On Text-based Personality Computing: Challenges and Future Directions

Qixiang Fang[1], Anastasia Giachanou[1], Ayoub Bagheri[1], Laura Boeschoten[1],
Erik-Jan van Kesteren[1], Mahdi Shafiee Kamalabad[1], and Daniel L. Oberski[1,2]

[1]Utrecht University, Utrecht, the Netherlands
[2]University Medical Center Utrecht, Utrecht, the Netherlands
{q.fang, a.giachanou, a.bagheri, l.boeschoten}@uu.nl
{e.vankesteren1, m.shafieekamalabad, d.l.oberski}@uu.nl

Abstract

Text-based personality computing (TPC) has gained many research interests in NLP. In this paper, we describe 15 challenges that we consider deserving the attention of the NLP research community. These challenges are organized by the following topics: personality taxonomies, measurement quality, datasets, performance evaluation, modelling choices, as well as ethics and fairness. When addressing each challenge, not only do we combine perspectives from both NLP and social sciences, but also offer concrete suggestions. We hope to inspire more valid and reliable TPC research.

1 Introduction

According to the APA Dictionary of Psychology (APA, 2022), personality refers to personality traits, which are "relatively stable, consistent, and enduring internal characteristics inferred from a pattern of behaviours, attitudes, feelings and habits in individuals". Knowledge about personality can be useful in many societal and scientific applications. For instance, it can help individuals choose learning styles (Komarraju et al., 2011) and occupations (Kern et al., 2019) suited for their personality; it can help clinical psychologists to better understand psychological disorders (Khan et al., 2005) and to deliver personalised treatment plans for mental health patients (Bagby et al., 2016); changes in personality can even help with early diagnosis of Alzheimer's (Robins Wahlin and Byrne, 2011) and Parkinson's disease (Santangelo et al., 2017).

Traditionally, personality assessment is based on self- and other-report questionnaires, which is labour- and time-intensive. Recently, however, automatic personality assessment based on user-generated data (e.g., texts, images, videos) and machine learning algorithms has become a popular alternative. This is known as **personality computing (PC)** (Phan and Rauthmann, 2021), among

many other names[1].

In this paper, we focus on evaluating PC research in NLP, where personality is primarily inferred from text, such as tweets and Reddit posts (Hosseinia et al., 2021), conversations (Mairesse and Walker, 2006) and speech transcriptions (Das and Das, 2017). We refer to such research as **text-based personality computing (TPC)**.

In TPC, on the one hand, we see an increasing number of datasets curated, complex deep-learning algorithms adopted, and (sometimes) high prediction scores achieved. On the other hand, we see relatively little discussion about open challenges and future research directions. For example, the presence of measurement error in questionnaire-based personality scores remains an un(der)addressed issue. Given that such scores are often used as the gold standard for training and validating TPC algorithms, we find it important to discuss related implications and remedies. Another relevant issue concerns how to reduce the risks of TPC research.

Therefore, in this paper, we reflect on current TPC research practices, identify open challenges, and suggest better ways forward. To spot such challenges, we conduct a literature search in the scope of the ACL Anthology[2]. While there are TPC papers published in other venues, we consider our selection from ACL Anthology a good representation of TPC research in NLP. Appendix A describes our search strategy and results in detail. In total, we find and review 60 empirical TPC papers, based on which we identify 15 challenges. They are organized by the following topics: personality taxonomies (§3), measurement quality (§4), datasets (§5), performance evaluation (§6), modelling choices (§7), as well as ethics and fairness (§8). We discuss each challenge and give concrete

---

[1]Personality recognition (Liu et al., 2017), personality identification (Das and Das, 2017), personality detection (Yang et al., 2021a) and personality prediction (Yamada et al., 2019)
[2]https://aclanthology.org/

suggestions where we also draw on broader NLP and social science literature. Furthermore, during our literature review, we identify 18 TPC datasets that can be (re)used by other researchers. We summarize them in Appendix B.

Note that our paper focuses on identifying and discussing TPC-related research challenges instead of providing a comprehensive overview of past TPC studies. For the latter, we refer you to the survey papers by Stajner and Yenikent (2020); Derakhshi et al. (2021); Mushtaq and Kumar (2023).

2 Background

2.1 Current TPC in a Nutshell

TPC concerns computing personality information from texts. This can be either a regression or a classification task, depending on whether the personality measurements are continuous or discrete. Supervised learning has been the predominant approach, relying on text datasets labelled with personality traits via either self-report or crowdsourced annotations. Among many others, log-linear models (Volkova et al., 2015), random forests (Levitan et al., 2016), GloVe embeddings with Gaussian processes (Arnoux et al., 2017), recurrent neural networks (Liu et al., 2017), convolutional neural networks (Majumder et al., 2017), support vector machines (Lan and Paraboni, 2018), ridge regression (He and de Melo, 2021), graphical networks (Yang et al., 2021b) and transformers (Kreuter et al., 2022) have been used. Another popular, psycho-linguistically motivated approach are dictionaries/lexicons (e.g., Oberlander and Nowson, 2006; Sinha et al., 2015; Das and Das, 2017). They are typically lists of curated terms that have pre-assigned weights associated with different personality traits. This allows researchers to compute personality scores from texts by simply matching a dictionary to the texts and aggregating the weights of matched words and phrases.

2.2 Related Work

We find three relevant TPC survey papers. The first one, Stajner and Yenikent (2020), not only summarizes previous research in TPC but also discusses three interesting issues: the difference between MBTI and the Big-5 (two most popular personality taxonomies, see §3), the difficulty in predicting MBTI from Twitter data, and ethical concerns about TPC. We find all three discussions necessary and helpful. Our paper not only provides our own (differing) perspectives on these issues, but also raises and discusses many others. The other two survey papers, Mushtaq and Kumar (2023) and Derakhshi et al. (2021), after reviewing prior TPC research, very briefly suggest some open challenges in TPC research. While the open challenges mentioned in these two papers (e.g., better data quality, data sharing, and ethics) partially overlap with our list of challenges, we adopt a more evidence-based approach to listing challenges, engage in much more thorough discussion, and offer concrete solutions. Furthermore, our list of challenges goes beyond those in these two papers (e.g., measurement error reduction, performance expectation, joint personality modelling).

We also find three additional papers (Bleidorn and Hopwood, 2019; Stachl et al., 2020; Phan and Rauthmann, 2021) concerning issues in general PC research. Our paper, in contrast, focuses on challenges specific to TPC research.

Lastly, TPC is closely related to other fields, such as automatic emotion recognition (Barriere et al., 2022), opinion mining (Hosseinia et al., 2021) and mental health prediction (Guntuku et al., 2018). These fields are all concerned with the computation of social science constructs (see §4).

3 Personality Taxonomies

A TPC research project typically starts by choosing a personality taxonomy, which is a descriptive framework for personality traits (John and Srivastava, 1999). Among the 60 papers we review, we find two prominent taxonomies: the Myers-Briggs Type Indicator (MBTI; 14 papers) and the Big-5 (45 papers).[3] Fifty of these papers adopt *either* the MBTI *or* the Big-5 but not both. This invites the first challenge:

Challenge 1 (C1): MBTI vs. Big-5

The MBTI originated from the theoretical work of Jung (1971) and was further developed by Briggs Myers and Myers (1995). It proposes four personality dimensions that characterize people's differences in perception and judgement processes: Extraversion/Introversion (E/I), Sensing/iNtuition (S/N), Thinking/Feeling (T/F), and Judgement/Perception (J/P). Individuals are classified into one of the two possible categories across each dimension (e.g., INFP and ESTJ).

---

[3] Also, Enneagram and HEXACO each appeared once.

