Inferring Personality from GitHub Communication Data: Promises & Perils

Version of June 21, 2020

Frenk C.J. van Mil

Inferring Personality from GitHub Communication Data: Promises & Perils

THESIS

submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE

in

COMPUTER SCIENCE

by

Frenk C.J. van Mil

born in Schipluiden, the Netherlands

Software Engineering Research Group
Department of Software Technology
Faculty EEMCS, Delft University of Technology
Delft, the Netherlands
www.ewi.tudelft.nl

Inferring Personality from GitHub Communication Data: Promises & Perils

Author: Frenk C.J. van Mil

Student id: 4464907

Email: f.c.j.vanmil@student.tudelft.nl

Abstract

Personality plays a significant role in our lives; it does not only influence what we think, feel, and do, but also affects what we say about what we think, feel, and do. In software engineering (SE), it might help in improving team composition through a combination of personalities within a team, and it could help explain work preferences and work satisfaction. Earlier studies in the field of software engineering have focused on extracting personality from developers with the use of questionnaires and automatic tools such as psycholinguistic tests. Psycholinguistic tests infer personality based on the words people use. As taking questionnaires is time-consuming, the interest in automated tools has grown. However, there is a lack of studies comparing different psycholinguistic models on actual SE data to validate to what extent these tools apply to software engineering.

In this study, we compare two well-established academical models proposed by Yarkoni [124] and Golbeck et al. [47] and the popular industrial model Personality Insights by IBM. We use the three models to infer personality from comments on open-source projects on GitHub and compare the found scores to a ground-truth obtained through a questionnaire among software developers.

In this study, we establish a baseline and compare three models on their performance to this baseline. We show three methods to perform almost equally when mean-centered, indicating the three methods may work on different scales. We show log-transformations to improve LIWC category scores found by reducing the effect of outliers and give recommendations for thirteen preprocessing steps to improve the inference on SE data. We found 600 to 1200 words per person to provide sufficient accuracy while remaining resource-aware and recommend a minimum of a hundred

words for all three methods. Furthermore, we do not find enough evidence for discrimination by all three methods for people proficient in English compared to those who found themselves non-proficient in English.

We find existing psycholinguistic models to be most useful for software engineering when used on a group or team level. When used on an individual level, one should take into account possible inaccuracies and consider the potentially harmful impact the misuse or misinterpretation of scores may have on an individual.

Thesis Committee:

Chair:	Prof. Dr. A. Zaidman, Faculty EEMCS, TU Delft
University supervisor:	Dr. A. Rastogi, Faculty EEMCS, TU Delft
Committee Members:	Dr. M. Aniche, Faculty EEMCS, TU Delft
	Dr. C. Hauff, Faculty EEMCS, TU Delft

Preface

Before you lies the Master Thesis project *"Inferring Personality from GitHub Communication Data: Promises & Perils"*, the basis of which is a comparison of three different psycholinguistic models on the automatic inference of personality on GitHub comments and a personality survey amongst software developers. The study is written to fulfill the graduation requirements for the Master Computer Science at Delft University of Technology (TU Delft), The Netherlands. From the period of November 2019 till June 2020, I engaged in the writing of this thesis.

During the period of research, many obstacles had to be overcome. Fortunately, both A. Rastogi and A. Zaidman stood ready to support me in this research with delightful insights and high standard feedback. The amount of time they took to guide me on my research, even with their busy schedules, was far beyond my expectation. For that, I owe you both my deepest gratitude.

Furthermore, I would like to thank all of the questionnaire participants. It was inspiring to see so many people interested and willing to invest time and effort to support this study. Without all the responses, I would not have been able to conduct this analysis.

I want to thank M. Aniche and C. Hauff for their time and willingness to be part of the thesis committee. I would also like to thank X. Zhang, for his cooperation during the data gathering process, and the members of the Software Engineering Research Group (SERG) for providing me with the needed resources to conduct this research.

Every study is a great adventure, and no research ever proves to be an uncomplicated journey. However, with all the great people I am surrounded with—academics, friends, family, and colleagues—every journey becomes a great team effort.

Frenk C.J. van Mil
Schipluiden, the Netherlands
June 21, 2020

Contents

List of Figures

Chapter 1

Introduction

Personality plays a significant role in our lives; it does not only influence what we think, feel, and do, but also affects what we say about what we think, feel, and do [124]. Personality is the combination of characteristics that evolved from biological and environmental factors [31].

There are different ways to assess personality. A popular approach among earlier studies is self-assessment through questionnaires. However, such an approach often suffers from low response rates [109]. Alternatives exist based on data-driven approaches, namely psycholinguistic tests. The words we choose, give away a lot about our personality [124] and can, in addition to that, be translated to personality profiles. Psycholinguistic tests rely on the interrelation between linguistic factors and psychological factors [62] and convert blocks of text into personality profiles (e.g., [124, 47, 48]).

Research on personality is not limited to the field of psychology. In software engineering (SE), it might help in improving team composition through a combination of personalities within a team [36, 46], and it could help explain work preferences [68] and work satisfaction [1].

Earlier studies in the field of software engineering have focused on the assessment of personality from developers through questionnaires [110], but also by extracting personality from text written by developers [17, 18, 86, 13, 75]. However, to the best of our knowledge, there is a lack of studies comparing different psycholinguistic models on actual SE data to validate to what extent these tools apply to software engineering, as also pointed out by many of these studies [17, 18, 86, 75]. To compare different personality inference models in SE, we focus on communication between software engineers during the development process. To approach this, we rely on comments publicly placed on the popular collaboration platform GitHub. Based on the comments of developers, we infer personality profiles with the use of psycholinguistic test methods. To assess the performance of the methods used, we need to understand how the data models should be formulated for software engineering specifically and how the data should be processed to extract only the information desired for the inference of personality.

A multitude of models for psycholinguistic inference exists. In this study, we compare some of the more established models covered in multiple studies. The first model is the method proposed by Yarkoni [124], an academic model based on long blog posts. The

second, an academic proposed by Golbeck et al. [47], based on Twitter posts. The third model is Personality Insights by IBM[1], an often used industrial model based on any text given.

1.1 Research Questions

In this study, we describe the process for psycholinguistic analysis for multiple models and compare these models for their accuracy and reliability. We test the models on developer communication data fetched from GitHub against personality scores obtained through a questionnaire. Based on the performance of inference on software engineering communication data, we put the choice of models on a firmer basis by answering the research question:

Main research question

How useful are psycholinguistic tests to infer developer personality from SE data?

Based on existing literature, we need to implement models to infer personality from text and compare their performance. For the psycholinguistic tests to be useful, they need to capture the personality of an individual rightfully. To properly analyze the comments, we should sanitize the text from unwanted influences. An example of taint is code snippets inside comments, which could lead the misinterpretation of context. To approach the problem, we propose some data sanitization steps and try to answer the following sub-question:

Research sub-question I

(RQ.I). Does data sanitization influence the outcome of personality inference tests?

For this study, we establish a ground-truth from personality questionnaires among software developers. Based on these results, we allow for a comparison between the used psycholinguistic methods. To answer the main research question, we compare the methods for their performance to the ground-truth and try to answer the sub-question:

Research sub-question II

(RQ.II). How well do the psycholinguistic models perform when compared to the ground-truth?

Psycholinguistic tests rely on the words written. However, not only the words themselves but also the number of words may influence the outcome. Intuitively the more words you provide, the more personality you could capture. It is, however, not always possible to capture all words, both in the availability of data, but also to keep in mind the processing power available. To approach this problem statement, we investigate the question:

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

Research sub-question III

(RQ.III). How much does the number of words in messages influence the reliability of personality inference?

In this study, we focus our research on the English language. On GitHub, English is the main language used. The users of GitHub, however, are not all natives in English. People with English as their second language show different frequency-related aspects[2] in their writing than people with English as their first language [73]. Hypothetically, this could mean native English and non-native English writings may have different behavior in terms of accuracy to the true personality. As we do not want the models to discriminate to a group of people (i.e., perform worse for one group compared to the other), we investigate the question:

Research sub-question IV

(RQ.IV). Does the English proficiency of a person impact the quality of personality inference?

In Figure 1.1, the illustration shows an overview of the data processing leading to personality profiles. From left to right, we fetch the user comments from different users around the world, where the comments are combined and preprocessed. The output of the preprocessing steps goes into the three psycholinguistic methods, resulting in three personality score profiles. In Chapter 2, we explain what these profiles mean. In the figure, we illustrate the possible external influences of English proficiency and the number of words on the way we write at the bottom.

1.2 Thesis outline

The remainder of this study is organized as follows.

- First, we introduce a background to the study (Chapter 2). The chapter introduces a definition of personality, personality in software engineering, and a more elaborate explanation of the challenges for this study.
- After some introductory background information, we explain the methodology of the report. We explain how the psycholinguistic models work, how the SE data is selected, and how the validation data is selected. Finally, we introduce some definitions and assumptions for the statistical comparisons used in succeeding chapters (Chapter 3).
- To assess performance in terms of accuracy, we make use of a ground-truth obtained through a questionnaire. To make sure this ground-truth represents all personality types, we investigate the distributions of scores (Chapter 4).

[2]Differences in frequency-related aspects in text refer to the frequency of words occurring. Someone with English as their second language may use particular words more or less often than those with English as their first language.

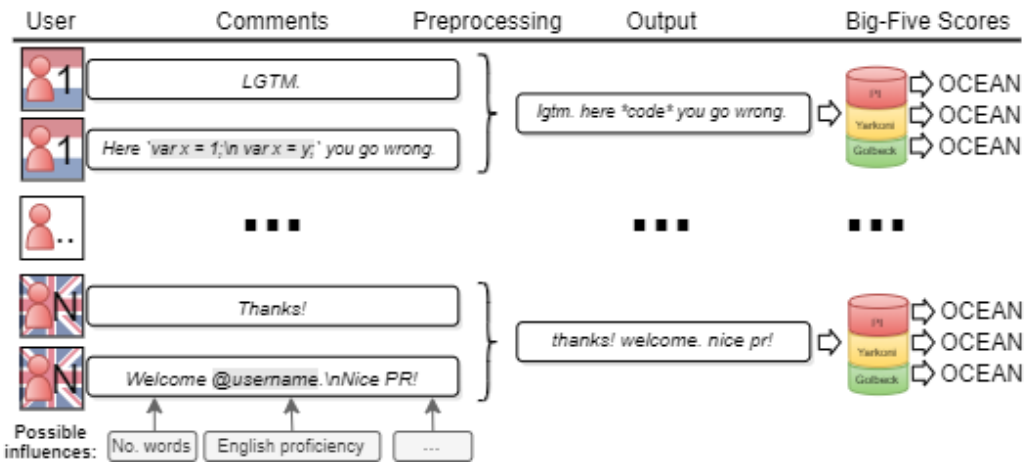


Figure 1.1: The general overview of data processing flow from user to Big-Five personality scores. From left to right: the user writes a comment. All comments are preprocessed and concatenated to a single text blob. The output text is converted to Big-Five personality scores using three different methods: Yarkoni, Golbeck, and PI. At the bottom of the image, we display possible external influences on the content of comments. How the preprocessing and personality score calculations work will be explained later in this study.

- For the psycholinguistic models to infer personality, each model requires some adaption of the text, i.e., *RQ.I*. As to make sure the preprocessing steps improve the process, we compare the theoretical and practical implications of the preprocessing steps (Chapter 5). In this chapter, we propose thirteen preprocessing steps for which all, but one, show to be sufficiently improving the outcome of the analyses, improve privacy, or improve the speed performance of processing.
- To assess the performance of psycholinguistic models, we compare the inferred personality scores against the ground-truth, i.e., *RQ.II* (Chapter 6). In this chapter, we find Yarkoni and PI to have reasonably similar performance and show all three methods to be reasonably similar but on different scales. Furthermore, we propose transformations on LIWC category scores to improve the personality prediction of the LIWC-based methods.
- The size of the text and the number of words may influence the reliability of the model, i.e., *RQ.III*. We check the effect of text size by testing different volumes of text on each of the methods (Chapter 7), and we construct recommendations on the number of words to use for each method.
- Another aspect excepted to influence the outcome of the personality analysis is the English proficiency of the individual, i.e., *RQ.IV*. As we are concerned for all software engineers and not only those that are proficient in English (or the other way around), we investigate if there is a differences for these groups (Chapter 8). In this chapter, we do find some differences in scores between proficient and non-proficient people and investigate possible reasons for this effect.

- Next, we discuss the found results of this study and the limitations and threats to the validity of the results. Building upon this, we extend the discussion with the ethics of automatic personality inference, covering the possible misinterpretation of results, misconduct, and possible negative impact misuse may have on an individual (Chapter 9).
- We then cover some related work to this study, where we look into similar studies in this field (Chapter 10).
- Finally, we outline the contributions of this research, followed by the conclusions and recommendations for future work (Chapter 11).

Source code and anonymized data used to obtain the results of this paper have been made available at `https://doi.org/10.5281/zenodo.3865341` and `http://doi.org/10.4121/uuid:6b648676-26f4-4eb1-89dc-050810909b3b`, respectively.

Chapter 2

Background

To understand personality inference from software engineering data, we must first understand how to represent personality. In the first section of this chapter, we explain a model to express personality and outline some alternative models. In the next section, we explain how personality can be measured. Finally, we explain how existing literature uses the methods for software engineering.

2.1 Personality

Personality is the pattern visible in the thinking, behaving, and feeling of an individual that tends to be consistent over time and across relevant situations [111]. The most widely used model to express personality is the Big Five Personality model (BFP). Based on the BFP, the choice of expression can contribute to a personality trait score. The Big Five personality framework consists of five personality traits, which gained recognition among trait psychologists regarding its validity and reliability [55, 81]. These traits are Openness to Experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (in short: OCEAN). By using the values obtained for each of these personality traits, a comparison between people is possible.
The model constitutes the following explanations for the traits:

- *Openness to Experience* — Characterized by intellectual curiosity, imagination, and open-mindedness. Close-minded people often have a narrow range of creativity and intellectual interest. Sometimes referred to as Intellect/Autonomy or Openness [111].
- *Conscientiousness* — Characterized by the preference of order, structure, persistence to a goal, and responsibility. Low conscientiousness is linked with the comfort with flexibility and spontaneity, but also with sloppiness and lack of reliability [111].
- *Extraversion/Extroversion* — 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 [111].

- *Agreeableness* — 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 [111].
- *Neuroticism* — Characterized by the tendency to experience negative emotions. Lower neuroticism is linked to emotional stability. Lower neurotic people tend to stay calm and resilient. Sometimes referred to as Emotional Stability [111].

A reason to research personality is the common belief that team diversity has a positive effect on team performance [91, 92]. However, establishing a team structure with multiple different personalities in a team is hard. The nature of this problem comes from the difficulty of assigning personality labels to persons. The idea of inferring personality from communication comes from the hypothesis that the personality structure of an individual is, in part, a function of the linguistic characteristics the individual shares with a group of people [53]. Similarities in linguistic characteristics form the basis for personality profiles relative to others with similar characteristics.

2.1.1 Alternative models

Alternatives to the BFP model exist. One widely used example in the industry is the MBTI model [16]. Around 2 million people in about 170 different countries take the MBTI test a year [98]. The model classifies people into 16 personality categories. Although the industry often uses the model, there is little scientific evidence to support its use [117]. Druckman and Bjork recommend not to use MBTI as an instrument for personality in career counseling until there is scientific evidence to support its validity [33].

Another example is HEXACO, which is essentially an extension of the BFP model with the addition of the Honesty-Humility trait. The predictive ability of HEXACO compared to BFP is in several instances similar [9]. However, to use HEXACO for psycholinguistic analyses, more research is needed for the creation of such models.

2.2 How is it measured?

As we have now defined personality in terms of BFP, we should now establish a way to assess personality. In this section, we cover two of the more popular methods: (self-assessment) questionnaires and psycholinguistic tests. We then explain how such practices are generally used to obtain personality scores for individuals. Other methods do exist, for example, personality inference from task performance [58]. However, such an approach requires access to the task performance information, which is not always available.

2.2.1 Questionnaire

A questionnaire is a manual assessment method in which you ask subjects to answer personality items that inquire about self-perceptions and retrospection on self-related events [5]. To elicit the desired information from questionnaires, the authors of questionnaires need to

be careful with the sentence types used. For example, items should be short, and multiple clauses and negations should be avoided [5].

Different questionnaire instruments exist; each differs in various aspects. For example, NEO-PI-R highlights particulars of a language or culture, HiPIC focuses on assessing specific age groups, NPQ emphasizes alternative media of communication, SIFFM emphasizes specific uses in the assessment process, or dyadic interactional perspective provides a different theoretical perspective [93]. The most frequently used instrumentation method for personality questionnaires is NEO-PI-R. However, NEO-PI-R is time-consuming; filling in the NEO-PI-R takes around 30-40 minutes. A shorter alternative to NEO-PI-R is BFI (Big Five Inventory) [64, 65, 63]. Filling in the BFI takes only 5-10 minutes. Although it has fewer questions, BFI still provides satisfyingly reliable and valid data [44, 3, 8]. More studies on personality in software engineering chose BFI as a ground-truth, five (or six if we count BFI-like) out fifteen studies mentioned in the literature study of Calefato et al. [18] used BFI as their ground-truth method. Other studies used customized questionnaires, NEO-FFI, or inferred personalities from earlier studies [18].

2.2.2 Psycholinguistic tests

The second popular approach to assessing personality is psycholinguistic tests. Research on this field has mainly focused on determining the underlying psychological representation and the processes that are required for both simple and complex words [45]. Different studies have already tried to assign personality labels to people through automatic composition. Automatic systems mostly work with psycholinguistic text analysis. The texts analyzed in such a case are the communication between the different team members (e.g., email traffic, direct message communication, or notes).

An example of such a psycholinguistic test is the academical model by Yarkoni [124]. Yarkoni relies on correlations between word categories and personality traits found from blog posts. Similarly, the studies of Golbeck et al. [47, 48] found predictive information of personality through the analysis of Facebook/Twitter posts. For the categorization of words, both studies used LIWC (Linguistic Inquiry and Word Count)[1] library. LIWC is a library that categorizes words in psychologically meaningful categories [116]. This tool works by counting the percentage of words reflecting different emotions, thinking styles, social concerns, and part of speech. A user-defined dictionary reflects each of these categories. The analyzed text is compared to this dictionary to assign percentages to each category. The findings of Pennebaker [88] support the idea of using a set of words to describe personality. According to Pennebaker, only a small group of words can already describe your personality. The smallest and stealthiest words in our vocabulary define something about our personality. An example given in his study is about narrative writers, where words like 'with' and 'together' often indicate the author having better social skills, having more friends, and the authors often rate themselves as more outgoing.

Earlier studies on personality among software developers used Personality Insights (PI)[2] [17, 86, 84, 86] to analyze the digital communication of developers. PI uses machine learn-

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

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

ing with an open-vocabulary method [106], meaning the system evolves. Such an open-vocabulary approach has earlier been shown to work both on small samples of text, like Tweets [7], as well as larger samples of text, like Facebook posts [106].

2.3 How are psycholinguistic tests used in software engineering and other fields?

Not only the method used for the inference of personality is essential, but also the source influences the outcome. Different types of data sources have already been studied for the deduction of Big Five personality scores for psychology in general, such as essays [77], emails [108], and social media [47, 48]. However, it is yet unknown in the field of psychology which data source would be most efficient in deriving Big Five personality scores [49], let alone in the field of software engineering. Often studies in personality inference show different results, possibly caused by environmental influences on the source of data. Some examples of such environmental controls are a restriction on the number of characters or words, the audience on the platform, and the indentations of the author (e.g., conducting code reviews). Even from the type of work, one could already infer some indications of personality. Software developers with independent tasks that require a certain degree of creativity tend to be introverts [19, 37], while developers who perform jobs requiring collaboration and leadership tend to be extroverts [102]. However, the inference of personality based on the intent of the work lies outside the scope of this study.

For SE, it is yet unknown how well existing methods perform on the inference of personality based on developer communication. Different methods have already been applied. Literature reviews have been conducted [12, 110], as well as inference from developer communications [17, 18, 86, 75]. In software engineering, studies used different sources, such as emails [17, 86], comments on code collaboration platforms [75], and posts on StackOverflow [13].

Chapter 3

Methodology

In the previous chapters, we identified how personality is defined and measured in the literature. In this chapter, we describe the methodology used to approach the research questions of this study. We first define which methods we will use for personality inference and explain their implementation. We then describe the selection process for the software engineering data on which we apply the techniques. Based on this data, the proposed methods could, in theory, give personality profiles of the developers analyzed. To verify and compare the results against a ground-truth, we conduct a self-assessment questionnaire to obtain personality scores. Finally, we define some statistical notions and assumptions that are used throughout this study to prevent repetition for all ensuing chapters.

3.1 Inference methods

To calculate the personality scores for each developer, we implemented three different models: the model by Yarkoni [124], the model by Golbeck et al. [47], and IBM Cloud Personality Insights[1]. The first two methods, we refer to as the methods `Yarkoni` and `Golbeck`, are two academical methods based on correlations between word categories (obtained with LIWC) and Big-Five personality types. The third method is a commercial application popular in the industry, making use of an unspecified machine learning model.

Yarkoni has been around for over ten years and has been used in a multitude of studies. The correlations found by Yarkoni are based on blog posts, for which most blog posts have relatively many words. In contrast, Golbeck is based on Twitter posts and, therefore, relies on much smaller volumes. Furthermore, the concise interactions on GitHub might come closer to Twitter than the usually long blog posts. The third method, Personality Insights, is a well-established method in the industry. Personality Insights is an open-vocabulary approach, trained on a much larger dataset than the academic models. The open-vocabulary approach, a data-driven approach, allows for new words, phrases, and topics to arise and be categorized [106]. The model of PI is not trained on a single text source, like Yarkoni and Golbeck. Because of the differences in the three models, different results may be expected from their analyses.

