Promises and Perils of Inferring Personality on GitHub

Frenk C.J. van Mil
Delft University of Technology
The Netherlands

Ayushi Rastogi
University of Groningen
The Netherlands
a.rastogi@rug.nl

Andy Zaidman
Delft University of Technology
The Netherlands
a.e.zaidman@tudelft.nl

ABSTRACT

Background: Personality plays a pivotal role in our understanding of human actions and behavior. Today, the applications of personality are widespread, built on the solutions from psychology to infer personality. **Aim:** In software engineering, for instance, one widely used solution to infer personality uses textual communication data. As studies on personality in software engineering continue to grow, it is imperative to understand the performance of these solutions. **Method:** This paper compares the inferential ability of three widely studied text-based personality tests against each other and the ground truth on GitHub. We explore the challenges and potential solutions to improve the inferential ability of personality tests. **Results:** Our study shows that *solutions for inferring personality are far from being perfect.* Software engineering communications data can infer individual developer personality with an average error rate of 41%. In the best case, the error rate can be reduced up to 36% by following our recommendations[1].

CCS CONCEPTS

• **Software and its engineering** → *Programming teams*; • **Human-centered computing** → *Natural language interfaces*; • **Social and professional topics** → Cultural characteristics; • **Computing methodologies** → Simulation evaluation.

KEYWORDS

Personality, Software Developer, Mining Software Repositories, LIWC, Personality Insights

ACM Reference Format:
Frenk C.J. van Mil, Ayushi Rastogi, and Andy Zaidman. 2021. Promises and Perils of Inferring Personality on GitHub. In *ACM / IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM) (ESEM '21), October 11–15, 2021, Bari, Italy.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3475716.3475775

[1]This work is based on the MSc thesis of Frenk van Mil [58]. The study data is available at [60], while all essentials scripts to replicate our work are available at [59].

ESEM '21, October 11–15, 2021, Bari, Italy

ACM ISBN 978-1-4503-8665-4/21/10…$15.00
https://doi.org/10.1145/3475716.3475775

1 INTRODUCTION

Personality is an indicator of how we think, feel, and do [62] with widespread applications. In software engineering, for example, individual developer personality is used to understand contribution patterns [44], work preferences [36], and work satisfaction [2], while collectively, it is used to improve team composition [19, 22].

There are two widely used methods to infer personality: questionnaire and psycholinguistic test. A questionnaire is a gold standard to measure personality (e.g., [53]) in which people are asked a series of questions, responses to which indicate personality. This approach, however, is time-consuming and relies heavily on the response rate. On the other hand, a psycholinguistic test fetches a sizeable amount of text written by a person to a 'model' to generate personality scores (e.g., [44]; also see Figure 1).

Psycholinguistic models are widely used in different contexts (e.g., software engineering [44] and social media platforms such as Twitter [23, 24]), often replacing its alternative questionnaire. In software engineering, models based on psycholinguistic tests are widely used to measure developer personality [9, 44].

While applied to the technical discussions in software engineering, these models are trained on casual conversations (e.g., essays and blog posts [62]). Therefore, how well these models infer developer personality is questionable. To assess the inferential ability of the models used for measuring developer personality in software engineering, this paper solicits answer to the following research question:

Do psycholinguistic tests (trained on different text source(s)) reliably infer developer personality from SE communications data?

We analyzed developer discussions on collaborative software projects at GitHub using three state-of-the-art and practice psycholinguistic models for inferring personality. We studied the Personality Insights tool developed by IBM Watson[2] and two models from academia: Golbeck et al. [23] designed for small text sizes such as Twitter posts and Yarkoni et al. [62] designed for longer texts such as blog posts. We also generated ground truth by conducting a questionnaire for a subset of developers to validate the personality inferences from the three models.

We answer the research question in terms of four sub-questions (see Figure 1 for an overview), exploring the characteristics of (1) the input text fetched into a model, (2) the model itself, and (3) the person whose personality is measured. First, we explored the influence of textual features that do not appear in a usual conversation and are otherwise a part of software engineering communications (e.g., technical jargon) or the syntax used on the platform the discussion ensues (e.g., markdown). Our first question is:

RQ1. Do characteristics of the software engineering communications data influence inferred personality?

[2]https://www.ibm.com/watson/services/personality-insights/

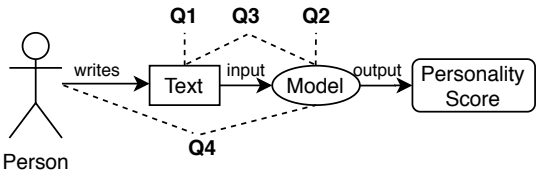

Figure 1: shows the working of the psycholinguistic test. Q1-4 points to the aspect studied by a research question.

Second, we compared models inferring personality to each other and to the ground truth to find:

RQ2. Is one model better than other in inferring developer personality from software engineering data?

Next, exploring the relations of the input written data to the model, we investigate:

RQ3. Does the reliability of inferred personality change with the size of the input text?

Finally, linking the characteristics of individuals to that of the model used for inferring personality, we investigate:

RQ4. Does English proficiency relate to the reliability of personality inferred from psycholinguistic tests?

Our study shows that psycholinguistic tests can be administered to infer developer personality from software engineering communications data with an error rate of 25-48%, with some exceptions. With our recommended data-sanitization steps, the error rates can be reduced (in our case, up to 36%). We observed that the three models perform comparably and optimally with an average word count between 600 to 1200 and that English proficiency influences the inferred personality traits.

This paper is organized into seven sections. Starting with an Introduction in Section 1, we describe the background information and related work in Section 2. Section 3 presents our study design followed by results and their discussion in Section 4 and 5 respectively. We highlight the limitations and threats to validity of our study in Section 6. Finally, in Section 7 we present our conclusions and directions for future work.

2 BACKGROUND AND RELATED WORK

There are many ways to express personality, of which the Big Five Personality model (or BFP) is the most widely used. The Big Five personality framework has gained recognition among trait psychologists for its validity and reliability [28, 39]. It comprises five traits, namely Openness to Experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (in short: OCEAN). Following is an explanation for the five personality traits:

- *Openness to Experience* is characterized by intellectual curiosity, imagination, and open-mindedness. It is also referred to as Intellect/Autonomy, or Openness [54]. Its opposite, close-minded people often have a narrow range of creativity and intellectual interest.
- *Conscientiousness* is characterized by the preference of order, structure, persistence to a goal, and responsibility. Low conscientiousness is linked with comfort, flexibility, and spontaneity but also sloppiness and lack of reliability [54].
- *Extraversion/Extroversion* is characterized by energy creation from external means, social engagement, and assertiveness. Highly extraverted people feel comfortable in social environments, experience positive emotions more often than introverted people [54].
- *Agreeableness* is characterized by the general concern for other's well-being and social harmony. Disagreeable people have less concern for the regard of others and social norms of politeness [54].
- *Neuroticism* is characterized by the tendency to experience negative emotions. Lower neuroticism is linked to emotional stability. Lower neurotic people tend to stay calm and resilient, also referred to as emotional stability [54].

Studies show that the personality structure of an individual is, in part, a function of the linguistic characteristics the individual shares with a group of people [27]. Further, Pennebaker [43] adds that the smallest and stealthiest words in our vocabulary define something about our personality; words like 'with' and 'together' often indicate the author having better social skills, having more friends, and the author usually rate themselves as more outgoing. With this as basis, automatic methods called 'psycholinguistic models' identify personality from a text by transforming the words used into personality scores.