In contrast, the Big-5 was developed based on the lexical hypothesis: *all important personality traits must have been encoded in natural language and therefore, analysis of personality-related terms should reveal the true personality taxonomy* (Goldberg, 1990). Independent research groups (Cattell, 1946; Goldberg, 1982; Costa and McCrae, 1992; Tupes and Christal, 1992) investigated this hypothesis. They identified numerous English terms that might describe inter-individual differences (e.g., warm, curious), asked participants to rate how well these terms describe them on numerical scales, and factor-analyzed the responses, which revealed five consistent dimensions of personality: Openness (O), Conscientiousness (C), Extraversion (E), Agreeableness (A), and Neuroticism (N).[4] Furthermore, each dimension includes six finer-grained sub-traits called *facets* (e.g., agreeableness includes trust, straightforwardness, altruism, compliance, modesty and tender-mindedness). Big-5 is the most widely accepted and researched taxonomy of personality traits in psychology (as opposed to the popularity of MBTI in non-academic settings like job interviews) (Phan and Rauthmann, 2021).

We recommend the Big-5 over MBTI for the following reasons:

First, the Big-5 is a more realistic and accurate personality taxonomy. It scores individuals along a continuous spectrum, which describes inter-individual differences more accurately and preserves more information (as opposed to MBTI's dichotomous approach). Also, the Big-5 includes facets, which allows for finer-grained analysis of personality. Facets are also more predictive of life outcomes (compared to dimensions; Mershon and Gorsuch 1988; Paunonen and Ashton 2001). A potential new TPC research direction can be to predict facets in addition to dimensions.

Second, Big-5 has a much stronger empirical basis than MBTI. Namely, it is grounded in large-scale quantitative analysis of natural language and survey data. Also, Big-5 questionnaires have undergone much more extensive development and validation processes than MBTI's. Consequently, many validated Big-5 questionnaires exist, which vary in length (15-240 items), the inclusion of facet measures, as well as the target populations (e.g., nationalities, professions, languages, and age groups).[5] This enables researchers to choose a questionnaire most appropriate given a population of interest, available resources (e.g., can you afford a longer questionnaire?) and research interests (e.g., are you interested in facets?). In comparison, MBTI is purely theory-driven, lacks empirical support and has officially only four questionnaires that have not been thoroughly tested (Pittenger, 1993; Nowack, 1996; Grant, 2013).[6] Therefore, compared to MBTI, Big-5 is a much more credible and flexible choice for research purposes.

Third, the Big-5 is rooted in natural language (i.e., the lexical hypothesis), suggesting that Big-5-related cues may be more present than do MBTI-related cues in text data. This conjecture is supported by Stajner and Yenikent (2021), who find either insufficient or mixed signals for MBTI dimensions in tweets and short essays.

Nevertheless, despite the many advantages of Big-5 over MBTI, we acknowledge that studying MBTI can still be useful given its popularity in non-academic settings (Lloyd, 2012). Furthermore, it is also important to mention that Big-5 is not without criticisms and other personality taxonomies may be preferred, which invites the next challenge.

C2: Beyond the Big-5

While earlier lexical studies of the English language revealed five core dimensions of personality, more recent analysis of both English and non-English languages (e.g., Italian, Dutch, German, Korean), based on larger sets of adjectives, has suggested the existence of a sixth core dimension: Honesty/Humility, giving rise to a new "Big-6" taxonomy (a.k.a. HEXACO) (Ashton and Lee, 2007). Therefore, we encourage TPC researchers to explore alternatives to the Big-5, in timely accordance with developments in personality psychology. For a comprehensive overview of (other) personality taxonomies, see Cervone and Pervin (2022).

4 Measurement Quality

Personality traits are latent, theoretical variables (a.k.a. social science constructs), which can not be directly or objectively observed. Examples are emotions, prejudice and political orientation. Thus,

---

[4]The Big-5 is also called OCEAN.

[5]E.g., the NEO Personality Inventory (McCrae and Costa, 1989), the Revised NEO-PI and the NEO Five Factor Inventory (Costa and McCrae, 1992), the Big-5 Inventory (BFI) (Benet-Martinez and John, 1998), BFI-2 (Soto and John, 2017), the Short 15-item Big-5 Inventory (Lang et al., 2011).

[6]Form M with 93 items, Form M self-scorable with 93 items, Form Q with 93 items and Step III Form with 222 items. See https://www.myersbriggs.org/.

personality traits are inherently difficult to measure. We can only approximate the true underlying personality trait scores from often noisy observations collected using personality instruments such as self- and other-report questionnaires. Due to the uncertainty in this approximation process, measurements for personality traits likely contain non-negligible error. This error is called **measurement error**, defined as the difference between an observed measurement and its true value (Dodge et al., 2003).

The presence of error in personality measurements can have negative consequences for TPC research. For instance, consider the case of having substantial measurement error in questionnaire-based personality scores. When a TPC model treats these measurements as the gold standard for training and validation, the measurement error will likely propagate to the predictions, rendering the model less helpful (or even harmful) especially for diagnostic or clinical purposes. The study by Akrami et al. (2019) lends support to this hypothesis, where the authors find TPC models to perform better on small datasets with low measurement error than on large datasets with high measurement error. Therefore, it is important that TPC researchers are aware of the presence and influence of measurement error and can deal with it. Unfortunately, none of the 60 TPC studies that we survey touch upon this issue, suggesting that this is an under-explored issue in TPC research. This observation inspires the next four challenges (3-6).

C3: Choose high-quality instruments

Collecting high-quality personality measurements begins with using high-quality instruments, be they questionnaires or models.[7] By high quality, we specifically mean high measurement quality. To determine the measurement quality of an instrument, it is important to understand the two components of measurement error: random error and systematic error, and how they relate to two quality criteria: reliability and validity.

Random error refers to random variations in measurements across comparable conditions, due to factors that cannot be controlled (Trochim et al., 2015). For instance, when someone completes a personality questionnaire twice, the responses may differ between the two attempts because the person misread a question in the second attempt. Random error is always present and unpredictable; it

can be reduced but not eliminated. In NLP systems, random error can be due to random data splitting (Gorman and Bedrick, 2019), stochastic algorithms (Zhou et al., 2020), and certain random processes in data annotations, such as sampling of annotators and random annotation mistakes (Uma et al., 2021).

In contrast, **systematic error** occurs due to factors inherent to an instrument (Trochim et al., 2015). For instance, a poorly constructed Big-5 questionnaire may contain an item that is used to measure neuroticism while in fact it does not. Consequently, anyone taking this questionnaire will get a biased estimate of their neuroticism. Thus, systematic error is foreseeable and often constant or proportional to the true value. As long as its cause is identified, systematic error can be removed. In NLP systems, systematic error can occur when spurious correlations (or "short-cuts", instead of causal relationships) are learned (Wang et al., 2022).

Reliability and **validity** are the two criteria to the measurement quality of an instrument. The former concerns the extent to which an instrument can obtain the same measurement under comparable conditions, while the latter concerns the extent to which an instrument captures what it is supposed to (Trochim et al., 2015). Random error reduces an instrument's reliability, while systematic error undermines its validity. Therefore, a high-quality instrument is a reliable and valid one. We describe below how we can find out about the validity and reliability of a personality instrument.

For personality questionnaires, especially of Big-5, it is relatively easy to determine their measurement quality because many corresponding validity and reliability studies exist (see van der Linden et al. 2010 for an overview). For model-based personality instruments, however, they rarely undergo comprehensive analysis of measurement quality. Typically, studies report the predictive performance of a personality model on some test data (using metrics like accuracy, recall, precision, F1, mean squared error, correlation). Such performance numbers can offer insight into the model's validity, assuming that the gold-standard personality measurements are low in error. However, the model's validity in the presence of substantial measurement error in the data, as well as the model's reliability, remains unclear. Therefore, we urge future researchers to also examine and report both the validity and reliability of a model-based personality instrument. In

---

[7]We consider prediction models and dictionaries/lexicons as personality instruments, just like questionnaires.

Challenge 5, we discuss how this can be done.

However, it is important to exercise caution when selecting a personality instrument based on validity and reliability information obtained from earlier studies, as these studies had limitations in terms of the populations they examined (especially concerning demographic and linguistic characteristics), the time frames and contexts in which they were conducted. In a new study, the population of interest, the time and context may differ from those of previous studies. Consequently, researchers must carefully evaluate the validity and reliability evidence of an existing instrument in light of the specifics of the new study.