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

3.1.1 LIWC

Both Yarkoni and Golbeck use LIWC, a (now) commercial text-analysis program that counts words in word categories. LIWC bases these word categories on psychological association [89, 90]. Yarkoni and Golbeck, however, do not use the same version of LIWC; Yarkoni uses the LIWC2001 library [89] while Golbeck uses the LIWC2007 library [90]. For both methods, the conversion method from LIWC category scores to personality scores follows the same principle. To calculate the personality scores, we take the dot product with a matrix of all category scores found by LIWC and a matrix with all significant correlations found by one of the studies (see example Figure 3.1). This process is repeated for all five BFP traits, resulting in five raw personality scores for a person.

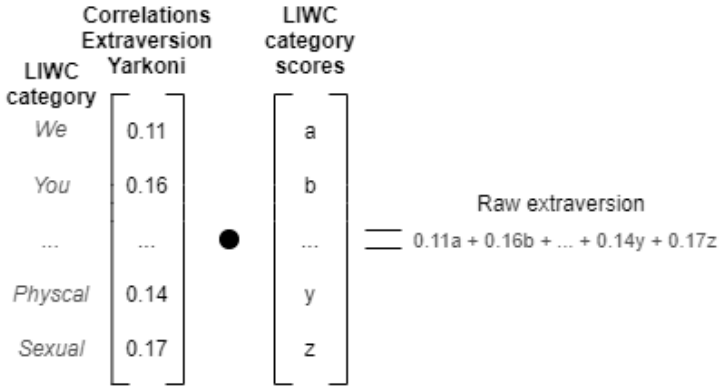


Figure 3.1: Example calculation of extraversion for Yarkoni. The raw extraversion score is the dot product of the $m \times 1$ matrix with fixed correlations found by Yarkoni and the $m \times 1$ matrix of LIWC category scores. The raw scores are later normalized (see Section 3.1.3)

3.1.2 Personality Insights

For the Personality Insights service, the application makes use of the IBM Watson Developer Cloud Python SDK[2]. The API connects with the IBM Cloud service using the version parameter `2017-10-13` (the current version of Personality Insights version as of this writing). Because the PI service is not backward compatible[3], a future version may break the current implementation of the application.

The application created for this study sends all comments of each user to the PI service and obtains personality profiles in return. We save both the percentile scores and the raw score of each personality trait. PI normalizes the percentile scores based on a sample population. A score of 0.99 for extraversion would mean that you score higher on extraversion than roughly 99% of the people. The raw score is the non-normalized score obtained from the service.

[2] `https://github.com/watson-developer-cloud/python-sdk`

[3] See `https://github.com/watson-developer-cloud/api-guidelines/#versioning` for versioning information

3.1.3 Normalization

Both the scores obtained through the LIWC-based methods and PI generate raw personality scores. For each method and trait, we normalize all raw scores using Mix-Max normalization between zero and one. Similarly to the PI percentile score, these normalized scores represent a relative score for the population taken in this study. Someone with a score of 1.0 in extraversion is (one of) the most extraverted person(s) in our sample set. A score of 0.5 indicates an average rating in the specific personality trait relative to our sample set.

3.2 Data selection

For the methods to work, we need to give input texts to analyze. For this study, we select communication data from GitHub[4] developers. To obtain data from the developers, we downloaded GitHub project data publicly available through GHTorrent [50]. GHTorrent is a dataset containing enriched development history from GitHub and has been collecting this data since 2012. We selected data according to the following selection criteria:

1. *Project selection criteria.* For this study, we selected the projects available through GHTorrent. For these projects, we filtered out all projects with less than 33 pull requests, giving us the top 3% of projects. The projects with less than 33 pull requests were assumed not to follow a strict development process with less communicative intentions amongst the developers. Projects deleted at the time of the creation of the dataset were left out, as we could not trace the comments back to GitHub anymore. Of all projects, we selected six programming languages (Java, JavaScript, Python, Ruby, Go, and Scala) to allow for potential context analysis.

2. *User selection criteria.* We merged all comments into a single string for each user. As the models need a sufficient number of words to work, single comments are unlikely to contain enough information to obtain a full personality profile. We omitted users with less than 100 words in total, as Personality Insights requires such a minimum [25]. As the number of words used may be part of the personality, we choose a minimum as low as possible. After the selection of the pull requests, we removed bots. On development platforms as GitHub, it is not unusual to see bots. Most of these bots generate very long traces of comments, where every sentence is nearly equal to the others. Bots were identified through manual analysis on the top 1% of people in terms of word count by identifying repetitive communication behavior or indicative statements. An example of such an indicative statement for a bot looks like:

   ```
    Hello contributor, thanks for submitting a PR for this project!
   I am the bot who triggers "standard-ci" builds for this project (...)
   ```

   We ended up removing 90 bots with their respective comments. For this study, we only select the people that share their email and location (e.g., country, city, or con-

[4] https://github.com

tinent) to allow contact for the ground-truth and the analysis on English proficiency. Finally, we end up with 8,436 projects and 4,081,957 comments from 28,337 different users.

3.3 Validation data

With the above-explained steps, we can now infer personality scores. By preprocessing the data and applying the models, personality scores can be inferred. Which preprocessing steps are required is investigated in Chapter 5. However, it is yet unknown how these models perform in terms of accuracy.

To validate the personality scores, we use questionnaires on a subset of people. The questionnaire can be found in Appendix A. The participants for the questionnaires are all identified from the complete dataset using the following procedure. Users are first filtered according to the minimum word rule of 100 words and linked with their usernames. To allow us to contact the users, we obtain the email addresses through the GitHub API.

Alongside the email addresses, we fetch the location of the user. Users can state their location in their GitHub profile, for those who did, we transform the location to a longitude and latitude value by using the Bing Map API Service[5]. The Bing Map service can translate most types of location names. Users are free to choose how to describe their location; the location could be their country, state, city, or street addresses. We omit those without a location or a location too ambiguous (e.g., 'The internet' or 'Earth' as location).

After the identification of locations, we plot the locations on a world map and create a K-Means cluster. For the K-Means clustering, we use six initial cluster centroid, one cluster centroid for each continent (excluding Antarctica). Figure 3.2 shows the resulting clustering. With the locations, we can draw a fair distribution of people from all around the world for the questionnaire. After the clustering, we take samples from each continent (cluster). Random undersampling on the majority class helps each continent to be equally represented. Through this sampling, we select 2,050 participants.

We send all invitations for the questionnaire through email. For each potential participant, we translate their location to a timezone. With the timezone, we could send emails on the time and day most likely to receive responses [41]. The campaign eventually resulted in a 13% response rate.

3.4 Statistics

In this section, we outline some fundamental definitions and notions regarding the statistical analyses with a description corresponding to each item. All succeeding chapters assume and build upon the mentioned definitions and notions below.

Scale

All personality scores reported in this study range between a scale of 0 to 1, this

[5] `https://www.bingmapsportal.com/Application`

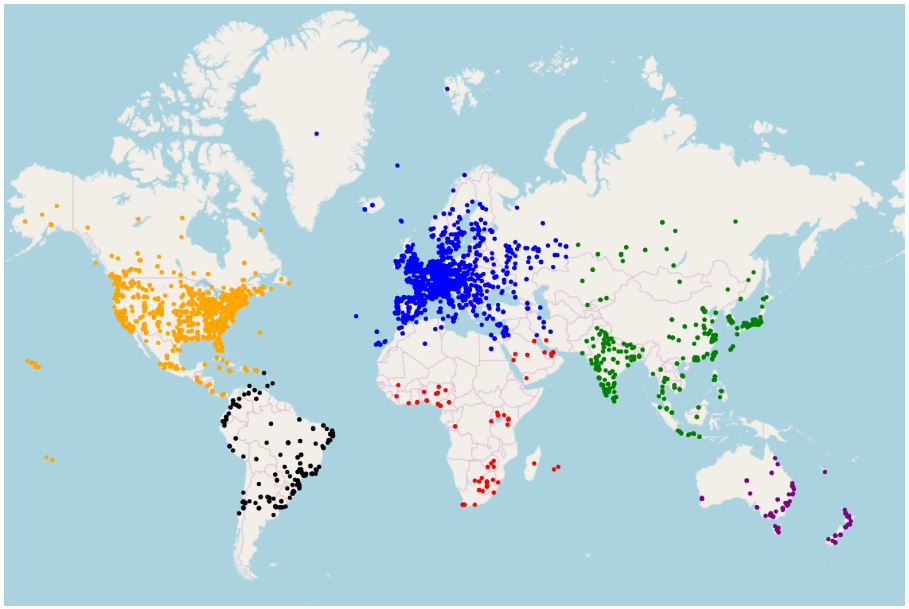


Figure 3.2: Initial clustering of all users with locations and email addresses on their GitHub profile. The coloring indicates each cluster, where the initial cluster points are set on the middle of each continent. The distance in longitude and latitude values are used as metric for the K-Means clustering algorithm. Please mind many points overlap.

includes the personality scores found with the questionnaire as well as the three psycholinguistic test methods (i.e., Yarkoni, Golbeck, and PI). The personality scores found are relative to the sample space taken for this study. Someone with an extraversion score of 0.5 has an average extraversion score relative to the 2,050 people in our dataset.

Significance

In this report, we deem p-values below 0.05 as significant and any values above or equal to 0.05 as insignificant.

Normality

In this report, we check normality in data with the Shapiro-Wilk test [107] and QQ-plots. In this definition, the notion of 'data' refers to a collection of personality scores. Data for which the Shapiro-Wilk test shows a significant non-normal distribution, non-normality is assumed. If and only if, both the Shapiro-Wilk test shows no significant non-normal distribution and the QQ-plots show a (near) normal distribution, we assume normality for the data. We check for normality for each method and personality trait individually.

Difference in means

We make use of multiple methods to compare differences in means. For normally distributed data, we apply the paired or unpaired Student t-test [115], depending on

the comparison required. For data deemed non-normally distributed, we use the non-parametric Wilcoxon signed-rank test or Wilcoxon sum rank test [123] for paired and unpaired comparisons, respectively.

Effect size

All effect sizes adhere to Cohen's interpretation for effect size [26] (see Table 3.1). Significance combined with the effect size form a basis for conclusions. Different effect size metrics are used in different scenarios. For each table we indicate which effect size metric is used. The used effect size metrics in this study are:

r — Effect size r with $r = \frac{Z}{\sqrt{N}}$[99], where Z describes the test statistic of the method for which the effect size is calculated and N the number of observations.

d, Cohen's d — Effect size d or Cohen's d [26]. Cohen's d assumes normality [74], if normality cannot be assumed we refrain from using Cohen's d and often settle with effect size r.

Cramér's V — Effect size in terms of Cramér's V [35].

Accuracy

We express accuracy toward the ground-truth in terms of MAE and RMSE. We chose not to go with correlations, as correlation methods are invariant to scale. The methods used may work on a different scale and are expected to show low correlations.

For the t-test, we use Cohen's d to obtain effect sizes. For the Wilcoxon signed-rank test, we cannot use Cohen's d, as it violates the assumption of normality [74]. For the Wilcoxon signed-rank test, the effect size r is used following formula $r = Z/\sqrt{N}$, with Z the statistical value found with the Wilcoxon signed-rank test and N the number of score pairs compared.

Table 3.1: The interpretations of effect size according to Cohen [26, 27]. The first column indicates the given naming to the effect size according to Cohen, the second column the values required for the given interpretation for Cohen's d effect size. The third column indicates the values for the r effect size. The last column indicates the interpretations for Cramér's V effect size with two degrees of freedom.

Interpretation	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

Chapter 4

Ground-truth on Personality

In this chapter, we investigate the personality scores obtained from the questionnaire, which works as ground-truth for this study. We asked all participants of the survey to answer questions according to the BFI scoring scheme [64, 65, 63] to obtain the Big Five personality scores. As the data functions as a ground-truth for personality, it is important to discover which personality types the dataset contains and, more importantly, if the data represents all personality types. Without a representation for all personality types, a bias toward an over-represented group may invalidate the results for an under-represented (or even absent) personality type.

Table 4.1: The distribution of personality traits for survey respondents. The first column indicates the Big-Five personality trait inferred from the BFI questionnaire. Min indicates the minimum personality scores, followed by the first quartile, median, mean, third quartile, and maximum personality scores found for each trait.

Trait	Min	Q1	Median	Mean	Q3	Max
Openness	0.4	0.61	0.7	0.71	0.8	1
Conscientiousness	0.17	0.53	0.64	0.63	0.72	1
Extraversion	0.09	0.41	0.53	0.54	0.69	0.97
Agreeableness	0.25	0.58	0.67	0.68	0.78	1
Neuroticism	0	0.28	0.41	0.42	0.53	1

4.1 Overview scores

To get a first impression of the scores found with the BFI questionnaire, we compare the overview given in Table 4.1 with the histograms in Figure 4.1. For most traits, we can observe a near-normal distribution. In this context, a normal distribution suggests that most people have a close to average score, while only a few people have an extreme score. 'Average' in this context means a trait score close to 0.5, while an 'extreme' rating indicates a score close to either one or zero.

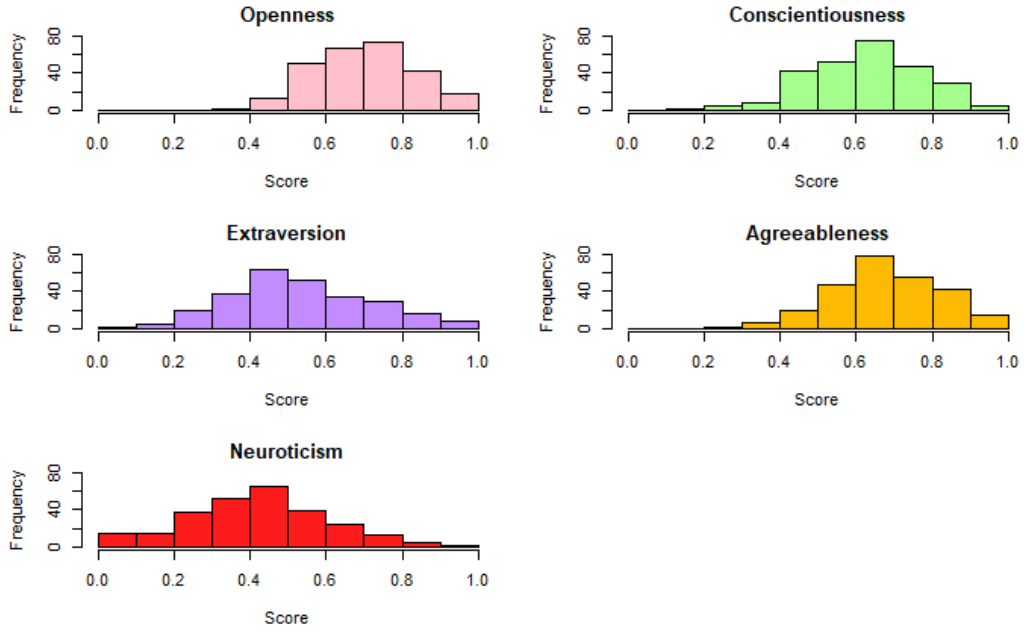


Figure 4.1: Histograms of all BFI personality trait scores found through the BFI questionnaire. On the y-axis the frequency of people with a score contained in the each bin and the x-axis showing the personality score for each Big-Five trait.

We can observe extraversion to have a near-normal distribution between 0.09 and 0.97. Neuroticism shows a representation of all values between 0 and 1, with a majority of scores below 0.5 ($IQR = 0.28 - 0.53$). When we look at the other three traits, we can observe mostly higher scores (above 0.5). Conscientiousness has values between 0.17 and 1.0, with an IQR of $0.53 - 0.72$ and an average of 0.64. Openness show to represent values between 0.4 and 1.0 but not for values below 0.4. Similarly, agreeableness has an average score of 0.68 and no values below 0.25, which suggests an over-representation of high ratings.

To summarize, extraversion shows represent all types to the least. Similarly, neuroticism shows all personality types with a slight over-representation of below-average scores. Openness, conscientiousness, and agreeableness show an over-representation of above-average scores and an under-representation (or absence) of below-average scores. Potentially this could mean the ground-truth introduces a bias toward the over-represented group of more open, conscientious, and agreeable people.

4.2 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. [17] 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 could potentially mean we show a similar population of software engineers.

Other explanations for the found scores, could fit with the finding of Licorish and MacDonell [75] that especially top members[1] of a project tend to score significantly higher on openness than other members. Although, the participants chosen for our dataset are not necessarily top members. In the study of Calefato et al. [17], they found developers with a higher value of openness and agreeableness to be more likely to be a contributor. As we only made use of open-source projects, this could mean that the people we covered in our dataset are higher in openness and agreeableness because they are more likely to contribute. Furthermore, we potentially introduce a bias in the responses, as people who volunteer to fill in surveys and consent to the processing of their data are generally more open and agreeable [40].

Another explanation for the higher scoring traits could be the finding of Kosti et al. [68]. They found that there exist two major groups among developers, formally called 'intense' and 'moderate.' These groups were based on the mean numerical values of the BFP traits. Those with higher scores on all five traits are in the 'intense' group, the others in the 'moderate' group. People in the 'intense' group prefer working in teams [68]. Hypothetically, this could mean that the group now analyzed could be over-represented by this 'intense' group due to the cooperative nature of GitHub or, more specifically, open-source projects. For extraversion and neuroticism, however, we do not necessarily see an over-representation of an 'intense' group (see Figure 4.1). If this theory is applicable for our sample, the 'intense' group only shows a slight over-representation of higher openness, conscientiousness, and agreeableness values.

Takeaway

The personality scores we use as the ground-truth for this study, show a near-normal distribution; the distribution mostly represents average personality, while the lowest and highest personalities are under-represented or sometimes not represented at all. Extraversion and neuroticism show to at least represent all personality types. Neuroticism, however, has a slight over-representation of below-average scores. Openness, conscientiousness, and agreeableness show an over-representation of above-average scores and an under-representation (or almost absence) of below-average scores. The ground-truth, therefore, may introduce a potential bias for over-represented groups of personality. The scores found for the ground-truth, do seem to fit with related work on software engineers.

[1] The developers making the most contributions.

Chapter 5

Does data sanitization influence the outcome of personality inference tests? (RQ.I)

To infer personality effectively, we should first sanitize the comments. The steps are added to either ease the processing, to make sure the context is interpreted rightfully, for model conformance, or to remove third-party influence. This chapter focuses on the effect of each of the preprocessing steps, ultimately, to give recommendations for each step. Each step should improve the outcome, not influence the outcome, or at least not worsen the outcome significantly. In the latter case, a property other than accuracy should improve (e.g., privacy).
Before we cover all thirteen preprocessing steps, we first investigate the possible effect of the removal of words, followed by an outline of the remainder of the chapter.

5.1 Effect of LIWC

For some of the preprocessing steps mentioned in this section, the steps taken remove content not contributing to personality scores. However, they still have some influence on the outcome of the LIWC methods (i.e., Yarkoni and Golbeck). For these methods, the scores for the categories are based on the percentage of the words compared to the total of words. Meaning that if some category contains two words in a message of a total of a hundred words, the category gains a score of 0.02 (i.e., 2%). Would we remove words not contributing to any category, we remain with the same numerator for each category but a lower denominator. With more words, this effect should be minimized. To demonstrate the idea we have the following unprocessed and processed comment:

Comment unprocessed

You could send me an email to email@address.com **Word count:** 10
1 2 3 4 5 6 7 8 9 10

Comment processed

You could send me an email to **Word count:** 7
1 2 3 4 5 6 7

Example 1 shows the unprocessed comment and example 2 the processed comment. With the email address, the total word count is ten and seven without. LIWC calculates the scores based on a dictionary per category. For example, the category score for the `You` category in the first example is 10.0(%), as $\frac{1}{10}th$ of the sentence belongs to the category. In the second example, the word `You` receives a score of 14.29(%), as it takes up $\frac{1}{7}th$ part of the sentence. From this example, it becomes clear that the number of words contained in the complete comment influences the whole of a category. Although it is clear there is an effect on the outcome of the analysis, it also becomes clear LIWC counts certain elements differently than you might expect (e.g., email addresses). More on this is made clear for each preprocessing step where applicable.

5.2 Preprocessing steps

In this chapter, we cover thirteen proposed preprocessing steps for data sanitization. The steps are 'sentences to lowercase', 'remove code blocks', 'remove quotes', 'conform white space', 'remove numbers', 'remove hashtags', 'remove IP-addresses', 'remove URLs', 'remove emails', 'remove double white space', 'remove space before punctuation', and 'remove images'. Most of these steps come from earlier studies on psycholinguistic models. For each preprocessing step, we investigate their need with the following plan.

1. First, we discuss the implementation of the step and reason its purpose along with an example of the implementation.
2. Second, we investigate the theoretical/expected influence. Often the theoretical influence for PI is unknown, as PI is a black-box method. Here we rely on their documentation. If deemed necessary (i.e., a theoretical influence is found or expected), we move on to step 3.
3. Third, we look into the actual implication on the scores by comparing the parsed results and unparsed results for each preprocessing step. We first check for point 3.a.

    a) We first investigate if the processing step changes the scores. If the differences are deemed significant, we move on to 3.b.

    b) Next, we compare the personality scores found with the preprocessing step enabled and disabled to the ground-truth.

4. Finally, we conclude on the found results of previous steps and outline our recommendation.

The theoretical and practical implications might differ per model used. For all thirteen preprocessing steps, we divide each step into four sub-sections: reasoning and implementation

of the preprocessing step, the theoretical influence, the practical influence, and a concluding remark. For each step, we compare the scores of all 28,337 people contained in the dataset with the preprocessing step enabled and with the step disabled. To compare accuracy, we only consider the 267 people in the ground-truth.
On page 51, we show an overview of all concluding remarks on preprocessing steps on page.