Many studies in software engineering (e.g., [10, 42, 44]) and beyond (e.g., [23]) have used psycholinguistic tests for inferring personality. In software engineering, one of the first studies showed that communications data can be used to infer personality [46]. Today, software communications data are used to infer developer personality (individually and as a group) to characterize developers and their participation patterns [10, 44], and its effect on project success [63], including intermediate steps such as pull request acceptance [30].

As the studies exploring the role of personality in software engineering continue to grow, it is crucial to understand the promises and perils of existing solutions and explore future directions, if necessary.

3 STUDY DESIGN

To investigate the inferential ability of psycholinguistic tests on software engineering communications, we compare the test scores from three state-of-the-practice models to each other and the ground truth. This section presents how we infer developer personality using psycholinguistic tests and gauge ground truth using a questionnaire. Next, we describe the data collected for analysis followed by the statistical tests to investigate each research question. A curated list of data and scripts used for analysis (in compliance with the GDPR) is available at respectively [60] (data) and [59] (scripts).

3.1 Psycholinguistic Tests

At the core of psycholinguistic tests are models that take written text as input and transform it to generate five numbers, representing the five personality traits (see Figure 1). In this study, we use three widely used models: two from academia (Yarkoni [62] and Golbeck et al. [23, 24]) and one from industry (IBM) - Personality Insights.[3]

[3]https://www.ibm.com/watson/services/personality-insights/

The two academic models are based on Linguistic Inquiry and Word Count tool (LIWC)[4]. LIWC calculates the percentage of words in a text indicating emotions and part of speech, among others [57] (also referred to as *word categories*). These word categories are correlated with personality traits; therefore, calculating a weighted sum of word categories and correlation coefficient indicates personality.

While the two models are structurally the same, they are different in the text sources for training. Yarkoni trained its model on long essays [62], while Golbeck trained it on short Tweets [23, 24]. These differences in text culminates into correlations, hence differences in the weights of the two models. Likewise, we can expect differences to culminate in software engineering text which is different from tweets and essays due to the use of markdown, SE-specific terms, formality, and structure of pull requests.

Unlike the two academic models, Personality Insights (or PI) is not trained on one type of text source and is expected to be more general-purpose compared to the two academic models. It uses an open vocabulary and a machine learning model that continues to learn new words, phrases, topics, and categories [52]. This model acts as a black box generating personality scores for a given text.

Preprocessing. The three models generate five real numbers each, indicating the five personality traits. However, these numbers do not have a meaning in themselves and show personality relative to a population. For example, imagine two people, Alice and Bob, with an extraversion score of 1.25 and -2.3. How do we interpret the two scores? How big is the extraversion score 1.25 compared to the extraversion score -2.3?

To make the inferred personality scores meaningful, we bring them to a comparable scale, assuming that the sample population represents all personality types. We transform the numeric scores for each separate personality trait using min-max normalization to values including and between 0 and 1. In the revised scale, extraversion score 0 refers to an introvert, relative to the studied population. Similarly, score 1 refers to an extrovert, and a score of 0.5 refers to an average personality trait with combined introvert and extrovert characteristics.

3.2 Questionnaire

We use personality traits inferred from a questionnaire (also a gold standard) as our ground truth. Currently, there are two widely used questionnaires for inferring personality: NEO-PI-R [16] and Big Five Inventory (BFI) [31–33]. We choose to use BFI driven by two factors. One, we intended to reach a wider audience since only then can we reliably interpret the inferential ability of psycholinguistic tests. Two, we realize that people are less likely to react to our questionnaire since answers to our questions indicate their personality, and they may have privacy concerns. BFI is the shorter of the two methods, as filling in the questionnaire takes 5-10 minutes, compared to the 30-40 minutes required for the NEO-PI-R questionnaire. Furthermore, BFI's reported reliability and validity are comparable to NEO-PI-R [4, 6, 21]. We hypothesize that by selecting a less time-intensive questionnaire, we can reduce the number of participants dropping out, enabling us to reach a wider audience. We further incentivized participation by offering an Amazon gift card of 25 USD as prize. To mitigate privacy concerns, we informed our prospective survey respondents on our objective, implementation details, and their right to withdraw, in compliance with GDPR[5] and as approved by the University Ethics Board via Data Privacy Impact Assessment.

Our survey comprises 44 questions indicating personality (Openness (10), Extraversion (8), Agreeableness (9), Conscientiousness (9), and Neuroticism (8)). An example question in our survey is: 'I am someone who is talkative.' The answer to each question is a value between 1 and 5, with 1 representing 'strongly disagree' and 5 - 'strongly agree.'

Personality is then calculated as a sum of the answers to the following question numbers for a given personality trait [8, 31, 32] (see the survey in replication package for detail).

- Openness: 5, 10, 15, 20, 25, 30, 35R, 40, 41R, 44
- Conscientiousness: 3, 8R, 13, 18R, 23R, 28, 33, 38, 43R
- Extraversion: 1, 6R, 11, 16, 21R, 26, 31R, 36
- Agreeableness: 2R, 7, 12R, 17, 22, 27R, 32, 37R, 42
- Neuroticism: 4, 9R, 14, 19, 24R, 29, 34R, 39

here, 'R' is a reversed score calculated as 6 - <score> .

In addition to the questions relating to personality, we also asked questions indicating proficiency in English: (1) 'English is my mother tongue' (2) 'I am fluent in written English', and (3) 'In what country did you spend most your youth?'. The first two questions accepted a three-point Likert scale ('yes,' 'no,' and 'maybe') as valid answers. We also informed the participants about the compliance of our investigation to the GDPR.

3.3 Data Collection

We select contributors to infer personality by mining their communication history relating to software development and can gather ground truth by running a questionnaire. For some contributors, we cannot infer personality. This includes contributors who did not communicate on GitHub, or their communication history was insufficient or unavailable. We analyze development activities on GitHub for our analysis [35], as is also used by the recent studies on personality (e.g., [30, 44]).

We collected information on the top 3% of projects with most development activities from GHTorrent [25] (version June 2019). We set a lower bound of 33 pull requests for project selection (same as the new dataset for pull request research [64]) since such projects are less likely to exhibit communicative intentions amongst developers. Projects deleted at the time of creating the dataset were left out, as we could not trace the comments back to GitHub anymore. Ultimately, we collected data from 8,436 projects developed in Java, JavaScript, Python, Ruby, Go, and Scala. The selected projects may have more contributors. However, we leave out contributors who did not write any comments or whose comments are not available for analysis.

For each selected contributor, we had at least one written comment (in a pull request, issue, or commit) and possibly many comments available in the selected projects. While each contributor comment can suggest personality, we combine all available sentences and present them as the input to the three models for improved reliability.

[4] http://liwc.wpengine.com/

[5] https://gdpr-info.eu/

Table 1: presents the distribution of survey responses continent-wise and English proficiency inferred from self-perceived fluency and mother tongue English.

	Fluent			Mother tongue		
	Yes	No	Maybe	Yes	No	Maybe
Europe	102	7	11	8	110	2
Asia	49	9	9	5	59	3
North-America	40	0	0	34	6	0
South-America	15	2	1	0	18	0
Africa	12	0	2	5	8	1
Total	218	18	23	52	201	6

Preprocessing. Before fetching the comments' data to the model, we manually analyzed the comments and found two outlier behavior. First, some contributors had written less than 100 words, even after combining the texts from all written comments. Since shorter texts are less likely to reliably infer personality (e.g., PI tool denies request with less than 100 words [13]), we removed contributors with shorter texts from our analysis. Second, we found some accounts with an extraordinarily large size of text (e.g., *lintr-bot*, a bot for static code analysis for R). Upon closer inspection of these comments, we found that texts in such accounts have repetitive statements and texts such as "`I am a bot`" suggesting that these accounts are not operated by humans but by bots. We manually identified and after inspection removed all such accounts based on keyword search 'bot' and otherwise large text size. Finally, we analyze 4,081,957 comments written by 28,337 contributors from 8,436 projects.