Once a personality instrument is selected, it is also important to cite the source of the instrument and report its validity and reliability information. Among the 60 reviewed papers, 5 use self-identified outcomes (like one's MBTI type or Big-5 scores mentioned in a tweet or user profile) where tracing down the instruments is impossible; 14 make use of proprietary instruments whose reliability and validity information is inaccessible to the public; among the 41 that use an existing personality instrument, 9 do not mention the specific instrument and only 4 report validity or reliability information based on previous studies.

C4: Further reduce measurement error by study design

Even when the best possible instrument is used, there can still be substantial measurement error that results from other design factors of a study, especially when questionnaires are used. Factors like questionnaire characteristics (e.g., the number of questions, visual layout, topics, wording) and data collection modes (e.g., online, in person) can affect the measurement quality of the responses (Biemer et al., 2013). Furthermore, factors related to respondents (e.g., inattention) can also affect measurement quality (Fleischer et al., 2015).

Therefore, when planning personality data collection using questionnaires, it can be beneficial to take into account different possible sources of measurement error. This helps to further reduce measurement error in questionnaire responses, in addition to using a valid, reliable questionnaire. For a comprehensive overview of factors that can influence measurement quality in questionnaires and the possible ways to control for them, we refer you to Callegaro et al. (2015) and Biemer et al.

(2013). We also encourage collaboration with survey methodology experts.

C5: Quantify measurement error

Now, assume that personality measurements have been collected. The next step is to quantify both random and systematic error.

For questionnaire-based measurements, we recommend using factor analysis, which is a type of latent variable model that relates a set of observed variables (e.g., personality questionnaire items) to some latent variables (e.g., one's true, underlying personality traits) (Oberski, 2016). Depending on the model specification and data characteristics, factor analysis can decompose the total variation in the observed variables into different sources: variation due to the latent personality traits, variation due to systematic factors like questionnaire characteristics and the time of data collection, and finally, the unexplained variation (at the item level and at the questionnaire level). Larger variation due to the underlying personality traits and lower variation due to systematic factors are desirable, because they indicate higher measurement validity (i.e., less systematic error). In contrast, more unexplained variation indicates more random error (i.e., lack of reliability). Based on this variance decomposition, estimators of reliability and validity can be derived. We refer you to Saris and Gallhofer (2014, Chapter 9-12) for an overview of various factor analysis strategies and estimators of reliability and validity.

For model-based measurements, the number of measurements per personality trait and person is typically limited to one. Factor analysis cannot be applied to such data because the factor model is mathematically non-identifiable (i.e., there is no unique solution). Therefore, different methods for quantifying random and systematic error (or equivalently, reliability and validity) are needed.

Random error is due to factors that cannot be controlled. Therefore, by varying the instrument or measurement condition along such factors, we can quantify the associated random error. In TPC models, one such factor is small variation in text (e.g., the use of singular vs. plural noun, which has not been linked to personality traits by prior research) that should not affect predictions. By introducing simple perturbations to the data, and comparing the new predictions with the ones based on the original data, we can gauge the degree of random error associated with this factor. This approach is analogous

to Ribeiro et al. (2020)'s invariance tests. Du et al. (2021) also showcase such reliability analyses for word embedding-based gender bias scores.

To quantify validity, apart from the usual performance metrics, we can check whether the predicted scores of different personality traits correlate with one another in expected ways. For instance, while the overall correlations should be low, as different personality traits are distinct constructs, some correlations should be more positive (e.g., between conscientiousness and agreeableness) than others (e.g., between openness and neuroticism). van der Linden et al. (2010) provide an overview of empirical correlations between personality traits across demographic groups that can be expected. To the best of our knowledge, no TPC study has assessed the correlations among predicted personality scores. Furthermore, it can be helpful to check whether the predicted personality scores relate to other constructs like emotions (if data is available) in expected ways. Such tests are conceptually similar to convergent and discriminant validity analyses in the social sciences (Stachl et al., 2020). Even in the presence of large measurement error in the data, they can be useful. For a more in-depth discussion of validity testing in machine learning and NLP, see Jacobs and Wallach (2021) and Fang et al. (2022).

C6: Correct for measurement error

Choosing high-quality personality instruments and reducing measurement error by design are likely the most important and effective ways to ensure high quality measurements. Once personality measurements have been collected, however, much less can be done about measurement error.

For questionnaire-based measurements, it might still be helpful to take a closer look at the results from factor analysis. For instance, do the measurements fit the assumed personality model (e.g., the Big-5)? If not, is it due to one or more questionnaire items that show unexpected relationships with the personality traits (e.g., the relationship is zero; the item correlates strongly with a different personality trait than expected)? If so, removing those problematic items can improve the validity of the personality measurements. Are there items with large unexplained variances? If so, removing them may increase reliability.

As for model-based measurements, if the personality models are proprietary or cannot be modified and retrained (e.g., due to lack of data or model details), then no correction for measurement error is possible. If retraining the model is possible, several techniques may help (see C12, C13 in §7).

5 Datasets

Across the 60 TPC studies, we find 41 unique datasets, which vary in terms of the personality taxonomy, instrument, type of text data, sample size, sample characteristics etc. Among them, however, only 18 are potentially accessible to other researchers (see Appendix B). Shareable datasets are key to advancing TPC research, as it leads to accumulation of data and allows for replication studies. This invites the next challenge:

C7: Construct shareable datasets

One obstacle to sharing TPC datasets is privacy preservation, as TPC datasets often contain identifiable information (e.g., names, locations, events) about data subjects. For instance, with social media posts, their authors can be easily found by using the content of the posts as search terms (Norman Adams, 2022). We suggest two ways to make data sharing more privacy-preserving.

First, data pseudonymization and anonymization techniques can be used. With pseudonymization, the data subjects can still be identified if additional information is provided. With anonymization, however, re-identification is impossible. Whether to pseudonymize or anonymize depends on many factors, such as the difficulty in data anonymization and the severity of re-identification. Nevertheless, for TPC datasets containing social media posts, anonymization is likely impossible. We refer you to Lison et al. (2021) for more information.

Second, we can replace texts with paraphrases or synthetic data. The latter aims to "preserve the overall properties and characteristics of the original data without revealing information about actual individual data samples" (Hittmeir et al., 2019). However, whether these strategies are effective enough remain an open research question in NLP.

C8: Finer-grained measurements

All the 18 shareable datasets we find include only aggregated measurements of personality traits. Namely, for MBTI, only the classification types (e.g., INFP and ESTJ; instead of scores on each questionnaire item) are available; for Big-5, only the aggregated scores (e.g., means across items) for the five dimensions. This makes it impossible for other researchers to investigate measurement

quality or train TPC models on the facet or item level. Even worse, some datasets provide no information about the personality instrument used. This especially concerns datasets that obtain personality labels from Twitter or Reddit based on the mention of MBTI or Big-5 information in a post or user profile (e.g., "INTJ"; "As an extravert...").

Other problematic treatments of aggregated personality measurements include further discretization, within-sample standardization or normalization to the target population. All leads to loss of information and limits the reusability of the measurements. Therefore, we suggest providing raw personality measurements, ideally on the item level.

C9: Include demographic information

We argue that the inclusion of demographic information (e.g., age, gender, education) can be important. Not only can this help researchers decide the appropriate personality instrument to use (in relation to the population of interest; see earlier discussion in §4: C3), it can also provide additional useful features for TPC models. Furthermore, researchers can make use of the demographic information to diagnose the model (e.g., whether the measurement quality or the model's prediction performance differs across demographic groups) (i.e., fairness). However, it is important to weigh the gain from including extra personal information against potential harm (see §8: C14).

6 Performance Evaluation

Across the 60 surveyed TPC papers, we identify two challenges related to performance evaluation:

C10: Use more appropriate and consistent performance metrics

11 out of the 60 studies model TPC as a regression task. Among them, 9 use Pearson's correlation between predicted personality trait scores and the true scores as the performance metric. However, correlation-based metrics can be misleading, as they register only ranks and do not reflect how accurate the predictions are on the original scale of the personality scores (Stachl et al., 2020).

