Future Generation Computer Systems 130 (2022) 114–127

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

SIMPA: Statement-to-Item Matching Personality Assessment from text☆

Matej Gjurković [a,*], Iva Vukojević [b,a], Jan Šnajder [a]

[a] University of Zagreb, Faculty of Electrical Engineering and Computing, Text Analysis and Knowledge Engineering Lab, Unska 3, 10000 Zagreb, Croatia

[b] University of Zagreb, Faculty of Humanities and Social Sciences, Department of Psychology, Ivana Lučića 3, 10000 Zagreb, Croatia

ARTICLE INFO

Article history:
Received 15 May 2021
Received in revised form 8 December 2021
Accepted 19 December 2021
Available online 23 December 2021

Keywords:
Text-based personality assessment
Natural language processing
Text analysis
Social media text
Realistic accuracy model
Personality prediction

ABSTRACT

Automated text-based personality assessment (ATBPA) methods can analyze large amounts of text data and identify nuanced linguistic personality cues. However, current approaches lack the interpretability, explainability, and validity offered by standard questionnaire instruments. To address these weaknesses, we propose an approach that combines questionnaire-based and text-based approaches to personality assessment. Our *Statement-to-Item Matching Personality Assessment* (SIMPA) framework uses natural language processing methods to detect self-referencing descriptions of personality in a target's text and utilizes these descriptions for personality assessment. The core of the framework is the notion of a trait-constrained semantic similarity between the target's freely expressed statements and questionnaire items. The conceptual basis is provided by the realistic accuracy model (RAM), which describes the process of accurate personality judgments and which we extend with a feedback loop mechanism to improve the accuracy of judgments. We present a simple proof-of-concept implementation of SIMPA for ATBPA on the social media site Reddit. We show how the framework can be used directly for unsupervised estimation of a target's Big 5 scores and indirectly to produce features for a supervised ATBPA model, demonstrating state-of-the-art results for the personality prediction task on Reddit.

1. Introduction

Personality refers to individual and stable differences in characteristic patterns of thinking, feeling, and behaving, referred to as *personality traits* [1]. These differences have been shown to correlate with many life outcomes, such as partner selection, job selection, religious or political inclination, and personal interests [2,3]. Personality is typically assessed with personality questionnaires [4]. These consist of a set of natural language statements, called *items*, which are positively or negatively associated with traits. For instance, "love to read challenging material" and "avoid philosophical discussions" are two such items that are associated positively and negatively, respectively, with intellect or openness to experience. Subjects respond to questionnaires on a Likert-type scale indicating the degree of agreement with the statement, i.e., how well these statements describe them. Widely used personality questionnaires, such as the NEO-PI-R [5] and the BFI [6], specifically measure five traits – extraversion, conscientiousness, agreeableness, neuroticism, and openness to experience – known as the Big 5 [7].

While personality questionnaires are a well-established instrument for assessing personality, research in personality psychology has also examined language as an alternate source of personality cues (e.g., [8,9]). The link between personality and language has long been acknowledged — the lexical hypothesis [10] posits a correlation between trait importance and the number of words in the language used to describe it, and the Big 5 traits were originally derived from latent structures of personality-relevant descriptions in language. As large amounts of user-generated text have become available online, more recent research has examined how these could be leveraged for personality assessment. In particular, social media, where users describe themselves and their personality while explaining their behavior, thoughts, or emotions, has been recognized as a valuable source for personality assessment. This has given rise to research that combines personality psychology and computer science in an attempt to automate personality assessment from large quantities of text. The potential of *automated text-based personality assessment* (ATBPA) is that it can not only efficiently analyze large

☆ This work has been supported in part by the Croatian Science Foundation under the project IP-2020-02-8671 PSYTXT ("Computational Models for Text-Based Personality Prediction and Analysis") and in part by the European Regional Development Fund under the grant KK.01.1.1.01.0009 DATACROSS.

* Corresponding author.
E-mail addresses: matej.gjurkovic@fer.hr (M. Gjurković), ivukojev@ffzg.hr (I. Vukojević), jan.snajder@fer.hr (J. Šnajder).

https://doi.org/10.1016/j.future.2021.12.014

amounts of text data but also identify nuanced personality cues that are often imperceptible to humans, such as linguistic style, while at the same time offering consistency well beyond human capabilities. For example, ATBPA avoids issues such as respondent fatigue, which limits the number of items that correlate with the quality of results [11], or respondents providing socially desirable responses [12]. This suggests that ATBPA methods could be used in a way that is complementary to standard personality assessment instruments.

Current ATBPA methods rely on natural language processing (NLP), which, in turn, relies heavily on machine learning (ML). The prevailing approaches are based on supervised ML using closed and open vocabulary [13,14]. In recent years the use of deep learning models has become the method of choice [15–17]. Supervised approaches learn to identify linguistic correlates of personality in text based on personality-labeled data. The obvious limitation of such models, however, is the need for labeled data. This poses a practical problem because there are few publicly available datasets, and those that are available have a number of deficits, such as a low number of users or texts per user or missing demographic data. While the lack of datasets certainly poses a practical challenge, ATBPA also suffers from more serious conceptual weaknesses. One fundamental weakness is the lack of explainability and validity. Until recently, these issues have not received much attention in ATBPA research, substantially limiting the use of ATBPA methods. However, another weakness is the exclusive focus on higher constructs of personality, mostly the domains (e.g., the Big 5's OCEAN), as opposed to more specific lower-level traits, such as aspects [18], facets [5], or nuances [19].

In this work, we propose a novel approach to ATBPA that aims to address the abovementioned weaknesses. It does so by being interpretable and explainable, providing more evidence for validity than existing ATBPA methods, and being able to output estimates for constructs from lower levels of the personality trait taxonomy. Conceptually, our approach combines questionnaire-based and text-based approaches to personality assessment. The existing ATBPA methods focus on predictive accuracy, which corresponds in psychology to convergent validity and constitutes only one source of validity evidence [20]. In contrast, questionnaires are expected to satisfy all psychometric properties, although they also have weaknesses, such as the time and effort required to fill out a questionnaire. The proposed method is an attempt to take the best of both: on the one hand we take items from personality questionnaires; on the other hand, we search for corresponding statements that the target has written in social media texts. The use of questionnaire items that are reliable and linkable to more specific personality traits provides background knowledge that enables interpretability and explainability. More concretely, matching the target's statements to questionnaire items provides interpretability (i.e., why the model decided a particular way), while items' pre-established links to traits provide explainability (i.e., why the decision makes sense). Furthermore, the use of statements avoids the ambiguity inherent in words. Words have been the primary unit of analysis of current ATBPA methods, but words may have different meanings in different contexts. Furthermore, they can be indicative of more than one trait, and their use generally depends on the topic of discussion. Our approach to ATBPA mitigates the ambiguity inherent in individual words by relying instead on entire statements. Once all of the target's statements corresponding to questionnaire items have been detected, they can be aggregated at any trait level (e.g., nuances, facets, aspects, dimensions), taking into account the polarity of the association (the item key) similar to the way scoring is done with standard personality questionnaires.

We develop the proposed approach into a framework, referred to as the *Statement-to-Item Matching Personality Assessment* (SIMPA). The framework operationalizes the idea of finding statements with self-reported descriptions of personality in the target's texts by matching these statements with items from a personality questionnaire. As the conceptual basis for SIMPA, we use the realistic accuracy model [21], which defines four sequential stages necessary for accurate personality judgments: cue relevance, availability, detection, and utilization. Adopting RAM allows us to draw parallels to the way a human judge would assess the personality of a target who has authored a text. In particular, starting from questionnaire items, which are relevant for traits (relevance), we take targets' statements from a particular source of text (availability), find the statements that correspond to items (detection), and use these detected statements for personality judgment (utilization). Unlike the original RAM, in which the judgment is based on one pass of the cues through the four stages, SIMPA allows for multiple passes via a feedback loop mechanism. The purpose of the feedback loop is to adapt the framework to a particular source of data (e.g., social media texts) by increasing its sensitivity in detecting the relevant cues.

On the technical side, the main challenge in SIMPA is to detect targets' statements corresponding to questionnaire items. A statement is a good match for a questionnaire item if the statement's semantic meaning is very similar to that of the item, thus making it indicative of the same trait. We conceptualize this idea with a notion of *trait-constrained semantic similarity*, which combines semantic similarity with knowledge about how a certain trait may be manifested. Determining this similarity automatically relates to several well-established NLP tasks, including paraphrase identification, semantic textual similarity, textual entailment, and natural language inference. Recent developments in NLP, particularly those based on deep representation learning, have produced remarkable progress in these tasks. We reap the benefits of this progress and use state-of-the-art (SOTA) NLP models that allow the matching of statements to items at accuracy levels that were impossible only a few years ago, allowing us to propose a framework that aims to provide interpretable, explainable, and valid personality assessment.

To demonstrate the feasibility of SIMPA, we present a simple proof-of-concept implementation for ATBPA on the social media site Reddit. We show how the framework can be used directly for unsupervised estimation of the target's Big 5 scores and indirectly to produce features for a supervised ATBPA model. We also investigate how extending the set of questionnaire items with more vernacular statements can improve the accuracy of personality assessments, and how the set of items can be extended even further by multiple passes through SIMPA's RAM stages. Our proof-of-concept implementation shows that even with many simplifying assumptions, SIMPA can achieve SOTA results in the personality prediction task.

In summary, the contributions of this work are twofold: we (1) define SIMPA, the first RAM-based framework for ATBPA that lays the groundwork for implementation of interpretable and explainable models for valid ATBPA, and (2) demonstrate the implementation of the SIMPA framework for ATBPA on Reddit. The encouraging results of the proof-of-concept implementation suggest that unsupervised models for ATBPA can be accurate, interpretable, and explainable, making them more attractive for both practical and research purposes than previous methods. Despite the encouraging results, this work represents only the first step in this direction, and we hope to encourage further research on the topic.