5.2.1 Sentences to lowercase

The preprocessing step essentially converts each word to its root. For example `'Hello'`, `'HELLO'`, and `'hello'` or even `'hElLO'` are treated as the same word. Other studies also applied this method [7, 21] or even extended this by stemming words [39, 101]. We decided not to go with stemming (although Porter Stemming is available in our application), as stemming might also remove essential information from a word.

Example unprocessed:		Example processed:
`Please read THIS carefully`	→	`please read this carefully`

If none of the models use the capitalization, we can optimize the subsequent steps to only consider 26 alphabetical lowercase characters instead of 52 characters. The optimization saves a significant amount of iteration steps for regexes.

Theoretical implication

The parsing of sentences to lowercase should not influence the outcome of most analyses. To the best of our knowledge, none of the methods consider casing. LIWC preprocesses the files by first removing all capitalization [89, 90]. As LIWC always removes capitalization, for Yarkoni and Golbeck, this step will not affect the outcome. For Personality Insights, we should check the practical implications to make sure the preprocessing step does not influence the outcome.

Practical implication

As stated, the prepossessing step cannot possibly influence the scores for the methods using LIWC. For PI, there are some differences. Although, at first sight, the difference does not appear to be large. Only 0.1% of people show a difference (see Table 5.1).
As we want to know if the processing step influences the outcome of the inference methods, we check for the differences in means. In Table 5.2, we observe no significant difference. Similarly, neuroticism shows no significant differences with the use of the t-test. Conscientiousness, extraversion, and agreeableness show a significant difference, but this difference is only negligible.
As we observe people with differences, we check for a change in RMSE and MAE when compared to the ground-truth. However, there are only two people with differences. For both these people, the change does not exceed 0.01 (RMSE and MAE). It is, however,

Table 5.1: Absolute differences in scores for PI with and without lower case parsing compared to the ground-truth. The first column shows the personality trait. Then from left to right: the maximum absolute difference observed (indicating the worst case value), the mean absolute difference (indicating an overall effect), and the percentage of people showing a difference with and without the preprocessing step.

Trait	Max abs. diff.	Mean abs. diff.	%People
PI Openness	0.02	0	0.1
PI Conscientiousness	0.07	0	0.1
PI Extraversion	0.09	0	0.1
PI Agreeableness	0.01	0	0.1
PI Neuroticism	0.05	0	0.1

Table 5.2: Paired Wilcoxon signed-rank test between the scores with the lower case preprocessing step enabled and disabled. The columns from left to right: the method and trait tested, the V-statistic of the test, the p-value, a boolean indicating significance for p-values below 0.05, a measure of effect size r, and an interpretation of the effect size r.

Trait	V	p.value	p<0.05	r	effect
PI Openness	-1.63	0.14	FALSE	-0.01	negligible

Table 5.3: Paired t-test and Cohen's d effect sizes between the scores with the lower case preprocessing step enabled and disabled. The columns from left to right: the method and trait tested, the V-statistic of the test, the p-value, a boolean indicating significance for p-values below 0.05, the effect size in terms of Cohen's d, and an interpretation of the effect size Cohen's d.

Trait	V	p.value	p<0.05	d	magnitude
PI Conscientiousness	3.34	0	TRUE	0	negligible
PI Extraversion	2.46	0.01	TRUE	0	negligible
PI Agreeableness	3.36	0	TRUE	0	negligible
PI Neuroticism	1.74	0.08	FALSE	0	negligible

hard to conclude from this ground-truth, as only two data points are not sufficient to safely generalize to the whole public.

Conclusion

We observed close to zero difference for all 28,337 people. For LIWC, the method does not influence the outcome. For PI, only 0.1% of people show some variation. However, for PI, we could not lay firm conclusions on the usage of the step. As the step does help to make the text used more consistent and also allows for improvement in performance for other preprocessing steps, we conclude this step to be sufficiently good.

5.2.2 Remove code blocks

Opposed to other fields, personality inference on the communication of developers may involve code blocks within the text. On GitHub, it is common to make use of code in comments. Often these code snippets are quotes from code or suggestions for alterations in code. For the first case, where code is quoted from someone else, it is evident that this piece of code does not contribute to the personality of the person quoting. In the second case, with the suggestion of a new piece of code, none of the methods captures any personality hidden inside code. Any personality hidden inside code lies outside the scope of this study. The primary purpose of this preprocessing step is to make sure the context fits with the interpretations of the models. Leaving in the code could drastically influence the outcome of the analyses used.

Below we show two examples of communication between two developers. In both instances, the authors integrate code in their comments.

```
Example 1 (unprocessed):
Why do you do `ArrayList<String> = new ArrayList<String>();`
instead of `List<String> = new ArrayList<>();`?

Example 2 (unprocessed):
You should definitely have a look at your usage of `String`,
as in most cases you could have simply used `char`.
```

In the above examples, the code is part of the sentence. In the first, we use an actual block of code to describe a problem in the code. The second example shows a slightly different scenario. The developer explains its concern with the use of datatypes. In the second example, the ticks could have been omitted, making *String* and *char* ordinary words.

In both examples, it becomes clear that the code influences the structure of the sentence. If code would remain, a keyword such as `new` (e.g., in Java) could influence the BFP obtained. To illustrate the necessity of overcoming this challenge, we draw an example of the effect on the method used by Yarkoni [124]. The keyword `new` is part of the `Time` category in the LIWC 2001 dictionary. As `Time` has a significant negative correlation with Openness ($p < .001$) and Agreeableness ($p < .01$) in Yarkoni [124], leaving the keyword in would influence the outcome. As to keep the text readable for the human reader, code blocks are replaced with a placeholder to keep the sentence structure but omit the contents of the code. For this study, the replacement is chosen to be `*code*` (which could be read as **code**). The placeholder makes it possible for the human reader to understand sentence structure. We remove the placeholders before any psycholinguistic analysis. Below we show the transformed examples.

Code blocks indicated with accent graves/ticks (‘) are considered a code block. In particular, we consider text surrounded with single ticks, double ticks, triple ticks, unclosed triple ticks, single ticks on two lines, and double ticks on multiple lines without two subsequent enters.

```
Example 1 (processed):
Why do you do code instead of code?

Example 2 (processed):
You should definitely have a look at your usage of code,
as in most cases you could have simply used code.
```

Theoretical implication

As also stated earlier, word categories of LIWC may contain the keywords reserved for programming languages. We showed earlier the keyword `new` to influence the Yarkoni method. Table 5.4 shows some more common keywords used in different popular programming languages that correlate with personality in Yarkoni and Golbeck. For each of the keywords, we show the correlations of Yarkoni [124] and Golbeck et al. [47].

Table 5.4: Examples of programming language reserved keywords and their correlation effects on Yarkoni and Golbeck. The columns from left to right: a keyword common in programming languages, the category the word belongs to in the LIWC2001 dictionary, the category the work belongs to in the LIWC2007 dictionary, the correlations this LIWC2001 category has with Yarkoni, and the correlations the LIWC2007 dictionary has with Golbeck. For the correlations, the first letter indicates the Big-Five personality trait, followed by the correlations value and asterisks indicating the significance of p-values found in the studies [124, 47]

Keyword	LIWC2001	LIWC2007	Yarkoni	Golbeck
class	Occup	Work	E(-.12)**	C(.33)*, O(.426)*
short	Space		O(-.11)**, A(.16)***	
super	Posemo,	Achiev	E(.1)*, O(-.15)***, A(.18)***	A(-.240)*
	Optim, Affect		O(.15)***, A(.18)***, O(-.12)**	A(-.240)*
if	Discrep,	Cogmech,	N(.13)**, O(-.12)**	C(-.244)*, C(-.292)*
	Preps	Discrep	C(-.13)**, O(.17)***	C(-.244)*, C(-.292)*
for	Preps		O(.17)***	
in	Preps, Incl		O(.17)***, O(.11)**, A(.18)***	
while	Time		O(-.22)***, A(.12)**	
with	Preps, Incl		O(.17)***, O(.11)**, A(.18)***	
try	Occup, Achieve,	Achiev	E(-.12)**, C(.14)***, C(-.11)**	A(-.240)*
	Present, Cogmech		O(-.16)***, N(.13)**	
new	Time		O(-.22)***, A(.12)**	
	$* = p < .05$	$** = p < .01$	$*** = p < .001$	

An example snippet based on the actual dataset is
```
```suggestion case class Action(id:  Int)```
```
. Here the word *class* is falsely identified as to be part of the categories `Occup` and the word *Action* is falsely classified in `Relative` and `Motion`. Although, *Action* might be considered right in this case if the class meant for actions in time, space, and motion. For the word *class*, it is more obvious that the category is wrong, as the class is not an occupation but rather a definition for object-oriented programming.
```

Many more examples exist. We should, therefore, check how much the removal of code blocks influences the outcome and if this change is for the better. The influence of these language-specific keywords on PI is hard to determine upfront, as the service does not publish any of its scoring models. However, we can reasonably assume PI to be influenced.

It is important to note that the method currently misses code not indicated with ticks. Moreover, we may miss the names of classes, paths, files, or variables. A more sophisticated approach, such as the method by Bacchelli et al. [11], possibly improves the parsing step. The method proposed uses island parsing to extract structured data from natural language documents. However, their process currently only works on Java-related materials, so it requires an extension before being deployed for personality inference on software engineering in general.

Practical implication

As the theory suggests, there should be substantial differences between the two datasets. From Table 5.5, we can observe there are indeed many differences found; for almost all traits, more than half of the people tested show a different score. The most notable differences are in Golbeck. For agreeableness, the maximum difference is even 0.75, with an average difference of 0.18.

Table 5.5: Absolute differences in scores for traits with and without code block parsing compared to the ground-truth. The first column shows the personality trait. Then from left to right: the maximum absolute difference observed (indicating the worst case value), the mean absolute difference (indicating an overall effect), and the percentage of people showing a difference with and without the preprocessing step.

Trait	Max abs. diff.	Mean abs. diff.	%People
PI Openness	0.6	0.02	80.2
PI Conscientiousness	0.45	0.03	82.3
PI Extraversion	0.47	0.02	85.5
PI Agreeableness	0.39	0.02	65.3
PI Neuroticism	0.51	0.01	52.1
Yarkoni Openness	0.35	0.01	85.6
Yarkoni Conscientiousness	0.48	0.02	92
Yarkoni Extraversion	0.41	0.02	92.5
Yarkoni Agreeableness	0.38	0.03	60.2
Yarkoni Neuroticism	0.41	0.01	55
Golbeck Openness	0.66	0.02	80.9
Golbeck Conscientiousness	0.82	0.52	100
Golbeck Extraversion	0.59	0.09	99.9
Golbeck Agreeableness	0.75	0.18	100
Golbeck Neuroticism	0.74	0.03	81.1

When looking into the differences in means in both datasets, both the Wilcoxon signed-rank test and the t-test show that all traits have a significant difference. However, most

Table 5.6: Paired Wilcoxon signed-rank test between the scores with the code block preprocessing step enabled and disabled. The columns from left to right: the method and trait tested, the V-statistic, the p-value, a boolean indicating significance for p-values below 0.05, a measure of effect size r, and an interpretation of the effect size r.

Trait	V	p.value	p<0.05	r	effect
PI Openness	129631384.5	0.58	FALSE	0	negligible
Golbeck Openness	258129882.5	0	TRUE	-Inf	large
Golbeck Conscientiousness	28321	0	TRUE	-Inf	large
Golbeck Extraversion	400157008	0	TRUE	-Inf	large
Golbeck Neuroticism	247236653.5	0	TRUE	-Inf	large

Table 5.7: Paired t-test and Cohen's d effect sizes between the scores with the code block preprocessing step enabled and disabled. The columns from left to right: the method and trait tested, the V-statistic, the p-value, a boolean indicating significance for p-values below 0.05, the effect size in terms of Cohen's d, and an interpretation of the effect size Cohen's d.

Trait	V	p.value	p<0.05	d	magnitude
PI Conscientiousness	122.11	0	TRUE	0.27	small
PI Extraversion	-73.48	0	TRUE	-0.14	negligible
PI Agreeableness	92.3	0	TRUE	0.18	negligible
PI Neuroticism	-33.71	0	TRUE	-0.07	negligible
Yarkoni Openness	40.49	0	TRUE	0.07	negligible
Yarkoni Conscientiousness	90.12	0	TRUE	0.15	negligible
Yarkoni Extraversion	-103.41	0	TRUE	-0.29	small
Yarkoni Agreeableness	77.37	0	TRUE	0.16	negligible
Yarkoni Neuroticism	62.66	0	TRUE	0.1	negligible
Golbeck Agreeableness	967.59	0	TRUE	2.05	large

personality traits only show a negligible or small effect. Golbeck, however, shows for all traits a large effect and Yarkoni extraversion a medium effect size. Would we compare the RMSE and MAE scores found with the different datasets (see Table 5.8), we can observe there are differences for which all but two traits are for the better on average. Only PI extraversion and Yarkoni extraversion worsen with 0.01 in terms of RMSE and MAE.

Conclusion

Theoretically, leaving in the code blocks does not make sense for the methods used for this study. Code, with the current approaches, is unlikely to contribute to personality. The keywords in code may have a different meaning than they are categorized to (cf. Table 5.4). From the results we found, removing code blocks mostly improves the predictions. Only for PI extraversion and Yarkoni extraversion, the RMSE and MAE values were slightly worse. The contribution of code to the personality inference is, however, more likely to be

Table 5.8: RMSE and MAE scores for traits with and without code block parsing when compared to the ground-truth. The lower the RMSE and MAE, the better. The columns from left to right: the method and personality trait (abbreviated), the RMSE for the scores with code block parsing enabled, the RMSE for the scores with code block parsing disabled, the MAE for the scores with code block parsing enabled, and the MAE for the scores with code block parsing disabled. Values lower than their counterpart are indicated in bold.

Trait	Enabled RMSE	Disabled RMSE	Enabled MAE	Disabled MAE
PI C.	**0.22**	0.24	**0.17**	0.20
PI E.	0.25	**0.24**	0.20	0.20
PI A.	**0.34**	0.36	**0.31**	0.33
Yarkoni C.	**0.20**	0.21	0.17	0.17
Yarkoni E.	0.23	**0.22**	0.18	0.18
Yarkoni A.	**0.21**	0.22	**0.18**	0.19
Golbeck O.	**0.13**	0.14	0.11	0.11
Golbeck C.	**0.20**	0.29	**0.16**	0.25
Golbeck E.	0.19	0.19	0.16	**0.15**
Golbeck A.	**0.16**	0.19	**0.13**	0.15
Golbeck N.	**0.38**	0.39	**0.32**	0.34

contributing positively by accident than to be a strong and consistent contribution. Both due to its theoretical need and the found results in practice, we conclude this step to be necessary. However, a better and more sophisticated method to extract code from text is required.

5.2.3 Remove quotes

On GitHub, one can quote others by using the greater-than symbol (>). The words used in a quote belong to someone else, meaning that the words within the quote do not contribute to the personality of the person quoting. The removal of quotes is, therefore, a necessary step to take. In the case of this preprocessing step, the effect size is not relevant. Leaving the quotes could give different personality scores but never for the better. This preprocessing step requires no further investigation on the practical implications nor its expected effect on the outcome.

```
Example unprocessed:                Example processed:
i saw you said:          →          i saw you said:
> this is a quote.
```

5.2.4 Conform white space

All three methods work by splitting the sentences on white space to obtain separate words. With the conform white space step, we translate all possible white space to plain spaces. The step replaces new line character `\n`, tab character `\t`, and carriage return `\r` with a

plain space. Please note that multiple consecutive white space characters lead to multiple consecutive plain spaces. A later preprocessing step could reduce these redundant spaces. As this step conforms to the model only to use plain space characters, all preceding processing steps can be optimized by only checking for spaces splitting words and characters. Furthermore, only information affecting the outcome of the analyses would remain.

Example unprocessed:		Example processed:
`one line.\nand the next.`	→	`one line._and the next`

Theoretical implication

The theoretical implications of Yarkoni and Golbeck are dependent on the working of LIWC. Neither LIWC2001 nor LIWC2007 capture the use of white space. Words are split and categorized on this white space; white space itself is discarded. For Personality Insights, the actual rules for white space are, to the best of our knowledge, never published. However, no difference in scores could be found between text with all white space types and text after the conformation of white space for any person. Therefore, we expect PI to ignore white space for the scoring. As there are no differences between the two data sets, no further practical investigation is required.

Conclusion

As white space parsing does not influence the outcome of the analyses and because the step can make the performance of the other preprocessing steps significantly faster, it is considered a required step for improving the overall performance of the application.

5.2.5 Remove numbers

It is not uncommon to include numbers in conversations. However, the methods we use do, for most traits, not interpret personality from these numbers. As we only want to keep data that influences the outcome of the analysis, we remove numbers. Many other studies also removed the numbers before doing linguistic analysis (e.g., [7, 104, 101, 2]).

Example unprocessed:		Example processed:
`it should be between 1 and 10,000.`	→	`it should be between and.`

Theoretical implication

Theoretically, removing numbers should not influence the analyses much. However, Yarkoni does use the `Number` category for extraversion and agreeableness. Furthermore, LIWC includes numbers in the word count. Meaning removing numbers decreases the total word count. In the example below, we can see that the number counts toward the total word count. Removing '1' from the sentence reduces the total word count to three.

Example with a number

Just made 1 commit **Word count:** 4
1 2 3 4

Practical implication

It is crucial to investigate how much the removal of numbers affects the outcome of the analyses. If it affects the result, we should verify it is for the better. Table 5.9 shows the absolute differences observed between the dataset with and without numbers parsing. The most notable differences can be found for Yarkoni, whereas we show the least differences for PI. Yarkoni extraversion even indicates a difference of 0.38 in the largest case and 0.3 for agreeableness. This is not unexpected, as the numbers directly contribute to the scores of these traits.

Table 5.9: Absolute differences in scores for traits with and without number parsing. The first column shows the personality trait. Then from left to right: the maximum absolute difference observed (indicating the worst case value), the mean absolute difference (indicating an overall effect), and the percentage of people showing a difference with and without the preprocessing step.

Trait	Max abs. diff.	Mean abs. diff.	%People
PI Openness	0.03	0	0.2
PI Conscientiousness	0.03	0	0.2
PI Extraversion	0.02	0	0.2
PI Agreeableness	0.02	0	0.1
PI Neuroticism	0.02	0	0.2
Yarkoni Openness	0.05	0	41
Yarkoni Conscientiousness	0.13	0.02	99.9
Yarkoni Extraversion	0	0	0
Yarkoni Agreeableness	0	0	0
Yarkoni Neuroticism	0.14	0	25.1
Golbeck Openness	0.05	0	14.6
Golbeck Conscientiousness	0.05	0	26.2
Golbeck Extraversion	0.05	0	4.2
Golbeck Agreeableness	0.09	0.02	99.5
Golbeck Neuroticism	0.14	0	7.6

As there are differences between the datasets, we check if these differences are significant. The results for the Wilcoxon signed-rank test (see Table 5.10) show only a large significant difference in means for Golbeck neuroticism. PI openness does show a significant difference, but this difference is only negligible. The t-test shows all, but PI agreeableness, to have a significantly small or negligible difference in means.

For these differences, we would like to know if they are for the better or at least if the scores do not get worse. We compare the scores found with both datasets to the ground-

Table 5.10: Paired Wilcoxon signed-rank test between the scores with the number preprocessing step enabled and disabled. The columns from left to right: the method and trait tested, the V-statistic, the p-value, a boolean indicating significance for p-values below 0.05, a measure of effect size r, and an interpretation of the effect size r.

Trait	V	p.value	p<0.05	r	effect
PI Openness	1203	0	TRUE	-0.03	negligible
Golbeck Neuroticism	2225005	0	TRUE	-Inf	large

Table 5.11: Paired t-test and Cohen's d effect sizes between the scores with the number preprocessing step enabled and disabled. The columns from left to right: the method and trait tested, the V-statistic, the p-value, a boolean indicating significance for p-values below 0.05, the effect size in terms of Cohen's d, and an interpretation of the effect size Cohen's d.

Trait	V	p.value	p<0.05	d	effect
PI Conscientiousness	-3.77	0	TRUE	0	negligible
PI Extraversion	3.95	0	TRUE	0	negligible
PI Agreeableness	-0.76	0.45	FALSE	0	negligible
PI Neuroticism	-5.08	0	TRUE	0	negligible
Yarkoni Openness	-129.41	0	TRUE	-0.05	negligible
Yarkoni Conscientiousness	654.57	0	TRUE	0.23	small
Yarkoni Extraversion	-190.12	0	TRUE	-0.25	small
Yarkoni Agreeableness	-89.31	0	TRUE	-0.07	negligible
Yarkoni Neuroticism	92.18	0	TRUE	0.03	negligible
Golbeck Openness	76.41	0	TRUE	0.05	negligible
Golbeck Conscientiousness	92.14	0	TRUE	0.08	negligible
Golbeck Extraversion	25.76	0	TRUE	0.01	negligible
Golbeck Agreeableness	496.96	0	TRUE	0.21	small

truth. In Table 5.12, we report all traits with a difference in either the RMSE or the MAE. Important to note is that for PI, the number of people in the ground-truth showing any difference is only one. The RMSE and MAE for Yarkoni conscientiousness improve with numbers parsing, whereas extraversion and agreeableness improve without. Golbeck conscientiousness, performs slightly better without number parsing, while neuroticism performs slightly better with number parsing.

Conclusion

Theoretically, number parsing could influence the outcome of the analyses. Especially for Yarkoni extraversion and agreeableness, the numbers directly contribute to personality. From a practical examination, we found the differences for PI only to be negligible. For Yarkoni, however, the differences are more substantial. Yarkoni extraversion and agreeableness both correlate with the `Number` category, which explains largely the differences found.