Some studies report mean squared error (MSE), which is arguably better than correlations because it quantifies the absolute difference between the predictions and the true values. However, MSE scores depend on the scale of the personality measurements and are not bounded, making interpretation and comparison (between studies) difficult.

Stachl et al. (2020) propose a better performance metric: $R^2 = 1 - RSS/TSS$, where RSS refers to the sum of squares of residuals and TSS the total sum of squares. R^2 has several benefits. First, it can be considered a normalised version of MSE, which has an upper limit of 1 (perfect agreement). Second, R^2 has a natural zero point, which occurs when the mean is used as the prediction. Third, when the model makes worse predictions than a simple mean baseline, R^2 becomes negative.

While two studies report R^2, their calculation of R^2 is unclear. The researchers may have calculated R^2 not based on the formula shown earlier, but by squaring Pearson's correlations. This would lead to always positive R^2, which can be misleading.

In the 33 studies that model TPC as a classification task, a more diverse set of metrics are used (i.e., accuracy, recall, precision, F1, and AUC, in macro, micro, weighted or unweighted forms). One problem, however, is that different studies report different metrics (sometimes, only one). This makes comparison across studies difficult. We encourage future researchers to report all common metrics for classification studies (like those mentioned above).

C11: Report performance expectation

While it is normal to optimize the prediction performance of TPC models, it is also important to set correct expectations: What kind of performance can we realistically expect? How accurate can personality predictions be when only (short) text is used? How good does the performance need to be for a particular system? Stajner and Yenikent (2021) make the first question even more relevant, as they find either few or mixed MBTI-related signals in typical text data used for MBTI prediction.

Quantifying measurement error in the personality scores used for modelling can help researchers to set clearer expectations about model performance, because high systematic error will limit the model's generalizability to new data, while high random error will result in unstable predictions.

Thus, setting expectations "forces" researchers to learn more about their data and to avoid unrealistic expectations that may lead to problematic research practices like cherry-picking results. Unfortunately, none of the 60 reviewed papers discusses performance expectation.

7 Modelling Choices

Most TPC studies model different personality traits (i.e., dimensions) separately. This strategy is unde-

sirable, because it ignores the correlations among personality traits that models can learn from. Modelling personality traits jointly may also help to prevent overfitting to a specific trait and thus learn more universal personality representations (Liu et al., 2019). Hence, the next challenge:

C12: Joint personality modelling

Out of the 60 studies, only 5 attempt at (some form of) joint modelling of personality traits. Yang et al. (2021a) implement a transformer-based model to predict MBTI types, where the use of questionnaire texts allows the model to infer automatically the relevant MBTI dimension, and hence removes the need for independent modelling of different MBTI dimensions. Gjurković and Šnajder (2018), Bassignana et al. (2020b) and Hosseinia et al. (2021) frame the prediction of MBTI types as a 16-class classification task (as there are in total 16 MBTI types), thereby using only one single model. Hull et al. (2021) apply "stacked single target chains" (Xioufis et al., 2016), which feeds the predictions of one personality trait back in as features for the prediction of the next trait(s).

Multitask learning may also be useful, which trains a model on multiple tasks simultaneously and thus might help to improve the generalizability of the model (Caruana, 1997). In addition, Stachl et al. (2020) suggest modifying a model's loss function such that the correlations between theoretically distinct constructs are minimised. Building on this idea, we can also specify the loss function in a way that it not only focuses on general prediction performance but also minimises the difference between the predicted covariance matrix and the observed covariance matrix of personality traits.

C13: Build on best modelling choices

As the field progresses, it is important to not only investigate new modelling ideas, but also accumulate knowledge about best modelling practices. We list below several empirically supported ideas that should not slip past the community's attention.

First, researchers should leverage the texts in personality questionnaires. Kreuter et al. (2022), Vu et al. (2020) and Yang et al. (2021a) find that incorporating personality questionnaire texts into model learning can lead to better personality predictions.

Second, when sample sizes are small, data augmentation and dimensionality reduction techniques are beneficial. Kreuter et al. (2022) show that using data augmentation to increase the training size of personality questionnaire items leads to better predictions. V Ganesan et al. (2021) show that PCA helps to overcome the problem of fine-tuning large language models with a small TPC dataset.

Third, incorporating personality-related variables (e.g., psycholinguistic features, emotions, user interests, demographics, opinions and brand preference) as features, if available, can improve personality predictions (Kerz et al., 2022; Cornelisse, 2020; Kishima et al., 2021; Li et al., 2022; Hosseinia et al., 2021; Yang et al., 2015).

8 Ethics and Fairness

Out of the 60 reviewed papers and the 2 additional survey papers, only 7 provide some reflection about ethics and none about fairness. This can be because ethics and fairness only became central in NLP recently. Nevertheless, the last two challenges:

C14: More ethical and useful TPC

As useful as TPC can be, it is important to ask whether gathering personal information like personality or computing them is necessary. This is especially relevant for research where TPC is only an intermediate step to another end such as opinion mining (Hosseinia et al., 2021), dialogue generation (Mairesse and Walker, 2008) and brand preference prediction (Yang et al., 2015). Such studies typically argue that the computed personality traits can be used as features for another task and that it leads to better task performance; however, they do not consider alternatives (e.g., replacing the prediction of personality traits with using lexical cues that are non-personal but still indicative of personality). Thus, we encourage researchers to justify PC and to find alternatives when PC is only a means.

Even when TPC can be justified, it is important to reduce potential harm. For instance, many TPC studies and datasets make use of public social media profiles for predicting personality traits. While this is often legal, no explicit consent for PC is obtained from the social media users, which makes using public social media data an ethically ambiguous issue (Norman Adams, 2022). Boeschoten et al. (2022) proposes a privacy-preserving data donation framework that may help to alleviate this problem, where data subjects can voluntarily donate their data download packages (e.g., from social media accounts) for research and give explicit consent.

To further increase the benefit of TPC, we can consider applying it to clinical, professional or educational settings, where (traditional) personality assessment has proven useful and relevant (e.g., personalised treatments; career recommendation; individualised learning). None of the 60 TPC studies in our review investigates these applications.

C15: Research on Fair TPC

Fairness research in machine learning concerns identifying and mitigating biases that may be present within a system, particularly towards specific groups (Mehrabi et al., 2021). For instance, a fair TPC model should exhibit equal predictive performance across different demographic groups. Among many others, biased training data is a significant factor contributing to algorithmic bias. From the perspective of measurement quality, the choice of a personality instrument that lacks equal validity and reliability across all demographic groups of interest can introduce variations in the quality of personality measurements among different groups. Consequently, these discrepancies can perpetuate algorithmic bias within the system.

Remarkably, none of the 60 TPC papers we survey address the topic of fairness. Therefore, there is a clear need for future research on fairness in the context of TPC.

9 Conclusion

In this paper, we review 60 TPC papers from the ACL Anthology and identify 15 challenges that we consider deserving the attention of the research community. We focus on the following 6 topics: personality taxonomies, measurement quality, datasets, performance evaluation, modelling choices, as well as ethics and fairness. While some of these topics (e.g., personality taxonomies and ethics) have been discussed elsewhere, we provide new perspectives. Furthermore, in light of these challenges, we offer concrete recommendations for future TPC research, which we summarise below:

- Personality taxonomies: Choose Big-5 over MBTI; Try modelling facets and using other taxonomies like HEXACO where appropriate.

- Measurement quality: Pay attention to measurement error in personality measurements, be they based on questionnaires or models; Try to reduce measurement error by design (e.g., choose higher-quality instruments; use

better data collection practices); Provide quality evaluation (i.e., validity and reliability) for any new (and also existing) approaches.

- Datasets: Make TPC datasets shareable, which should also contain fine-grained personality measurements and descriptions of the target population;

- Performance evaluation: Report a diverse set of performance metrics; Report R^2 for a regression task.