The rest of the article is organized as follows. In Section 2, we describe background and related work. Sections 3 and 4 describe

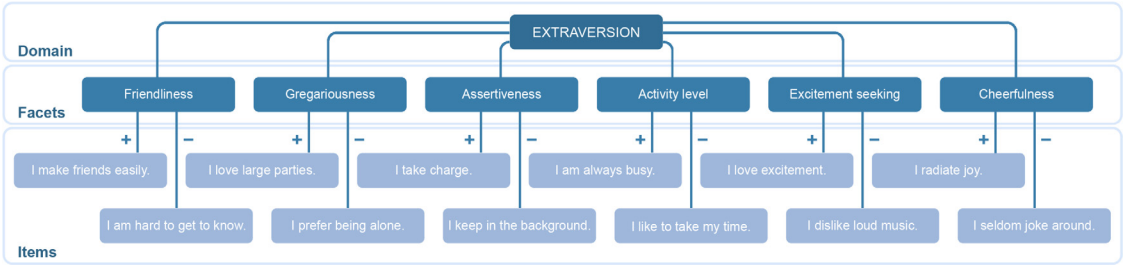


Fig. 1. Trait hierarchy for the extraversion domain in the IPIP-NEO questionnaire. Questionnaire items are linked to facets, which are grouped into domains. The links between the items and the facets are positively (+) or negatively (−) keyed.

SIMPA through the four RAM stages and the proof-of-concept implementation, respectively. In Section 5, we review the limitations and challenges of the proof-of-concept implementation and the SIMPA framework more generally and discuss options for future work. Section 6 presents the conclusion.

2. Background and related work

2.1. Personality models

The choice of personality traits to be assessed presupposes the choice of a personality model. There are numerous personality models based on different personality theories and derived using different methodologies. Hierarchical models with five [5,22] or six personality domains [23] are the most widely used in personality psychology research. The five domains, known as the Big 5 [5], include extraversion, conscientiousness, agreeableness, neuroticism (or emotional stability [22]), and openness to experience (or intellect [22]), while the HEXACO model [23] adds modesty/humility as the sixth domain. Some models additionally identify hierarchical trait structures within each domain consisting of aspects, facets, and nuances. More precisely, each domain is composed of two aspects [18]. Facets are narrower traits than aspects, and depending on the model, they are grouped into aspects [18] or directly into domains [5]. The narrowest form of traits is nuances, which correspond to groups of conceptually redundant questionnaire items or even individual items [19]. For instance, a person with an average score on extraversion may be very high on the assertiveness facet but very low on the cheerfulness facet. This person might be likely to take charge in a group assignment but remain serious during the assignment, whereas if the scores were reversed, the person would wait for others to take the lead but be the group's jokester.

Personality questionnaires are the most commonly used personality instrument [4]. Such questionnaires consist of a set of items — concrete, manifest variables that serve to measure traits (nuances, facets, aspects, domains), i.e., latent variables or constructs. Every item is labeled with the trait it measures, and a "key" indicates the polarity of the trait it measures ("+ " for the polarity of the trait's name, "−" for the opposite polarity; see Fig. 1). Because questionnaire items were originally constructed to fit into predetermined personality domains with the highest possible internal consistency, it is possible that some nuances did not make the cut, and there are ongoing efforts to construct more items in an attempt to include more nuances [24]. Many items are made publicly available in the largest repository of personality items, IPIP [25], which currently contains 3320 items.[1] Some of the items measure the Big 5 domains and their corresponding facets, such as those of the IPIP-NEO inventory [25]. Fig. 1 shows an example of the trait hierarchy for the extraversion domain from the IPIP-NEO inventory.

[1] https://ipip.ori.org/.

Our SIMPA framework is inspired by items and the rating process of personality questionnaires as well as the possibility of assessing traits at different levels of trait hierarchy. We use questionnaire items to find statements in the text of a target person that correspond to what the target's answer to that item would be had the questionnaire been administered to the target. Items are thus used as a prop in detecting the target's freely expressed statements that validly indicate personality traits. Unlike existing ATBPA approaches, SIMPA allows judgments to be made at multiple hierarchical levels of traits. The framework also enables the broadening of the set of items and nuances.

2.2. The realistic accuracy model (RAM)

An alternative to self-reports of personality is assessment based on observer reports, either other-reports (in which judges are usually closely related to the target) or zero-acquaintance judgments (observational studies in which raters are generally strangers to the target). In this case, it becomes important to understand how people make accurate judgments of other persons' traits [21]. A comprehensive conceptualization of this process is offered by the realistic accuracy model (RAM) proposed by Funder [21]. The model stipulates that the process from cue to accurate personality judgment evolves in four stages: the cue must be (1) relevant and (2) available, after which it must be (3) detected and ultimately (4) utilized by the judge. The target primarily influences the success of the relevance and availability stages, while the judge is mainly responsible for the success of the detection and utilization stages. The model has thus far been used to investigate how people judge the personality of others based on in-person (e.g., [26]), videotaped (e.g., [27]), or online behavioral cues (e.g., [28]). It has also been considered in the context of evaluating ATBPA methods (e.g., [9,29]). However, to the best of our knowledge, there is no ATBPA method that is specifically modeled on RAM.

The RAM model offers useful insights into validity problems that plague existing ATBPA methods [30]. Validity is concerned with whether an instrument measures the construct of interest [31]. In practice, validity is demonstrated using pieces of evidence that support the proposed interpretations of the obtained scores [32]. In current ATBPA approaches, validity is often endangered by using questionnaire personality scores as the ground truth, a lack of validity evidence based on content, and limited generalizability [20,30]. Most ATBPA studies thus far have used supervised ML models and thus human judgments (questionnaire scores from self-reports or other-reports) as the ground truth of personality. Because human judgments are prone to errors (e.g., response biases and cognitive fallacies), utilizing such judgments as the ground truth will make these errors propagate to predicted scores. Furthermore, as noted by Tay et al. [30], using questionnaire scores as the ground truth for both training and evaluating supervised prediction models invokes problematic circularity, which renders predictive accuracy useless for gauging

model validity. Another problem is the lack of validity evidence based on content, which relates to the fact that ATBPA methods often rely on linguistic signals that are only proxies of personality traits (e.g., topical interests) rather than the true, stable, and enduring patterns of personality. Finally, applying ATBPA to data harvested from a specific social media platform limits the generalizability of the findings to other platforms because the characteristics of each particular platform affect how personality traits are expressed by users, which in turn affects the personality judgments.

Unlike existing work on ATBPA, the SIMPA approach we propose is deliberately modeled after RAM with the aim of ensuring interpretability and validity, thereby addressing the three abovementioned limitations of existing ATBPA methods. Another interesting point of divergence from existing methods is that SIMPA uses elements of both self-reports and observational studies because it relies on items from self-report questionnaires to detect responses to these items in observed data.

2.3. Automated text-based personality assessment

Bleidorn and Hopwood [20] described the evolution of ATBPA in three generations. The first-generation ATBPA studies [33,34] provided initial insights into how personality is manifested in text, often based on small samples of authors and texts and using simple correlations or feature weights. The second generation is characterized by studies [15,35] aimed at improving predictive power for much larger samples, enabling larger statistical power with self-reported personality scores from validated instruments (e.g., Facebook myPersonality dataset [36]). Finally, the third generation of research [9,30,37] focuses on the different sources of validity and reliability and the added value of ATBPA in comparison to traditional personality assessment methods. Our work contributes to third-generation ATBPA methods.

An orthogonal perspective on ATBPA methods is offered by the RAM framework [21]. Specifically, the main points of differentiation between ATBPA methods from all three generations involve the types of relevant textual cues they use and the source of textual data. Prior work has considered a wide range of relevant cues (called "features" in ML parlance), which can be broadly categorized as features of content, style, or a combination of both. The most commonly used content features are words, phrases, word categories, and topics from a prespecified list (closed-vocabulary approaches) or extracted from the text itself (open-vocabulary approaches) [13,33]. In contrast, stylistic features capture the linguistic style of the text and typically include subword-level features (e.g., character ngrams), punctuation and special symbols (e.g., emojis, exclamation marks), and discourse-level indicators (e.g., readability indices, cohesion metrics, type-token ratios). The more recent ATBPA methods rely on deep learning representations, often in combination with the abovementioned linguistic features [17,38,39]. The second important factor is the source of the data. ATBPA has been used on emails [40], essays [41], forums [42], and, more recently, social media platforms such as Twitter [43], Facebook [9,15], and Reddit [44,45]. The source of data directly drives cue availability and hence the availability of the different types of features. For instance, emojis as cues are not available in great numbers in business emails, unlike in texts from social media. The SIMPA framework is agnostic to the choice of data source while also providing a means to detect source-specific relevant cues by the use of a feedback loop mechanism.

Conceptually, the works most similar to ours are that of Vu et al. [46] and Yang et al. [47]. The approaches they propose also relate the targets' text to questionnaire items by attempting to directly predict how targets will respond to items based on their texts. This is in contrast to our approach, where we find the most similar statements that targets expressed about themselves. The key difference, however, is that we propose a general framework for personality assessment that is agnostic to technical implementation and the intended application.

Issues of validity and reliability have thus far not received much attention in ATBPA research, and limited work incorporates standard practices of personality psychology research, such as checking for different sources of validity (e.g., content, discriminant) and reliability (i.e., test–retest) [9,30,37,48]. In ML, validity relates to model interpretability and explainability. Interpretability concerns the extent to which a cause and effect can be observed in a model, while explainability reflects the extent to which the inner workings of a model can be explained in human terms [49,50]. ATBPA models should be both interpretable and explainable because we certainly expect a decision on a characteristic of human subjects to be fair, supported by solid evidence, and easily explained if needed. The use of text as a source of personality cues makes meeting these requirements even more challenging because the use of language is generally influenced by both the author's personality and sociodemographics, such as age, gender, and cultural background. Existing ATBPA methods, however, are too easily swayed by the differences in distributions of some of these factors in the data sample, causing the models' predictions to be based on spurious associations between linguistic features and personality traits. The lack of interpretability and the existence of confounding variables can lead to inconclusive or biased results and can even pose ethical challenges. Addressing these challenges is particularly important as personality-aware models are becoming an integral part of widely used services such as dialog systems [51] and recommender systems [52].

Our framework satisfies the increased need for interpretable and explainable models that provide more evidence for validity than existing ATBPA methods. Furthermore, unlike other approaches, the implementation of our framework can be used both in an unsupervised and a supervised setup. It is also suited for the extraction of personality-indicative cues from large amounts of text per author and for the use of these cues to obtain judgments at different levels of the trait hierarchy.

3. The SIMPA framework

The Statement-to-Item Matching Personality Assessment (SIMPA) framework is based on the idea of matching a target's statements to questionnaire items or similar trait-relevant statements. This matching is performed between a pair of statements: a trait-indicative statement and a trait-relevant statement. We define a *trait-indicative statement* (TIS) as a statement extracted from the target's texts that can serve as a cue for personality assessment, and we define a *trait-relevant statement* (TRS) as a statement with known and valid links to a particular trait. The set of TRSes initially comprises only the questionnaire items but, as we discuss below, may subsequently be extended to include TISes deemed relevant for a particular trait.