Table 5.12: RMSE and MAE scores for traits with and without number parsing when compared to the ground-truth (not reporting traits without differences in MAE or RMSE). The lower the RMSE and MAE, the better. The columns from left to right: the method and personality trait (abbreviated), the RMSE for the scores with number parsing enabled, the RMSE for the scores with number parsing disabled, the MAE for the scores with number parsing enabled, and the MAE for the scores with number parsing disabled. Values lower than their counterpart are indicated in bold.

	Enabled RMSE	**Disabled RMSE**	**Enabled MAE**	**Disabled MAE**	**people**
Yarkoni C.	**0.2**	0.22	**0.17**	0.18	255
Yarkoni E.	0.23	0.23	0.19	**0.18**	229
Yarkoni A.	0.22	**0.21**	0.18	0.18	93
Golbeck C.	0.19	0.19	0.16	**0.15**	53
Golbeck N.	0.38	0.38	**0.32**	0.33	15

Compared to the ground-truth, some traits improve with the parsing step, while other traits get worse. Different studies have implemented a number parsing step [7, 104, 101, 2]. The RMSE and MAE suggest a possibility for personality inference from the use numbers. We decided to keep the preprocessing step, but disable it for Yarkoni extraversion and agreeableness. Yarkoni extraversion and agreeableness base their scores on the correlations with numbers. The other personality traits follow the same principle as other studies.

5.2.6 Remove hashtags

On GitHub, people use hashtags in multiple ways. Probably the most common way is to reference issues with a hashtag followed by the issue number (see example 1). Another way of using a hashtag would be comparable to Twitter. One could, for example, decide to express their intentions (see example 2). For issue references, it is unlikely the models extract any personality rightfully. However, in the case of a hashtag followed by a word, the word could express personality. Therefore, the step removes issue references and keeps the words preceded by a hashtag. The method is comparable to the technique used by Carducci et al. [21], where they remove the hashtags, followed by a step removing all non-existent words. Their definition of 'non-existent' is the absence of a word in their trained models. Halicioglu et al. [52] and Arnoux et al. [7] used a similar approach by simply removing only the hashtag symbol. In our case, LIWC ignores words that are non-existent in their vocabulary. PI, on the other hand, might learn new words. An extra advantage of removing hashtags is the reduced need for storage.

```
Example 1 unprocessed:
go to #1 for information   →   Example 1 processed:
                               go to for information

Example 2 unprocessed:
today is #release time!    →   Example 2 processed:
                               today is release time!
```

Theoretical implication

Removing issue references has some influence on the outcome of the analyses. All three methods interpret all characters preceding a hashtag as a word, which influences the total word count (cf. section 5.1). The actual implication of this should be investigated.

Practical implication

To check if the preprocessing step affects the outcome, we compare the scores with all preprocessing steps enabled and the scores with all, but the hashtag preprocessing step enabled. Table 5.13 shows the differences in scores found for both datasets. In this table, it is visible there is not much difference; the largest observed difference is PI conscientiousness with 0.14.

Table 5.13: Absolute differences in scores for traits with and without hashtag parsing. The first column shows the personality trait. Then from left to right: the maximum absolute difference observed (indicating the worst case value), the mean absolute difference (indicating an overall effect), and the percentage of people showing a difference with and without the preprocessing step.

Trait	Max abs. diff.	Mean abs. diff.	%People
PI Openness	0.12	0	0.8
PI Conscientiousness	0.14	0	1.1
PI Extraversion	0.09	0	0.8
PI Agreeableness	0.1	0	0.7
PI Neuroticism	0.07	0	0.8
Yarkoni Openness	0.02	0	1.2
Yarkoni Conscientiousness	0.04	0	2.7
Yarkoni Extraversion	0.38	0.01	91.4
Yarkoni Agreeableness	0.3	0	31
Yarkoni Neuroticism	0.03	0	2.8
Golbeck Openness	0.02	0	23.8
Golbeck Conscientiousness	0.03	0	2.1
Golbeck Extraversion	0.01	0	0.8
Golbeck Agreeableness	0.11	0	3
Golbeck Neuroticism	0.05	0	1.3

Both the Wilcoxon-signed rank test (see Table 5.14) and the t-test (see Table 5.15) all show significant differences (except for PI neuroticism) but all these differences are negligible or small. In terms of RMSE and MAE compared to the ground-truth, there are even no differences to report up to two decimal points.

Conclusion

If we look at the differences reported, we observe only slight differences caused by the different word counts. However, this effect is minimal. Especially for people with higher

Table 5.14: Paired Wilcoxon signed-rank test between the scores with the hashtag preprocessing step enabled and disabled. The columns from left to right: the method and trait tested, the V-statistic, the p-value, a boolean indicating significance for p-values below 0.05, a measure of effect size r, and an interpretation of the effect size r.

Trait	V	p.value	p<0.05	r	effect
PI Openness	8171	0	TRUE	-0.03	negligible
Golbeck Openness	430357	0	TRUE	-Inf	large
Golbeck Conscientiousness	2145.5	0	TRUE	-0.14	small
Golbeck Extraversion	22233	0	TRUE	-0.07	negligible
Golbeck Neuroticism	63932.5	0	TRUE	-0.1	negligible

Table 5.15: Paired t-test and Cohen's d effect sizes between the scores with the hashtag preprocessing step enabled and disabled. The columns from left to right: the method and trait tested, the V-statistic, the p-value, a boolean indicating significance for p-values below 0.05, the effect size in terms of Cohen's d, and an interpretation of the effect size Cohen's d.

Trait	V	p.value	p<0.05	d	magnitude
PI Conscientiousness	-5.77	0	TRUE	0	negligible
PI Extraversion	-3.05	0	TRUE	0	negligible
PI Agreeableness	-3.53	0	TRUE	0	negligible
PI Neuroticism	-0.38	0.7	FALSE	0	negligible
Yarkoni Openness	11.09	0	TRUE	0	negligible
Yarkoni Conscientiousness	-24.23	0	TRUE	0	negligible
Yarkoni Extraversion	202.93	0	TRUE	0.27	small
Yarkoni Agreeableness	74.46	0	TRUE	0.06	negligible
Yarkoni Neuroticism	25.96	0	TRUE	0	negligible
Golbeck Agreeableness	-10.37	0	TRUE	0	negligible

word counts, this effect becomes smaller. Because this preprocessing step removes content seemingly irrelevant for the personality of the author and reduces the storage needs, we consider this step to be sufficiently good.

5.2.7 Remove @-references

On GitHub, an @-reference means a reference to another developer to gain their attention. As our models only capture the content of a reference and not the act of referencing, it is a candidate for removal. The preprocessing step removes the full mention. This step is similar to the preprocessing step of Halicioglue et al. [52] for sentiment analysis and Alamsyah et al. [2] who removed usernames before personality measurement.

Table 5.16: RMSE and MAE scores for traits with and without hashtag parsing when compared to the ground-truth (only reporting traits with a difference in RMSE or MAE). The lower the RMSE and MAE, the better. Values lower than their counterpart are indicated in bold. The columns from left to right: the method and personality trait (abbreviated), the RMSE for the scores with hashtag parsing enabled, the RMSE for the scores with hashtag parsing disabled, the MAE for the scores with hashtag parsing enabled, and the MAE for the scores with hashtag parsing disabled.

Trait	Enabled RMSE	Disabled RMSE	Enabled MAE	Disabled MAE
Yarkoni E.	**0.23**	0.24	**0.18**	0.19

Theoretical implication

Theoretically, the removal of @-references could have some implications on the outcome of the analyses. @-References get counted toward the total word count. In the example below, we show how LIWC counts the @-references.

Comment Example @-Reference

@ username could you check this? **Word count:** 5
1 2 3 4 5

The actual theoretical implication for PI is unknown, as the implementation of PI is mostly unknown to the public.

Practical implication

To check the practical implications of the parsing of @-references, we compare the dataset with all preprocessing steps enabled and a dataset with @-reference parsing disabled. If we look at Figure 5.1, we can see that for PI extraversion, the densities are close but not entirely the same, meaning the removal of @-references does seem to influence the outcome of at least some scores. This holds for many of the other personalities.

To check if there are differences, we first compare the scores of people in both datasets. In Table 5.17 we can see the differences found. For some traits, these differences can become reasonably large. Golbeck neuroticism even shows a maximum difference of 0.49, which is a different personality type. At the same time, this trait is affected by the least amount of people. For Yarkoni conscientiousness, the step affects almost everyone (93.9% of people).

As we want to know if the preprocessing step influences the outcome of the analysis, we check for differences in means of the datasets. From these tests, we find significant differences for all traits. However, most differences are only small or negligible, with an exception to PI openness and Golbeck neuroticism, where the effect size is large.

For the people that showed a significant difference, we would like to know if these differences are for the better. We do this by comparing to the scores obtained with the ground-truth. In Table 5.20 we report all traits that showed a difference in RMSE or MAE. From this table, all traits improve with the new parsing step or remain the same. Interestingly, we find no difference in RMSE and MAE for Golbeck neuroticism. This might be

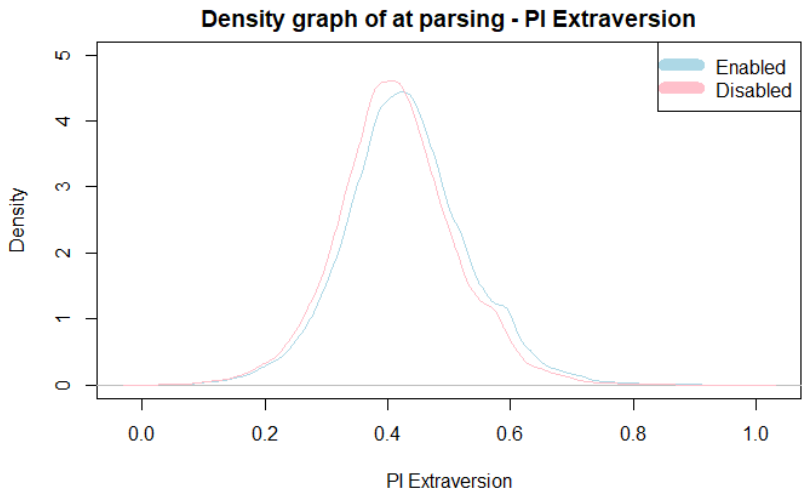


Figure 5.1: PI extraversion density of the dataset with all preprocessing steps enabled (referred to as 'Enabled') and the dataset without the @-reference preprocessing step (referred to as 'Disabled'). The graph shows a difference in scores observable between the two datasets holding the same people.

caused by the low number of people (only 30 people) in the ground-truth with a change in score with the @-reference parsing.

Conclusion

Although we found some significant differences, most of these differences are only small or negligible and large for PI openness and Golbeck neuroticism. Where a discrepancy is found, in terms of RMSE and MAE, it is for the better. Also, it is unlikely that an @-reference with the current implementations of the methods used will have a sound improving effect on the inference of personality. Although the actual use of @-references may indicate personality, the models only capture the content. As of these findings, we consider this step to be necessary for the methods used to improve the interpretation of context.

5.2.8 Remove IP-addresses

In the context of programming, someone possibly talks about IP-addresses. From a manual analysis, we often see conversations like `You should connect to 127.0.0.1`. In this context, the IP-address is unlikely to carry any personality. It could, however, be deemed personally identifiable information in legislation, like the European GDPR [85, 34]. As we expect the IP-address not to contribute to personality, we remove them. By removing the IP-addresses, we reduce the risk of sending sensitive data to PI, reducing the risk of exposure to a third-party. In our implementation, we consider both IPv4 and IPv6 addresses.

Example unprocessed:		Example processed:
`the address should be 127.0.0.1`	→	`the address should be`

Table 5.17: Absolute differences in scores for traits with and without @-parsing. The first column shows the personality trait. Then from left to right: the maximum absolute difference observed (indicating the worst case value), the mean absolute difference (indicating an overall effect), and the percentage of people showing a difference with and without the preprocessing step.

Trait	Max abs. diff.	Mean abs. diff.	%People
PI Openness	0.15	0.01	49.2
PI Conscientiousness	0.16	0.01	79.7
PI Extraversion	0.1	0.02	96.6
PI Agreeableness	0.09	0.01	70.5
PI Neuroticism	0.12	0	26.2
Yarkoni Openness	0.08	0	35.3
Yarkoni Conscientiousness	0.16	0.01	93.9
Yarkoni Extraversion	0.38	0.01	79.7
Yarkoni Agreeableness	0.3	0.01	42.6
Yarkoni Neuroticism	0.13	0.01	61.8
Golbeck Openness	0.05	0	18.4
Golbeck Conscientiousness	0.1	0	18.6
Golbeck Extraversion	0.05	0.01	87.6
Golbeck Agreeableness	0.11	0	35.1
Golbeck Neuroticism	0.49	0	13.7

Table 5.18: Paired Wilcoxon signed-rank test between the scores with the @-reference preprocessing step enabled and disabled. The columns from left to right: the method and trait tested, the V-statistic, the p-value, a boolean indicating significance for p-values below 0.05, a measure of effect size r, and an interpretation of the effect size r.

Trait	V	p.value	p<0.05	r	effect
PI Openness	96229750.5	0	TRUE	-Inf	large
Golbeck Neuroticism	6461343.5	0	TRUE	-Inf	large

Theoretical implication

Theoretically, removing IP-addresses from the text should not influence the outcome of the analyses too much. Both Yarkoni and Golbeck do not use the `AllPunc` or `Period` category scores of LIWC. For PI, the theoretical effect is unknown. We do know that LIWC counts the numbers in IP-address toward the word counts, so removing them has some influence on the average scores of categories. In the example below, we show how the local host IP-address counts toward the total word count.

Example IP-address

127.0. 0. 1:8888
1 2 3 4

Word count: 4

Table 5.19: Paired t-test and Cohen's d effect sizes between the scores with the at reference preprocessing step enabled and disabled. The columns from left to right: the method and trait tested, the V-statistic, the p-value, a boolean indicating significance for p-values below 0.05, the effect size in terms of Cohen's d, and an interpretation of the effect size Cohen's d.

Trait	V	p.value	p<0.05	d	magnitude
PI Conscientiousness	-35.48	0	TRUE	-0.03	negligible
PI Extraversion	308.15	0	TRUE	0.14	negligible
PI Agreeableness	-106.83	0	TRUE	-0.05	negligible
PI Neuroticism	-59.82	0	TRUE	-0.03	negligible
Yarkoni Openness	66.11	0	TRUE	0.03	negligible
Yarkoni Conscientiousness	330.23	0	TRUE	0.14	negligible
Yarkoni Extraversion	123.55	0	TRUE	0.16	negligible
Yarkoni Agreeableness	8.18	0	TRUE	0.01	negligible
Yarkoni Neuroticism	181.42	0	TRUE	0.08	negligible
Golbeck Openness	-20.76	0	TRUE	-0.01	negligible
Golbeck Conscientiousness	43.72	0	TRUE	0.04	negligible
Golbeck Extraversion	-419.35	0	TRUE	-0.21	small
Golbeck Agreeableness	-104.98	0	TRUE	-0.05	negligible

Table 5.20: RMSE and MAE scores for traits with and without at reference parsing when compared to the ground-truth (only reporting traits with a difference in MAE or RMSE). The lower the RMSE and MAE, the better. The columns from left to right: the method and personality trait (abbreviated), the RMSE for the scores with at reference parsing enabled, the RMSE for the scores with at reference parsing disabled, the MAE for the scores with at reference parsing enabled, the MAE for the scores with at reference parsing disabled, and the number of people in the ground-truth showing a difference. Values lower than their counterpart are indicated in bold.

Trait	Enabled RMSE	Disabled RMSE	Enabled MAE	Disabled MAE	people
PI O.	0.18	0.18	**0.14**	0.15	122
PI E.	**0.24**	0.25	**0.19**	0.2	249
Yarkoni C.	**0.2**	0.21	**0.17**	0.18	246
Yarkoni E.	0.23	0.23	**0.18**	0.19	205
Golbeck E.	**0.19**	0.2	0.16	0.16	219
Golbeck A.	**0.16**	0.17	0.13	0.13	98

We do, however, need to investigate the live numbers to say something about the actual implications of the preprocessing step and the significance of the influence.

Practical implication

From a first glance, there are not many differences to observe between the personality traits with all preprocessing steps enabled and without IP-address preprocessing. In Table 5.21, the most differences can be found for Yarkoni extraversion. This is undoubtedly caused by

the `Numbers` category of LIWC, which correlates for -0.12 with extraversion [124]. On a general level, the influence of IP parsing appears to be small and unlikely to influence the outcome.

Table 5.21: Absolute differences in scores for traits with and without IP address parsing. The first column shows the personality trait. Then from left to right: the maximum absolute difference observed (indicating the worst case value), the mean absolute difference (indicating an overall effect), and the percentage of people showing a difference with and without the preprocessing step.

Trait	Max abs. diff.	Mean abs. diff.	%People
PI Openness	0.11	0	0.6
PI Conscientiousness	0.07	0	0.7
PI Extraversion	0.05	0	0.6
PI Agreeableness	0.05	0	0.5
PI Neuroticism	0.06	0	0.4
Yarkoni Openness	0.04	0	0.2
Yarkoni Conscientiousness	0.05	0	2.5
Yarkoni Extraversion	0.37	0.02	95.5
Yarkoni Agreeableness	0.3	0	32.3
Yarkoni Neuroticism	0.02	0	0.2
Golbeck Openness	0.02	0	24.9
Golbeck Conscientiousness	0.06	0	1.4
Golbeck Extraversion	0.01	0	0.1
Golbeck Agreeableness	0.06	0	0.2
Golbeck Neuroticism	0.08	0	0.1

Table 5.22: Paired Wilcoxon signed-rank test between the scores with the IP address preprocessing step enabled and disabled. The columns from left to right: the method and trait tested, the V-statistic, the p-value, a boolean indicating significance for p-values below 0.05, a measure of effect size r, and an interpretation of the effect size r.

Trait	V	p.value	p<0.05	r	effect
PI Openness	-6.48	0	TRUE	-0.04	negligible
Golbeck Neuroticism	0	0.93	FALSE	0	negligible

To assess if there are differences in means for the datasets, we apply the Wilcoxon signed-rank test [123] for the non-normal traits (see Table 5.22) and apply a paired t-test [115] for the normally distributed traits (see Table 5.23). Most LIWC-based traits do not show a significant difference. The exceptions are Yarkoni conscientiousness and extraversion (due to the `Number` category) and Golbeck conscientiousness (due to the `Colon` category). For PI, all but neuroticism show a significantly negligible difference.

If there are differences, we would like to know if these differences improve the scores or at least do not worsen the outcome. We do this by comparing the scores found with both

Table 5.23: Paired t-test and Cohen's d effect sizes between the scores with the IP address preprocessing step enabled and disabled. The columns from left to right: the method and trait tested, the V-statistic, the p-value, a boolean indicating significance for p-values below 0.05, the effect size in terms of Cohen's d, and an interpretation of the effect size Cohen's d.

Trait	V	p.value	p<0.05	d	magnitude
PI Conscientiousness	-4.5	0	TRUE	0	negligible
PI Extraversion	-7.73	0	TRUE	0	negligible
PI Agreeableness	-2.38	0.02	TRUE	0	negligible
PI Neuroticism	1.22	0.22	FALSE	0	negligible
Yarkoni Openness	0	1	FALSE	0	negligible
Yarkoni Conscientiousness	24.35	0	TRUE	0	negligible
Yarkoni Extraversion	287.91	0	TRUE	0.37	small
Yarkoni Agreeableness	89.93	0	TRUE	0.07	negligible
Yarkoni Neuroticism	-1.04	0.3	FALSE	0	negligible
Golbeck Openness	-94.22	0	TRUE	-0.07	negligible
Golbeck Conscientiousness	-16.31	0	TRUE	0	negligible
Golbeck Extraversion	0.54	0.59	FALSE	0	negligible
Golbeck Agreeableness	1.13	0.26	FALSE	0	negligible

Table 5.24: RMSE and MAE scores for traits with and without IP address parsing when compared to the ground-truth (only reporting traits with a difference in RMSE or MAE). The lower the RMSE and MAE, the better. Values lower than their counterpart are indicated in bold. The columns from left to right: the method and personality trait (abbreviated), the RMSE for the scores without IP addresses, the RMSE for the scores with IP addresses, the MAE for the scores with IP address parsing enabled, the MAE for the scores with IP address parsing disabled, and the number of people in the ground-truth showing a difference.

Trait	Enabled RMSE	Disabled RMSE	Enabled MAE	Disabled MAE	people
Yarkoni E.	**0.23**	0.24	**0.18**	0.19	246

datasets against the ground-truth. When comparing these scores with the RMSE and MAE, we can only observe Yarkoni extraversion to have an improvement in RMSE (see Table 5.24). All other traits do not show a difference to two decimals or did not have enough people in the ground-truth with differences.

Conclusion

From the above, we observe the preprocessing step to influence some of the personality traits. However, none of these differences are significant, and if significant, the difference is only negligible or small. As IP-addresses are rather work-related, and we do not expect them much an expression of personality, we consider this step relevant. Furthermore, under privacy regulations, the IP-address could be regarded as personally identifiable information. Thus, the removal of IP-addresses in this analysis reduces the privacy risks for the data

subject[1]. We, therefore, consider this step sufficiently good to adopt.

5.2.9 Remove URLs

Like many other platforms, GitHub allows for sharing links. As the link itself is unlikely to carry personal information, we remove the URLs. One could think of a way to extract personality from URLs, for example, by checking where to it refers. However, this does not lie in the scope of this study. Many other studies also removed URLs before their personality-related analysis [52, 101, 21, 2, 7].

Theoretical implication

Theoretically, the psycholinguistic models should not get much information from URLs. However, the words used in the URL are, to the least, categorized by LIWC. LIWC splits each of the sections in the URL. The example below shows how LIWC interprets an URL.