Survey. From the 28,337 contributors, we chose a representative sub-sample worldwide, accounting for the observed regional differences in personality traits [34, 50]. Using K-Means clustering, we created six clusters on the world map, one centroid for each continent (except Antarctica). The location information for these contributors is inferred using the Bing Map API service[6], in combination with the information available on GHTorrent [25]. We randomly selected 2,050 participants from the six clusters, equally from each cluster for whom we could infer email addresses. We invited 2,050 participants to answer our research question.

To boost the survey response rate, we send customized emails to contributors at 10:00 AM in their timezone [20]. In the end, 267 people filled our questionnaire (a 13% response rate). A detailed description of the demographics of our survey respondents is presented in Table 1. Our participants over-represent Europe and the USA, but we have a representation of each continent.

3.4 Statistical tests

To answer the four questions, we need information on: (1) do there exist differences in the personality scores calculated using two methods? (2) if a difference exists, how big is the effect? and (3) overall accuracy of a method.

To check whether there exists a statistically significant difference in the personality scores inferred using two methods, we use the Student t-test [56] or the Wilcoxon signed-rank test [61]. When the data is (near) normally distributed, we use the Student t-test and Wilcoxon signed-rank test otherwise. We use a combination of the Shapiro-Wilk test and Q-Q plots to infer the non-normality of the data.

Table 2: presents effect size and its interpretation

Effect size	d	r	Cramér's V
negligible	<0.2	<0.1	<0.07
small	<0.5	<0.3	<0.21
medium	<0.8	<0.5	<0.35
large	≥ 0.8	≥ 0.5	≥ 0.35

If there are statistically significant differences in the distribution of personality scores inferred using any two models, we calculate effect size to indicate the actual differences in personality scores. We report effect size r [48], Cohen's d [14, 15], or Cramérs V [18] depending on data distribution and choice of method (see Table 2 for interpreting effect sizes). Cohen's d is used for normally distributed data, and we use r otherwise [14, 15]. Cramérs V is used for nominal data.

Finally, we report the accuracy of a method compared to the ground truth. We report accuracy in Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). MAE considers each error equally, and by calculating an average of the errors, it gives an average impression of the overall error. RMSE, in contrast, takes the square root of the average squared error, which means that more significant errors count more than minor errors. Collectively, the two metrics present an overall and nuanced view of errors with values closer to zero, indicating similarity in scores. In this paper, we report MAE and both MAE and RMSE in the technical report [58]. All statistical tests are conducted in R, details on which are available in the replication package.

3.5 Characteristics of SE communications data

RQ1. Do characteristics of the software engineering communications data influence inferred personality?

Psycholinguistic tests for inferring personality are trained on natural language text. When applied to GitHub communications data, the written text has two other sources of variability: (1) technical words relating to software engineering, and (2) GitHub flavored language (i.e., markdown markup language[7]). So, suppose the model proposed by Golbeck et al. were to infer personality. It can incorrectly take exclamation mark (used to integrate images as `![...](...)`) as an indicator for high conscientiousness, high neuroticism, and low openness to experience [23].

To systematically identify factors that can potentially influence the inferred personality score, we looked into four directions. First, we identified preprocessing steps used by the previous studies in software engineering [3, 5, 11, 49, 51]. Second, we searched for terms in LIWC dictionary that can be incorrectly represented in software engineering text. We looked for extreme/outlier personality scores to identify such elements and propose preprocessing steps based on our findings. Third, we investigated the influence of elements that raises privacy concerns (e.g., emails and IP addresses - European GDPR [17, 41]). Finally, we investigate the effect

[6] https://www.bingmapsportal.com/Application

[7] https://daringfireball.net/projects/markdown/

of removing elements for improved efficiency and reduced storage requirements.

We identified twelve preprocessing steps indicating (a) the influence of software engineering on the written text, (b) platform-specific features (e.g., Markdown), and (c) identifying/personal information. We identified removing (1) numbers [3, 5, 49, 51], (2) hashtag [5, 11], (3) URLs [3, 5, 11, 49], and (4) @-references [3] from previous studies. Relating to the terms incorrectly represented in software engineering, we identified removing (5) quotes, (6) code blocks, and (7) images as part of Markdown characteristics. For privacy reasons, we explore removing (8) email addresses and (9) IP addresses. Finally, for efficiency, we propose removing (10) upper case, (11) variants of white space (e.g., \r, \n, and \t) replacing it with a single white space, and (12) double white space and space before punctuation. Some or all of these steps can potentially influence the inferential ability of the psycholinguistic models for inferring personality in software engineering.

Approach. To gauge the impact of a preprocessing step, we compute personality scores with and without the step. If we find statistically significant differences in the distribution of scores (for paired observations), we explore whether the scores are better with or without the preprocessing step. Next, we compare the scores to the ground truth to see if the proposed change is for the better. We report the accuracy of the inferred personality scores using Mean Absolute Error and Root Mean Squared Error.

3.6 Comparison of models

RQ2. Is one model better than other in inferring developer personality from software engineering data?

We compare the scores inferred from each model to the ground truth and report accuracy in MAE and RMSE. But before we answer this question, we process the data in two ways. One, we applied the appropriate preprocessing steps identified in the previous research question to the input written text. Two, we manually investigate the distribution of scores in the three models to identify any need for transformation to bring the data from the three models to the same scale. We apply mean-centering such that the new mean for each score is zero [29]. Finally, we compare the scores inferred from each model (original and mean-centered) to the ground truth.

3.7 Size of input text

RQ3. Does the reliability of inferred personality change with the size of the input text?

Golbeck is trained on Twitter messages of size 50 to 5724 words [23], while Yarkoni is trained on blog posts with at least 50,000 words [62]. The text size mentioned above refers to the concatenation of available Twitter messages and blog posts for an author. PI requires all text to have at least a hundred words and allows up to an estimated 42,000 words[8].

To identify the optimal text size for inferring personality scores, we compare subsets of the same text (100, 600, 1200, and 3000 words) to each other. Our choice of sizes is inspired by PIs analysis of optimal text size [13]. In our dataset, we have 4,346 contributors who had written at least 3000 words are candidate for analysis. We select the first 'n' words for each text size to identify differences in personality score inferred using different text sizes. If vast differences exist, we compare the accuracy of the increased text size to the ground truth.

[8]PI allows for JSON formatted input of $250KB = 250 * 1024B = 256,000B$ in size. ASCII encoding uses 8 bits (=1 byte) per character. If we take an average of 5.1 letters per English word [1], with one space after each word, this gives an estimated $256,000/(5.1+1) \approx 42,000$ words.

3.8 English proficiency

RQ4. Does English proficiency relate to the reliability of personality inferred from psycholinguistic tests?

We explore English proficiency in two ways. Our first definition links *English mother tongue*, often referred to as the child's native or first acquired language [40], to English proficiency. Second, we explore *fluency*, referring to one's ability to express oneself easily, as an indication of English proficiency. The first definition is objective and captures the sub-population who innately speak English. The second definition is somewhat subjective but caters to the sub-population who acquired English proficiency by other means.

Survey responses to the question on English proficiency classified contributors into three groups: (1) English proficiency -yes, (2) English proficiency - no, and (3) English proficiency - maybe. We left the option 'maybe' since it showed considerable overlap with the yes and no groups, as inferred from the English Proficiency Index. [9] We compared the personality scores for the groups (yes and no) to each other to identify if there exist differences in inferred personality. We use unpaired tests for this question (i.e., Wilcoxon summed rank test and unpaired t-test). If there exist differences, we calculate its extent (by comparing it to ground truth) to find the influence of English proficiency on inferred personality.