Fig. 2 presents an overview of the SIMPA framework. The framework is based on RAM, which breaks down the personality judgments process into four stages: cue relevance, availability, detection, and utilization. The relevant cues are the TISes, which must be available in sufficient quantity in the source text for accurate judgments to be possible. The detection stage involves matching TISes to TRSes based on the notion of trait-constrained semantic similarity. Finally, in the utilization stage the detected TISes are assigned relevance scores, which are aggregated to produce trait-level scores.

Adopting RAM as the basis for SIMPA also helps in identifying the weak spots in the pipeline, i.e., how much and what kind of error is propagated from one stage to the next. In particular,

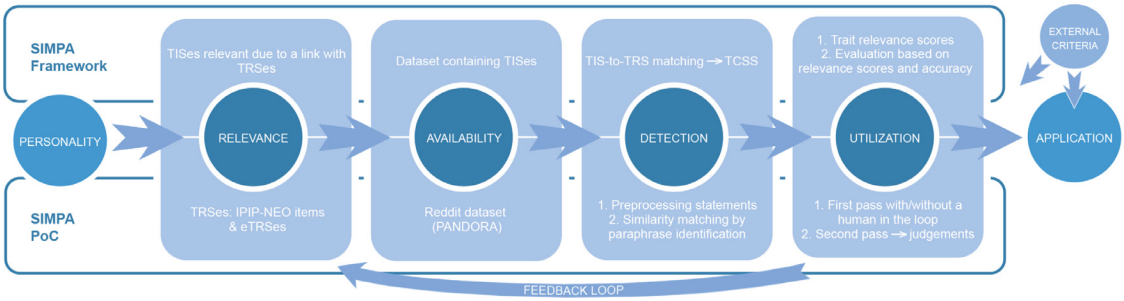


Fig. 2. A schematic showing the SIMPA framework (Section 3) and SIMPA's proof-of-concept (Section 4) split into four stages, relevance, availability, detection, and utilization, accompanied by a feedback loop and the application of obtained personality judgments. SIMPA thus conveys the idea of the (iterative) flow of relevant textual cues (TISes) through all stages, resulting in personality judgments that can be used in different applications.

we distinguish between two types of errors in SIMPA: *errors of omission* and *errors of commission*. The former involve not detecting a relevant cue or not linking a relevant cue to a trait, while the latter involve detecting an irrelevant cue or linking a relevant cue to an incorrect trait. Both types of errors can decrease the accuracy of personality judgment. We design our framework to allow for mitigation of these errors by incorporating a feedback loop mechanism that enables multiple passes through the four RAM stages to acquire data-specific TRSes, thereby increasing the likelihood of detecting TISes.

We next describe the four RAM stages of the SIMPA framework in more detail.

3.1. Relevance

Relevance is the starting point of RAM; for a person to accurately judge someone's traits, the person who is being judged must do something related to that trait [21]. For example, an extroverted person needs to have thoughts and emotions or take actions related to extroversion; otherwise, it would be difficult, if not impossible, to judge the person as extroverted. In SIMPA, we consider relevance to be an intrinsic property of a TIS and further assume that TIS relevance can only be established by appealing to external knowledge that serves as evidence of TIS validity. This is in contrast to most ATBPA research to date, in which cue relevance is equated with feature weights of a supervised model trained to predict questionnaire scores as labels, which leads to low generalizability and interpretability and endangers the validity of such methods. SIMPA sidesteps this problem by matching TISes to TRSes, where TRSes are questionnaire items carefully crafted and validated by psychologists to target specific personality traits. The rationale is that if a target uses a statement that conveys the same meaning as a questionnaire item, then the judge can consider this statement a valid personality cue supported by content-based validity evidence. In other words, a TIS is relevant for a trait by virtue of it being a good match to the TRS linked to this trait.

We distinguish between two broad types of TISes: *self-concepts of personality* and *personality manifestations*. Statements of self-concepts of personality are descriptive statements in which the target explicitly defines himself or herself in terms of personality (e.g., "I am a disciplined person"). In contrast, personality manifestations are not as self-categorizing as self-concepts and can further be divided into *trait references* and *individual acts*. We define a trait reference as a statement in which a speaker makes a generalizable reference to emotions, cognition, or behaviors known to be signals of personality (e.g., "I never procrastinate" as a cue for high self-discipline). In contrast, individual acts are statements pertaining to personality-indicative actions that the target describes having performed (referred to as *act reports*; e.g., "I finished all of my chores before noon" as a cue for high self-discipline) or which the target has performed (referred to as *act observations*; e.g., "I will answer you after work" as a cue for high self-discipline).

TISes offer at least three principled advantages over other textual cues used in previous work on ATBPA: they are (1) explicit and strong signals of personality, (2) less prone to contextual meaning alterations, and (3) domain-independent. Some TISes (notably, self-concepts of personality and trait references) are very explicit and strong signals of personality. However, even the less explicit TIS types correspond to TRSes and hence pertain to targets' self-references. TISes as cues are less prone to meaning alterations because they usually include sufficient context to reliably determine the semantic relationship between a TIS and a TRS, i.e., how well a TIS replicates a TRS. For example, the statement "I like to party" is less prone to meaning alteration than the word "party", which can also be part of the sentence "I'm not somebody who likes to party". Finally, TISes are, to a certain extent, domain-independent; that is, the same TISes appear across different social media domains, although their distribution certainly depends on the characteristics of the social media platform. We next discuss how this and other aspects affect TIS availability.

3.2. Availability

The relevant cues need to be made available to the judge to be used for trait judgment. In ATBPA, the availability of relevant cues primarily depends on the source of textual data, which ideally will maximize the quantity and quality of relevant cues. In SIMPA, this amounts to maximizing the likelihood of finding relevant TISes for a given set of TRSes. The availability of TISes is affected by several factors related to the communication environment, the target, and the trait.

When social media texts are used as a source of data, environment-related factors encompass a range of characteristics pertaining to the communication platform (e.g., reliance on textual communication, constraints on textual format, topical focus, synchronicity of communication) and its users (e.g., anonymity, demographics) [53]. For instance, on Twitter and Reddit, users mostly communicate via text messages, while on Instagram and TikTok, communication is mostly visual. Furthermore, some platforms impose a limit on message length (e.g., Twitter), while others impose no such limit (e.g., Reddit). User anonymity on the platform may also affect TIS availability; users tend to provide more information about themselves online when they believe they are anonymous [54,55]. Another important characteristic is whether the platform is topically focused, which may affect the distribution of available TISes. For instance, LinkedIn is typically used for work-related purposes; thus, one can expect work-related TISes to be more readily available on LinkedIn than on a topically diverse platform such as Reddit. The topical focus can

be linked to another factor that affects TIS availability, namely, the need for self-introduction, because discussions of certain topics require users to provide more background information about themselves. For instance, in discussions about personal relationships, users will typically provide more personality-related descriptions of themselves. TIS availability is also affected by the different social norms on the platforms as well as within their various subcommunities. For instance, in some communities (e.g., Facebook groups of first-time parents), an important part of interaction is encouraging each other, which can increase the availability of the TISes of the individual act type. Finally, the choice of platform can also affect the demographics of users, which in turn can be linked to personality differences as well as TIS availability [30]. For instance, the average Reddit user is a young male [56], which can increase the availability of TISes related to this age or gender group (e.g., "I enjoy being part of a loud crowd").

Other factors that affect TIS availability are target-related factors, such as social desirability and the situations in which targets find themselves. Some people are more prone to expressing themselves in a socially desirable manner than others [12]. This affects personality assessment [12]; consequently, one can expect more socially desirable TISes to be available for such targets. Additionally, certain situations evoke certain traits [57]; thus, situations that arise in online communication may affect cue availability. For example, being accused of lying can trigger cues for the cooperation facet that are either positively or negatively keyed.

TIS availability also depends on the traits of the target person. Generally, traits associated with many cues that are highly visible and frequently available are judged more accurately than those associated with less visible and less likely available cues [26]. In the case of TISes, it is reasonable to expect that not all traits, or both keys of all traits, have the same likelihood of being manifested in online text communication. It might be that it is inherently difficult for a person to be aware of a particular trait and consequently express it as a TIS. A case in point is self-consciousness: persons low on this trait are not bothered with how others perceive their behavior, and one could expect fewer negatively keyed TISes for self-consciousness than positively keyed TISes. Other trait-related availability factors may involve the social undesirability of some traits (e.g., negatively keyed self-discipline) or even bragging about desirable traits.

The above identified factors that affect TIS availability may also interact. For instance, anonymity can encourage users to talk about different, even shameful subjects and thus lower the social desirability effect [58]. The social network Reddit is a good example of a topically diverse platform on which users are anonymous, which might allow them to be more flexible about their behavioral choices (e.g., to show their weaknesses).

3.3. Detection

The detection stage in RAM is simply about the judge noticing relevant and available personality cues [59]. In SIMPA, the goal of the detection stage is to detect as many TISes as possible in the target's text. This is accomplished by attempting to match each target's statement against each statement from the set of TRSes (initially, the questionnaire items). The challenge here is the lexical gap between TISes and TRSes: TISes are vernacular statements extracted from naturalistic texts (e.g., "awkward situations are my game"), whereas TRSes are carefully worded statements that transmit the core meaning of emotional, cognitive, or behavioral patterns related to personality (e.g., "I am not bothered by difficult social situations" from IPIP-NEO). Because of this gap, relying only on verbatim matching of TISes against TRSes would result in many errors of omission.

In SIMPA, we operationalize TIS-to-TRS matching as the similarity between the two statements: if a statement is identical or highly similar to a given TRS, then it is likely a TIS and relevant for the same trait as that particular TRS. Note that the similarity in question is a graded notion that includes but also goes beyond the semantic equivalence of the two statements (i.e., a paraphrase relation). More precisely, the similarity between a TIS and a TRS must (1) cover semantic similarity between the two statements and (2) ensure that both statements are indicative of the same underlying trait. For example, the statement "I just hate crowded places" is both highly similar to the TRS "I avoid crowds" and indicative of the same extraversion facet (gregariousness) with the same key. On the other hand, "I work in a crowded place" is merely similar to that TRS but not indicative of the same trait. Conversely, the statement "I don't like being in shopping malls on Saturdays" is indicative of the same trait but is semantically less similar to the TRS in question. We call this type of similarity, which covers both semantic similarity and indicativeness of the same trait, *trait-constrained semantic similarity* (TCSS).

