-consciousness. *Consciousness and cognition*, 22(4):1239–1252.

Elizabeth A Sharpless. 1985. Identity formation as reflected in the acquisition of person pronouns. *Journal of the American Psychoanalytic Association*, 33(4):861–885.

Mark Siderits, Evan Thompson, and Dan Zahavi. 2013. *Self, no self?: Perspectives from analytical, phenomenological, and Indian traditions*. OUP Oxford.

Mauricio Sierra and German E Berrios. 2000. The cambridge depersonalisation scale: a new instrument for the measurement of depersonalisation. *Psychiatry research*, 93(2):153–164.

Almog Simchon, Britt Hadar, and Michael Gilead. 2023. A computational text analysis investigation of the relation between personal and linguistic agency. *Communications Psychology*, 1(1):23.

Vinitra Swamy, Syrielle Montariol, Julian Blackwell, Jibril Frej, Martin Jaggi, and Tanja Käser. 2025. Intrinsic user-centric interpretability through global mixture of experts. In *The Thirteenth International Conference on Learning Representations*.

Shogo Tanaka. 2018. What is it like to be disconnected from the body?: A phenomenological account of disembodiment in depersonalization/derealization disorder. *Journal of Consciousness Studies*, 25(5-6):239–262.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Erico Tjoa and Cuntai Guan. 2020. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE transactions on neural networks and learning systems*, 32(11):4793–4813.

R. Uno and S. Imaizumi. 2025. Sensing minimal self in a sentence that involves the speaker. Preprint available at OSF.

Kush R Varshney and Homa Alemzadeh. 2017. On the safety of machine learning: Cyber-physical systems, decision sciences, and data products. *Big data*, 5(3):246–255.

Vivitha Vijayan and Sharvari Govilkar. 2019. A survey on author profiling techniques. *International Journal of Computer Sciences and Engineering*, 7:1065–1069.

Caroline Wang, Bin Han, Bhrij Patel, and Cynthia Rudin. 2023. In pursuit of interpretable, fair and accurate machine learning for criminal recidivism prediction. *Journal of Quantitative Criminology*, 39(2):519–581.

Theodore EA Waters and Robyn Fivush. 2015. Relations between narrative coherence, identity, and psychological well-being in emerging adulthood. *Journal of personality*, 83(4):441–451.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808*.

Han-xue Yang, Han-yu Zhou, Simon SY Lui, and Raymond CK Chan. 2024. Interoception in mental disorders: from self-awareness to interventions. *Proceedings of the European Academy of Sciences and Arts*, 3.

Dan Zahavi. 2008. *Subjectivity and selfhood: Investigating the first-person perspective*. MIT press.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

A Preliminary Experiments

To explore the feasibility of Self-aspect classification in natural language, we conducted a preliminary study focused on the Social Self (SS; “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” Caporusso et al., 2024), a subcomponent of our proposed ontology. We selected this category due to its relatively balanced presence in the dataset and its high inter-annotator agreement during annotation.

A.1 Dataset and Annotation

We used a publicly available dataset of 1,473 diary sub-entries (Li and Parikh, 2019), which we augmented with binary annotations for SS. Annotation combined manual labelling and automated classification using three versions of Gemma2 (Team et al., 2024)—personalised with psychological and phenomenological expertise. Inter-annotator agreement was assessed via Cohen's Kappa: 0.80 between human annotators, and 0.84–0.89 between human and model annotators.

A.2 Experimental Setup

We trained and evaluated six models using 10-fold cross-validation, combining three different classifiers—SVM, Logistic Regression (LR), and Naïve Bayes (NB)—with two types of feature representations. The first type comprised learned features, specifically TF-IDF weighted unigrams and bigrams. The second relied on predefined features derived from the LIWC-22 lexicon, specifically those previously identified as correlated with the Social Self aspect (Caporusso et al., 2024). Text preprocessing included converting all text to lowercase, removing punctuation, and applying z-score normalisation to the LIWC-derived features to ensure comparability across feature scales. To interpret the trained models, we employed feature importance techniques tailored to each algorithm: linear SVM coefficients for SVM, SHAP values for Logistic Regression, and permutation importance for Naïve Bayes.

A.3 Results

The best-performing model was the SVM trained on LIWC features, achieving a precision of 0.81 (STD = 0.03), recall of 0.82 (STD = 0.02), and F1-score of 0.81 (STD = 0.03) across 10 folds. It consistently outperformed all other models. Models using learned features (TF-IDF) performed slightly worse overall, with the SVM on learned features achieving an F1-score of 0.73 (STD = 0.04) and particularly lower recall. Statistical analysis confirmed the significance of these differences via a Friedman test (statistic = 44.26, $p < 0.001$) and pairwise Wilcoxon signed-rank tests (adjusted $p = 0.03$ for several comparisons). Feature importance analyses identified intuitive and interpretable markers of Social Self, including "we", social referents, affect terms, and pronoun use, aligning with prior findings and theoretical expectations.

A.4 Implications and Limitations

This pilot study demonstrates that interpretable models trained on psychologically grounded features can reliably identify expressions of Social Self in everyday texts. It also confirms the utility of a hybrid human-LLM annotation pipeline, especially in early dataset development. However, several limitations emerged. Performance is currently limited to binary classification of a single Self-aspect. The current study also relies on English-language data, which restricts immediate generalisability.