- Modelling choices: Make use of their psychometric properties when modelling personality traits (e.g., use joint modelling; modify the loss function to preserve the covariance information); For even better predictions, try incorporating personality questionnaire texts, applying data augmentation and dimensionality reduction techniques, as well as incorporating more personality-related variables.

- Ethics and fairness: Avoid unnecessary TPC; Apply TPC to clinical, professional and educational settings; Investigate fairness.

- Lastly, engage in (interdisciplinary) research work with survey methodologists, psychologists, and psychometricians.

We hope that our paper will inspire better TPC research and new research directions.

Limitations

Our paper has some limitations. First, we do not give detailed instructions about techniques that we recommended (e.g., factor analysis, synthetic data generation). We rely on our readers' autonomy to acquire the necessary information (that is specific to their research projects) by further reading our recommended references. Second, we only survey TPC papers included in the ACL Anthology, despite other TPC papers existing outside this venue. While this means that the challenges we identified might be specific to these papers, we believe they are still a good representation of the TPC research done in NLP. Lastly, we limit our discussion to text data. It would be beneficial for future research to also discuss challenges facing PCs based on other types of data (e.g., images, behaviours, videos), which may offer additional insights to TPC research.

Statement of Ethics and Impact

Our work provides a critical evaluation of past TPC research, where we present 15 open challenges that we consider deserving the attention of the research community. For each of these challenges, we offer concrete suggestions, thereby hoping to inspire higher-quality TPC research (e.g., more valid and reliability personality measurements, better datasets, better modelling practices). We also discuss issues related to ethics and fairness (see §8). We hope to see more ethical and fair TPC research.

Acknowledgements

References

Nazar Akrami, Johan Fernquist, Tim Isbister, Lisa Kaati, and Björn Pelzer. 2019. Automatic extraction of personality from text: Challenges and opportunities. *2019 IEEE International Conference on Big Data (Big Data)*, pages 3156–3164.

Khalid Al Khatib, Michael Völske, Shahbaz Syed, Nikolay Kolyada, and Benno Stein. 2020. Exploiting personal characteristics of debaters for predicting persuasiveness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7067–7072. Association for Computational Linguistics.

American Psychological Association APA. 2022. Personality. In *APA Dictionary of Psychology*. Retrieved November 09, 2022 from https://dictionary.apa.org/personality.

Pierre-Hadrien Arnoux, Anbang Xu, Neil Boyette, Jalal Mahmud, Rama Akkiraju, and Vibha Sinha. 2017. 25 tweets to know you: A new model to predict personality with social media. In *International Conference on Web and Social Media*.

Michael C Ashton and Kibeom Lee. 2007. Empirical, theoretical, and practical advantages of the hexaco model of personality structure. *Personality and social psychology review*, 11(2):150–166.

R Michael Bagby, Tara M Gralnick, Nadia Al-Dajani, and Amanda A Uliaszek. 2016. The role of the five-factor model in personality assessment and treatment planning. *Clinical Psychology: Science and Practice*, 23(4):365.

Valentin Barriere, Shabnam Tafreshi, João Sedoc, and Sawsan Alqahtani. 2022. WASSA 2022 shared task: Predicting empathy, emotion and personality in reaction to news stories. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 214–227. Association for Computational Linguistics.

E Bassignana, M Nissim, V Patti, et al. 2020a. Personality: A novel youtube-based corpus for personality prediction in italian. In *7th Italian Conference on Computational Linguistics, CLiC-it 2020*, volume 2769, pages 1–7. CEUR-WS.

Elisa Bassignana, Malvina Nissim, and Viviana Patti. 2020b. Matching theory and data with personal-ITY: What a corpus of italian YouTube comments reveals about personality. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 11–22. Association for Computational Linguistics.

Veronica Benet-Martinez and Oliver P John. 1998. Los cinco grandes across cultures and ethnic groups: Multitrait-multimethod analyses of the big five in spanish and english. *Journal of personality and social psychology*, 75(3):729.

Joan-Isaac Biel and Daniel Gatica-Perez. 2012. The youtube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs. *IEEE Transactions on Multimedia*, 15(1):41–55.

Paul P Biemer, Robert M Groves, Lars E Lyberg, Nancy A Mathiowetz, and Seymour Sudman. 2013. *Measurement errors in surveys*, volume 548. John Wiley & Sons.

Wiebke Bleidorn and Christopher James Hopwood. 2019. Using machine learning to advance personality assessment and theory. *Personality and Social Psychology Review*, 23(2):190–203.

Laura E. Boeschoten, Jef Ausloos, Judith E. Möller, Theo B. Araujo, and Daniel L. Oberski. 2022. A framework for privacy preserving digital trace data collection through data donation. *Computational Communication Research*.

Isabel Briggs Myers and Peter B Myers. 1995. Gifts differing: Understanding personality type. *Davis-Black Publishing*.

Mario Callegaro, Katja Lozar Manfreda, and Vasja Vehovar. 2015. *Web survey methodology*. Sage.

Rich Caruana. 1997. Multitask learning. *Machine learning*, 28(1):41–75.

Raymond Bernard Cattell. 1946. *Description and measurement of personality*. World Book Company.

Daniel Cervone and Lawrence A Pervin. 2022. *Personality: Theory and research*. John Wiley & Sons.

Joran Cornelisse. 2020. Inferring neuroticism of twitter users by utilizing their following interests. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 1–10. Association for Computational Linguistics.

Paul T Costa and Robert R McCrae. 1992. Professional manual: revised neo personality inventory (neo-pi-r) and neo five-factor inventory (neo-ffi). *Odessa, FL: Psychological Assessment Resources*, 61.

John Culnan, Seongjin Park, Meghavarshini Krishnaswamy, and Rebecca Sharp. 2021. Me, myself, and ire: Effects of automatic transcription quality on emotion, sarcasm, and personality detection. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 250–256. Association for Computational Linguistics.

Kumar Gourav Das and Dipankar Das. 2017. Developing lexicon and classifier for personality identification in texts. In *Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017)*, pages 362–372. NLP Association of India.

Ali Reza Feizi Derakhshi, Mohammad-Reza Feizi-Derakhshi, Majid Ramezani, Narjes Nikzad-Khasmakhi, Meysam Asgari-Chenaghlu, Taymaz Akan, Mehrdad Ranjbar-Khadivi, Elnaz Zafarni-Moattar, and Zoleikha Jahanbakhsh-Naghadeh. 2021. Text-based automatic personality prediction: a bibliographic review. *Journal of Computational Social Science*, 5:1555 – 1593.

Yadolah Dodge, David Cox, and Daniel Commenges. 2003. *The Oxford dictionary of statistical terms*. Oxford University Press on Demand.

Yupei Du, Qixiang Fang, and Dong Nguyen. 2021. Assessing the reliability of word embedding gender bias measures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10012–10034.

Qixiang Fang, Dong Nguyen, and Daniel L. Oberski. 2022. Evaluating the construct validity of text embeddings with application to survey questions. *EPJ Data Science*, 11:1–31.

Avi Fleischer, Alan Mead, and Jialin Huang. 2015. Inattentive responding in mturk and other online samples. *Industrial and Organizational Psychology*, 8:196 – 202.

Pascale Fung, Anik Dey, Farhad Bin Siddique, Ruixi Lin, Yang Yang, Yan Wan, and Ho Yin Ricky Chan. 2016. Zara the supergirl: An empathetic personality recognition system. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 87–91. Association for Computational Linguistics.

Soumitra Ghosh, Dhirendra Maurya, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Team IITP-AINLPML at WASSA 2022: Empathy detection, emotion classification and personality detection. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 255–260. Association for Computational Linguistics.

Alastair Gill, Carsten Brockmann, and Jon Oberlander. 2012. Perceptions of alignment and personality in generated dialogue. In *INLG 2012 Proceedings of the Seventh International Natural Language Generation Conference*, pages 40–48. Association for Computational Linguistics.

Matej Gjurković, Mladen Karan, Iva Vukojević, Mihaela Bošnjak, and Jan Snajder. 2021. PANDORA talks: Personality and demographics on reddit. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 138–152. Association for Computational Linguistics.