In general, the semantic similarity between two statements, as perceived by humans, depends on the contexts of both statements [60]. However, we conceptualize TCSS as a context-agnostic similarity measure that corresponds to the semantic similarity of two statements when they are used in their "typical" or "default" contexts. Even with this restriction, we note that the TCSS should capture various aspects of human and expert knowledge, including linguistic knowledge (natural language semantics and pragmatics), common-sense knowledge and inference, and sociocultural knowledge relevant for assessing personality traits. TCSS can thus be ideally conceived as measuring the Likert scale agreement level with a questionnaire item (a TRS) when the response is expressed as a natural language statement (a TIS). In other words, TCSS should ideally be high when the target's statement expresses agreement with the TRS and low otherwise. Providing a more concrete definition of the TCSS turns out to be difficult because of the many interacting linguistic phenomena involved. While a more precise characterization could in principle be derived by formalizing the desired properties of the TCSS as a similarity measure (e.g., that similarity should revert in case of negation, or that it should decrease if the statement is specialized or epistemically hedged), we leave this intriguing direction for future work.

Determining the degree of semantic similarity between two natural language statements in the presence of a lexical gap is a well-studied problem in NLP, where it has been variously framed as paraphrase detection, semantic textual similarity, and recognizing textual entailment [61–63]. Solving these tasks successfully requires language understanding capabilities that cover a wide range of linguistic phenomena, such as negation, synonymy and antonymy, semantic relations, and lexical entailment relations, as well as logical and common-sense inference [64,65]. Recent research on deep representation learning [66,67] has led to the development of highly accurate models for these tasks (e.g., [68]). Such models yield SOTA results and can be used off-the-shelf for this stage of the SIMPA framework because they are likely to do a good job of minimizing errors of both omission and commission. However, because these models were not specifically designed for TCSS, some errors will inevitably slip through. While this could be mitigated by adapting the existing SOTA models to the TCSS task (for example, by transfer learning [69]), the adaptation would not be trivial, and it would require a dataset labeled with TCSS scores.

The choice of the NLP model for TCSS directly determines the number of errors of omission and commission. An error of omission occurs when a relevant cue is not detected, which we expect to primarily occur when the model fails to assign a

high TCSS score to a vernacular statement that is semantically equivalent to a particular TRS. In contrast, an error of commission occurs when the model detects irrelevant statements as relevant by assigning it an overly high TCSS score. We expect this to happen in cases when very similarly worded statements hold different meanings. Obvious examples are statements containing negations or antonyms; for example, "I love art" might incorrectly match "I don't like art". In this case, the TIS is not indicative of the same trait as the particular TRS to which it is similar, yet it is still trait-indicative. This suggests that errors of commission can be further divided into two types: (1) *informative errors*, which occur when the TIS is not a good match to an existing TRS, but it is still trait-indicative to a judge, and as such, may even be considered a new TRS and (2) *noninformative errors*, which do not provide any useful trait-indicative information to a judge (e.g., a TIS "I feel welcomed" matched to the TRS "I make people feel welcome"). In the utilization stage of SIMPA, we leverage commission errors of the former type to iteratively improve the cue detection process.

3.4. Utilization

The last RAM stage in the SIMPA framework is utilization, where the detected TISes produced by a target are used for an accurate judgment of the target's personality traits. The main challenges here stem from the fact that the relevance of a TIS for a particular trait is a matter of degree, that several TISes may act as a cue for a single trait and hence their relevance must be combined somehow, that cue relevance from lower-level traits combine into cue relevance for higher-level traits, and that errors from earlier RAM stages now must be taken into account if the judgments are to be as accurate as possible.

With this in mind, utilization in SIMPA proceeds in two steps: (1) for each target, we recursively determine the relevance scores of cues for traits on all levels of the trait hierarchy, starting from TISes at the lowest level and moving up to the level of trait domains and taking into account both the relevance and quantity of cues available for the lower-level traits, and (2) deciding whether to make another pass through the RAM stages via the feedback loop mechanism based on the obtained relevance scores as well as external criteria. These steps account for the fact that personality traits are hierarchical constructs that comprise nuances on the bottom level that are linked to facets, which are linked to aspects and finally to trait domains (cf. Fig. 1). Moreover, these steps allow insights into the contribution of each lower-level trait because the judgment of traits at higher levels are estimated based on aggregate judgment for traits at lower levels.

Relevance scoring. We define the relevance score of a cue for a trait as the output of a scoring function that aggregates (e.g., by calculating the sum, the mean, or a weighted mean) the relevance scores of the cues for the traits at lower levels of the trait hierarchy. The lowest level is that of TRSes, and the relevance score for each TRS is obtained by aggregating the relevance scores of the matching TISes, which we define as their TCSS score (cf. Section 3.3). Intuitively, the relevance of a cue corresponding to a single TRS is determined by the combined evidence of TISes that match this TRS, taking into account the strength of that match as indicated by the TCSS score. This relevance score of TRSes may then be further aggregated to produce the relevance score of cues associated with traits at higher levels (e.g., facets or aspects) to which these TRSes are linked. In this manner, the cues' relevance scores can be propagated all the way up the hierarchy to the level of trait domains. More importantly, the judge can choose the hierarchical level of the traits for which she wishes to obtain relevance scores.

In addition to the relevance scores associated with lower-level traits, a relevance score generally depends on the context, which we take to include the quantity of cues as well as extralinguistic context (e.g., social norms in the subcommunity, demographic factors). The quantity is important here because the number of cues affects the confidence of judgment; the more TISes a judge has for judging each trait of each target, the more certain she will be in her judgment. For example, if TISes detected for a target include the statements "I have read Nietzsche", "I have read a scientific paper on that topic", and "I just finished a Stephen Hawking's book", a judge can be more certain about the target being high on the intellect facet than if only one of these TISes were detected. However, low cue quantity can be compensated for by high relevance. For instance, a single TIS matching the TRS "I love to read challenging material" is a strong enough cue to judge the target as high on intellect.

Feedback loop. To improve the accuracy of judgments in the presence of a lexical gap between TISes and TRSes (cf. Section 3.3), SIMPA extends RAM by adding a feedback loop from the utilization stage back to the availability stage. This makes it possible to iterate through the four RAM stages until a certain criterion has been met. More precisely, the feedback loop serves two intertwined purposes: (1) to adapt TIS detection to the source-text language by expanding the set of TRSes with TISes and (2) to increase the confidence in judgments in cases when the desired confidence criteria are not met, such as the minimum number of detected TISes per trait at some level of the trait hierarchy or the minimum level of the relevance score. The criterion for whether to loop back is generally a function of relevance scores and judgment accuracy. The desired level of accuracy is determined by external criteria, e.g., validity evidence based on correlations with other variables assessing the same trait (e.g., self-reports).

Prior to looping back, the set of TRSes is extended with those TISes detected in the text that carry sufficient trait-relevant information to serve as TRSes. The intuition here is that a larger set of TRSes in the next iteration of the RAM pipeline will detect more relevant and available cues. Specifically, there are two cases when a TIS can be promoted to a TRS: (1) TISes with a sufficiently high TCSS, which are essentially paraphrases of a TRS, and (2) TISes that are the result of an informative error from the detection stage (cf. Section 3.3), i.e., TISes that are highly indicative of a trait, but do not match existing TRSes well. The first case can be seen as domain adaptation because it makes it possible to extend the set of TRSes with statements specific to the source text. The second case can be seen as a way to cover new nuances of certain traits. The enlarged set of TRSes is then used for the next pass through the RAM stages, where the available and relevant cues are again detected as TISes and utilized to determine the relevance scores. The loop can be initiated multiple times until no new TISes can be promoted to TRSes or the desired confidence criteria are met.

4. Proof-of-concept implementation

This section describes our proof-of-concept implementation of the SIMPA framework for ATBPA on the social media platform Reddit. We choose Reddit specifically because its characteristics may facilitate the availability of TISes. The implementation uses questionnaire items from personality questionnaires (IPIP-NEO) as the relevant background knowledge (TRSes). However, we also investigate whether additional TRSes compiled by an expert and adapted to the language of Reddit can increase the relevance and quantity of the detected cues. The detection of relevant cues (TISes) via TIS-to-TRS matching is accomplished using a SOTA NLP model for paraphrase detection, serving as a proxy for TCSS computation. We also introduce a number of simplifying assumptions in the utilization stage. We make two passes through the pipeline to demonstrate the utility of the feedback loop mechanism for adapting the TRSes to the language used on Reddit. After the

second pass, we utilize the detected TISes to obtain the cue relevance scores for the Big 5 traits. Finally, we use these scores in two different applications: as raw estimates for Big 5 scores to showcase the use of SIMPA for unsupervised ATBPA and as features of a supervised ATBPA model, where we establish a new SOTA result for personality prediction on Reddit.

4.1. Reddit as a source of data

Reddit[2] is a popular social media platform with an estimated 430 million users.[3] Several characteristics make Reddit stand out among other social media sites as a valuable source of data for personality research: (1) user anonymity, which encourages users to express their thoughts more freely, (2) topical diversity, because Reddit comprises more than two million subcommunities called subreddits in which many aspects of personality can be expressed, and (3) the high quality and quantity of text per user, which increases the likelihood of the availability of personality-indicative cues.

For our proof-of-concept implementation, we use PANDORA [45], a dataset with self-reported personality scores of Reddit users acquired by mining their texts for disclosures of personality questionnaire results. Users' self-reported Big 5 scores are highly valuable because they provide convergent validity evidence for new ATBPA methods and enable a direct comparison of their predictive accuracy. The total number of users with self-reported Big 5 scores in PANDORA is $n = 1,608$. These users wrote a total of 1.3M comments, consisting of a total of 14.3M sentences. The extraction and aggregation of TISes from so much text for each user is an open challenge. However, we hypothesize that this is not a hindrance for our approach; on the contrary, SIMPA can only work better with more data because more data increases the availability of TISes.

4.2. Relevance

The SIMPA framework uses questionnaire items such as TRSes, which have been carefully crafted and validated by domain experts to target specific personality traits (cf. Section 3.1). The proof-of-concept implementation uses 300 items from the IPIP-NEO questionnaire [25]. To facilitate TIS-to-TRS matching, we convert each item to a self-referencing sentence by adding the pronoun "I" at the beginning of the item (e.g., "often feel blue" becomes "I often feel blue"). We have determined that doing so improves the detection of relevant cues.