Comment Example URL

https://www.tudelft.nl/en/education/programmes/ **Word count:** 7
1 2 3 4 5 6 7

The above example indicates that by removing the URLs, the average category scores might vary. For PI, we do not know how the service handles URLs. How much the step affects the outcome of the analyses needs further investigation.

Practical implication

In Table 5.25, we show the differences found for each of the methods and traits. All of the traits and methods prove to be affected by this parsing method. Yarkoni agreeableness and Golbeck neuroticism even show an average difference of 0.04.

To check if the differences are significant to the whole, we compare the differences in means for the scores with URLs to scores with all URLs removed. In Table 5.26, the results of the Wilcoxon signed-rank test show a significantly large difference for all traits. In Table 5.27, showing the T-test results, all traits are significantly different in means, but with negligible effect. Only Yarkoni agreeableness, Golbeck conscientiousness, and Golbeck agreeableness have a medium effect.

As there are medium and large differences in the dataset, we know the preprocessing step does influence the outcome of the analyses. To verify this step improves or at least not worsens the result, we compare the RMSE and MAE scores against the ground-truth. Table 5.28 lists all traits with a difference in RMSE or MAE with and without URL parsing. We can observe that in almost all cases, the preprocessing step does improve the scores. Only in the case of Yarkoni openness and neuroticism, the RMSE scores of the URL dataset are better. At the same time, the MAE scores do not change for these two traits. The worsening toward the ground-truth of the RMSE is only 0.01, so the impact is not substantial. As it

[1]The data subject is the person who wrote the comments. In other words, the person being processed.

Table 5.25: Absolute differences in scores for traits with and without URL parsing. The first column shows the personality trait. Then from left to right: the maximum absolute difference observed (indicating the worst case value), the mean absolute difference (indicating an overall effect), and the percentage of people showing a difference with and without the preprocessing step.

Trait	Max abs. diff.	Mean abs. diff.	%People
PI Openness	0.44	0	20.9
PI Conscientiousness	0.35	0	24.9
PI Extraversion	0.39	0	20.8
PI Agreeableness	0.25	0.01	60.7
PI Neuroticism	0.33	0	18.8
Yarkoni Openness	0.29	0.01	51.3
Yarkoni Conscientiousness	0.23	0.01	82.5
Yarkoni Extraversion	0.28	0.01	64.4
Yarkoni Agreeableness	0.4	0.05	99.9
Yarkoni Neuroticism	0.24	0.01	65.3
Golbeck Openness	0.25	0.01	87.7
Golbeck Conscientiousness	0.17	0.01	78.8
Golbeck Extraversion	0.36	0	32.1
Golbeck Agreeableness	0.24	0.03	98.8
Golbeck Neuroticism	0.56	0.04	83

Table 5.26: Paired Wilcoxon signed-rank test between the scores with the URL preprocessing step enabled and disabled. The columns from left to right: the method and trait tested, the V-statistic, the p-value, a boolean indicating significance for p-values below 0.05, a measure of effect size r, and an interpretation of the effect size r.

Trait	V	p.value	p<0.05	r	effect
PI Openness	14542504	0	TRUE	-Inf	large
Golbeck Openness	2483381.5	0	TRUE	-Inf	large
Golbeck Neuroticism	274722034	0	TRUE	-Inf	large

is unlikely for the URLs to contribute to the personality, this worsening effect compared to the ground-truth may be caused on accident.

Conclusion

Removing URLs from the text does influence the outcome of the analyses. Especially for Yarkoni, the differences become apparent. With the theory and context, it is unlikely the worsening of scores toward the ground-truth is caused by actually missing on personality but rather on accident. Many other studies also believe the URLs do not contribute to personality in linguistic analyses [52, 101, 21, 2, 7]. If we also consider the scores obtained with the comparison to the ground-truth, it seems that, in most cases, the scores improve

Table 5.27: Paired t-test and Cohen's d effect sizes between the scores with the URL preprocessing step enabled and disabled. The columns from left to right: the method and trait tested, the V-statistic, the p-value, a boolean indicating significance for p-values below 0.05, the effect size in terms of Cohen's d, and an interpretation of the effect size Cohen's d.

Trait	V	p.value	p<0.05	d	magnitude
PI Conscientiousness	41.86	0	TRUE	0.02	negligible
PI Extraversion	-9.29	0	TRUE	-0.01	negligible
PI Agreeableness	123.52	0	TRUE	0.06	negligible
PI Neuroticism	-13.41	0	TRUE	-0.01	negligible
Yarkoni Openness	-85.89	0	TRUE	-0.09	negligible
Yarkoni Conscientiousness	104.73	0	TRUE	0.11	negligible
Yarkoni Extraversion	143.92	0	TRUE	0.21	small
Yarkoni Agreeableness	453.93	0	TRUE	0.74	medium
Yarkoni Neuroticism	131.33	0	TRUE	0.13	negligible
Golbeck Conscientiousness	-192.02	0	TRUE	-0.43	small
Golbeck Extraversion	-45.94	0	TRUE	-0.06	negligible
Golbeck Agreeableness	417.8	0	TRUE	0.51	medium

with the removal of URLs. We, therefore, conclude this step to enhance the personality inference or at least not to worsen significantly.

5.2.10 Remove emails

Emails, in the context of our study, unlikely hold any personality. Emails may even contain non-existent words. As we want our models only to consider words that may influence the personality, we should remove emails. Besides, an email may be considered personally identifiable information, as with the GDPR [34]. By removing emails, we reduce the risk of re-identification by third party applications, like PI.

```
Example unprocessed:                    Example processed:
contact me on email@address.com.   →    contact me on.
```

Theoretical implication

Theoretically, removing emails should not impact the outcome of the analyses. The words chosen inside the email address are an identifier of a person, who is not necessarily the author of the comment. With the removal of emails, the total word count varies and, thus, the category scores may vary slightly (see example below). For PI, the effect of removing emails mostly depends on the word splitting method used.

Table 5.28: RMSE and MAE scores for traits with and without URL parsing when compared to the ground-truth (only reporting traits with a difference in MAE or RMSE). The lower the RMSE and MAE, the better. Values lower than their counterpart are indicated in bold. The columns from left to right: the method and personality trait (abbreviated), the RMSE for the scores with URL parsing enabled, the RMSE for the scores with URL parsing disabled, the MAE for the scores with URL parsing enabled, the MAE for the scores with URL parsing disabled, and the number of people in the ground-truth showing a difference.

Trait	Enabled RMSE	Disabled RMSE	Enabled MAE	Disabled MAE	people
PI C.	0.22	0.22	**0.17**	0.18	59
PI A.	**0.34**	0.35	**0.31**	0.32	172
Yarkoni O.	0.16	**0.15**	0.13	0.13	127
Yarkoni C.	**0.2**	0.21	**0.17**	0.18	214
Yarkoni E.	**0.23**	0.24	**0.18**	0.19	250
Yarkoni A.	**0.21**	0.25	**0.18**	0.22	259
Yarkoni N.	0.22	**0.21**	0.18	0.18	167
Golbeck C.	**0.2**	0.21	0.16	0.16	210
Golbeck N.	**0.38**	0.4	**0.32**	0.35	208

Example email

email@example.com
1 2 3

Word count: 3

Practical implication

To investigate the practical implications, we compare personality scores with all preprocessing steps enabled to scores with all preprocessing, but email parsing enabled. First, we investigate if there are differences in each of the personality traits and methods. In Table 5.29 we display the found results. From this table, it becomes apparent that the step only affects a small group of people. The highest number of people is 43 for Yarkoni extraversion. The largest difference is 0.09 for PI extraversion. For most traits, the average difference must be below 0.005 which is negligible.

To check if these differences are significant, we check for the differences in means (see Tables 5.30 and 5.31). From Table 5.30, we can see all (but Golbeck extraversion) to be significantly different in means, although all differences are small. Similarly, for Table 5.31, all (but PI neuroticism) are significantly different, but with a negligible difference.

None of the 267 people in the ground-truth showed a difference with the parsing method enabled and disabled. Therefore, no comparison on the ground-truth is possible.

Conclusion

For all personality traits, we did not see evident differences in means and medians. Some traits show significant difference in means, but all differences were negligible to small. We

Table 5.29: Absolute differences in scores for traits with and without email parsing. The first column shows the personality trait. Then from left to right: the maximum absolute difference observed (indicating the worst case value), the mean absolute difference (indicating an overall effect), and the percentage of people showing a difference with and without the preprocessing step.

Trait	Max abs. diff.	Mean abs. diff.	%People
PI Openness	0.05	0	0.1
PI Conscientiousness	0.03	0	0.1
PI Extraversion	0.09	0	0.1
PI Agreeableness	0.03	0	0.1
PI Neuroticism	0.07	0	0.1
Yarkoni Openness	0.01	0	0
Yarkoni Conscientiousness	0.01	0	0.1
Yarkoni Extraversion	0.38	0.01	90.3
Yarkoni Agreeableness	0.3	0	32.2
Yarkoni Neuroticism	0.01	0	0.1
Golbeck Openness	0.02	0	24.9
Golbeck Conscientiousness	0.01	0	0
Golbeck Extraversion	0.01	0	0
Golbeck Agreeableness	0.07	0	0.1
Golbeck Neuroticism	0.02	0	0.1

Table 5.30: Paired Wilcoxon signed-rank test between the scores with all preprocessing steps and the scores without email parsing. The columns from left to right: the method and trait tested, the V-statistic, the p-value, a boolean indicating significance for p-values below 0.05, a measure of effect size r, and an interpretation of the effect size r.

Trait	V	p.value	p<0.05	r	effect
PI Openness	3.27	0	TRUE	-0.02	small
Golbeck Openness	2.53	0	TRUE	-0.02	small
Golbeck Conscientiousness	-3.61	0	TRUE	-0.02	small
Golbeck Extraversion	2.11	0.07	FALSE	-0.01	small
Golbeck Neuroticism	2.5	0.01	TRUE	-0.01	small

do not have enough people in the ground-truth to make a comparison on accuracy. As we could not find apparent differences between the scores of both datasets, we conclude that the email parsing does not improve nor worsen the scores found with any of the methods used. Emails are, with the current approach, unlikely to contribute to personality. The removal of email addresses does improve the privacy implications for the data subject. We, therefore, recommend the preprocessing step.

Table 5.31: Paired t-test and Cohen's d effect sizes between the scores with all preprocessing steps and the scores without email parsing. The columns from left to right: the method and trait tested, the V-statistic, the p-value, a boolean indicating significance for p-values below 0.05, the effect size in terms of Cohen's d, and an interpretation of the effect size Cohen's d.

Trait	V	p.value	p<0.05	d	effect
PI Conscientiousness	4.65	0	TRUE	0	negligible
PI Extraversion	1.46	0.14	FALSE	0	negligible
PI Agreeableness	4.9	0	TRUE	0	negligible
PI Neuroticism	-2.69	0.01	TRUE	0	negligible
Yarkoni Openness	-1.15	0.25	FALSE	0	negligible
Yarkoni Conscientiousness	-2.89	0	TRUE	0	negligible
Yarkoni Extraversion	5.77	0	TRUE	0	negligible
Yarkoni Agreeableness	-3.18	0	TRUE	0	negligible
Yarkoni Neuroticism	2.13	0.03	TRUE	0	negligible
Golbeck Agreeableness	-1.19	0.23	FALSE	0	negligible

5.2.11 Remove double white space

The removal of double white space is to make the text better readable to the human reader, as well as to reduce the space required for storing the data. The extra spaces are unlikely to hold any information for the models used and can, therefore, safely be removed. This step mostly cleans up the spaces that are left by previous preprocessing steps.

```
Example unprocessed:                    Example processed:
double_ spaces _are _removed    →       double spaces are removed
```

Theoretical implication

Theoretically, removing extra white space cannot influence any of the outcomes of the analyses. If all methods split words based on white space, they cannot extract new words from double white space. In the example above, at most four words can be extracted, independent of the number of spaces.

Practical implication

Although theoretically there should not be any influence, we observe some differences with and without the use of double white space parsing. If we look at Table 5.32, we can observe that PI is influenced for at most 34 out of 28,337 people with a maximum difference of 0.01. All other traits and methods have fewer people with a difference with all a maximum of 0.01. With this observation, it is already quite unlikely the difference is significant. From a manual inspection on the people that have a difference in Yarkoni and Golbeck, it becomes apparent that the splitting method of LIWC causes the differences. Without the removal of double white space, LIWC misses some words from categorization. An example found is

'`mr.␣␣␣`' compared to '`mr.`'. In the first case, `mr` is not categorized in `Social`, while for the second case it is. This problem seems to be a bug in the implementation of LIWC. For PI, a similar problem could be possible, but due to it being a black-box, we cannot point to the exact reason.

Table 5.32: Absolute differences in scores for traits with and without double white space parsing. Not reporting traits without a difference. The first column shows the personality trait. Then from left to right: the maximum absolute difference observed (indicating the worst case value), the mean absolute difference (indicating an overall effect), and the percentage of people showing a difference with and without the preprocessing step rounded to one decimal.

Trait	Max abs. diff.	Mean abs. diff.	%People
PI Openness	0.01	0	0.1
PI Conscientiousness	0.01	0	0.1
PI Extraversion	0.01	0	0.1
PI Agreeableness	0.01	0	0.1
PI Neuroticism	0.01	0	0
Yarkoni Openness	0.01	0	0
Yarkoni Conscientiousness	0.01	0	0
Yarkoni Extraversion	0.01	0	0
Yarkoni Agreeableness	0.01	0	0
Yarkoni Neuroticism	0.01	0	0
Golbeck Openness	0.01	0	0
Golbeck Conscientiousness	0.01	0	0

To check if there the differences in the datasets are significant, we check for differences in means (see Tables 5.33 and 5.34). From all results, we observe none of the traits to show a significant effect ($p > .05$).

We could not draw any conclusion on differences with the ground-truth, as only one person showed a variation in a trait. For this person, only PI conscientiousness and extraversion improved with 0.01, and PI neuroticism worsened with 0.01. All other traits remained the same.

Conclusion

From all information found, due to a possible bug in LIWC, some influence is seen for some people. Due to the black-box nature of PI, we could not pinpoint the cause of changes in PI. We can, however, safely conclude that this step does not influence the outcome or at least does not worsen it (significantly). As it does clean up the data by making it more readable and reduces the space required for storage, we conclude it to be a sufficiently good preprocessing step to take.

Table 5.33: Paired Wilcoxon signed-rank test between the scores with the double white space preprocessing step enabled and disabled. The columns from left to right: the method and trait tested, the V-statistic, the p-value, a boolean indicating significance for p-values below 0.05, a measure of effect size r, and an interpretation of the effect size r.

Trait	V	p.value	p<0.05	r	effect
PI Conscientiousness	190350973	0.69	FALSE	0	small
PI Extraversion	189193460	0.89	FALSE	0	small
PI Agreeableness	190508671	0.86	FALSE	0	small
PI Neuroticism	184465051.5	0.97	FALSE	0	small
Yarkoni Openness	185216276	0.87	FALSE	0	small
Yarkoni Conscientiousness	185918154	0.96	FALSE	0	small
Yarkoni Extraversion	170957066	0.99	FALSE	0	small
Yarkoni Agreeableness	179679398.5	0.9	FALSE	0	small
Yarkoni Neuroticism	186601164	0.97	FALSE	0	small

Table 5.34: Paired t-test and Cohen's d effect sizes between the scores with the double white space preprocessing step enabled and disabled. The columns from left to right: the method and trait tested, the V-statistic, the p-value, a boolean indicating significance for p-values below 0.05, the effect size in terms of Cohen's d, and an interpretation of the effect size Cohen's d.

Trait	V	p.value	p<0.05	d	effect
PI Openness	0.01	0.99	FALSE	0	negligible
Golbeck Openness	0	1	FALSE	0	negligible
Golbeck Conscientiousness	0	1	FALSE	0	negligible

5.2.12 Remove space before punctuation

Similarly to the removal of double white space, the removal of space before punctuation (e.g., dots and commas) is mostly because of the need for clean up caused by other preprocessing steps. E.g., removing a word before a comma may leave a space. This preprocessing step is to reduce the space required for the storage and to make the data better readable. The punctuation considered are dots, commas, exclamation marks, semicolons, and question marks.

```
Example unprocessed:                          Example processed:
space before punctuation_, is removed_.   →   space before punctuation, is removed.
```

Theoretical implication

Theoretically, removing white space before punctuation should not influence any of the methods used. Methods split on the white space characters, and punctuation does not count toward a word count. Therefore, splitting a sentence like '`thnx . lgtm !`' should not be interpreted differently by any of the methods than '`thnx. lgtm!`'.

Practical implication

In practice, the assumption seems to be true. For both LIWC-based methods and PI, there are no differences in the scores with all preprocessing steps and the scores with spaces before punctuation. For PI, as the service is taxing for each call, we only check for ten different text bodies ranging from 100 to 7020 words. For each of these texts, we intentionally added spaces before each punctuation to force the effect to the most extreme. For none of these calls, we find a difference. Although this is not an extensive search, it is highly unlikely there will ever be a difference in scores with this preprocessing step.

Conclusion

With all findings, we conclude the preprocessing step not to influence the outcome of any of the personality scores. It does, however, reduce the storage needs and improves readability of comments. We, therefore, conclude this step to be a sufficiently good preprocessing step.

5.2.13 Remove images

Within a GitHub discussion, it is possible to insert images using the markdown format `![image_path](alt_text)`, where `image_path` is the path to the image and `alt_text` is the alternative text when the image is not available. To make sure the context is interpreted rightfully by the methods used, we remove the images. Markdown images always follow the same pattern: an exclamation mark followed by square brackets with the file path in between, followed by round brackets and an optional alternative text in between.

```
Example unprocessed:                                  Example processed:
the illustration:  ![path/image.png](alt)     →       the illustration:
```

Theoretical implication

Theoretically, the images could influence the outcome of all three methods in many ways. The path to the image, the alt text, and even the symbols used may impact the linguistic analysis. For Golbeck, the exclamation mark also influences the outcome of openness, conscientiousness, and neuroticism. The exclamation mark is falsely identified as being part of the sentence context. Without this step, people with many images may find themselves classified as highly neurotic. Similarly, any trait correlating with the `Parenth` or `Exclam` category of LIWC may be influenced. Furthermore, the words contained in the image contribute to the word count. In the example below, we show how LIWC counts the terms in these cases.

Comment Example Number

`![image.png](image unavailable)`
1 2 3 4

Word count: 4

To know how much images influence the outcome of the inference of all three methods, we must investigate the differences in scores when applied.

Practical implication

To check if the image parsing affects the outcome, we first investigate the absolute differences between the dataset with all preprocessing steps and the scores without image parsing. In Table 5.35, we can observe differences for all traits. The most notable differences are in PI extraversion and Golbeck extraversion and neuroticism. For Golbeck extraversion, this change is mostly due to the `Parenth` category. For neuroticism, the change is primarily due to the `Exclam` category. Golbeck openness and conscientiousness also use `Exclam`. However, as they have correlations with more categories than neuroticism, they are relatively less influenced by `Exclam`.

Table 5.35: Absolute differences in scores for traits with and without image parsing. The first column shows the personality trait. Then from left to right: the maximum absolute difference observed (indicating the worst case value), the mean absolute difference (indicating an overall effect), and the percentage of people showing a difference with and without the preprocessing step to one decimal

Trait	Max abs. diff.	Mean abs. diff.	%People
PI Openness	0.23	0	3.2
PI Conscientiousness	0.34	0	4.2
PI Extraversion	0.27	0.01	99
PI Agreeableness	0.12	0	2.7
PI Neuroticism	0.22	0	4.6
Yarkoni Openness	0.11	0	2.1
Yarkoni Conscientiousness	0.06	0	2.3
Yarkoni Extraversion	0.11	0	2.9
Yarkoni Agreeableness	0.09	0	2.1
Yarkoni Neuroticism	0.09	0	2.3
Golbeck Openness	0.3	0	27
Golbeck Conscientiousness	0.07	0	3.6
Golbeck Extraversion	0.29	0.02	97.6
Golbeck Agreeableness	0.22	0	2.2
Golbeck Neuroticism	0.72	0.05	83.7

Table 5.36: Paired Wilcoxon signed-rank test between the scores with the image preprocessing step enabled and disabled. The columns from left to right: the method and trait tested, the V-statistic, the p-value, a boolean indicating significance for p-values below 0.05, a measure of effect size r, and an interpretation of the effect size r.

Trait	V	p.value	p<0.05	r	effect
PI Openness	259438.5	0	TRUE	-0.04	negligible
Golbeck Openness	2525628	0	TRUE	-Inf	large
Golbeck Neuroticism	276104425.5	0	TRUE	-Inf	large

As there are many differences, we would like to investigate if there are significant dif-

ferences in the means of the datasets. In Tables 5.36 and 5.37 we outline the results of the Wilcoxon signed-rank test and the t-test, respectively. From these tests, all (but Yarkoni agreeableness) show a significant difference in means. Most traits, however, only show a negligible or small effect. The only exceptions are Golbeck openness and neuroticism, which show a large effect size.

Table 5.37: Paired t-test and Cohen's d effect sizes between the scores with the image preprocessing step enabled and disabled. The columns from left to right: the method and trait tested, the V-statistic, the p-value, a boolean indicating significance for p-values below 0.05, the effect size in terms of Cohen's d, and an interpretation of the effect size Cohen's d.