Matej Gjurković and Jan Šnajder. 2018. 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*, pages 87–97. Association for Computational Linguistics.

Lewis R Goldberg. 1982. From ace to zombie: Some explorations in the language of personality. *Advances in personality assessment*, 1:203–234.

Lewis R Goldberg. 1990. An alternative" description of personality": the big-five factor structure. *Journal of personality and social psychology*, 59(6):1216.

Kyle Gorman and Steven Bedrick. 2019. We need to talk about standard splits. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 2786–2791.

Adam Grant. 2013. Goodbye to mbti, the fad that won't die. *Psychology Today*, 9.

Sharath Chandra Guntuku, Salvatore Giorgi, and Lyle Ungar. 2018. Current and future psychological health prediction using language and socio-demographics of children for the CLPysch 2018 shared task. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 98–106. Association for Computational Linguistics.

Xiaoli He and Gerard de Melo. 2021. Personality predictive lexical cues and their correlations. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 514–523. INCOMA Ltd.

Markus Hittmeir, Andreas Ekelhart, and Rudolf Mayer. 2019. On the utility of synthetic data: An empirical evaluation on machine learning tasks. In *Proceedings of the 14th International Conference on Availability, Reliability and Security*, pages 1–6.

Marjan Hosseinia, Eduard Dragut, Dainis Boumber, and Arjun Mukherjee. 2021. On the usefulness of personality traits in opinion-oriented tasks. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 547–556. INCOMA Ltd.

James R. Hull, Valerie Novak, C. Anton Rytting, Paul Rodrigues, Victor M. Frank, and Matthew Swahn. 2021. Personality trait identification using the russian feature extraction toolkit. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 583–592. INCOMA Ltd.

Ritsuko Iwai, Daisuke Kawahara, Takatsune Kumada, and Sadao Kurohashi. 2020a. Acquiring social knowledge about personality and driving-related behavior. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2306–2315. European Language Resources Association.

Ritsuko Iwai, Daisuke Kawahara, Takatsune Kumada, and Sadao Kurohashi. 2020b. Development of a japanese personality dictionary based on psychological methods. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3103–3108. European Language Resources Association.

Abigail Z Jacobs and Hanna Wallach. 2021. Measurement and fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 375–385.

Michel-Pierre Jansen, Khiet P. Truong, Dirk K.J. Heylen, and Deniece S. Nazareth. 2020. Introducing MULAI: A multimodal database of laughter during dyadic interactions. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4333–4342. European Language Resources Association.

Oliver P John and Sanjay Srivastava. 1999. The big-five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin and O. P. John, editors, *Handbook of personality: Theory and research*, page 102–138. Guilford Press.

Carl G Jung. 1971. *Psychological Types*. Princeton University Press, Princeton, New Jersey.

Koichi Kamijo, Tetsuya Nasukawa, and Hideya Kitamura. 2016. Personality estimation from Japanese text. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, pages 101–109, Osaka, Japan. The COLING 2016 Organizing Committee.

Onno Kampman, Elham J. Barezi, Dario Bertero, and Pascale Fung. 2018. Investigating audio, video, and text fusion methods for end-to-end automatic personality prediction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 606–611. Association for Computational Linguistics.

Margaret L Kern, Paul X McCarthy, Deepanjan Chakrabarty, and Marian-Andrei Rizoiu. 2019. Social media-predicted personality traits and values can help match people to their ideal jobs. *Proceedings of the National Academy of Sciences*, 116(52):26459–26464.

Elma Kerz, Yu Qiao, Sourabh Zanwar, and Daniel Wiechmann. 2022. Pushing on personality detection from verbal behavior: A transformer meets text contours of psycholinguistic features. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 182–194. Association for Computational Linguistics.

Amir A Khan, Kristen C Jacobson, Charles O Gardner, Carol A Prescott, and Kenneth S Kendler. 2005. Personality and comorbidity of common psychiatric disorders. *The British Journal of Psychiatry*, 186(3):190–196.

Ryota Kishima, Kazuyuki Matsumoto, Minoru Yoshida, and Kenji Kita. 2021. Construction of MBTI personality estimation model considering emotional information. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 262–269. Association for Computational Lingustics.

Meera Komarraju, Steven J Karau, Ronald R Schmeck, and Alen Avdic. 2011. The big five personality traits, learning styles, and academic achievement. *Personality and individual differences*, 51(4):472–477.

Maria Koutsombogera and Carl Vogel. 2018. Modeling collaborative multimodal behavior in group dialogues: The MULTISIMO corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).

Anne Kreuter, Kai Sassenberg, and Roman Klinger. 2022. Items from psychometric tests as training data for personality profiling models of twitter users. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 315–323. Association for Computational Linguistics.

Alex Lan and Ivandré Paraboni. 2018. Definite description lexical choice: taking speaker's personality into account. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).

Frieder R Lang, Dennis John, Oliver Lüdtke, Jürgen Schupp, and Gert G Wagner. 2011. Short assessment of the big five: Robust across survey methods except telephone interviewing. *Behavior research methods*, 43(2):548–567.

Sarah Ita Levitan, Yocheved Levitan, Guozhen An, Michelle Levine, Rivka Levitan, Andrew Rosenberg,

and Julia Hirschberg. 2016. Identifying individual differences in gender, ethnicity, and personality from dialogue for deception detection. In *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*, pages 40–44. Association for Computational Linguistics.

Bin Li, Yixuan Weng, Qiya Song, Fuyan Ma, Bin Sun, and Shutao Li. 2022. Prompt-based pre-trained model for personality and interpersonal reactivity prediction. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 265–270. Association for Computational Linguistics.

Pierre Lison, I. Pilán, David Sánchez, Montserrat Batet, and Lilja Øvrelid. 2021. Anonymisation models for text data: State of the art, challenges and future directions. In *ACL*.

Fei Liu, Julien Perez, and Scott Nowson. 2016. A recurrent and compositional model for personality trait recognition from short texts. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, pages 20–29. The COLING 2016 Organizing Committee.

Fei Liu, Julien Perez, and Scott Nowson. 2017. A language-independent and compositional model for personality trait recognition from short texts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 754–764. Association for Computational Linguistics.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. In *ACL*.

John B Lloyd. 2012. The myers-briggs type indicator® and mainstream psychology: analysis and evaluation of an unresolved hostility. *Journal of Beliefs & Values*, 33(1):23–34.

Kim Luyckx and Walter Daelemans. 2008. Personae: a corpus for author and personality prediction from text. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. European Language Resources Association (ELRA).

Veronica Lynn, Niranjan Balasubramanian, and H. Andrew Schwartz. 2020. 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*, pages 5306–5316. Association for Computational Linguistics.

François Mairesse and Marilyn Walker. 2006. Automatic recognition of personality in conversation. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 85–88. Association for Computational Linguistics.

François Mairesse and Marilyn Walker. 2007. PERSONAGE: Personality generation for dialogue. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 496–503. Association for Computational Linguistics.

François Mairesse and Marilyn Walker. 2008. Trainable generation of big-five personality styles through data-driven parameter estimation. In *Proceedings of ACL-08: HLT*, pages 165–173. Association for Computational Linguistics.

Navonil Majumder, Soujanya Poria, Alexander Gelbukh, and E. Cambria. 2017. Deep learning-based document modeling for personality detection from text. *IEEE Intelligent Systems*, 32:74–79.

Maxim Makatchev and Reid Simmons. 2011. Perception of personality and naturalness through dialogues by native speakers of american english and arabic. In *Proceedings of the SIGDIAL 2011 Conference*, pages 286–293. Association for Computational Linguistics.

RR McCrae and PT Costa. 1989. The neo-pi/neo-ffi manual supplement. *Odessa, FL: Psychological Assessment Resources.*

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6):1–35.

Bryan Mershon and Richard L Gorsuch. 1988. Number of factors in the personality sphere: Does increase in factors increase predictability of real-life criteria? *Journal of Personality and Social Psychology*, 55(4):675.