4.3. Availability

In SIMPA, the availability of TISes depends on several factors related to the environment, the target, and the trait (cf. Section 3.2). Environment-related factors make Reddit especially suitable for ATBPA: anonymity (which mitigates the effect of social desirability and increases the need for self-introduction), the number of users, the number of texts per user, and topical diversity, all of which have a positive association with the expected number of TISes. To gain a sense of cue availability in the data, we approximate the number of candidates for the TISes as the number of sentences containing the pronoun "I". We find almost 5.2M such sentences, which amounts to 36.3% of sentences in the dataset. This relatively high ratio suggests that texts on Reddit may indeed contain a high number of TISes.

4.4. Detection

TIS detection involves matching the target's statements to TRSes by means of TCSS (cf. Section 3.3). We implement this in two steps: (1) text preprocessing and (2) similarity matching.

In preprocessing, we first use Sentencizer from Spacy [70] to split the comments of each Reddit user into sentences, treating each sentence as a possible TIS candidate. While clauses would perhaps better correspond to TRSes, we use entire sentences to simplify and speed up preprocessing because syntactic parsing would otherwise be required. Moreover, most of the SOTA NLP models were trained on sentence-level data. However, we expect this simplification to lead to an increase in errors of both omission and commission. We also filter the sentences to those containing the pronoun "I".

Once we obtain a set of sentences for all targets, we proceed to the similarity matching step. We choose to operationalize TCSS as paraphrase detection and semantic textual similarity, to be able to use off-the-shelf pretrained models for these tasks. This simplification allows us to avoid model fine-tuning, which would require a dataset labeled with correct TIS-to-TRS matches. It does, however, also lead to an increase in errors of commission (because some TISes will be incorrectly matched to a TRS) and errors of omission (because some TISes will not be correctly matched to a TRS). While there exist a multitude of NLP models for these tasks, we test three distinct and commonly used models implemented in the SentenceTransformer package [68]: (1) semantic textual similarity models based on Siamese networks [68], (2) Komnios word2vec model [71], and (3) the RoBERTa based paraphrase detection model [68]. We choose these three models because they are conceptually different and hence may yield different performances on our task.

Using these models, and in line with current practice in NLP [68], we compute the similarity (the TCSS score) between a TIS candidate and a TRS as follows. We encode (i.e., "embed") each TIS as a vector in a high-dimensional vector space derived so that semantically similar TISes have similar vectors. We then compute the semantic similarity across all TIS–TRS pairs simply as the cosine between their corresponding vectors. We store these similarities in a matrix whose rows correspond to the target's sentences and whose columns correspond to TRSes. For each target (Reddit user), we build one such matrix for each of the three NLP models. Thus, each sentence from target's comments will be associated with a distribution of similarities with all TRSes for each of the NLP models. However, for the sake of simplicity, we consider as matches only the TRSes with the highest similarity scores for each sentence. In effect, this means that we allow each TIS to match only to a single TRS.

The above-described implementation of the detection stage, albeit simplified, still involves a number of design decisions. The two main design decisions pertain to the NLP model used for matching and the threshold on cosine similarity. To investigate this further, we sampled 100 sentences with the highest similarity scores for each model and qualitatively analyzed their performance. This analysis provided two key insights: (1) the threshold for the similarity score directly influences the relevance and the quantity of TISes passing that threshold, and (2) the similarity scores for different models vary greatly across different TRSes. In particular, the paraphrase detection model exhibited the most stable performances in terms of relevance and quantity across different TRSes at a fixed similarity threshold, compared to the other two models. The paraphrase detection model was also the best of the three in capturing the TCSS and overall had a better ratio of informative to noninformative errors of commission. To further simplify the implementation, we decided to use only the paraphrase detection model.

[2] https://www.reddit.com.

[3] https://backlinko.com/reddit-users#reddit-statistics.

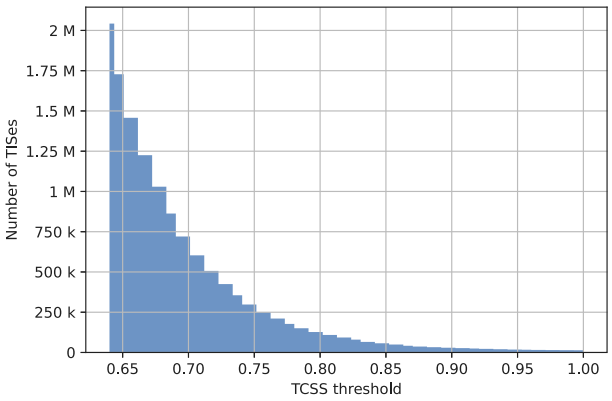


Fig. 3. The number of TISes detected at different TCSS thresholds.

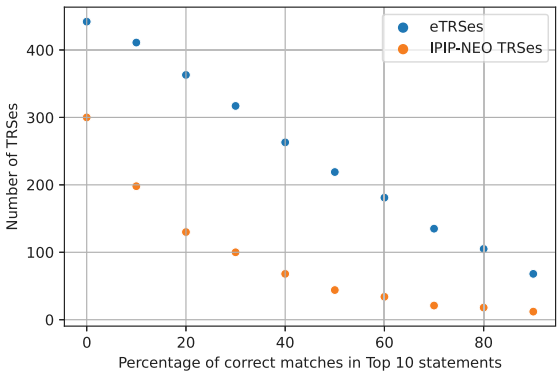


Fig. 4. The number of TRSes (IPIP-NEO TRSes in orange and eTRSes in blue) for the different proportions of correct matches in the top 10 TIS candidates with the highest TCSS. For example, there are 100 IPIP-NEO TRSes with more than 30% correctly matched TISes among the 10 top-ranked TISes ranked by the TCSS.

4.5. Utilization

Recall that the goal of utilization in SIMPA is to estimate relevance scores for a given target based on the detected TISes and their TRS counterparts (cf. Section 3.4). Thus, the first step is to calculate relevance scores for detected TISes based on the TCSS and the context. We experiment with two scenarios for choosing which TISes to use and how to calculate their relevance score. For the first scenario, we introduce the simplifying assumption that the relevance score is equal to the TCSS, effectively ignoring any context of TISes. We then select a threshold and proclaim all statements with a relevance score above that threshold to be TISes. Fig. 3 shows the number of TISes as a function of the similarity threshold, showing that the choice of the threshold considerably affects the number of TISes detected.

The above scenario simplifies both the detection (by using semantic similarity as a proxy for TCSS) and utilization (by equating the relevance score with TCSS). This raises the question of how much the accuracy of judgments could be improved if more sophisticated methods were used in these two stages. We investigate this in the second scenario by including in the loop a human expert (personality psychologist). For each TRS in the set of original 300 IPIP-NEO items, the expert was asked to annotate 20 statements from the dataset that were most similar to it according to the cosine similarity.

The annotation scheme used is essentially a binary scheme that distinguishes between correct and incorrect matches, with the latter further subcategorized into six types, five informative errors and one noninformative error. This gives a total of seven categories: (1) correct match (TIS is of the same level of generality and the same polarity as the TRS), (2) the same level of generality but opposite polarity, (3) less general and the same polarity, (4) less general and opposite polarity, (5) points to the average score item, (6) other item/facet of the same domain, and (7) other (a noninformative error). As an example, consider the TRS "I'm always prepared". Examples of corresponding TISes for the seven categories are as follows: (1) "I'm always prepared for whatever comes my way", (2) "I'm never prepared", (3) "I am prepared", (4) "I came unprepared", (5) "I'm never fully prepared, but I'm not unprepared either", (6) "I always arrange things in order", and (7) "I prepared a meal".

Fig. 4 shows the number of TRSes for different proportions of correct matches in the top 10 TIS candidates with the highest TCSS, serving as a proxy for the quality of TRSes. There is a considerable difference in the quality of the detected TISes per TRS. Only 44 out of 300 TRSes based on IPIP-NEO have more than 50% correct matches. Table 1 shows the proportions of annotated TIS candidates (top 20 statements matched to TRSes by means of TCSS) considering all annotation categories per personality domain. We observe marked differences in the proportions of correct matches and informative and noninformative errors across the personality domains detected with IPIP-NEO TRSes. These differences might be due to the skewed trait distributions in PANDORA [45] or simply because in some domains there are fewer IPIP-NEO TRSes that match to TISes from social media texts. Adapting TRSes to social media language could be a way of mitigating the latter.

Table 1
The proportions of annotated TIS candidates (top 20 statements matched to the IPIP-NEO TRSes and eTRSes by means of TCSS) broken down by annotation categories for each Big 5 domain (O — openness; C — conscientiousness; E — extraversion; A — agreeableness; N — neuroticism). Categories (columns): OK — a correct match; G− — statement of the same level of generality as the TRS but opposite polarity; LG+ — less general statement with the same polarity; LG− — less general with the opposite polarity; Avg — statements that point to the average TRS expression, SD — statements relevant for other items/facets of the same domain; NOK — a noninformative error.

Domains	TRSes		Informative errors					
		OK	G−	LG+	LG−	Avg	SD	NOK
		1	2	3	4	5	6	7
O	IPIP-NEO	24.0	7.7	21.2	5.0	1.1	3.8	37.2
	eTRS	36.2	4.5	16.1	4.0	2.6	2.3	34.3
C	IPIP-NEO	13.2	1.8	27.9	4.5	0.3	2.8	49.5
	eTRS	33.6	5.2	22.9	5.1	1.2	2.4	29.6
E	IPIP-NEO	21.2	3.4	25.1	5.6	0.6	0.8	43.2
	eTRS	33.9	5.7	23.5	4.6	4.0	1.1	27.2
A	IPIP-NEO	18.9	2.8	17.8	5.9	0.2	1.4	53.1
	eTRS	33.5	5.0	20.1	5.3	0.6	2.5	33.0
N	IPIP-NEO	22.8	1.9	35.2	4.3	1.2	1.0	33.5
	eTRS	29.1	3.6	32.7	4.8	2.8	0.6	26.3

In both scenarios, we make two passes through the RAM stages, and we expand the initial set of TRSes before making the second pass. In the first scenario, we use a threshold of .7 to decide which TISes to promote to TRSes, while in the second scenario we promote only the TISes that the expert has labeled as correctly matched to a TRS. In both cases, we link the new TRSes to traits of the TRSes to which they matched. After the second pass, we proclaim the statements to be TISes if they have similarity scores above the threshold of .8. As in the first step, we link TISes to traits linked to the most similar TRS.