Trait	**V**	**p.value**	**p<0.05**	**d**	**effect**
PI Conscientiousness	14.11	0	TRUE	0	negligible
PI Extraversion	395.1	0	TRUE	0.13	negligible
PI Agreeableness	10.38	0	TRUE	0	negligible
PI Neuroticism	-10.55	0	TRUE	0	negligible
Yarkoni Openness	-17.23	0	TRUE	0	negligible
Yarkoni Conscientiousness	-17.68	0	TRUE	0	negligible
Yarkoni Extraversion	21.29	0	TRUE	0.01	negligible
Yarkoni Agreeableness	1.64	0.1	FALSE	0	negligible
Yarkoni Neuroticism	18.23	0	TRUE	0	negligible
Golbeck Conscientiousness	-21.56	0	TRUE	-0.01	negligible
Golbeck Extraversion	-272.6	0	TRUE	-0.33	small
Golbeck Agreeableness	-10.38	0	TRUE	0	negligible

As there are traits that are primarily influenced by the preprocessing step, we need to investigate if this change is for the better. We compare the scores to the ground-truth. Table 5.38 shows the traits for which there is a different RMSE or MAE when the parsing step is enabled or disabled. For both traits, the dataset with the image preprocessing enabled shows lower RMSE and MAE values compared to the dataset without the image preprocessing.

Table 5.38: RMSE and MAE scores for traits with and without image parsing when compared to the ground-truth (only showing traits with a difference in MAE or RMSE). The lower the RMSE and MAE, the better. Values lower than their counterpart are indicated in bold. The columns from left to right: the method and personality trait (shortened), the RMSE for the scores with image parsing enabled, the RMSE for the scores with image parsing disabled, the MAE for the scores with image parsing enabled, the MAE for the scores with image parsing disabled, and the number of people affected also contained in the ground-truth.

	Enabled RMSE	**Disabled RMSE**	**Enabled MAE**	**Disabled MAE**	**people**
PI E.	**0.24**	0.25	**0.19**	0.2	258
Golbeck N.	**0.38**	0.41	**0.32**	0.36	207

Conclusion

In theory, the images can be misconstrued. E.g., Golbeck neuroticism assumes the exclamation mark to be part of the sentence. However, the exclamation mark is merely a syntax chosen for Markdown images. Similarly, the parentheses of the Markdown syntax influences the outcome of the scores found for Golbeck extraversion. From the theory, along with the results found for the RMSE and MAE scores, we conclude this step to improve the outcome of the inference step.

Takeaway

Thirteen preprocessing steps have been proposed and shown to improve the outcome, not to influence the outcome, or at least not worsen the outcome significantly. The only exception is number parsing. Here we decided to disable the preprocessing step for Yarkoni extraversion and agreeableness. To summarize the findings:

- Preprocessing steps to improve the models:
  `Code blocks`, `Quotes`, `URLs`, `Numbers`[a], `Images`
- Preprocessing steps not affecting the outcome significantly, but improving other properties (e.g., privacy or storage needs):
  `Lower cases`[b], `White space`, `IP-addresses`, `Double white space`, `Spaces before punctuation`
- Parsing methods that do influence the outcome but with a theoretical need:
  `@-References`, `Hashtags`, `Emails`

Based on this conclusion, we apply the preprocessing steps on all input of the psycholinguistic methods for the remainder of this document. Only for Yarkoni extraversion and agreeableness, the numbers parsing is disabled.

[a] With the exception of Yarkoni extraversion and agreeableness.
[b] No influence is observed for LIWC. For PI, there is only a small difference.

Chapter 6

How well do the psycholinguistic models perform when compared to the ground-truth? (RQ.II)

Each of the methods for personality inference works differently and, thus, we may expect different outcomes. We use the preprocessing steps found in the previous chapter to obtain personality scores from GitHub comments. In this chapter, we aim to get a general sense of the performance of the models by investigating the scores generated through inference. We first look into the distribution of the inferred scores among the models and compare the scores between the models to investigate possible differences in the models. Based on these observations, we could conclude if models perform similarly or differently. Then, we compare the scores of the inferred personality scores to the ground-truth to assess the performance for each of the models. Finally, we explain how we use transformations on category scores of LIWC-based methods to improve the predicted personality scores.

6.1 Comparison of models

Table 6.1 shows an overview of the scores found through linguistic analysis with all three methods. If the inferred scores in our dataset can be generalized to all people in the world, someone with a score of 0.5 has an average personality for that trait. With this in mind, we can observe from Table 6.1 that the openness scores of all methods are fairly high with a mean of 0.62-0.7. Meaning the people in our dataset, on average, have a higher openness score than the average person. For PI and Yarkoni, the average scores for conscientiousness ($\Delta = .02$)[1], extraversion ($\Delta = .03$), and neuroticism ($\Delta = .04$) are reasonably close. Agreeableness scores, on the other hand, are significantly lower for PI ($M = 0.38$) than for Yarkoni ($M = 0.53$), $V = 1e7, p < 0.05$, with a large effect size.

If we look at the differences in means between the methods, we find several notable differences. Golbeck shows a significantly large difference with all traits of PI. For all traits, except neuroticism, these average scores found are significantly higher than for PI. Golbeck

[1]The symbol Δ, in this context, means the 'difference' between both methods.

neuroticism, however, shows a significantly lower mean ($M = 0.12$) than PI neuroticism ($M = 0.5$) and Yarkoni neuroticism ($M = 0.46$). For Golbeck neuroticism, the third quartile is even 0.17. Such a low third quartile would mean that if the inferred scores are relative to all people in the world, the vast majority of people in our dataset are amongst the one-fifth of the least neurotic people in the world. PI, Yarkoni, and also the ground-truth (cf. Chapter 4) suggest the majority of people in the dataset should belong to the group of averagely neurotic people (with a mean close to 0.5).

Yarkoni and PI, on average, seemingly agree more on scores than with Golbeck. However, due to Min-Max normalization used, Golbeck may only be different in scaling caused by outlier behavior. Due to the linear behavior of Min-Max normalization, one large outlier could trigger all other values to become lower in scale.

Table 6.1: The distribution of all inferred scores for each method. In the first column, the method used and the Big-Five personality trait is given. From left to right, for each trait, the minimum, first quartile, median, mean, third quartile, and maximum score are displayed.

Trait	**Min**	**Q1**	**Median**	**Mean**	**Q3**	**Max**
PI Openness	0	0.58	0.64	0.62	0.68	1
PI Conscientiousness	0	0.43	0.49	0.5	0.56	1
PI Extraversion	0	0.37	0.43	0.43	0.49	1
PI Agreeableness	0	0.31	0.37	0.38	0.44	1
PI Neuroticism	0	0.46	0.5	0.5	0.54	1
Yarkoni Openness	0	0.64	0.68	0.68	0.73	1
Yarkoni Conscientiousness	0	0.47	0.52	0.52	0.57	1
Yarkoni Extraversion	0	0.38	0.4	0.4	0.42	1
Yarkoni Agreeableness	0	0.49	0.53	0.53	0.56	1
Yarkoni Neuroticism	0	0.41	0.46	0.46	0.51	1
Golbeck Openness	0	0.68	0.7	0.7	0.72	1
Golbeck Conscientiousness	0	0.74	0.75	0.75	0.77	1
Golbeck Extraversion	0	0.58	0.6	0.6	0.62	1
Golbeck Agreeableness	0	0.71	0.74	0.74	0.78	1
Golbeck Neuroticism	0	0.05	0.11	0.12	0.17	1

In Figure 6.1, the histograms illustrate the distributions of scores with bins of size 0.1. From this illustration, we can see some disagreement between the methods.
On openness, all three methods seem to agree more compared to other traits. For conscientiousness, PI and Yarkoni show a reasonably similar distribution but a much more different distribution than Golbeck. For extraversion, Yarkoni and Golbeck both put most people in the same bins, although on a different scale. PI, however, seems to classify people more evenly over five bins. For agreeableness, Golbeck and Yarkoni again seem to place people on a relatively similar distribution but with a different scale. PI and Yarkoni neuroticism appear to have a distribution reasonably alike while the distribution of Golbeck is more diverse. With an IQR of $0.05 - 0.17$, Golbeck expects almost all people to have a low neurotic score.

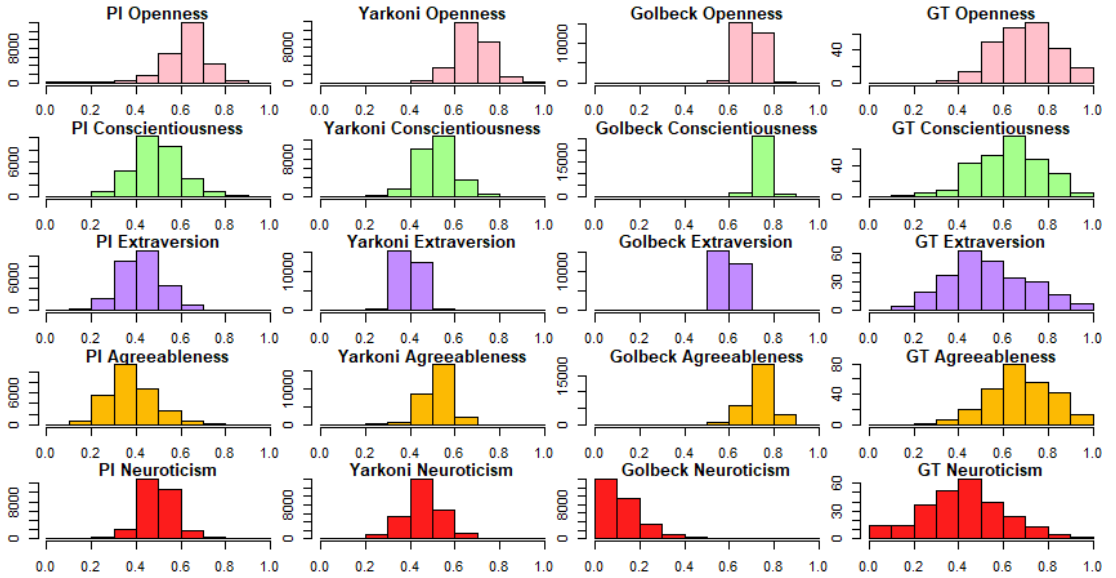


Figure 6.1: Histograms of all inferred scores and the methods used compared to the ground-truth distributions. Each column depicts the Big-Five personality trait, while each row depicts each method used (i.e., PI, Yarkoni, Golbeck, and the ground-truth). The y-axis shows the frequency of people for each score and the x-axis the personality score.

6.2 Comparison against ground-truth

To get a general sense of the accuracy of the models, we compare the inferred scores against the ground-truth. In Figure 6.3, we show the MAE (Mean Absolute Error) scores with and without mean centering[2]. For MAE, the lower the value, the better. The mean-centered scores show the models to be more alike in terms of accuracy than without mean centering. With mean centering, you partly remove the difference in scaling, putting the average MAE values between 0.1 and 0.15. Without mean centering, especially PI agreeableness and Golbeck neuroticism stand out with average MAE values above 0.3.
Interestingly, PI agreeableness without mean centering is one of the most inaccurate predictors when compared to the ground-truth. Yet, with mean-centering, it is one of the most accurate predictors. This indicates some models and traits work on a different scale. Without mean centering, for most trait predictions P holds approximately $P \pm 0.3$ to $P \pm 0.4$ with 95% confidence. This, however, means that making conclusions on an individual's personality profile may be wrong. For example, someone given an extraversion score of 0.8 given by any of the methods, its actual score may very well be 0.5 or even 0.4. The person could be in a group of more extraverted, or the group of averagely extraverted people. PI agreeableness shows to be predicted within around ± 0.6 from the ground-truth and Golbeck neuroticism scores even within around ± 0.7, with 95% confidence. With mean-centering, the scale for each trait and each method may be different. To remain consistent on scaling

[2]Mean centering means that the mean of all scores is subtracted from each score.

throughout this study, we continue without mean-centering to keep a scaling between zero and one.

6.3 Transformations

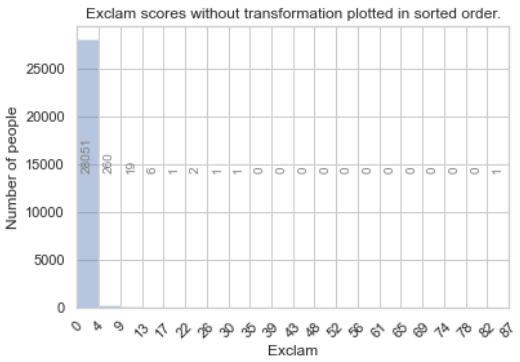
Originally, with the application of Yarkoni and Golbeck, we found the MAE and RMSE scores to be reasonably high for some traits. After an investigation into the problem, we found several outlier scores in LIWC categories. To illustrate the problem of outliers, we show the scores found for `Exclam`[3] with LIWC for the people in our dataset. From this illustration it becomes apparent that some people receive significantly higher values for `Exclam` than others (see Figure 6.2a). As the category scores are multiplied with the correlations found by Yarkoni [124] and Golbeck et al. [47], outlier scores can significantly affect other scores when normalized.

In this study, we apply log-transformations to reduce the effect of outliers (see Figure 6.2b). For example, the unnormalized personality scores for Golbeck neuroticism change from a range of $0.00 - 27.52$ to $0.00 - 1.74$, with medians of 0.17 and 0.19, respectively. As the range becomes much smaller with transformations, the effect of outliers is reduced on normal behavior. In the current implementation of the study, we apply transformations on `Exclam`, `Social`, `hear`, `you`, and `Pronoun`. These five categories showed the most extreme outlier behavior. The effects on MAE (see Figure 6.4) and RMSE show the reduction of outliers improves the scores toward the ground-truth on average. With the five transformations, the MAE for Golbeck agreeableness nearly halved. All scores for LIWC-based methods reported in this study used the mentioned transformations. We recommend future studies to find optimal transformations for all LIWC categories that show outlier behavior.

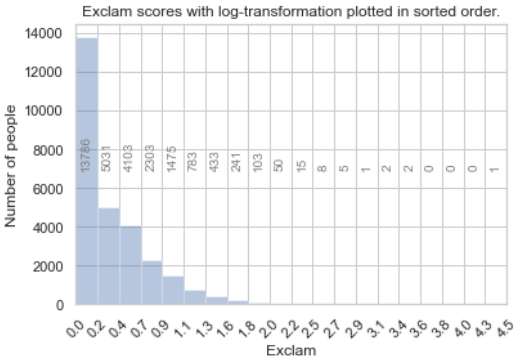
Takeaway

PI and Yarkoni show a more similar distribution of scores than with Golbeck. Golbeck shows higher average scores for all traits, except for neuroticism. However, with mean-centered scores, the performance of all three methods get closer to each other. Indicating the methods may work on different scales.
With the use of transformations on LIWC category scores, we can improve all scores reducing the effect of large outliers on the scaling of all scores when normalized. We recommend future studies to apply transformation methods to all category scores of LIWC showing outlier behavior and to find the most optimal transformation methods.
After the transformations, most traits can be predicted within ± 0.3 to ± 0.4 from the ground-truth with 95% confidence. PI agreeableness and Golbeck neuroticism show scores to be predicted within around ± 0.6 and ± 0.7 from the ground-truth, with 95% confidence. With the fault margin for all traits, one must be careful to take the resulting scores for an individual only as an initial indicator of personality rather than a truth value.

[3] This category score indicates the usage of exclamation marks in messages.



(a) Bar chart of Exclam untransformed values with on the x-axis the Exclam scores and the y-axis the number of people in a bin. Note the high Exclam values at the very right of the plot.



(b) Bar chart of Exclam log-transformed values with on the x-axis the Exclam scores and the y-axis the number of people in a bin. Note the lower Exclam values on the x-axis compared to Figure 6.2a.

Figure 6.2: Exclam scores with and without log-transformation applied. The numbers in gray in the middle of the plot indicate the number of people in a bin.

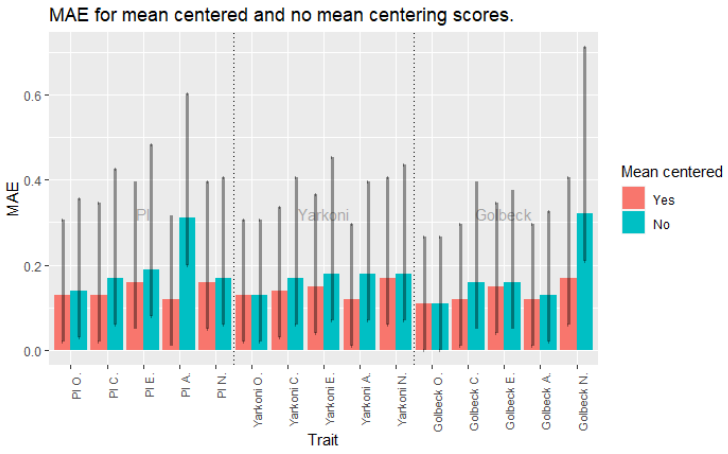


Figure 6.3: MAE scores for all inference methods and personality traits. On the x-axis, the personality inference methods and all Big-Five traits. On the y-axis, the MAE score found when compared to the ground-truth. The graph shows that after mean-centering, the methods obtain accuracy scores closer to each other than without mean-centering. The gray bars indicate the 95% confidence interval. The lower the MAE value, the better the accuracy toward the ground-truth.

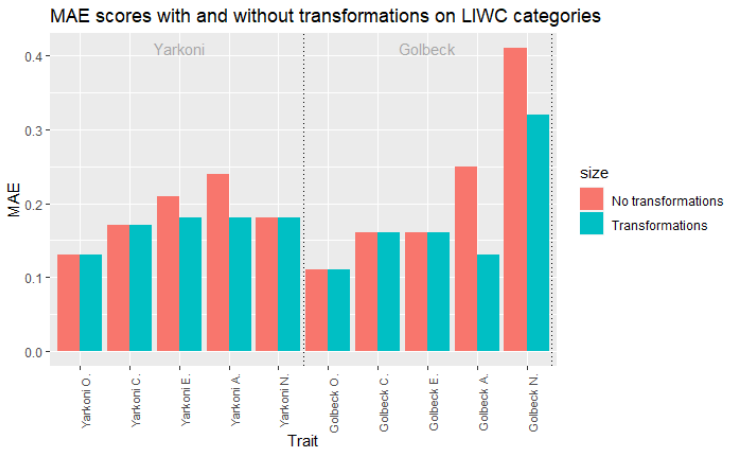


Figure 6.4: MAE values for scores without transformation and with transformation on LIWC categories Exclam, Social, hear, you, and Pronoun. The lower the MAE value, the better the accuracy toward the ground-truth.

Chapter 7

How much does the number of words in messages influence the reliability of personality inference? (RQ.III)

With psycholinguistic tests, the words analyzed influence the resulting personality score. However, the number of words fed into the analysis also influences the outcome. For example, if the model infers personality from only a hundred words, another hundred words of the same person could potentially lead to a different interpretation than for the first hundred words. Adding as many words as possible would mitigate this problem. However, adding more words also means the resources required increase. If, after several words, adding more words does not change the outcome of the psycholinguistic test, any additional terms are a waste of time, energy, and computational power. We, therefore, need some recommendation on the number of words required for the psycholinguistic tests to maximize the accuracy of the models, while minimizing the waste on resources.

In this chapter, we first identify if there are theoretical reasons for a particular minimum or maximum recommended number of words for each method. We then take subsets of comments per user to compare the influence of the number of words on the outcome of the three analyses methods. Finally, where possible, we give recommendations on the number of words to use for each method.

7.1 Number of words in theory

In this study, we make a comparison between three distinct methods. Yarkoni used blog posts to extract LIWC2001 category scores and reported the correlations found with personality [124]. Golbeck used LIWC2007 category scores but based its correlations on Twitter posts [47]. Yarkoni used much longer texts with at least 50,000 words [124], while Golbeck used small Twitter posts with a total concatenation word count ranging from 50 to 5724 [47]. The style of writing and the length of the text may be different for blog posts compared to Twitter posts. The third method, PI, used to be based on LIWC but now uses an open-vocabulary approach [106], meaning that instead of having a fixed correlation for categories, it can learn new words. Earlier studies have shown that such an open-vocabulary

approach works for both larger and smaller text sizes [7]. PI is a general-purpose service. PI is not trained on a particular source of text but is trained on Twitter posts, blog posts, or any other text provided to the service.

As the methods trained on different text sizes, we also expect them to perform differently on different text sizes. For both PI and Yarkoni, we expect it to perform better with as much text as possible. However, at some point, the improvement in performance flattens after round 3000 words for PI [25] as well as for Yarkoni [124]. Golbeck [47] does not report such an upper bound.

Opposed to the other two methods, PI has an upper bound; PI truncates the input to the service to a maximum size of 250KB (JSON or HTML formatting not included) [25]. If we consider ASCII encoding and an average word size of 5.1 letters [15], this would mean an upper bound of around 42,000 words[1]. Table 7.1 below shows an overview of the above.

Table 7.1: Overview of the methods used for the answering of RQ.III. The table lists for each method, the technique, text sizes in the number of words used in the study or mentioned by the service, and the type of text used to train/build the model on.

Method	Technique	Text sizes (in words)	Text type
Yarkoni	LIWC2001	>50,000	Blog posts
Golbeck	LIWC2007	50-5724	Tweets
PI	Open-vocabulary	>100 *Minimum for the service to work* >600 *Within 3% best MAE the service can return* >1,200 *Within 2% best MAE the service can return* >3,000 *Maximum precision service can provide* >40,000 *Truncation to around 40,000 words*	Any

As the different models trained on different input sizes, the number of words can influence the effectivity of the analysis. If we only compare Yarkoni [124] and Golbeck [47], Yarkoni trained on blogs with even an average of 115,423 words each while Golbeck trained on Twitter posts. Given that a single Tweet historically could only have at most 140 characters (now this limit is 280 characters) [118], the nature of such messages is different than for blog posts where the number of words is not limited.