4 RESULTS

RQ1. Do characteristics of the software engineering communications data influence inferred personality?

Eliminating platform-specific features (e.g., quotes and code block) improved inferred personality scores. Other factors did not affect personality scores but otherwise improved efficiency (e.g., double white space) and lowered privacy concerns (e.g., email address).

Our investigation of twelve preprocessing steps shows either improvement in the inferred personality score or improvements in efficiency and reduced privacy concerns without influencing the inferred personality score. Table 3 presents the maximum percentage improvements in the inferred personality traits (calculated as mean absolute error) when a preprocessing step is applied. The reported scores represent the maximum for the five personality traits (OCEAN) applied on the three models. Table 3 also shows the percentage of the text population that is affected by the preprocessing step. Remember that the characteristics of the text do not change with the choice of model or personality trait. We report findings from 11 out of 12 preprocessing steps. We do not present results from removing quoted text since this text is written by someone else. We do not consider the quoted text as a representation of the analyzed person's personality.

[9]https://www.ef.com/wwen/epi/

Table 3: Reports maximum percentage improvement in MAE and population affected for all traits for a preprocessing.

Preprocessing step	Improvement MAE (%)	Population (%)
Code blocks	<36%	<100%
Remove URLs	<18.2%	<99.9%
Remove images	<11.1%	<99%
Remove numbers	<9.1%	<99.9%
Remove @-ref	<6.7%	<93.9%
Remove IP	<5.3%	<95.5%
Remove hashtags	<5%	<91.4%
Lowercase parsing	0%	<0.1%
Whitespace parsing	0%	0%
Remove emails	0%	<0.1%
Remove spaces	0%	0%

We found that removing code blocks, quotes, images, URLs, and numbers improves the models' accuracy. The first three factors – code blocks, quotes, images – are platform and software engineering-specific features, and as expected, their removal improves inferred personality scores. We recommend to always remove quotes and code blocks as they can reflect another person's personality. Likewise, by removing the markdown text indicative of images, which otherwise can be linked with personality, we reduce the chances of misclassification. Removing URLs also improved our inference, a factor identified in prior studies [3, 5, 11, 49]. We believe that the improvement is attributed to the words in the URL which are otherwise misclassified as a part of the communication.

Another factor identified in the existing studies is the removal of numbers [3, 5, 49, 51]. Generally, our analyses show that removing numbers improves the inferred personality scores, except for Yarkoni's extraversion and agreeableness. For Yarkoni's extraversion and agreeableness, we observed that removing numbers made the inference less accurate. See the technical report for details [58]. On further investigating Yarkoni's model, we found that the word category *Number* correlates to extraversion and agreeableness [62], and hence the observation. 'Number' has no relation to any other personality trait in the other two models. Golbeck does not use it and there is no effect on PI.

Other preprocessing steps did not influence the personality score but improved another property. We found that removing the upper casing and spaces before punctuation and conformance of white space did not influence inferred personality (refer to Table 3) but improved processing speed and reduced storage needs. The remaining factors: hashtags [5, 11] and @-references/usernames [3] (identified in the literature) and email addresses and IP addresses (found in this study), did not influence personality score but reduced privacy concerns by eliminating personally identifiable information.

For the remainder of this study, we applied all twelve preprocessing steps identified above on the input text by removing them from the text. The only exceptions are Yarkoni's extraversion and agreeableness for which we retained the word category 'number'.

RQ2. Is one model better than other in inferring developer personality from software engineering data?
The three models perform comparably when brought to the same scale.

Figure 2 presents the distribution of personality scores for each of the three models and the ground truth. Figure 2 shows that PI and Yarkoni have a similar distribution compared to Golbeck. Generally, Golbeck has a high average personality score indicating big outliers, except for neuroticism.

Figure 2 also indicates a similar distribution in personality scores, albeit at different scales. Therefore, before comparing any two models, we mean-centered the scores. We observed that while the differences among the three methods are still statistically significant, the differences in MAE and RMSE decrease to have a negligible effect when mean-centered. The differences among the models with and without mean-centering is shown in Figure 3. This suggests that the three models perform comparably but at different scales. A detailed comparison of the differences among the models with and without mean centering is available in the technical report [58].

Table 4: Maximum absolute error in inferring personality scores at 90%, 95%, and 99% confidence intervals.

	PI					Yarkoni					Golbeck				
	O	C	E	A	N	O	C	E	A	N	O	C	E	A	N
90%	.28	.36	.39	.51	.36	.25	.32	.41	.40	.37	.22	.31	.32	.26	.61
95%	.34	.43	.46	.54	.41	.29	.38	.48	.45	.43	.25	.38	.37	.31	.68
99%	.42	.53	.55	.61	.50	.38	.46	.56	.52	.50	.29	.51	.44	.42	.77

When on the same scale, we found that most traits can be inferred with a 25-48% error rate from the ground truth with 95% confidence, except Golbeck neuroticism (68%) and PI agreeableness (54%) (see Table 4). Table 4 also shows error rates at 90% and 99% confidence. With error margins this high, one must be cautious in interpreting personality scores, particularly with the two extreme cases: PI agreeableness and Golbeck neuroticism. For these two traits, the high error rate renders the results meaningless.

When inferring individual personality, in the worst case, this implies that a person deemed high on agreeableness can actually be low on agreeableness. This problem, however, can somewhat subside when seen collectively (as is in the case of group or team personality). In case of group personality, aggregation of scores can reduce the effect of error margins contingent on the choice of aggregation technique (e.g., median instead of mean). This group personality are the personal characteristics or qualities shared by the members of the group, and the group personality composition has been observed to influence group effectiveness in domains outside of software engineering [26]. We believe that personality inferred from the three psycholinguistic tests will be a better representation of the team than individuals.

Other than the above, we also observed that transformations reduced the effect of large outliers, improving the scores. For future research, we recommend applying transformation(s) to minimize the effects of outliers.

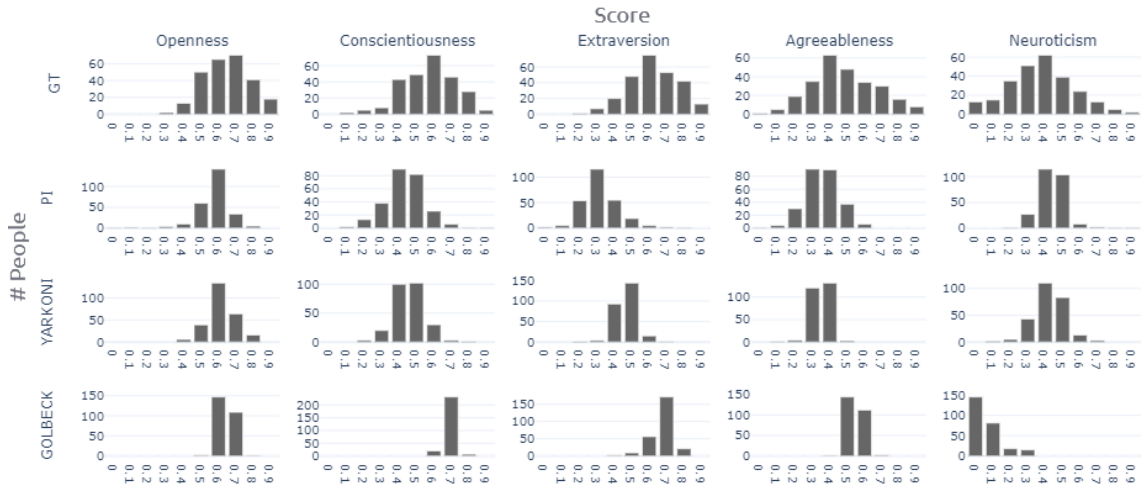

Figure 2: presents the personality traits distribution for the three psycholinguistic tests. Row 1 shows the ground truth (GT) in comparison to PI, Yarkoni and Golbeck. Each histogram presents the distribution of personality scores for a single trait.