After obtaining the TISes, we proceed to calculate the cue relevance scores for traits at different levels of the trait hierarchy. Here we introduce another simplification and fix the values of the relevance scores of all TISes to 1. We then sum all TIS scores into facet scores for each key (±1); finally, we do the same for

the domain traits. To obtain the relative expressions of traits of the targets in the dataset, we calculate the relative percentile score for each target based on other targets' scores. We do this by calculating the proportions of positively and negatively keyed scores for each trait, sorting targets by their proportion score, and then assigning percentile scores based on these proportion scores for each target and each trait.

4.6. Extending TRSes

The results of expert-based annotations in the second scenario showed that many of the original IPIP-NEO items often lead to a lower rate of correctly matched TISes. We therefore investigate whether the set of TRSes comprising IPIP-NEO items could be extended with additional expert-crafted statements. We refer to the set of additional items as eTRS. An expert psychologist compiled eTRSes based on the interpretations of items that form IPIP-NEO facets and what one would expect people who scored high or low on each of those facets to write in a comment on a social media platform such as Reddit to describe themselves. The eTRSes can be categorized into three types: (1) paraphrases of an exact IPIP-NEO item (e.g., "I don't try to be successful" for the IPIP-NEO item "I am not highly motivated to succeed"), (2) new items linked to an existing nuance (e.g., "I don't try hard" for a possibly existing nuance comprising "I do just enough work to get by" and "I put little time and effort into my work"), and (3) items that refer to a new nuance (e.g., "I don't care if I win or lose" as an item representing a new nuance for indifference to winning). In crafting the eTRSes, the expert was guided by the following principles: an eTRS should be (1) a strong signal of a facet, (2) more vernacular than standard items, (3) general (specific only when specificity makes the cue more relevant, e.g., "I hit stuff when I'm angry"), (4) worded to increase the likelihood that an NLP model relying on semantic similarity will correctly match TISes to it (i.e, avoid the use of metaphors, be concise), and (5) negated not merely through negations (e.g., "I'm always late" instead of "I'm never on time" for low self-discipline). As a result, eTRSes are mostly in the form of trait references and self-concepts of personality (cf. Section 3.1). With eTRSes written as trait references, we expect to mostly cover trait reference TISes and, to a lesser degree, individual acts and self-concepts (e.g., an eTRS "I read a lot of nonfiction" can link to an equivalent self-reference TIS, an act report TIS "I've read the best nonfiction book of 2020", an act observation TIS "I recommend you read more nonfiction", and a self-concept TIS "I'm a nonfiction kind of person"). The eTRSes worded as self-concepts are expected to match self-concept TISes, but they may also match trait references and, to a lesser extent, individual acts (e.g., an eTRS "I'm a creative person" might match to a self-reference TIS "I do a lot of creative DIY"). Note that eTRSes of the self concept type are entirely new compared to IPIP items.

We further refine eTRSes to improve their ability to detect correctly matched TISes. We do this in two steps by looking more closely at the semantic similarity (1) between the eTRSes and (2) between the eTRSes and a set of statements from a held-out portion of PANDORA comprising comments of targets without Big 5 ground truth scores ($n = 8684$). In the first step, we consider similarity scores between eTRSes and add to the set of eTRSes semantically highly similar sentences linked to two different facets or domains. This reduces the informative error in detecting TISes relevant to other TRSes or facets of the same domain and improves the detection of correct matches. In the second step, we take eTRSes and a set of sentences from the held-out Reddit dataset and rank those sentences by cosine similarity. We then expand the set of existing TRSes with those sentences if they had the same meaning but were worded differently. We removed or improved existing TRSes that were not an excellent hit for a trait (i.e., to which most similar sentences were not a good hit), considering both semantic similarity and whether the TRSes correctly linked to the presupposed traits.

The result was a list of 453 new items pertaining to Big 5 domains and 30 IPIP-NEO facets. The total effort to create and refine the eTRSes was about 100 h, and the annotation took about 40 h. The obtained list of eTRSes is preliminary. Although, the subsequent results (e.g., correlations between personality scores based on the matches with these items and gold-labeled scores) suggest that the list is valid, it should be validated with more scrutiny in further research.

We compare the relevance and quantity of the TISes detected using TRSes from the original IPIP-NEO items with that of TISes detected using eTRSes. The relevance is estimated based on annotations of detected TISes that are matched to TRSes in both sets (cf.) and the quantity of the number of TISes at different TCSS thresholds. Fig. 4 shows that eTRSes outperform IPIP-NEO items in terms of the number of correctly matched TISes. For instance, only 44 out of 300 (14.7%) IPIP-NEO items had 50% or more correct matches in the top 10 most similar statements, while the same was true for 219 out of 453 (48.3%) eTRSes. Table 1 shows a more detailed comparison of the relevance of detected TIS-es for IPIP-NEO and eTRSes. In addition to detecting more correct TISes, eTRSes yield fewer noninformative errors and more balanced proportions between correct matches, informative errors, and noninformative errors than IPIP-NEO TRSes. These results suggest that creating TRSes that are better suited for the way personality is expressed in language on social media is a worthy effort.

4.7. Application

We consider two applications of the SIMPA framework for the ATBPA task: (1) an unsupervised setup in which we use percentile scores from the utilization stage as estimates for targets' Big 5 scores and (2) a supervised setup in which we extend a regression model for text-based personality prediction with features based on relevance scores from the utilization stage. The latter application achieves SOTA performance on the PANDORA Reddit dataset [45].

Unsupervised setup. In the first application, we use percentile scores for each of the Big 5 traits obtained as aggregate relevance scores (cf. Section 4.5). We test whether estimations of Big 5 traits based on such a simple implementation of the relevance scoring function provide validity evidence by comparing them with self-reported Big 5 scores of targets in PANDORA.

As in the proof-of-concept implementation, we simplify the implementation and fix the values of some parameters. Specifically, we use only the TRSes that have at least one correct match in the top 10 of the most similar sentences. We also set a threshold of TCSS at 0.6. Finally, we consider only those targets with more than 10 detected TISes for at least one Big 5 trait because the number of detected TISes was highly correlated between different traits of the same target ($r > .85$). Fig. 5 shows the number of TISes detected across all targets, broken down by positive and negative keys for all Big 5 facets. We observe an imbalance between the numbers of positively and negatively keyed TISes for most facets and traits. For example, for a facet of cheerfulness, there are six times more positively than negatively keyed TISes. This discrepancy may indicate the higher availability of TISes of a particular key, the imbalance in the distribution of targets' personality traits in the dataset, or insufficient coverage of TRSes for one of the two keys of certain facets.

Table 2 shows the Pearson correlation coefficients between the estimated and ground truth Big 5 scores for TISes matched to

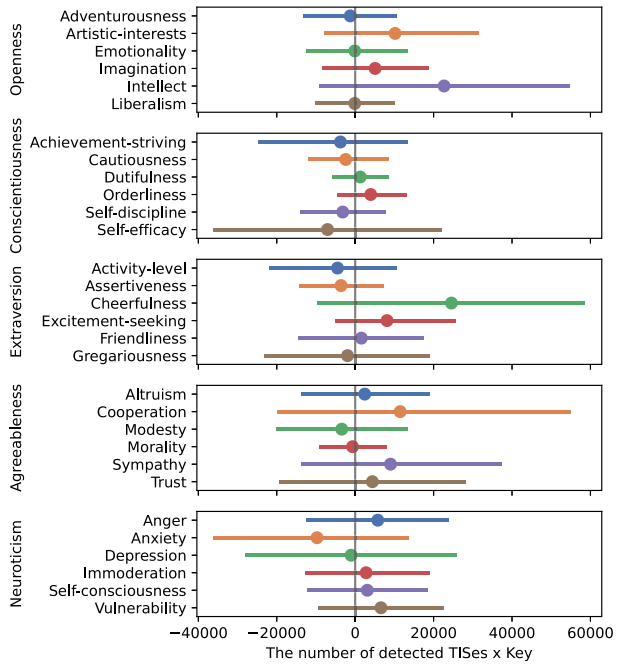


Fig. 5. The number of detected TISes per domain and facet across all targets. The dots represent the difference between the total number of positively and negatively keyed TISes for each facet. The lines show the range of the number of positively and negatively keyed TISes.

Table 2
Pearson correlation coefficients between percentile estimates obtained using different sets of TRSes and ground truth scores for the Big 5 domains (O – openness; C – conscientiousness; E – extraversion; A – agreeableness; N – neuroticism). Significant correlations ($p < .05$) are shown in bold.

TRSes	Domains	O	C	E	A	N
IPIP-NEO TRS	O	.099	.055	−.020	.136	**.203**
	C	**.231**	.179	**.227**	**.190**	**−.220**
	E	.066	.113	**.285**	.075	.013
	A	.003	.010	.053	**.204**	−.088
	N	−.088	**−.186**	−.146	**−.253**	**.175**
eTRS	O	**.143**	.045	.074	.084	−.061
	C	−.006	**.160**	.018	.052	−.104
	E	.040	.064	**.241**	.034	−.122
	A	.012	.054	.067	**.154**	.073
	N	.036	−.025	−.003	−.072	.121
Combined	O	**.141**	.013	.067	**.165**	.066
	C	**.126**	.102	**.102**	−.003	−.089
	E	.070	**.075**	**.226**	**.138**	−.041
	A	.027	−.021	**.106**	**.185**	.036
	N	−.025	−.064	−.063	−.086	**.106**

TRSes based on IPIP-NEO items, eTRSes, and the combination of both. Because of the abovementioned constraints, the sample size of targets is smaller than the starting $n = 1,608$ targets with self-reported Big 5 scores. More specifically, the number of targets per experiment is 155 for IPIP-NEO TRSes, 280 for eTRSes, and 399 if we consider both sets of TRSes. The results show that even with many simplifications introduced in the implementation of the detection and utilization stages of the SIMPA framework (cf. Sections 4.4 and 4.5), there is evidence for both convergent and discriminant validity. Concretely, using only eTRSes we manage to achieve significant positive correlations for all traits of interest (convergent validity) in addition to neuroticism and no significant correlations between traits (discriminant validity). However, the magnitude of the effect sizes shows that there is much room for improvement.

Table 3
Pearson correlation coefficients between predicted Big 5 scores and ground truth scores for PANDORA best performing prediction models with and without SIMPA TIS-based features. Significant differences in correlations ($p < .05$) are shown in bold.