J. Mitchell. 2017. (mbti) myers-briggs personality type dataset. Data retrieved from Kaggle, https://www.kaggle.com/datasets/datasnaek/mbti-type.

Sumiya Mushtaq and Neerendra Kumar. 2023. Text-based automatic personality recognition: Recent developments. In *Proceedings of Third International Conference on Computing, Communications, and Cyber-Security: IC4S 2021*, pages 537–549. Springer.

Nicholas Norman Adams. 2022. 'scraping'reddit posts for academic research? addressing some blurred lines of consent in growing internet-based research trend during the time of covid-19. *International journal of social research methodology*, pages 1–16.

K Nowack. 1996. Is the myers briggs type indicator the right tool to use? performance in practice. *American Society of Training and Development, Fall*, 6.

Jon Oberlander and Scott Nowson. 2006. Whose thumb is it anyway? classifying author personality from weblog text. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 627–634.

10873

Daniel Oberski. 2016. Mixture models: Latent profile and latent class analysis. In *Modern statistical methods for HCI*, pages 275–287. Springer.

Sampo V Paunonen and Michael C Ashton. 2001. Big five factors and facets and the prediction of behavior. *Journal of personality and social psychology*, 81(3):524.

James W Pennebaker and Laura A King. 1999. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296.

Le Vy Phan and John F Rauthmann. 2021. Personality computing: New frontiers in personality assessment. *Social and Personality Psychology Compass*, 15(7):e12624.

David J Pittenger. 1993. Measuring the mbti. . . and coming up short. *Journal of Career Planning and Employment*, 54(1):48–52.

Daniele Pizzolli and Carlo Strapparava. 2019. Personality traits recognition in literary texts. In *Proceedings of the Second Workshop on Storytelling*, pages 107–111. Association for Computational Linguistics.

Barbara Plank and Dirk Hovy. 2015. Personality traits on Twitter—or—How to get 1,500 personality tests in a week. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 92–98, Lisboa, Portugal. Association for Computational Linguistics.

Víctor Ponce-López, Baiyu Chen, Marc Oliu, Ciprian Corneanu, Albert Clapés, Isabelle Guyon, Xavier Baró, Hugo Jair Escalante, and Sergio Escalera. 2016. Chalearn lap 2016: First round challenge on first impressions-dataset and results. In *European conference on computer vision*, pages 400–418. Springer.

Daniel Preoţiuc-Pietro, Johannes Eichstaedt, Gregory Park, Maarten Sap, Laura Smith, Victoria Tobolsky, H. Andrew Schwartz, and Lyle Ungar. 2015. The role of personality, age, and gender in tweeting about mental illness. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 21–30. Association for Computational Linguistics.

Ricelli Ramos, Georges Neto, Barbara Silva, Danielle Monteiro, Ivandré Paraboni, and Rafael Dias. 2018. Building a corpus for personality-dependent natural language understanding and generation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).

Francisco Rangel, Paolo Rosso, Martin Potthast, Benno Stein, and Walter Daelemans. 2015. Overview of the 3rd author profiling task at pan 2015. In *CLEF*, page 2015. sn.

Philip Resnik, Anderson Garron, and Rebecca Resnik. 2013. Using topic modeling to improve prediction of neuroticism and depression in college students. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1348–1353. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Sherry Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist. In *ACL*.

Tarja-Brita Robins Wahlin and Gerard J Byrne. 2011. Personality changes in alzheimer's disease: a systematic review. *International journal of geriatric psychiatry*, 26(10):1019–1029.

Alexandra Roshchina, John Cardiff, and Paolo Rosso. 2012. Evaluating the similarity estimator component of the TWIN personality-based recommender system. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 4098–4102. European Language Resources Association (ELRA).

Sougata Saha, Souvik Das, and Rohini Srihari. 2022. Stylistic response generation by controlling personality traits and intent. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 197–211. Association for Computational Linguistics.

Gabriella Santangelo, Fausta Piscopo, Paolo Barone, and Carmine Vitale. 2017. Personality in parkinson's disease: Clinical, behavioural and cognitive correlates. *Journal of the Neurological Sciences*, 374:17–25.

Willem E Saris and Irmtraud N Gallhofer. 2014. *Design, evaluation, and analysis of questionnaires for survey research*. John Wiley & Sons.

Farhad Bin Siddique, Onno Kampman, Yang Yang, Anik Dey, and Pascale Fung. 2017. Zara returns: Improved personality induction and adaptation by an empathetic virtual agent. In *Proceedings of ACL 2017, System Demonstrations*, pages 121–126. Association for Computational Linguistics.

Priyanka Sinha, Lipika Dey, Pabitra Mitra, and Anupam Basu. 2015. Mining HEXACO personality traits from enterprise social media. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 140–147. Association for Computational Linguistics.

Christopher J Soto and Oliver P John. 2017. The next big five inventory (bfi-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of personality and social psychology*, 113(1):117.

Clemens Stachl, Florian Pargent, Sven Hilbert, Gabriella M Harari, Ramona Schoedel, Sumer Vaid,

10874

Samuel D Gosling, and Markus Bühner. 2020. Personality research and assessment in the era of machine learning. *European Journal of Personality*, 34(5):613–631.

Sanja Stajner and Seren Yenikent. 2020. A survey of automatic personality detection from texts. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6284–6295. International Committee on Computational Linguistics.

Sanja Stajner and Seren Yenikent. 2021. Why is MBTI personality detection from texts a difficult task? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3580–3589. Association for Computational Linguistics.

Chanchal Suman, Aditya Gupta, Sriparna Saha, and Pushpak Bhattacharyya. 2020. A multi-modal personality prediction system. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 317–322. NLP Association of India (NLPAI).

Edward Tighe and Charibeth Cheng. 2018. Modeling personality traits of filipino twitter users. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 112–122. Association for Computational Linguistics.

Jackson Tolins, Kris Liu, Michael Neff, Marilyn Walker, and Jean Fox Tree. 2016a. A verbal and gestural corpus of story retellings to an expressive embodied virtual character. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3461–3468. European Language Resources Association (ELRA).

Jackson Tolins, Kris Liu, Yingying Wang, Jean E. Fox Tree, Marilyn Walker, and Michael Neff. 2016b. A multimodal motion-captured corpus of matched and mismatched extravert-introvert conversational pairs. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3469–3476. European Language Resources Association (ELRA).

William M. K. Trochim, James P. Donnelly, and Kanika Arora. 2015. *Research Methods: The Essential Knowledge Base*. Cengage Learning, Boston.

Ernest C Tupes and Raymond E Christal. 1992. Recurrent personality factors based on trait ratings. *Journal of personality*, 60(2):225–251.

Ana Sabina Uban, Berta Chulvi, and Paolo Rosso. 2021. Understanding patterns of anorexia manifestations in social media data with deep learning. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 224–236. Association for Computational Linguistics.

Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Adithya V Ganesan, Matthew Matero, Aravind Reddy Ravula, Huy Vu, and H. Andrew Schwartz. 2021. Empirical evaluation of pre-trained transformers for human-level NLP: The role of sample size and dimensionality. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4515–4532. Association for Computational Linguistics.

Dimitri van der Linden, Jan te Nijenhuis, and Arnold B. Bakker. 2010. The general factor of personality: A meta-analysis of big five intercorrelations and a criterion-related validity study. *Journal of Research in Personality*, 44(3):315–327.

Ben Verhoeven and Walter Daelemans. 2014. CLiPS stylometry investigation (CSI) corpus: A dutch corpus for the detection of age, gender, personality, sentiment and deception in text. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3081–3085. European Language Resources Association (ELRA).

Ben Verhoeven, Walter Daelemans, and Barbara Plank. 2016. TwiSty: A multilingual twitter stylometry corpus for gender and personality profiling. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1632–1637. European Language Resources Association (ELRA).

Giuseppe Vettigli and Antonio Sorgente. 2021. EmpNa at WASSA 2021: A lightweight model for the prediction of empathy, distress and emotions from reactions to news stories. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 264–268. Association for Computational Linguistics.