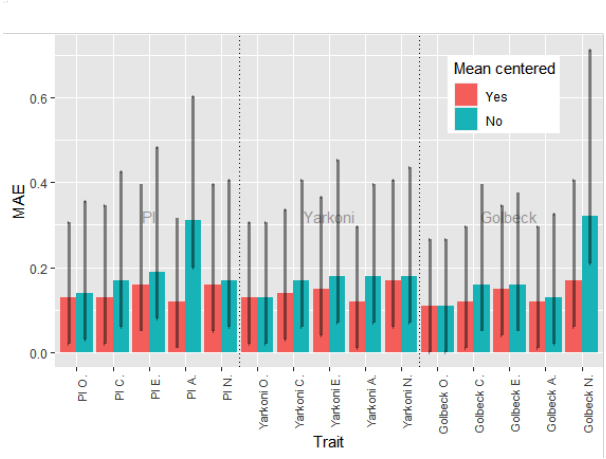

Figure 3: MAE scores for each personality trait with and without mean-centering

RQ3. Does the reliability of inferred personality change with the size of the input text?
Inferential ability of PI and Yarkoni is optimal at 600-1200 words and 600 words for Golbeck. Text with less than 100 words gives unreliable estimates, while beyond 3000 words, we expect no further improvements.

For each model, we compared personality scores with increasing text size: 100, 600, 1200, and 3000 (similar to the research of PI [13]). We compared personality inferences of text size 100 to text size 600, 600 to 1200, and 1200 to 3000. The optimal text size is the one after which there are no significant improvements in the inferred personality scores with the increasing text size. Table 5 presents the optimal text size for each personality trait of the three models. Generally, the inferred personality scores increased with text size. However, for Yarkoni conscientiousness and Golbeck neuroticism, we did not find a consistent pattern and hence no optimal text size.

Table 5: Optimal text size for personality inferences

	PI	Yarkoni	Golbeck
O	600	100	3000
C	1200	?	600
E	1200	1200	600
A	3000	3000	600
N	1200	600	?

Our comparison of personality scores inferred with text sizes 100, 600, 1200, and 3000 words show that an optimal text size for PI and Yarkoni mostly ranges from 600 to 1200 words. After 1200 words, we did not see significant improvements in personality score. For Golbeck, the optimal text size is 600 words. This is a lower word count compared to the previous estimates by PI [13] and Yarkoni [62] that shows no significant improvements in MAE after 3000 words. Golbeck does not provide such an estimate, but we have no reason to believe it needs more than 3000 words.

RQ4. Does English proficiency relate to the reliability of personality inferred from psycholinguistic tests?
Fluency in English influences openness to experience scores. Other traits are inferred more accurately for non-fluent people by PI and Yarkoni and by Golbeck for fluent people.

We observed that, depending on the model, English proficiency is linked to the differences in personality traits. Generally, PI and Yarkoni generate somewhat less accurate scores for fluent people in comparison to the ground truth. Golbeck, on the contrary, generates less accurate scores for non-fluent people (refer to the technical report for all the scores [58]). Specifically, PI agreeableness and

Table 6: presents traits for which inferred personality changes with the choice of model and its effect size.

	Fluency			Mother tongue		
		MAE			MAE	
Trait	Difference effect size	Yes	No	Difference effect size	Yes	No
PI A	small	.32	.23	-	-	-
Yarkoni N	medium	.18	.13	-	-	-
Golbeck O	large	.13	.16	medium	.13	.11

Yarkoni neuroticism present worse scores for people fluent in English by a small and medium amount, respectively (see Table 6). In the case of Golbeck openness to experience, we observe significant differences relating to fluency and medium differences for mother tongue in favor of fluent people.

Table 7: Mean openness scores for each method and the ground truth for fluent and non-fluent people and people with and without English as their mother tongue.

	Fluent		Mother tongue	
Method	Yes	No	Yes	No
PI	0.63	0.6	0.63	0.63
Yarkoni	0.68	0.64	0.68	0.67
Golbeck	0.7	0.67	0.71	0.7
Ground truth	0.71	0.64	0.74	0.7

That said, openness to experience generally changes with fluency for all the three models, with less fluent people being inferred as less open to experience (see Table 7 for details). We observed a similar pattern for the ground truth. The mean score of fluent people ($M = 0.71$) is significantly lower than the mean of non-fluent people ($M = 0.65$), $V = 6.55$, $p < 0.05$. This could indicate that the differences found for openness are introduced by the population, not by the methods used. In addition, mother tongue English is linked to higher openness to experience than the sub-population whose mother tongue is not English, except for PI, which shows the same (refer Table 7).

This finding in itself is not surprising, as several studies have shown the link of cultural background [47] and geographical location [50] to the openness to experience, which can be linked to English proficiency. Earlier studies found lower openness scores for the people in East-Asia than for the people in Europe [34, 50]. Mak and Tran [38] further added that Asians with high English proficiency (in terms of fluency) are reportedly more open to experience.

Our ground-truth, however, did not show a significant difference in the means among Europe ($M = 0.70$), Asia ($M = 0.69$) and America ($M = 0.73$). This also applied to the three methods, which did not significantly differ in the inferred personality scores between continents. We do not have information about the culture and demographics of our participants to study its influence.

Our findings imply that when inferring personality using psycholinguistic tests, we can incorrectly infer personality scores for two reasons: differences in the background of the participants (including English proficiency) and the limitations of the model itself.

5 DISCUSSION AND IMPLICATIONS

Do psycholinguistic tests (trained on different text source(s)) reliably infer developer personality from SE communications data? *Partially yes*

Our study shows that irrespective of the choice of psycholinguistic test (with different text size requirements) and the application of proposed preprocessing steps, personality traits can be inferred with an average error rate of 41% at 95% confidence. While we found that twelve out of the thirteen preprocessing steps can improve personality inference up to 36% (in the best case), personality traits such as Golbeck neuroticism can have an error rate up to 68%.

With error margins this high, individual personality inferences are far from an accurate depiction. Notably, the high error rates found here corroborate with the existing research inferring personality using psycholinguistic tests in software engineering [44], but also broadly [37]. Other than the limits of the psycholinguistic tests, our study further highlights the role of English proficiency. The current personality traits make an implicit assumption that the written text is only a reflection of author. In reality, proficiency in English - an individual characteristic, can influence the written text and hence the inferred score.

With such fault margins, we urge people to be careful with concluding from the inferred personality scores. Therefore, a peril of all covered proposed psycholinguistic methods, currently, is their ability to predict all personality types accurately for individuals. A such, we should be very much aware that the inferred personality can be very different from the actual personality. Therefore, we issue a stark warning that this approach should not be used for making decisions relating to individuals. In the case of personality inference on an individual level, one must be careful to use the scores as an indicator, not a truth value. While considering the possible error, the personality scores can be effectively used for team formations or group-related research. However, we do expect that the approach can work better when analyzing the personality at a group level. When personality is measured for a group, depending on the choice of aggregation techniques (e.g., median instead of mean), psycholinguistic models can offer a reasonable estimate.

Next, we present recommendations on choosing a model and optimally inferring developer personality from software communications data.