Domains	PANDORA-best	PANDORA-best+SIMPA
Openness	.265	.285
Conscientiousness	.273	.304
Extraversion	.387	**.458**
Agreeableness	.270	.287
Neuroticism	.283	.312

Supervised setup. In the supervised setup, we use the currently best-performing model on the PANDORA dataset [45] and extend it with TIS-based features. The model is a Ridge regression using Tf–Idf weighted unigrams and predictions based on MBTI and Enneagram classifiers as features.[4] We declare as TIS all sentences that have a TCSS of at least 0.6 when matched to the initial set of TRSes consisting of both IPIP-NEO TRSes and eTRSes. We also include TISes that are matched with a TCSS above that threshold to the expanded set of TRSes after one iteration of the feedback loop based on correctly matched statements as validated by a human expert. As model features, we use the outputs of the relevance scoring function computed as sums of positively and negatively keyed TISes for all Big 5 facets and domains. We then apply PCA (with 10 principal components) separately on the raw relevance scores matrix and on a row-normalized scores matrix for all targets. This enables us to obtain a fixed set of 20 dense features for every target in the dataset, alleviating the problem of TIS sparsity. To allow for a direct comparison of the results, we adopt the same cross-validation procedure and the same folds as in [45].

Table 3 shows the Pearson correlation coefficients between the predicted Big 5 scores and ground truth scores for all targets ($n = 1,608$) of the original PANDORA best-performing models and our SIMPA-based extension of this model. Our final model, using only 20 additional TIS-based features, achieves new SOTA results on all five Big 5 traits. A statistically significant improvement ($p = 0.018$, two-tailed Steiger's test for dependent correlations [73]) is achieved for the extraversion trait, reaching a correlation of .458. These results are encouraging for two reasons. First, they show that the TIS-based features are complementary to word unigrams and features based on predictions for other personality models. Second, considering the many simplifications introduced in the detection and utilization stages, these results suggest that significant room remains for further improvements in predictive accuracy.

5. Discussion

The proposed SIMPA framework relies on detecting TISes via their match with TRSes, both expressed in natural language. Making this work on noisy social media data requires solutions to novel and nontrivial NLP tasks. In the proof-of-concept implementation, we demonstrated the viability of the approach and introduced several simplifying assumptions along the way. More precisely, in the detection stage, we simplified the preprocessing and modeled TISes as sentences, whereas clauses would be more appropriate. We operationalized the TCSS computation as a paraphrase detection task, making TIS-to-TRS matching generally agnostic to the trait-constraint requirement. We also decomposed the matching into two steps (embedding sentences

[4] The Myers–Briggs Type Indicator (MBTI) model [72] and Enneagram are type-based personality models. While widely used in the general public, they are discredited by most personality psychologists.

followed by computing the cosine similarity), although doing this jointly could be more efficient. In the utilization stage, the relevance scoring function was oversimplified: the TCSS score was used as the relevance score of TISes, and only the maximum TCSS score was considered rather than the whole distribution of similarities available to the judge. Furthermore, the base rates of expected TISes per TRS, TRSes per facet, and facet at the domain level were not accounted for, as the weights were fixed at 1.

While viable, the proof-of-concept implementation is too simplistic to be useful in real applications. Future implementations should consider building on more sophisticated models that are better suited to the complexity of the task. Efforts should focus on the relevance scoring function and the TCSS. Both of these tasks are novel, and their difficulty remains unknown, but developing models that approach human-level performance will likely require substantial effort. The relevance scoring function should encompass extralinguistic contextual information, while an effective TCSS implementation would have to take into account discourse-level phenomena, such as co-references and ambiguities resolvable beyond the sentence level. A promising approach to this end is to train a TCSS model using metric learning [74,75] on an expert-annotated dataset, similar to the work of Reimers and Gurevych [68].

Regarding the overall SIMPA framework, its limitations and applicability have yet to be thoroughly tested. However, certain challenges and opportunities for future work are already evident. The main challenge stems from the fact that, compared to self-reports where each participant responds to every item, our method does not cover TISes for all TRSes for each participant. This impairs content validity evidence, which concerns the cues being construct-relevant and fully covering the construct (e.g., TISes should be relevant for a facet and should cover all facets' nuances) [32]. The latter will be difficult to accomplish in our framework. On the other hand, the framework facilitates evidence based on content in the sense of cues being construct-relevant, due to the information it provides on the match between cues and valid items.

Future work could investigate how the framework can be adopted to different domains and its psychometric properties. Domain adaptation could be explored by using the data available from multiple domains. Similarly, the psychometric property of reliability could be evaluated by comparing the framework's performance on data from different time periods. This test–retest setting further leads to a possible framework extension for longitudinal research. Another interesting direction for future work is the application of the framework for assessing constructs other than personality. Any construct for which questionnaire items have been developed, could in principle, be assessed using SIMPA — for instance, attitudes, mental health, or even symptoms of physical health. This points to opportunities for the framework's use in different research areas.

6. Conclusion

The Statement-to-Item Matching Personality Assessment (SIMPA) framework combines questionnaire-based and text-based approaches to personality assessment. It uses natural language processing methods to detect self-referencing descriptions of personality in targets' texts and utilizes these for personality assessment. To the best of our knowledge, this is the first statement-level text-based personality assessment approach that focuses on ensuring interpretability, explainability, and validity, and the first framework modeled directly after the realistic accuracy model (RAM) of personality judgment. We demonstrated the feasibility of SIMPA via a proof-of-concept implementation for text-based personality assessment on Reddit. We used it to directly obtain estimates of the targets' Big 5 scores, which we correlated to ground truth Big 5 scores based on self-reports, achieving positive correlations for the traits of interest and demonstrating convergent validity evidence. The second application of SIMPA was a supervised setup in which we expanded a regression model and achieved SOTA results with features based on SIMPA's personality judgments. Future work should aim to build more sophisticated NLP models as solutions for new and complex NLP tasks posed by this framework. Another interesting direction for future work is the use of SIMPA for assessing constructs other than personality, i.e., any construct measurable with questionnaire items.

CRediT authorship contribution statement

Matej Gjurković: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization, Software. **Iva Vukojević:** Conceptualization, Investigation, Data curation, Visualization, Writing – original draft. **Jan Šnajder:** Conceptualization, Writing – original draft, Investigation, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

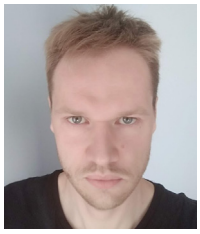
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] D.C. Funder, Accurate personality judgment, Dir. Psychol. Sci. 21 (2012) 177–182.
[2] L.M. Larson, P.J. Rottinghaus, F.H. Borgen, Meta-analyses of big six interests and big five personality factors, J. Vocat. Behav. 61 (2002) 217–239.
[3] C.J. Soto, How replicable are links between personality traits and consequential life outcomes? The life outcomes of personality replication project, Psychol. Sci. 30 (2019) 711–727.
[4] R.J. Larsen, D.M. Buss, Personality Psychology: Domains of Knowledge About Human Nature, McGraw-Hill Publishing, 2009.
[5] J.P.T. Costa, R.R. McCrae, Domains and facets: Hierarchical personality assessment using the revised neo personality inventory, J. Personal. Assess. 64 (1995) 21–50.
[6] O.P. John, E.M. Donahue, R.L. Kentle, Big five inventory, J. Personal. Soc. Psychol. (1991).
[7] L.R. Goldberg, Language and Individual Differences: The Search for Universals in Personality Lexicons, Beverly Hills, 1981, pp. 141–165.
[8] Y.R. Tausczik, J.W. Pennebaker, The psychological meaning of words: LIWC and computerized text analysis methods, J. Lang. Soc. Psychol. 29 (2010) 24–54.
[9] G. Park, H.A. Schwartz, J.C. Eichstaedt, M.L. Kern, M. Kosinski, D.J. Stillwell, L.H. Ungar, M.E. Seligman, Automatic personality assessment through social media language., J. Personal. Soc. Psychol. 108 (2015) 934.
[10] G.W. Allport, H.S. Odbert, Trait-names: A psycho-lexical study, Psychol. Monogr. 47 (1936) i.
[11] M. Galesic, M. Bosnjak, Effects of questionnaire length on participation and indicators of response quality in a web survey, Public Opin. Q. 73 (2009) 349–360.
[12] R.R. Holden, J. Passey, Socially desirable responding in personality assessment: Not necessarily faking and not necessarily substance, Pers. Individ. Differ. 49 (2010) 446–450.
[13] A.H. Schwartz, J.C. Eichstaedt, M.L. Kern, L. Dziurzynski, S.M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M.E. Seligman, et al., Personality, gender, and age in the language of social media: The open-vocabulary approach, PLoS One 8 (2013) e73791.
[14] J.W. Pennebaker, R.L. Boyd, K. Jordan, K. Blackburn, The Development and Psychometric Properties of LIWC2015, Technical Report, University of Texas at Austin, 2015.
[15] V. Lynn, N. Balasubramanian, H.A. Schwartz, Hierarchical modeling for user personality prediction: The role of message-level attention, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 5306–5316, Online.
[16] D. Xue, L. Wu, Z. Hong, S. Guo, L. Gao, Z. Wu, X. Zhong, J. Sun, Deep learning-based personality recognition from text posts of online social networks, Appl. Intell. (2018) 1–15.