Svitlana Volkova, Yoram Bachrach, Michael Armstrong, and Vijay Sharma. 2015. Inferring latent user properties from texts published in social media. In *AAAI Conference on Artificial Intelligence*.

Huy Vu, Suhaib Abdurahman, Sudeep Bhatia, and Lyle Ungar. 2020. Predicting Responses to Psychological Questionnaires from Participants' Social Media Posts and Question Text Embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1512–1524, Online. Association for Computational Linguistics.

Xuan-Son Vu, Lucie Flekova, Lili Jiang, and Iryna Gurevych. 2018. Lexical-semantic resources: yet powerful resources for automatic personality classification. In *Proceedings of the 9th Global Wordnet Conference*, pages 172–181. Global Wordnet Association.

Tianlu Wang, Diyi Yang, and Xuezhi Wang. 2022. Identifying and mitigating spurious correlations for improving robustness in nlp models. In *NAACL-HLT*.

Zhiyuan Wen, Jiannong Cao, Ruosong Yang, Shuaiqi Liu, and Jiaxing Shen. 2021. Automatically select emotion for response via personality-affected emotion transition. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5010–5020. Association for Computational Linguistics.

Eleftherios Spyromitros Xioufis, Grigorios Tsoumakas, William Groves, and Ioannis P. Vlahavas. 2016. Multi-target regression via input space expansion: treating targets as inputs. *Machine Learning*, 104:55–98.

Kosuke Yamada, Ryohei Sasano, and Koichi Takeda. 2019. Incorporating textual information on user behavior for personality prediction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 177–182. Association for Computational Linguistics.

Chao Yang, Shimei Pan, Jalal Mahmud, Huahai Yang, and Padmini Srinivasan. 2015. Using personal traits for brand preference prediction. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 86–96. Association for Computational Linguistics.

Feifan Yang, Tao Yang, Xiaojun Quan, and Qinliang Su. 2021a. Learning to answer psychological questionnaire for personality detection. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1131–1142. Association for Computational Linguistics.

Tao Yang, Feifan Yang, Haolan Ouyang, and Xiaojun Quan. 2021b. Psycholinguistic tripartite graph network for personality detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4229–4239. Association for Computational Linguistics.

Mohammadzaman Zamani, Anneke Buffone, and H. Andrew Schwartz. 2018. Predicting human trustfulness from facebook language. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 174–181. Association for Computational Linguistics.

Xiang Zhou, Yixin Nie, Hao Tan, and Mohit Bansal. 2020. The curse of performance instability in analysis datasets: Consequences, source, and suggestions. In *EMNLP*.

A   Literature Search and Overview

To search for relevant TPC studies in the ACL Anthology, we use the following search terms: "personality", "personality prediction", "personality computing" and "personality recognition". The search date is August 29, 2022. We also sort the results by "relevance" and "year of publication", separately, as these two options return different search results. For all the returned papers, we check their titles, abstracts and (when necessary) main texts, and retain those that fall under one of the following categories:

Dataset Type-A   Papers focusing on introducing new datasets for TPC. Some also benchmark TPC models against new datasets. There are 7: Luyckx and Daelemans (2008); Verhoeven and Daelemans (2014); Plank and Hovy (2015); Verhoeven et al. (2016); Bassignana et al. (2020b); Gjurković et al. (2021); Barriere et al. (2022).

Dataset Type-B   Papers that introduce datasets for non-TPC goals (e.g., multimodal PC, natural language generation, modelling laughter in conversations) but happen to include personality and text data that can be used for TPC. There are 5: Tolins et al. (2016a,b); Koutsombogera and Vogel (2018); Ramos et al. (2018); Jansen et al. (2020).

TPC as End   Empirical papers where TPC is the main goal. There are 31: Mairesse and Walker (2006); Resnik et al. (2013); Sinha et al. (2015); Fung et al. (2016); Kamijo et al. (2016); Levitan et al. (2016); Liu et al. (2016); Das and Das (2017); Liu et al. (2017); Siddique et al. (2017); Gjurković and Šnajder (2018); Kampman et al. (2018); Tighe and Cheng (2018); Vu et al. (2018); Zamani et al. (2018); Pizzolli and Strapparava (2019); Yamada et al. (2019); Cornelisse (2020); Iwai et al. (2020b); Lynn et al. (2020); Suman et al. (2020); Culnan et al. (2021); Hull et al. (2021); Kishima et al. (2021); V Ganesan et al. (2021); Yang et al. (2021a,b); Ghosh et al. (2022); Kerz et al. (2022); Kreuter et al. (2022); Li et al. (2022).

TPC as Means   Empirical papers where TPC is not the main goal but used for further purposes like emotion detection, natural language generation and sociolinguistic analysis. There are 17: Mairesse and Walker (2007, 2008); Makatchev and Simmons (2011); Gill et al. (2012); Roshchina et al. (2012); Preoţiuc-Pietro et al. (2015); Yang et al. (2015); Guntuku et al. (2018); Lan and Paraboni (2018);

Source	Name	Size	Taxonomy	Text Type	Language
Barriere et al. (2022)	WASSA 2022	2655	Big-5	News and essays	en
Bassignana et al. (2020a)	Personal-ITY	1048	MBTI	YouTube comments	it
Biel and Gatica-Perez (2012)	Youtube Personality	442	Big-5	YouTube vlogs transcripts	en
Gjurković et al. (2021)	PANDORA*	10288	Big-5 & MBTI	Reddit posts	en, es, fr, it, de, pt nl, eo, sv, pl
Gjurković and Šnajder (2018)	Reddit9k*	9111	MBTI	Reddit posts	en
Jansen et al. (2020)	MULAI*	26	Big-5	Conversation transcripts	en
Koutsombogera and Vogel (2018)	MULTISIMO*	49	Big-5	Conversation transcripts	en
Mitchell (2017)	Kaggle	8600	MBTI	Forum posts	en
Pennebaker and King (1999)	Essays	2479	Big-5	Essays	en
Plank and Hovy (2015)	Twitter	1500	MBTI	Twitter posts	en
Ponce-López et al. (2016)	ChaLearn V1	10000	Big-5	YouTube vlogs transcripts	en
Ramos et al. (2018)	b5*	1082	Big-5	Facebook status, referring expressions and scene descriptions	pt
Rangel et al. (2015)	PAN 2015*	300	Big-5	Twitter posts	en, es, it, nl
Tolins et al. (2016b)	Personality Dyads*	6	Big-5	Conversation transcriptions	en
Tolins et al. (2016a)	Storytron Retellings*	44	Big-5	Conversation transcriptions	en
Verhoeven and Daelemans (2014)	CSI	697	Big-5 & MBTI	Essays and reviews	nl
Verhoeven et al. (2016)	TWISTY	18168	MBTI	Twitter posts	de, it, nl, fr, pt, es
Wen et al. (2021)	PELD	711	Big-5	TV show dialogs	en

Table 1: Overview of Shareable Datasets. *Data available upon simple request (e.g., via a form on the dataset hosting website). Abbreviations of languages are based on ISO 639-1 Code.

Al Khatib et al. (2020); Iwai et al. (2020a); He and de Melo (2021); Hosseinia et al. (2021); Uban et al. (2021); Vettigli and Sorgente (2021); Wen et al. (2021); Saha et al. (2022).

Note that we exclude papers that mean PC as the computing of personality disorders or personal profiles. In total, we find and review 60 TPC papers. Also, see Appendix B for 18 shareable TPC datasets that we find via these 60 papers.

B Overview of Shareable TPC Datasets

See Table 1.

A   For every submission:

☑ A1. Did you describe the limitations of your work?
Left blank.

☑ A2. Did you discuss any potential risks of your work?
Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Left blank.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B   ☒ Did you use or create scientific artifacts?

Left blank.

☐ B1. Did you cite the creators of artifacts you used?
No response.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C   ☒ Did you run computational experiments?

Left blank.

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
No response.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☐ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
No response.

☐ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
No response.

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
No response.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.