What model should I choose? *Any of the three*

Our study shows that no one model is better than others in inferring developer personality from software engineering data, except when looking for specific personality traits. Irrespective of the choice of text sources (e.g., tweets vs. blogs), or its count (single data source vs. multiple data sources), the personality inferred from the three models are similar, with some exceptions. We recommend that with an appropriate amount of text size, all the models perform similarly. This is particularly helpful now that PI is deprecated [10] - a widely used option recently.

[10] https://cloud.ibm.com/docs/personality-insights?topic=personality-insights-about#about

How to infer developer personality optimally? *Clean the data, choose any of the three models with optimal text size and interpret the findings in relation to author's English proficiency and known error margins.*

We recommend a three-step process starting with preprocessing the software communications data based on the 12 steps identified in this study. Next, apply any model or choose a model that reportedly works best for a specific personality trait of interest. For instance, studies interested in personality trait neuroticism should avoid Golbeck's model (see Table 4). Depending on the choice of model, choose a minimum text size for optimally reliable scores. For example, when using Golbeck, 600 words will suffice while use 600-1200 words for Yarkoni. Ultimately, once the personality scores are found, it is essential to interpret them within the specific context of the person whose traits are inferred and the error margins when using psycholinguistic tests.

This brings us to the next question that while we can infer developer personality from software communications data,

... should we apply psycholinguistic tests to software engineering data to infer developer personality? *Yes and No*

We will answer this question through two sub-questions: (1) Can we apply? and (2) Should we apply? The answer to the first question is yes. We can apply psycholinguistic tests to infer developer personality at scale, but we need to be aware of the % error rates. Despite the error rates found here, many studies in software engineering have demonstrated the applications of inferring developer personality. Recent studies in software engineering have shown how developer personality can reflect in their contributions [10, 44], pull request acceptance [30] and project success [63].

The second aspect of this question is 'should we apply?', and the answer here is tricky. On the one hand, studies on personality have improved our understanding of development practices and software development, in general. The other extreme is the perpetual harm that studies of this kind can bring to an individual and community. For instance, filtering candidates for hiring merely based on personality traits can not only lead to false conclusions but also discrimination against certain personality types [55]. Another factor is mental health. Studies show that bipolar disorder can influence personality inference [12], since these inferences are a snapshot in time. Alternatively, a wrong judgement on personality (e.g., during hiring process) can have a backlash on the mental health.

Can psycholinguistic tests perform better on software engineering text? *Maybe*

This question too can be divided into two parts: (1) Can existing models perform better? and (2) Can we design optimal psycholinguistic models specific to software engineering?

We have no reasons to believe how and why existing models can perform better. We used models used in academia and industry; these models are trained on different sized text (small for tweets and long for essays), different sources (single vs. multiple), and trained once vs. constant learning. Despite these variabilities in the model, we did not find any model performing better than the other. Therefore, we believe that existing models cannot perform better. One possibility can be to rethink how we separate natural text from software engineering text, as suggested by Bachelli et al. [7].

That said, we can design psycholinguistic models specific to software engineering. The solutions proposed in this study optimize personality inference on syntactic elements such as removing markdown. Software communications data, however, have semantics that work differently in software engineering than a usual conversation. For example, a term such as 'cookies' is assigned to the word category 'bio,' which has a special meaning in the software engineering context. The term 'cookie' can also mean the food we understand from everyday life, but less likely in a software engineering context. With the current research, it is not evident whether designing an optimal psycholinguistic model specific to software engineering will help. More work is required to substantiate the claim.

5.1 Implications

Research: Our study presents the promises and perils of mining GitHub communications data for inferring developer personality. These findings can serve as a guideline for future research building on psycholinguistic tests for understanding a software engineering phenomenon. Further, by highlighting the limits of psycholinguistic tests for inferring personality in software engineering, our study opens up avenue for next steps (see Section 7 for future work).

Practice: Our study shows the practical usability of the existing solutions and the ethical concerns that one should consider prior to use. Further, our study offers a frame of reference on (1) how to infer personality scores optimally (based on the existing psycholinguistic models)? and (2) how to interpret it?

Education: Our study shows how signals (in our case derived from software communications data) can infer complex concepts, such as personality and means of doing it.

5.2 Comparison to related work

If we consider earlier studies on personality among software engineers, the found scores are not unexpected. In the study by Calefato et al. [9] the mean openness found among their participating developers (M = 0.79) is reasonably close to the mean openness found for this study (M = 0.71). Similarly, their mean for conscientiousness (M = 0.6) and agreeableness (M = 0.64) are close to our means for conscientiousness (M = 0.63) and agreeableness (M = 0.68). As earlier studies show similar distributions of scores, this indicates that our extraction process for personality traits from text works reasonably well.

In terms of the accuracy of automatically extracting personality traits from text, we compare our work against the large-scale analysis of personality in software engineering by Calefato et al. [10]. In their comparison of personality models, they observe an accuracy ranging from 40–70%, which is globally in line with our own results.

6 LIMITATIONS AND THREATS TO VALIDITY

Construct validity. Our analysis is as good as the ground truth. We used a lightweight questionnaire - which is the closest we had to the gold standard with scalability. Another choice we made relates to the written text. We selected the first 'n' number of words written by a person to gauge their personality. Had we chosen the middle 'n' words or the last 'n' words, our findings could have been different, but mostly since studies show that personality evolves over time, although slowly [9, 45]. Another related factor is the size of the text. We selected contributors whose written text (at least 100 words) is available for analysis. This, on the one hand, defines the limit of our approach. On the other hand, it can systematically exclude some personality traits. The same argument applies to surveys, where the respondents may have self-selection bias.

Internal validity. Our study presents inferred personality scores by applying preprocessing steps. While we systematically identify these steps, we might have missed steps that do not fall in our purview (e.g., SHA references or emojis). These words do not contribute to personality scores but add to word count, thereby likely influencing the inference. Another factor that can potentially inflate the reported percentage improvements is the normalization of personality scores. While we normalize the data to help understand the personality scores, change in a person's personality score after preprocessing can influence the normalized scores of others. This is a necessary trade-off, but we advise our readers to remember this potential side-effect while gauging the potential of preprocessing steps.

External validity. Finally, the usefulness of our results is as good as the sub-populations to which it apply. We have no reasons to believe why and how software communications data on other platforms will differ from Github (other than the differences in the features of the platform), suggesting that our findings should be generalizable. Also, we believe that the factors found important here should apply to other platforms, although their relative relevance can change. We will need more studies to substantiate these claims.

To counter self-selection, we have performed random undersampling on the majority class to ensure that each continent is equally represented. Through this sampling, we select 2,050 participants accounting for the observed regional differences in personality traits [34, 50].

7 CONCLUSIONS AND FUTURE WORK

This paper comes as a guideline for inferring personality using software engineering communications data. By comparing the personality scores inferred from three state-of-the-practice models to the ground truth, we provide recommendations on the promises and perils of mining GitHub communications data for inferring personality scores. We identify 12 preprocessing steps that improve personality inferences, yet the average error rate is 41% with 95% confidence. We also identify optimal text size for reliable personality inferences and recommend choosing any of the three models with some exceptions. Finally, we highlight the role of English proficiency and error margins while interpreting personality scores.

Knowing the limits of the existing solutions, future research should take one of the two possible directions. One, propose a solution specific to software engineering (e.g., process the software engineering text to resemble natural conversations or modify the model to perform optimally on software engineering communications data). Alternatively, look for personality cues in places other than text (e.g., software code and activity patterns) as indicators of personality.

ACKNOWLEDGMENTS

We thank all survey participants and Xunhui Zhang for technical support. This research was partially funded by the Dutch science foundation NWO through the Vici "TestShift" grant (No. VI.C.182.032).