7.2 Number of words in practice

To check the effect of size, we take all people from the dataset that have at least 3,000 words, giving us 4,346 people. We then take their comments and create four different categories. The first category, we call `data_3000`, contains all comments. The second category, we call `data_1200`, contains the first 1200 words of all people. The third and fourth categories have the same principle, where we take the first 600 and 100 words and call the datasets `data_600` and `data_100`, respectively. Important to note to this approach is a possible bias

[1] 250KB=250*1024B. ASCII encoding is 8 bits per character. 5.1 letters per word with one space after each word.

toward more expressive people, where less expressive people are less likely to have at least 3,000 words in the dataset. We partially mitigate this problem by taking as many comments of a person as possible, which increases the likelihood of someone having a total of at least 3,000 words. We base the number of words chosen for each subset on the recommended number of words by PI [25].

We are interested in the differences in scores between these datasets and their performance. We first focus on the differences between the datasets. Although the datasets partly contain the same words, they do not have the same volumes. To check for differences in the datasets, we first identify if the datasets follow a normal distribution. Table 7.2 shows an overview of the findings for each personality trait, method, and dataset.

To check if there are differences between the datasets, we tested for differences in means. In Tables 7.3, 7.4, and 7.5 the effect sizes for each personality trait and dataset is outlined for PI, Yarkoni, and Golbeck, respectively. The rows represent personality traits. The effect sizes indicated with an asterisk were found with the Student t-test [115] and those without asterisk with the Wilcoxon signed-rank test [123]. Only the cells that report significance in either of the methods show the effect size. Otherwise, it shows a dash.

Table 7.2: Normality for each personality trait, method, and dataset. Each cell indicated with an *'x'*, we find the trait to be normally distributed. The top row indicates for which dataset we check normality. The second row indicates the method tested. The first column indicates the personality trait abbreviated to OCEAN form.

	data_100			**data_600**			**data_1200**			**data_3000**		
	PI	**Yark.**	**Golb.**	**PI**	**Yark.**	**Golb.**	**PI**	**Yark.**	**Golb.**	**PI**	**Yark.**	**Golb.**
O	*x*	*x*	*x*		*x*			*x*			*x*	
C	*x*	*x*	*x*	*x*	*x*	*x*	*x*	*x*	*x*	*x*	*x*	*x*
E	*x*	*x*	*x*	*x*	*x*	*x*	*x*	*x*	*x*	*x*	*x*	*x*
A	*x*	*x*	*x*	*x*		*x*	*x*		*x*	*x*		*x*
N	*x*	*x*			*x*			*x*			*x*	

7.2.1 Comparison with ground-truth

It is generally hard to determine which personality score is the best representation of a person. For this study, we use the BFI [64, 65, 63] to assess a ground-truth. However, a drawback of surveys is the required responses from participants, which is generally low. To our survey, we received 267 responses. For the comparison, we only consider people with more than 3000 words. Only 30 people have at least this amount of total words in their comments. However, we may still draw some conclusions on them or at least get some sense of their performances.

Figure 7.1 and 7.2 illustrate the RMSE and MAE values of each personality trait and method, respectively. From the RMSE figure, we can observe that all three methods, for most traits, have at least some outliers where the scores are far from the ground-truth. Unexpectedly, the RMSE and MAE values for `dataset_100` are not always the highest and are sometimes even the lowest of the four datasets. There is no real consistent pattern for each method or trait. Based on these results alone, it is hard to draw any conclusion on the ideal

Table 7.3: Effect sizes of difference in means for PI. Methods indicated with a * are results of the t-test and without are results of the Wilcoxon signed-rank test. We compare the means between each dataset and personality trait. If there is no significant difference, a dash is displayed.

PI					
		data_100	**data_600**	**data_1200**	**data_3000**
data_100	**O**	-	**large**	medium	medium
	C	-	*negligible**	*small**	medium*
	E	-	*negligible**	*negligible**	medium*
	A	-	*small**	*small**	**large***
	N	-	**large**	**large**	**large**
data_600	**O**	**large**	-	medium	*small*
	C	*negligible**	-	*small**	medium*
	E	*negligible**	-	*small**	medium*
	A	*small**	-	*negligible**	medium*
	N	**large**	-	**large**	medium
data_1200	**O**	medium	medium	-	*small*
	C	*small**	*small**	-	*small**
	E	*negligible**	*small**	-	*small**
	A	*small**	*negligible**	-	medium*
	N	**large**	**large**	-	*small*
data_3000	**O**	medium	*small*	*small*	-
	C	medium*	medium*	*small**	-
	E	medium*	medium*	*small**	-
	A	**large***	medium*	medium*	-
	N	**large**	medium	*small*	-

number of words for each method. However, if we combine the results of Tables 7.3, 7.4, and 7.5 with Figures 7.1 and 7.2, we could observe some interesting hypotheses. To make a comparison between the number of words required for a method, we compare three dataset pairs: (data_100,data_600), (data_600,data_1200), and (data_1200,data_3000).

7.2.2 Personality Insights

For PI, we could observe openness, conscientiousness, and extraversion to show no more difference when adding more than 600 words. Similarly, agreeableness does not change much between data_600 and data_1200 but does become worse in RMSE and MAE for data_3000. Neuroticism shows no large differences after 1200 words. For PI, picking the number of words between 600 and 1200 seems logical when combined with the guidelines given by IBM (cf. Table 7.1).

7.2.3 Yarkoni

For Yarkoni, openness does not seem to be influenced much by the number of words. All differences are either non-significant or only negligible in effect. Conscientiousness, on the other hand, is inconsistent in RMSE and MAE. Here data_100 receives much better RMSE and MAE values. Extraversion shows only negligible differences after 1200 words, whereas

Table 7.4: Effect sizes of difference in means for Yarkoni. Methods indicated with a * are results of the t-test and without are results of the Wilcoxon signed-rank test. We compare the means between each dataset and personality trait. If there is no significant difference, a dash is displayed.

Yarkoni					
		data_100	**data_600**	**data_1200**	**data_3000**
data_100	**O**	-	*negligible**	*negligible**	*negligible**
	C	-	**large***	**large***	**large***
	E	-	**large***	**large***	**large***
	A	-	*small*	*small*	**large**
	N	-	**large***	**large***	**large***
data_600	**O**	*negligible**	-	-	*negligible**
	C	**large***	-	medium*	*small**
	E	**large***	-	medium*	**large***
	A	*small*	-	**large**	**large**
	N	**large***	-	*negligible**	*small**
data_1200	**O**	*negligible**	-	-	*negligible**
	C	**large***	medium*	-	**large***
	E	**large***	medium*	-	*negligible**
	A	*small*	**large**	-	**large**
	N	**large***	*negligible**	-	medium*
data_3000	**O**	*negligible**	*negligible**	negligible*	-
	C	**large***	*small**	**large***	-
	E	**large***	**large***	*negligible**	-
	A	**large**	**large**	**large**	-
	N	**large***	*small**	medium*	-

the RMSE and MAE do not seem to change after 600 words. Agreeableness has much more change in RMSE and MAE and shows large differences between `data_600` and `data_1200` and between `data_1200` and `data_3000`. For neuroticism, the differences between 100-600 are large, but 600-1200 are only negligible. From the RMSE and MAE values, adding more words does not necessarily improve results. Yarkoni shows quite inconsistent results between the traits, but the results do shed some light on a possibly desired recommendation of around 600 to 1200 words.

7.2.4 Golbeck

For Golbeck openness, the effect of adding more words decreases after 600 words. From the RMSE and MAE values, adding more words than 600 does not seem to improve anymore. Conscientiousness shows only negligible differences after 1200 words for which the RMSE and MAE seem at least sub-optimal. Extraversion shows large differences between 100-600 and becomes small and medium after. The RMSE and MAE are reasonably stable, suggesting a possible minimum of 600-1200 could be reasonable. Agreeableness does only show negligible change after 600 words. Neuroticism seems to fluctuate more. For each comparison, we can observe a medium to a large difference. With all results above, the results suggest around 600 words per person to be a possible guideline for Golbeck.

Table 7.5: Effect sizes of difference in means for Golbeck. Methods indicated with a * are results of the t-test and without are results of the Wilcoxon signed-rank test. We compare the means between each dataset and personality trait. If there is no significant difference, a dash is displayed.

Golbeck					
		data_100	**data_600**	**data_1200**	**data_3000**
data_100	**O**	-	**large**	**large**	**large**
	C	-	**large***	**large***	**large***
	E	-	**large***	medium*	**large***
	A	-	**large***	**large***	**large***
	N	-	medium	-	**large**
data_600	**O**	**large**	-	medium	**large**
	C	**large***	-	medium*	*small**
	E	**large***	-	*small**	*negligible**
	A	**large***	-	*negligible**	*small**
	N	medium	-	**large**	*small*
data_1200	**O**	**large**	medium	-	medium
	C	**large***	medium*	-	*small**
	E	medium*	*small**	-	medium*
	A	**large***	*negligible**	-	*negligible**
	N	-	**large**	-	**large**
data_3000	**O**	**large**	**large**	medium	-
	C	**large***	*small**	*small**	-
	E	**large***	*negligible*	medium*	-
	A	**large***	*small**	*negligible**	-
	N	**large**	*small*	**large**	-

7.2.5 Recommendation

To make optimal use of the psycholinguistic models, we would like to establish a recommendation on the lower and upper bound of the number of words to analyze per person. We do know for PI that the bare minimum of words is 100 for the service to give at least a result. For LIWC, we can reasonably expect a similar lower bound to exist for significant results. A person with only 20 words can have an increase/decrease in an LIWC categorical score of almost 0.04 (4%) with the addition of a single word. In contrast, a person with 1000 words has a maximum increase/decrease in the category score of around 0.001 (0.1%). How much this contributes to the final personality score, however, depends on the full population and correlations of each method.

From this reasoning, a minimum of a hundred words used for all methods in this study is a reasonable minimum. Taking around 600 words shows to be a promising lower bound for all methods. Important to note, adding more words to the analyses does, in most cases, decrease deviation and improve confidence (see Table C.1 in Appendix C.1). Therefore, adding more words does increase the confidence in scores found. For PI, one should also keep in mind the recommended minima for PI (i.e., a minimum of 600 words to a maximum

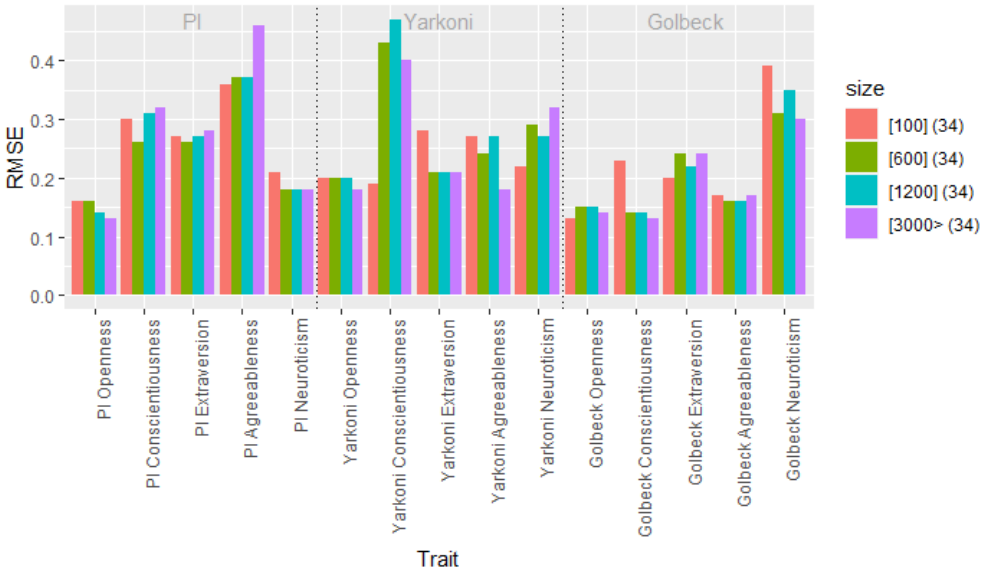


Figure 7.1: RMSE scores between the inferred personality scores and ground-truth scores. The figure shows on the x-axis the personality trait and method and on the y-axis the RMSE value found. The lower the RMSE value the better. The bars indicate, from left to right, the RMSE scores for data_100, data_600, data_1200, and data_3000.

of around 3000 words) [25]. Similarly, for Yarkoni, we recommend an upper bound of 3000 words, based on the findings of the Yarkoni [124]. Golbeck does not provide such a maximum, but it is unlikely for Golbeck to improve after 3000 words.

Takeaway

For all three methods, we strongly recommend a bare minimum of a hundred words. One should keep in mind that the personality inferred is otherwise likely unreliable. In the case of PI, the service will even deny the profile request.
A more precise recommendation on the number of words for PI and Yarkoni seems to lie around 600-1200 words and for Golbeck around 600. Earlier studies have shown, for PI and Yarkoni, the improvement in terms of MAE seems to reduce after 3000 words. Golbeck does not provide such a maximum, but it is unlikely for Golbeck to improve after 3000 words.

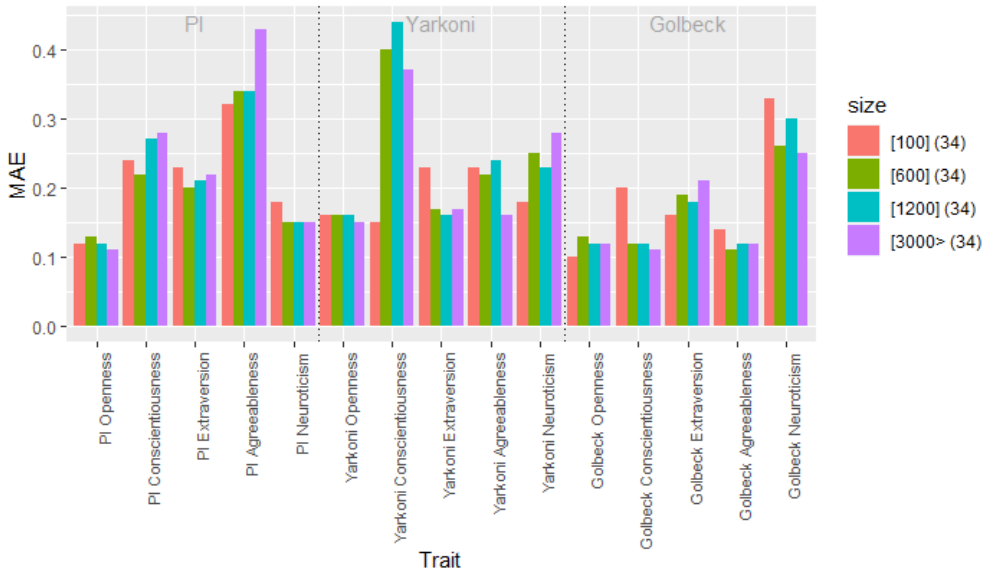


Figure 7.2: MAE scores between the inferred personality scores and ground-truth scores. The figure shows on the x-axis the personality trait and method and on the y-axis the MAE value found. The lower the MAE value the better. The bars indicate, from left to right, the MAE scores for data_100, data_600, data_1200, and data_3000.

Chapter 8

Does the English proficiency of a person impact the quality of personality inference? (RQ.IV)

The words we use influence the outcome of the personality inference tests. Here we assume our personality changes the words we choose. However, the language we use or even the languages we know influence the words we pick [73, 76]. In this study, we focus on the use of English. In the dataset, we have people from English speaking countries, as well as non-English speaking countries. For the latter group, English is likely their second language (or even third or fourth). It is well established that word recognition in the second language can be affected by the native language [73]. Hypothetically, the outcome of the personality inference methods could be influenced depending on whether someone is a native writer or a non-native writer. For example, people with English as their second language show to have different frequency-related aspects[1] compared to monolinguals [73]. In the case of the LIWC methods, a different frequency on certain words leads to different category scores and, therefore, different personality scores. As to make sure the psycholinguistic tests are useful for everyone, we investigate if the models do not discriminate on the English proficiency of a group.

In this chapter, we first outline three essential definitions for the comparison of proficient English writing against non-proficient writings. To compare the two groups of people, we then describe how we used the selection process for participants of the ground-truth to reduce biases. Finally, we investigate the possible influence of proficiency on the outcome of the three methods used.

8.1 Definitions

Before we can compare the use of language and its possible effect on the outcome of linguistic analysis, there are multiple definitions to consider:

[1]The term 'frequency-related aspects' refers to the frequency of particular words appearing in a sentence or story.

Def. 1: Bilingual

A bilingual is someone who uses two or more languages (or dialects) in their everyday lives [51].

This above definition is regardless of where, when, and how they learned and use those languages [70].

Def. 2: Native speaker

According to the definition of Lee [71] someone is a native speaker if he adheres to the following six features:

1. The individual acquired the language in early childhood [38, 80] and maintains the use of the language [80].
2. The individual has intuitive knowledge of the language [38, 113].
3. The individual can produce fluent and spontaneous discourse [38].
4. The individual is communicatively competent in the language [38] and can communicate within different social settings [113].
5. The individual identifies with or is identified by a language community [38].
6. The individual does not have a foreign accent using the language [32].

A native language is a language in which someone is a native speaker. With the above definition of a native speaker, someone growing up in a bilingual home could have multiple native languages. Furthermore, for this study, we focus on writing instead of speaking, which leaves some features unacquainted.

Def. 3: Mother tongue (or mother language)

The term mother tongue is often referred to as the child's native or first acquired language [83].

With the above definition of mother tongue, only the first learned language can be the mother tongue language. However, if someone grows up in a bilingual home, learning multiple languages at the same time, it is sometimes hard to define the 'first' language learned. In this case, we assume the person to have multiple mother tongue languages.

In our dataset, we only consider English words. Given the definition of bilingualism, everyone in our dataset with enough[2] English words may be bilingual if English is not their mother tongue or native language. Many people in our dataset are likely bilingual, as more than half of the world population is believed to be bilingual [6]. It is, however, hard to determine whether a language is used in everyday lives and sometimes even more challenging to automatically determine if the language used is the only language known by

[2]People with more than 100 words in all their comments after all preprocessing.

the person. One way to determine if someone uses English in everyday life would be to analyze all comments made over time by the users. However, this would still not create a clear picture of ordinary lives as the primary language used on GitHub is English.
Besides knowing a language, there might also be a difference in proficiency in reading, writing, hearing, and speaking English. For this study, we focus on writing only, but making this distinction may not always be straightforward.

8.2 Identification of participants

In our study, we approached people for a questionnaire to obtain a ground-truth. To mitigate the chance that the set of people is over-represented by a group of either native or non-native English people, we first assumed their English proficiency and randomly sampled in groups almost equal sizes for each classification. In this section, we describe how we assumed a preliminary English proficiency, followed by a description of the selection of participants for the questionnaire.

8.2.1 English Proficiency Index

For the classification of English proficiency, we made use of the English Proficiency Index (EPI) created by EF[3]. EF created a ranking of non-English speaking countries on their English proficiency. The proficiency scores of the EPI all originate from the English courses provided by EF. Averages for the countries were weighed against the populations of each country to obtain an index score. This approach does assume that the test takers are from a uniform distribution from the whole country. However, people curious about learning English might obtain different scores than those who are not interested in an English course. Furthermore, countries with low usage of the Internet may be underrepresented and biased toward the rich or the more educated community.

On GitHub, people can choose to share their location on their profile. Based on this location, we could assign their EPI classification. People are free to fill in the country, city, street, but also their continent or even words like ‘world’ and ‘the Internet’ are commonly found occurrences. We could only assign a classification to people with a country contained in the EPI.

Figure 8.1 shows a rough location of the GitHub developers in our dataset on the world map. From the world map, it becomes apparent that the classes `Very low` and `Low` are quite underrepresented compared to the other three classes. Given that the people we find work on GitHub, where the primary language of communication is English, we suspect people in the underrepresented classes of being more proficient in English than the EPI suggests. EPI bases its classification on a sample that should represent the majority population of a country. Our participants may only represent the group of people that work in English, which is for `Very low` and `Low` countries unlikely the same group of people. Anyhow, we can use this classification to get a first impression of the background of people and their possible English proficiency.

[3] `https://www.ef.com/wwen/epi`

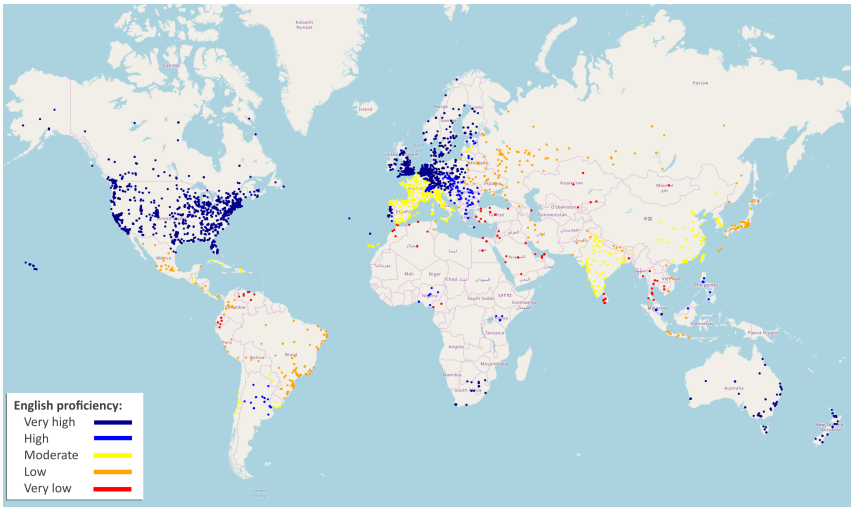


Figure 8.1: World map with locations of GitHub users and their corresponding EF English Proficiency classification. The colors on the map indicate the English Proficiency as indicated in the legend. Please note many points overlap and, due to privacy concerns, the points are not the exact locations.

8.2.2 Questionnaire ground-truth

Instead of relying on the assumption of the English proficiency of participants, we ask the volunteers to classify their English knowledge. We ask if they believe in having English as their mother tongue and if they believe to be fluent in English writing. We choose against the full definition of 'native English' and instead rely on the sub-definition 'fluency,' as the full description of 'native English' may be too strict and is, in addition to that, more prone to misinterpretation by participants. We did inform the users about the given definitions of mother tongue and fluency. Yet, people could have a different interpretation of the meanings.