[17] Y. Mehta, N. Majumder, A. Gelbukh, E. Cambria, Recent trends in deep learning based personality detection, Artif. Intell. Rev. 53 (2020) 2313–2339.
[18] C.G. DeYoung, L.C. Quilty, J.B. Peterson, Between facets and domains: 10 aspects of the big five, J. Personal. Soc. Psychol. 93 (2007) 880.
[19] R.R. McCrae, A more nuanced view of reliability: Specificity in the trait hierarchy, Pers. Soc. Psychol. Rev. 19 (2015) 97–112.
[20] W. Bleidorn, C.J. Hopwood, Using machine learning to advance personality assessment and theory, Pers. Soc. Psychol. Rev. 23 (2019) 190–203.
[21] D.C. Funder, On the accuracy of personality judgment: a realistic approach, Psychol. Rev. 102 (1995) 652.
[22] L.R. Goldberg, An alternative description of personality: The big-five factor structure, J. Personal. Soc. Psychol. 59 (1990) 1216.
[23] K. Lee, M.C. Ashton, Psychometric properties of the HEXACO personality inventory, Multivar. Behav. Res. 39 (2004) 329–358.
[24] R. Mõttus, T. Bates, D.M. Condon, D. Mroczek, W. Revelle, Leveraging a more nuanced view of personality: Narrow characteristics predict and explain variance in life outcomes, 2017, PsyarXiv.
[25] L.R. Goldberg, A Broad-Bandwidth, Public Domain, Personality Inventory Measuring the Lower-Level Facets of Several Five-Factor Models, Tilburg, The Netherlands, 1999, pp. 7–28.
[26] T.D. Letzring, S.M. Wells, D.C. Funder, Information quantity and quality affect the realistic accuracy of personality judgment, J. Personal. Soc. Psychol. 91 (2006) 111.
[27] D.R. Carney, C.R. Colvin, J.A. Hall, A thin slice perspective on the accuracy of first impressions, J. Res. Personal. 41 (2007) 1054–1072.
[28] S.D. Gosling, A.A. Augustine, S. Vazire, N. Holtzman, S. Gaddis, Manifestations of personality in online social networks: Self-reported facebook-related behaviors and observable profile information, Cyberpsychology Behav. Soc. Netw. 14 (2011) 483–488.
[29] W. Youyou, M. Kosinski, D. Stillwell, Computer-based personality judgments are more accurate than those made by humans, Proc. Natl. Acad. Sci. 112 (2015) 1036–1040.
[30] L. Tay, S.E. Woo, L. Hickman, R.M. Saef, Psychometric and validity issues in machine learning approaches to personality assessment: A focus on social media text mining, Eur. J. Pers. 34 (2020) 826–844.
[31] D. Borsboom, G.J. Mellenbergh, J. Van Heerden, The concept of validity, Psychol. Rev. 111 (2004) 1061.
[32] A.E.R. Association, A.P. Association, N.C. on Measurement in Education, Standards for Educational and Psychological Testing, American Educational Research Association, Washington D.C, 2014.
[33] J.W. Pennebaker, L.A. King, Linguistic styles: Language use as an individual difference, J. Personal. Soc. Psychol. 77 (1999) 1296.
[34] S. Argamon, S. Dhawle, M. Koppel, J. Pennebaker, Lexical predictors of personality type, in: Proceedings of the Joint Annual Meeting of the Interface and the Classification Society, 2005, pp. 1–16.
[35] M. Kosinski, S.C. Matz, S.D. Gosling, V. Popov, D. Stillwell, Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines., Am. Psychol. 70 (2015) 543.
[36] M. Kosinski, D. Stillwell, T. Graepel, Private traits and attributes are predictable from digital records of human behavior, Proc. Natl. Acad. Sci. 110 (2013) 5802–5805.
[37] V. Kulkarni, M.L. Kern, D. Stillwell, M. Kosinski, S. Matz, L. Ungar, S. Skiena, H.A. Schwartz, Latent human traits in the language of social media: An open-vocabulary approach, PLoS One 13 (2018).
[38] Y. Mehta, S. Fatehi, A. Kazameini, C. Stachl, E. Cambria, S. Eetemadi, Bottom-up and top-down: Predicting personality with psycholinguistic and language model features, in: 2020 IEEE International Conference on Data Mining, ICDM, 2020, pp. 1184–1189, http://dx.doi.org/10.1109/ICDM50108.2020.00146.
[39] A. Kazameini, S. Fatehi, Y. Mehta, S. Eetemadi, E. Cambria, Personality trait detection using bagged SVM over BERT word embedding ensembles, 2020, http://arxiv.org/abs/2010.01309[arXiv:2010.01309].
[40] J. Oberlander, A.J. Gill, Language with character: A stratified corpus comparison of individual differences in e-mail communication, Discourse Process. 42 (2006) 239–270.
[41] K. Luyckx, W. Daelemans, Personae: A corpus for author and personality prediction from text, in: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), 2008, pp. 2981–2987.
[42] F. Iacobelli, A.J. Gill, S. Nowson, J. Oberlander, Large scale personality classification of bloggers, in: Affective Computing and Intelligent Interaction, Springer, 2011, pp. 568–577.
[43] P.-H. Arnoux, A. Xu, N. Boyette, J. Mahmud, R. Akkiraju, V. Sinha, 25 tweets to know you: A new model to predict personality with social media, in: Proceedings of the Eleventh International AAAI Conference on Web and Social Media, 2017, pp. 472–475.
[44] M. Gjurković, J. Šnajder, Reddit: A gold mine for personality prediction, in: Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, Association for Computational Linguistics, New Orleans, Louisiana, USA, 2018, pp. 87–97.
[45] M. Gjurković, M. Karan, I. Vukojević, M. Bošnjak, J. Šnajder, PANDORA Talks: Personality and demographics on Reddit, in: Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media, Association for Computational Linguistics, 2021, pp. 138–152.
[46] H. Vu, S. Abdurahman, S. Bhatia, L. Ungar, Predicting responses to psychological questionnaires from participants' social media posts and question text embeddings, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, 2020, pp. 1512–1524, Online.
[47] F. Yang, T. Yang, X. Quan, Q. Su, Learning to answer psychological questionnaire for personality detection, in: Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 1131–1142.
[48] P. Novikov, L. Mararitsa, V. Nozdrachev, Inferred vs traditional personality assessment: Are we predicting the same thing?, 2021.
[49] L.H. Gilpin, D. Bau, B.Z. Yuan, A. Bajwa, M. Specter, L. Kagal, Explaining explanations: An overview of interpretability of machine learning, in: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics, DSAA, IEEE, 2018, pp. 80–89.
[50] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, G.-Z. Yang, XAI—Explainable artificial intelligence, Science Robotics 4 (2019).
[51] S. Roller, E. Dinan, N. Goyal, D. Ju, M. Williamson, Y. Liu, J. Xu, M. Ott, E.M. Smith, Y.-L. Boureau, J. Weston, Recipes for building an open-domain chatbot, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics, 2021, pp. 300–325, Online.
[52] S. Dhelim, N. Aung, M.A. Bouras, H. Ning, E. Cambria, A survey on personality-aware recommendation systems, Artif. Intell. Rev. (2021).
[53] L.A. McFarland, R.E. Ployhart, Social media: A contextual framework to guide research and practice, J. Appl. Psychol. 100 (2015) 1653.
[54] A.N. Joinson, Self-disclosure in computer-mediated communication: The role of self-awareness and visual anonymity, Eur. J. Soc. Psychol. 31 (2001) 177–192.
[55] W.-B. Chiou, Adolescents' sexual self-disclosure on the internet: Deindividuation and impression management, Adolescence 41 (2006).
[56] M. Duggan, A. Smith, 6% of online adults are Reddit users, Vol. 3, Pew Internet & American Life Project, 2013, pp. 1–10.
[57] J.J. Denissen, M. Luhmann, J.M. Chung, W. Bleidorn, Transactions between life events and personality traits across the adult lifespan., J. Personal. Soc. Psychol. 116 (2019) 612.
[58] L. Berdychevsky, G. Nimrod, Let's talk about sex: Discussions in seniors' online communities, J. Leis. Res. 47 (2015) 467–484.
[59] T.D. Letzring, D.C. Funder, The realistic accuracy model, in: The Oxford Handbook of Accurate Personality Judgment, 2019.
[60] E. Pavlick, T. Kwiatkowski, Inherent disagreements in human textual inferences, Trans. Assoc. Comput. Linguist. 7 (2019) 677–694.
[61] I. Dagan, B. Dolan, B. Magnini, D. Roth, Recognizing textual entailment: Rational, evaluation and approaches, J. Nat. Lang. Eng. 4 (2010).
[62] I. Androutsopoulos, P. Malakasiotis, A survey of paraphrasing and textual entailment methods, J. Artificial Intelligence Res. 38 (2010) 135–187.
[63] E. Agirre, D. Cer, M. Diab, A. Gonzalez-Agirre, SemEval-2012 task 6: A pilot on semantic textual similarity, in: *SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Vol. 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012), 2012, pp. 385–393.
[64] P. LoBue, A. Yates, Types of common-sense knowledge needed for recognizing textual entailment, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011, pp. 329–334.
[65] E. Cabria, B. Magnini, Decomposing semantic inference, in: Linguistic Issues in Language Technology, Vol. 9, 2014-Perspectives on Semantic Representations for Textual Inference, 2014.
[66] Y. Bengio, A. Courville, P. Vincent, Representation learning: A review and new perspectives, IEEE Trans. Pattern Anal. Mach. Intell. 35 (2013) 1798–1828.
[67] Y. Goldberg, Neural network methods for natural language processing, Synth. Lect. Hum. Lang. Technol. 10 (2017) 1–309.
[68] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using Siamese BERT-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, 2019, pp. 3973–3983.
[69] S. Ruder, M.E. Peters, S. Swayamdipta, T. Wolf, Transfer learning in natural language processing, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials, 2019, pp. 15–18.
[70] M. Honnibal, I. Montani, S. Van Landeghem, A. Boyd, Spacy: Industrial-strength natural language processing in python, 2020.

[71] A. Komninos, S. Manandhar, Dependency based embeddings for sentence classification tasks, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, San Diego, California, 2016, pp. 1490–1500.
[72] I.B. Myers, M.H. McCaulley, A.L. Hammer, Introduction to Type: A Description of the Theory and Applications of the Myers-Briggs Type Indicator, Consulting Psychologists Press, 1990.
[73] J.H. Steiger, Tests for comparing elements of a correlation matrix, Psychol. Bull. (1980) 245–251.
[74] P. Neculoiu, M. Versteegh, M. Rotaru, Learning text similarity with siamese recurrent networks, in: Proceedings of the 1st Workshop on Representation Learning for NLP, Association for Computational Linguistics, Berlin, Germany, 2016, pp. 148–157.
[75] E. Hoffer, N. Ailon, Deep metric learning using triplet network, in: A. Feragen, M. Pelillo, M. Loog (Eds.), Similarity-Based Pattern Recognition, Springer International Publishing, Cham, 2015, 84–92.



Matej Gjurković is a Ph.D. student and a research associate at the University of Zagreb, Faculty of Electrical Engineering and Computing (FER). He has been involved in several national and EU projects as a research engineer. His research interests focus on using natural language processing methods for personality prediction and analysis based on social media text.

Iva Vukojević is a Ph.D. candidate in psychology at the University of Zagreb, Faculty of Humanities and Social Sciences, and a research associate at the University of Zagreb, Faculty of Electrical Engineering and Computing (FER). Her research interests revolve around exploring human behavior in a text-based digital footprint form, focused on personality analysis and prediction from social media text.

Jan Šnajder received his Ph.D. in computing from the University of Zagreb, Faculty of Electrical Engineering and Computing (FER) in 2010. From 2016 he is holding the position of an associate professor at FER at the Department of Electronics, Microelectronics, Computer and Intelligent Systems. His research interests focus on natural language processing and machine learning. He has been the principal investigator of several projects funded by the Croatian Science Foundation.