REFERENCES

[1] [n.d.]. https://www.wolframalpha.com/input/?i=averageenglishwordlength
[2] S. Acuña, M. Gómez, J. Hannay, N. Juristo, and D. Pfahl. 2015. Are team personality and climate related to satisfaction and software quality? Aggregating results from a twice replicated experiment. *Information and Software Technology* 57 (01 2015), 141–156. https://doi.org/10.1016/j.infsof.2014.09.002
[3] A. Alamsyah, M. F. Rachman, C. S. Hudaya, R. P. Putra, A. I. Rifkyano, and F. Nurwianti. 2019. A Progress on the Personality Measurement Model using Ontology based on Social Media Text. In *2019 International Conference on Information Management and Technology (ICIMTech)*, Vol. 1. 581–586. https://doi.org/10.1109/ICIMTech.2019.8843817
[4] B. Alansari. 2016. The Big Five Inventory (BFI): Reliability and validity of its Arabic translation in non clinical sample. *European Psychiatry* 33 (2016), S209 – S210. https://doi.org/10.1016/j.eurpsy.2016.01.500
[5] P. Arnoux, A. Xu, N. Boyette, J. Mahmud, R. Akkiraju, and V. Sinha. 2017. 25 Tweets to Know You: A New Model to Predict Personality with Social Media. *CoRR* abs/1704.05513 (2017). arXiv:1704.05513 http://arxiv.org/abs/1704.05513
[6] B. J. Arterberry, M. P. Martens, J. M. Cadigan, and D. Rohrer. 2014. Application of Generalizability Theory to the Big Five Inventory. *Personality and individual differences* 69 (Oct. 2014), 98–103. https://doi.org/10.1016/j.paid.2014.05.015
[7] Alberto Bacchelli, Tommaso Dal Sasso, Marco D'Ambros, and Michele Lanza. 2012. Content classification of development emails. In *2012 34th International Conference on Software Engineering (ICSE)*. IEEE, 375–385.
[8] V. Benet and O. 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 (10 1998), 729–50. https://doi.org/10.1037//0022-3514.75.3.729
[9] F. Calefato, G. Iaffaldano, F. Lanubile, and B. Vasilescu. 2018. On Developers' Personality in Large-scale Distributed Projects: The Case of the Apache Ecosystem. In *Proceedings of the 13th International Conference on Global Software Engineering (ICGSE '18)*. ACM, New York, NY, USA, 92–101. https://doi.org/10.1145/3196369.3196372
[10] Fabio Calefato, Filippo Lanubile, and Bogdan Vasilescu. 2019. A large-scale, in-depth analysis of developers' personalities in the Apache ecosystem. *Information and Software Technology* 114 (2019), 1–20. https://doi.org/10.1016/j.infsof.2019.05.012
[11] G. Carducci, G. Rizzo, D. Monti, E. Palumbo, and M. Morisio. 2018. TwitPersonality: Computing Personality Traits from Tweets Using Word Embeddings and Supervised Learning. *Information (Switzerland)* 9 (5 2018). https://doi.org/10.3390/info9050127
[12] Chun-Hao Chang, Elvis Saravia, and Yi-Shin Chen. 2016. Subconscious Crowdsourcing: A feasible data collection mechanism for mental disorder detection on social media. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 374–379.
[13] IBM Cloud. 2019. IBM Cloud Docs Personality Insight. https://cloud.ibm.com/docs/services/personality-insights
[14] J. Cohen. 1988. *Statistical power analysis for the behavioral sciences*. Routledge, New York, USA. https://doi.org/10.4324/9780203771587
[15] J. Cohen. 1992. Statistical power analysis. *Current directions in psychological science* 1, 3 (1992), 98–101.
[16] Paul T Costa Jr and Robert R McCrae. 2008. *The Revised NEO Personality Inventory (NEO-PI-R)*. Sage Publications, Inc.
[17] Council of European Union. 2016. Council regulation (EU) no 2016/679. https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng
[18] H. Cramér. 1999. *Mathematical methods of statistics*. Vol. 43. Princeton university press.
[19] F Q. B. da Silva, A. C. C. França, M. Suassuna, L. M. R. de Sousa Mariz, I. Rossiley, R. C. G. de Miranda, T. B. Gouveia, C. V. F. Monteiro, E. Lucena, E. S. F. Cardozo,

and E. Espindola. 2013. Team building criteria in software projects: A mix-method replicated study. *Information and Software Technology* 55, 7 (2013), 1316–1340. https://doi.org/10.1016/j.infsof.2012.11.006

[20] K. S. Faught, D. Whitten, and K. W. Green Jr. 2004. Doing Survey Research on the Internet: Yes, Timing Does Matter. *Journal of Computer Information Systems* 44, 3 (2004), 26–34. https://doi.org/10.1080/08874417.2004.11647579 arXiv:https://www.tandfonline.com/doi/pdf/10.1080/08874417.2004.11647579

[21] A. Fossati, S. Borroni, D. Marchione, and C. Maffei. 2011. The Big Five Inventory (BFI): Reliability and Validity of its Italian Translation in Three Independent Nonclinical Samples. *European Journal of Psychological Assessment - EUR J PSYCHOL ASSESS* 27 (01 2011), 50–58. https://doi.org/10.1027/1015-5759/a000043

[22] A. Gilal, J. Jaafar, M. Omar, S. Basri, and A. Izzatdin. 2016. Balancing the Personality of Programmer: Software Development Team Composition. *Malaysian Journal of Computer Science* 29 (03 2016). https://doi.org/10.22452/mjcs.vol29no2.5

[23] J. Golbeck, C. Robles, M. Edmondson, and K. Turner. 2011. Predicting Personality from Twitter. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*. IEEE, Boston, MA, USA, 149–156. https://doi.org/10.1109/PASSAT/SocialCom.2011.33

[24] J. Golbeck, C. Robles, and K. Turner. 2011. Predicting personality with social media. *Conference on Human Factors in Computing Systems - Proceedings* (1 2011), 253–262. https://doi.org/10.1145/1979742.1979614

[25] G. Gousios. 2013. The GHTorrent dataset and tool suite. In *Proceedings of the 10th Working Conference on Mining Software Repositories (MSR '13)*. IEEE Press, Piscataway, NJ, USA, 233–236. http://dl.acm.org/citation.cfm?id=2487085.2487132

[26] Terry Halfhill, Eric Sundstrom, Jessica Lahner, Wilma Calderone, and Tjai M. Nielsen. 2005. Group Personality Composition and Group Effectiveness: An Integrative Review of Empirical Research. *Small Group Research* 36, 1 (2005), 83–105. https://doi.org/10.1177/1046496404268538 arXiv:https://doi.org/10.1177/1046496404268538

[27] R. V. Hamilton. 1957. A Psycholinguistic Analysis of some Interpretive Processes of Three Basic Personality Types. *The Journal of Social Psychology* 46, 2 (1957), 153–177. https://doi.org/10.1080/00224545.1957.9714317 arXiv:https://doi.org/10.1080/00224545.1957.9714317

[28] Ong Hee. 2014. Validity and Reliability of the Big Five Personality Traits Scale in Malaysia. *International Journal of Innovation and Applied Studies* (04 2014).

[29] Dawn Iacobucci, Matthew J. Schneider, Deidre L. Popovich, and Georgios A. Bakamitsos. 2016. Mean centering helps alleviate "micro" but not "macro" multicollinearity. *Behavior Research Methods* 48, 4 (01 Dec 2016), 1308–1317. https://doi.org/10.3758/s13428-015-0624-x

[30] Rahul N Iyer, S Alex Yun, Meiyappan Nagappan, and Jesse Hoey. 2019. Effects of personality traits on pull request acceptance. *IEEE Transactions on Software Engineering* (2019).

[31] O. John, L. Naumann, and C. Soto. 2008. *Paradigm shift to the integrative big five trait taxonomy: History, measurement, and conceptual issues.* Guilford Press, 114–158.

[32] O. P. John, E. M. Donahue, and Kentle R. L. 1991. The "Big Five" Inventory - Versions 4a and 54. *Journal of Personality and Social Psychology* (1991). https://doi.org/10.1037/t07550-000

[33] O. P. John and S. Srivastava. 1999. The Big Five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research* 2, 1999 (1999), 102–138.

[34] P. J. Kajonius. 2017. Cross-cultural personality differences between East Asia and Northern Europe in IPIP-NEO. *International Journal of Personality Psychology* 3, 1 (2017), 1–7.

[35] Eirini Kalliamvakou, Georgios Gousios, Kelly Blincoe, Leif Singer, Daniel M. German, and Daniela Damian. 2014. The Promises and Perils of Mining GitHub. In *Proceedings of the 11th Working Conference on Mining Software Repositories (MSR 2014)*. Association for Computing Machinery, New York, NY, USA, 92–101. https://doi.org/10.1145/2597073.2597074

[36] M. V. Kosti, R. Feldt, and L. Angelis. 2014. Personality, emotional intelligence and work preferences in software engineering: An empirical study. *Information and Software Technology* 56, 8 (2014), 973–990. https://doi.org/10.1016/j.infsof.2014.03.004

[37] Adam DI Kramer, Jamie E Guillory, and Jeffrey T Hancock. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences* 111, 24 (2014), 8788–8790.

[38] A. S. Mak and C. Tran. 2001. Big five personality and cultural relocation factors in Vietnamese Australian students' intercultural social self-efficacy. *International Journal of Intercultural Relations* 25, 2 (2001), 181–201. https://doi.org/10.1016/S0147-1767(00)00050-X

[39] R. McCrae and P. Costa. 1987. Validation of the five factor model of personality across instruments and observers. *Journal of personality and social psychology* 52 (02 1987), 81–90. https://doi.org/10.1037/0022-3514.52.1.81

[40] D. Mizza. 2014. The First Language (L1) or Mother Tongue Model Vs. The Second Language (L2) Model of Literacy Instruction. *Journal of Education and Human Development* 3 (01 2014). https://doi.org/10.15640/jehd.v3n3a8

[41] Opinion GDPR June 2007. Opinion 4/2007 on the concept of personal data. Article 29 Data Protection Working Party.

[42] O. H. P. Pabón, F. A. González, J. Aponte, J. E. Camargo, and F. Restrepo-Calle. 2016. Finding Relationships between Socio-Technical Aspects and Personality Traits by Mining Developer E-mails. In *2016 IEEE/ACM Cooperative and Human Aspects of Software Engineering (CHASE)*. IEEE, Austin, TX, USA, 8–14. https://doi.org/10.1109/CHASE.2016.010

[43] J. Pennebaker. 2011. The secret life of pronouns. *New Scientist - NEW SCI* 211 (9 2011), 42–45. https://doi.org/10.1016/S0262-4079(11)62167-2

[44] Ayushi Rastogi and Nachiappan Nagappan. 2016. On the personality traits of GitHub contributors. In *2016 IEEE 27th International Symposium on Software Reliability Engineering (ISSRE)*. IEEE, 77–86.

[45] A. Rastogi and N. Nagappan. 2016. On the Personality Traits of GitHub Contributors. In *2016 IEEE 27th International Symposium on Software Reliability Engineering (ISSRE)*. 77–86.

[46] Peter C Rigby and Ahmed E Hassan. 2007. What can oss mailing lists tell us? a preliminary psychometric text analysis of the apache developer mailing list. In *Fourth International Workshop on Mining Software Repositories (MSR'07: ICSE Workshops 2007)*. IEEE, 23–23.

[47] J. Rolland. 2002. *The Five-Factor Model of Personality Across Cultures.* 7–28. https://doi.org/10.1007/978-1-4615-0763-5_2

[48] R. Rosenthal, H. Cooper, and L. Hedges. 1994. Parametric measures of effect size. *The handbook of research synthesis* 621, 2 (1994), 231–244.

[49] S. Sagadevan, N. H. A. H. Malim, and M. H. Husin. 2015. Sentiment Valences for Automatic Personality Detection of Online Social Networks Users Using Three Factor Model. *Procedia Computer Science* 72 (2015), 201–208. https://doi.org/10.1016/j.procs.2015.12.122

[50] D. Schmitt, J. Allik, R. McCrae, V. Benet, J. Veríssimo, and U. Reips. 2007. The geographic distribution of Big Five personality traits. *Journal of Cross-Cultural Psychology* 38 (03 2007), 173–212. https://doi.org/10.1177/0022022106297299

[51] M. Schoonvelde, G. Schumacher, and B. Bakker. 2019. Friends With Text as Data Benefits: Assessing and Extending the Use of Automated Text Analysis in Political Science and Political Psychology. *Journal of Social and Political Psychology* 7 (2019), 124–143. Issue 1. https://doi.org/10.5964/jspp.v7i1.964

[52] H. Schwartz, J. Eichstaedt, M. Kern, L. Dziurzynski, S. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. Seligman, and L. Ungar. 2013. Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PloS one* 8 (09 2013), e73791. https://doi.org/10.1371/journal.pone.0073791

[53] A. S. Sodiya, H. Longe, A. Onashoga, A. Oludele, and L. Omotosho. 2007. An Improved Assessment of Personality Traits in Software Engineering. In *InSITE 2007: Informing Science + IT Education Conference.* https://doi.org/10.28945/3164

[54] C. Soto. 2018. *Big Five personality traits.* SAGE Publications, Thousand Oaks, California, US, 240–241.

[55] Eugene F Stone-Romero. 2005. Personality-based stigmas and unfair discrimination in work organizations. *Discrimination at work: The psychological and organizational bases* (2005), 247–272.

[56] Student. 1908. The Probable Error of a Mean. *Biometrika* 6, 1 (1908), 1–25. http://www.jstor.org/stable/2331554

[57] Y. Tausczik and J. Pennebaker. 2010. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology* 29 (3 2010), 24–54. https://doi.org/10.1177/0261927X09351676

[58] Frenk C.J. van Mil. 2020. *Inferring Personality from GitHub Communication Data: Promises & Perils.* Master's thesis. Delft University of Technology.

[59] Frenk C.J. van Mil. 2020. Scripts to reproduce "Promises and Perils of Inferring Personality on GitHub". https://doi.org/10.5281/zenodo.3865341.

[60] Frenk C.J. van Mil. 2020. Supplementary data for the Master Thesis: Inferring Personality from GitHub Communication Data: Promises and Perils. https://data.4tu.nl/articles/dataset/Supplementary_data_for_the_Master_Thesis_Inferring_Personality_from_GitHub_Communication_Data_Promises_Perils/12702809/1

[61] F. Wilcoxon. 1945. Individual Comparisons by Ranking Methods. *Biometrics Bulletin* 1, 6 (1945), 80–83. http://www.jstor.org/stable/3001968

[62] T. Yarkoni. 2010. Personality in 100,000 Words: A large-scale analysis of personality and word use among bloggers. *Journal of research in personality* 44 (6 2010), 363–373. https://doi.org/10.1016/j.jrp.2010.04.001

[63] Seonghu Yun. 2020. *Personality Traits of GitHub Maintainers and Their Effects on Project Success.* Master's thesis. University of Waterloo.

[64] Xunhui Zhang, Ayushi Rastogi, and Yue Yu. 2020. On the Shoulders of Giants: A New Dataset for Pull-based Development Research. In *Proceedings of the 17th International Conference on Mining Software Repositories.* 543–547.