Furthermore, the first point of contact with the participants is in English. Therefore, we already assume they know enough English to communicate with us. Possibly, the people who react are always at least one of the three definitions used; the person is at least a bilingual, a native English speaker, or someone with English as their mother tongue. An advantage of this approach is that we do not need to assume the English proficiency of the persons, which would otherwise ignore the individual's background (e.g., education and culture). For the classification, we ask the participants to indicate their fluency in writing specifically. The options available were: 'Yes', 'No', and 'Maybe'.

8.3 Results

In this section, we compare the scores of the different models and look for a possible influence of self-reported fluency and mother tongue on the outcome of the analyses.

8.3.1 Fluency

For the classification of fluency, we initially divide the responses into three different groups: Yes, No, and Maybe. However, due to inconsistent results found for the Maybe-group, we decided to remove this group from the analysis. In Appendix C.2.1, the results of the Maybe-group can still be found. In this classification, we assume the Yes group to be fluent in English writing, and the No group non-fluent.

From the two groups, we could observe the differences in means not to be consistent for one classification specifically. All traits found to be significantly different are reported in Tables 8.1 and 8.2. PI agreeableness shows a significant difference in means, but the effect size is only small. Yarkoni neuroticism shows a medium and Golbeck openness a large significant difference in means. Yarkoni neuroticism shows significantly lower scores for the Yes-group ($M = 0.47$) compared to the No-group ($M = 0.53$). For Golbeck openness, the Yes-group shows significantly higher scores ($M = 0.7$) compared to the No-group ($M = 0.67$).

Table 8.1: All traits showing significant difference in means using Wilcoxon summed rank test based on fluency. The first column indicates the personality trait and method checked. Then from left to right: the V-statistic of the Wilcoxon summed rank test, the p-value, a boolean indicating significance, the r effect size, and the corresponding effect size interpretation.

Column	V	p.value	p<0.05	r	effect
PI A.	1259	0.03	TRUE	-0.15	small

Table 8.2: All traits showing significant difference in means using the unpaired t-test based on fluency. The first column indicates the personality trait and method checked. Then from left to right: the V-statistic of the t-test, the p-value, a boolean indicating significance, the Cohen's d effect size, and the corresponding effect size interpretation.

Trait	V	p.value	p<0.05	d	effect
Yarkoni N.	-2.57	0.02	TRUE	-0.73	medium
Golbeck O.	4.23	0	TRUE	1.09	large

Using the chi-squared contingency table test (see Table 8.3), we could observe a significant association between the fluency and the personality scores of PI openness, extraversion, and agreeableness, Yarkoni openness, and Golbeck openness. Furthermore, irrelevant of the method used, English fluency influences openness scores. For all these traits found significant, the effect size is medium.

Finally, to get an impression of the influence of fluency on the outcome of the personality methods, we compare the scores against the questionnaire ground-truth (see Figure 8.2). The results of the PI and Yarkoni analyses seem to be worse for the Yes-group in most cases, where the difference in MAE is at most 0.1. This could potentially mean, the inferred scores for fluent people are, on average, less accurate compared to non-fluent people. On the contrary, Golbeck shows worse results for non-fluent people compared to fluent

Table 8.3: Chi-squared contingency table test with all significant associations between fluency and personality scores. The first column indicates the method and trait tested. Then from left to right: the chi-squared value, the p-value, a boolean indicating significance, the Cramér's V effect size, and Cohen's interpretation Cramér's V effect size with two degrees of freedom [26, 27].

Trait	X2	p.value	p<0.05	Cramér's V	effect
PI Openness	20.51	0.01	TRUE	0.29	medium
PI Extraversion	17.48	0.01	TRUE	0.27	medium
PI Agreeableness	24.16	0	TRUE	0.32	medium
Yarkoni Openness	16.09	0	TRUE	0.26	medium
Golbeck Openness	15.81	0	TRUE	0.26	medium

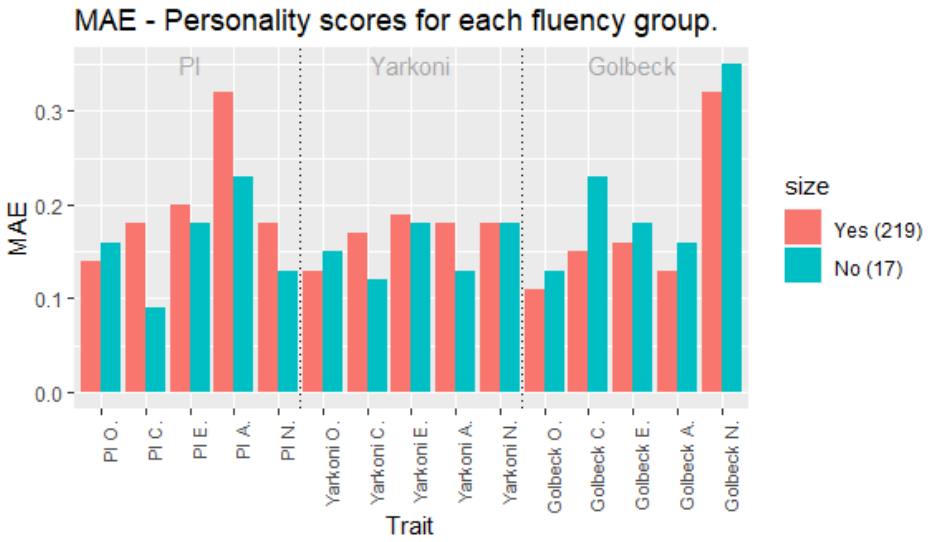


Figure 8.2: Influence of Fluency on the MAE scores when compared to the ground-truth. We used random oversampling on the No group to reduce the potential bias toward the over-represented Yes group. On the x-axis, we show the personality traits and methods. On the y-axis, the found MAE values. The lower the MAE value, the better the performance compared to the ground-truth. 'Yes' (219) depicts the 219 people in the Yes-group for fluency and 'No' (219) the 219 in the No-group.

people for all traits in terms of RMSE and MAE. For the traits where we find significant differences in means, the differences in MAE for Yarkoni neuroticism ($\Delta = 0.00$) and Golbeck openness ($\Delta = 0.02$) is small. For PI agreeableness, the difference is more substantial ($\Delta = 0.10$), meaning the prediction capabilities of PI for fluent people could be worse than for non-fluent people.

From Table 8.3, openness seems to be influenced by fluency for all methods. However, the MAE does not show large differences. Other more notable differences are for PI agreeableness, where the chi-squared test shows a medium association with fluency, and the MAE shows a higher MAE score for the Yes-group compared to the No-group. On average, fluent people receive slightly worse scores for PI agreeableness. We could not find the same

influence on agreeableness for the other methods, indicating PI may discriminate slightly for fluent people.

PI conscientiousness shows no significant difference in means or significant associations with fluency, albeit showing a higher MAE value for the Yes-group compared to the No-group. Yarkoni neuroticism does show a large difference in means for the Yes-group compared to the No-group, although there does not seem to be a large difference in MAE scores.

In summary, all methods, for at least some of the traits, are influenced by fluency. However, from the above results, the associations with fluency seem to be only small for all but PI agreeableness. We find a significantly medium difference in means for Yarkoni neuroticism and large for Golbeck openness. For PI agreeableness, the effect size is only small. The most considerable difference ($\Delta = 0.1$) in MAE can be found for PI agreeableness, where fluent people receive slightly worse predictions.

8.3.2 Mother tongue

Similar to fluency, we compare the influence of mother tongue on the resulting scores. If we compare the Yes- and No-group, we observe only for Golbeck openness a significant medium difference in the means (see Table 8.4), where the Yes-group has a significantly higher mean ($M = 0.71$) than the No-group ($M = 0.7$). All other traits did not show any significance. Furthermore, we could only find a significantly small association between Golbeck openness and mother tongue using the chi-squared test.

Table 8.4: All traits showing significant difference in means using the unpaired t-test based on mother tongue. The first column indicates the personality trait and method checked. Then from left to right: the V-statistic of the t-test, the p-value, a boolean indicating significance, the Cohen's d effect size, and the corresponding effect size interpretation.

Trait	**V**	**p.value**	**p<0.05**	**d**	**effect**
Golbeck O.	3.5	0	TRUE	0.43	medium

Table 8.5: Chi-squared contingency table test with all significant associations between mother tongue and personality scores. The first column indicates the method and trait tested. Then from left to right: the chi-squared value, the p-value, a boolean indicating significance, the Cramér's V effect size, and Cohen's interpretation Cramér's V effect size with two degrees of freedom [26, 27].

Trait	**X2**	**p.value**	**p<0.05**	**Cramér's V**	**effect**
Golbeck Openness	10.09	0.01	TRUE	0.2	small

When we compare Golbeck openness to the ground-truth, the impact of the mother tongue is reasonably small. The Yes-group shows an RMSE of 0.15 and MAE of 0.13, compared to an RMSE of 0.13 and MAE of 0.11 for the No-group. This would mean that, on average, people having English as their mother tongue receive slightly worse predictions.

Together with the comparison to the ground-truth, we find no real indication of the influence of the mother tongue for most traits and methods. Only Golbeck openness shows a difference in means, as well as a significant medium association with the mother tongue.

For most traits and methods, we do not find any significant differences between the scores found. Hypothetically, this may suggest that someone less proficient in English can develop a similar writing style to proficient people. Techniques exist to recognize non-native English text writings [23, 72], so differences may still be present in text. The order of words and combinations of words often found close together, for example, do not influence the outcome of the personality inference tests of this study but do influence the outcome of native classification [23].

8.3.3 Openness discussion

With both the analysis for fluency and mother tongue, we often find an influence on openness. For fluency, all three methods showed a lower average openness score for the No-group compared to the Yes-group. It is, however, unknown if fluency merely causes this influence or if it may also be caused by the geographical location [103] or cultural background [97]. If we compare the mean scores for fluent and non-fluent people in the ground-truth dataset, we notice a similar observation (see Table 8.6). 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 the differences found for openness are introduced by the population, not so much by the methods used.

Table 8.6: Mean openness scores for each method and the ground-truth for fluent and non-fluent people and people with or without English as their mother tongue. The columns from left to right: the method used to obtain the openness score, mean openness scores for fluent people, mean openness scores for non-fluent people, and two columns with mean openness scores for people with and without English as their mother tongue.

Method	Fluent	Non-fluent	Mother tongue	No mother tongue
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

Possibly this difference is introduced by demographics or culture. For example, earlier studies found lower openness scores for people in East-Asia than for people in Europe [66, 103]. Mak and Tran [78] found for Asians with a high English proficiency (in terms of fluency) a positive correlation with openness values. In certain cultures, adjectives, generally used as a major support in taxonomic approaches, does not cover the domain sufficiently to allow for identification [97]. In our ground-truth, however, we do not observe a significant difference in means between Europe ($M = 0.70$) and Asia ($M = 0.69$), and America ($M = 0.73$) and Asia nor do we find any significant differences in the inferred scores of the

methods between continents. We do, however, not have information about the culture and demographics of our participants to prove their influence.

8.4 Possible threats to validity

The method used to compare the English proficiency of the participants does not come without threats to validity. The preliminary assumption of people's English ability using EPI for the selection of participants does not come without risk. As people are free to leave out their locations, there might be a bias toward a particular personality type. People who are more willing to express themselves might also be more likely to share their location. Although Wang and Stefanone [121] did not find a significant indication of higher extraversion or narcissism among Facebook users willing to share their location, other studies did find a correlation with higher neuroticism and the willingness to share accurate information on their Facebook profile [4, 100]. If the ground-truth scores found, however, are a good representation of the personality of the participants, the studies [4, 100] partly contradict with the ratings found for our study. We would otherwise expect an over-representation of higher neurotic in our dataset. On the contrary, the neuroticism we find ranges from zero to one, with a mean of 0.42 and a median of 0.41. As the mean values are below 0.5, on average, we find reasonably low neurotic scores.

Another possible bias introduced with the location provided through GitHub is the difference between the current position of the individual and the country the person spent most of its youth. In the questionnaire, we ask people in which country they spent most of their youth. For 57 out of 267 people, the country of youth was different than the nation provided through GitHub. Of these 57 people, ten people did not see a change in their EPI category, six people filled in a country that has no EPI score, 18 received a lower EPI rating, and 23 received a higher EPI rating.

Furthermore, the interpretation of the participant is essential and may introduce a bias toward certain personality types. The choice between all three options may be influenced by, e.g., modesty, confidence, or pride. To mitigate this, we instructed people there is no right or wrong answer to the question. Furthermore, people sometimes give misleading answers to provide a favorable representation of themselves [67].

Takeaway

English proficiency does show to influence the outcome of psycholinguistic test methods. All openness scores show to be affected in fluency, while only Golbeck shows differences in openness with the mother tongue. Furthermore, PI extraversion and agreeableness and Yarkoni neuroticism show to be affected by English fluency. According to earlier literature, cultural or demographic differences may cause differences in openness. The differences in MAE toward the ground-truth are reasonably small. Only for PI agreeableness, we observe fluent people receive slightly worse scores than non-fluent people. PI and Yarkoni mostly receive slightly worse scores for fluent people. On the contrary, Golbeck receives slightly worse accuracy scores for non-fluent people.

Chapter 9

Discussion

In this section, we first discuss our results for each research question, eventually contributing to the answering of the main research question. Next, we discuss the possible limitations and threats to the validity of this study. Finally, we discuss the ethical side of automatic personality inference. We discuss the possible misinterpretation of scores, the potential misuse of the methods, and the harmful impact usage could have on an individual.

9.1 Results

In Chapter 4, we defined a ground-truth based on personality scores found with a BFI questionnaire among developers. All Big Five personality traits showed a near-normal distribution for which most personality types are represented. However, people with low openness, conscientiousness, or agreeableness scores are under-represented. Similarly, high scores for neuroticism are under-represented. The scores, however, show to partially comply with earlier work on personality among software engineers [17, 75, 68] and with the use of questionnaires in general [40].

9.1.1 Does data sanitization influence the outcome of personality inference tests? (RQ.I)

For RQ.I, we found twelve out of thirteen preprocessing steps to improve the outcome, not to influence the outcome, or at least not worsen the outcome significantly. In case the outcome is not influenced or at least not worsened significantly, another property is improved (e.g., privacy concerns or storage requirements). The only exception is number-parsing. For this step, we decided to disable the step for Yarkoni extraversion and agreeableness. For these traits, the numbers directly correlate with the personality traits [124]. We identified the following steps as necessary: the removal of code blocks, quotes, URLs [7, 104, 101, 2], numbers (except for Yarkoni extraversion and agreeableness), and images. Furthermore, we find recommendations for steps to improve on privacy, the speed performance of succeeding preprocessing steps, or storage needs while not affecting the outcome significantly. We recommend to conform white space and to remove IP-addresses, casing [7, 21, 39, 101], IP-addresses, double (subsequent) white space, and space before punctuation. Finally, we

recommend three methods that may influence the outcome but come with a strong theoretical need. We recommend the removal of @-references [52, 2], hashtags [21, 52, 7], and emails.

For some of the preprocessing steps recommended, the need is introduced by GitHub Markdown (e.g., images and @-references). On other communication platforms for developers, such as email, there may not be a need for such preprocessing steps.

9.1.2 How well do the psycholinguistic models perform when compared to the ground-truth? (RQ.II)

In Chapter 6, we showed how we could use transformations on LIWC category scores to reduce the effect of outliers on the normalization process of the LIWC-based methods. We only applied transformations on the categories identified as most affected by outliers (e.g., `Exclam`). After the transformation of scores, we found the distributions of inferred scores of PI and Yarkoni to be more similar than to the distribution of scores of Golbeck. Golbeck showed for all traits, except neuroticism, a higher score on average. Golbeck neuroticism showed most people to have a relatively low-lying neurotic score, with a mean of 0.12 and an IQR of $0.05 - 0.17$.
Furthermore, we compared the inferred scores from all three methods to the ground-truth. Overall, all traits (except PI agreeableness and Golbeck neuroticism) can be predicted with an MAE around $|\pm 0.3|$ to $|\pm 0.4|$ with 95% confidence. In other words, the models promise to predict within an error of 0.3 to 0.4 for around 95% of people. For PI agreeableness and Golbeck neuroticism, however, we find MAE values of $|\pm 0.6|$ to even $|\pm 0.7|$ from the ground-truth with 95% confidence. In other words, 5% of our respondents receive an error of 0.6 or higher for PI agreeableness and for Golbeck neuroticism almost 9% of people.

9.1.3 How much does the number of words in messages influence the reliability of personality inference? (RQ.III)

Concerning RQ.III, we suggest for all three models a required minimum of a hundred words. PI even rejects any request otherwise [25]. Neither LIWC [89, 90] nor Yarkoni [124] or Golbeck et al. [47, 48] state such a minimum required number of words. LIWC does not reject any processes below a hundred words, but one should keep in mind the effect of a single word on the resulting category score to become more substantial with fewer words.
From our results, we observed 600 to 1200 words to be at least sub-optimal for PI and Yarkoni and around 600 for Golbeck. The results, however, do not state that taking 600 to 1200 words always derives the best results. We did not find evidence for a maximum number of words, but for PI and Yarkoni, after 3000 words, the improvement reduces as the differences in MAE reduce [25, 124].

9.1.4 Does the English proficiency of a person impact the quality of personality inference? (RQ.IV)

Regarding RQ.IV, the analysis demonstrates a potential effect of the English proficiency on the outcome of the psycholinguistic tests. To investigate the influence of English proficiency, we examined the differences in scores possibly caused by fluency in English and having English as a mother tongue. All methods show a difference in openness for fluent people. Only Golbeck shows a difference for the mother tongue. The differences, however, are also visible in the ground-truth (cf. Table 8.6). According to earlier literature, cultural or demographic differences may cause the contrast in openness. We did, however, not obtain enough information about culture to prove this statement.
PI and Yarkoni mostly receive slightly worse scores for fluent people. On the contrary, Golbeck receives slightly worse accuracy scores for non-fluent people. For PI agreeableness, we find indications of possible discrimination toward fluent people, whereas the models show to perform better on non-fluent people. However, these results might be biased due to an under-representation of non-fluent people. With the use of random oversampling on the No group, we tried to mitigate this potential issue.

9.1.5 How useful are psycholinguistic tests to infer developer personality from SE data?

Finally, the results of the aforementioned research question culminate in answering the main research question. To make sure the psycholinguistic test methods are as useful as possible, we have shown transformations and preprocessing steps to improve the outcome of the inference methods. Furthermore, we demonstrated a recommended number of words, for which the models prove to be most effective, making them more useful. Finally, for psycholinguistic models to be deemed useful, they should not discriminate against people, but instead, work for everyone.

We find, for most traits, no firm indication of discrimination toward people more proficient or less proficient in English for any of the three methods. The only exception to this seems to be PI agreeableness, where fluent people appear to receive slightly worse predictions than non-fluent people.

With all enhancing steps we propose, the methods promise to give, on average, a personality score with a fault margin less than 0.2 for most traits. However, would we consider a 95% confidence interval, fault margins as large as 0.3 to even 0.7 for individuals are possible. 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. 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.

9.2 Threats to validity

In this section, we explain which limitations and threats may affect the validity of our work. For each limitation or threat, we elaborate on how we mitigated it, why it was left in, or we give a direction on how to implement the mitigation.

9.2.1 Limitations

In our study, we find some possible limitations, which we describe in the following section.

Implementation of Golbeck

The implementation of Golbeck is not a perfect replica of the study of Golbeck et al. [47]. In their application, the authors suggest three additional non-LIWC features, specifically for Tweets.

1) Hashtags in the context of GitHub are references to issues, while hashtags in Tweets are referrals to topics of interest. As mitigation, one could only count any hashtags in GitHub comments followed by letters, leaving out all issue references.
2) The number of words per Tweet is limited by a maximum of 280 characters per Tweet [118]. In our dataset, the most considerable comment is 203,381 characters long (after preprocessing). As we multiply the word counts the fixed correlation value of Golbeck [47], the outcome for this feature can become relatively high compared to other features found for Golbeck. Normalization or transformation on GitHub comment lengths between 1 and 280 could be a possible solution. However, the resulting numbers should follow a similar distribution to the number of words in the comments of Twitter to mimic the values found by Golbeck et al. [47].
3) Finally, the links per tweet did not seem to improve the outcome. It is likely the URLs are used with different intentions on GitHub than on Twitter.

Hypothetically, leaving out these three features could negatively impact the scores of openness (for the number of hashtags), extraversion (for the words per Tweet), and conscientiousness (for the links per Tweet). However, according to our findings for *RQ.I*, the removal of URLs and hashtags seemingly improves the outcome of the Golbeck method.

Personality influence on population

Particularly for the approach used in this study, we rely on GitHub users that participated in discussions. Individuals with higher levels of extraversion and neuroticism are more likely to share their experiences online [87]. This, however, contradicts with the values we find for neuroticism in our ground-truth, where our ground-truth shows and under-representation of highly neurotic people. For a questionnaire, due to its voluntary basis, the responses may be biased toward more open or agreeable groups of people [124]. We may miss upon personality groups less inclined to volunteer on giving away personal information. Furthermore, we may miss out on programmers who tend to discuss more in person or leave out of

most, or even all discussion. Models used for the inference of personality may, therefore, introduce bias due to the nature of personality types. In our ground-truth, we find under-representations for lower openness, conscientiousness, and agreeableness, and high neurotic values. Without a representation for all personality types, a bias toward an over-represented group may invalidate the results for an under-represented (or even absent) personality type. We tried to mitigate this threat by inviting as many people as possible. Furthermore, we made our intentions and approach clear to remain transparent. However, even with larger population sizes and clear intentions, we may still not be able to rule out the limitation entirely.

Interpretation of context

The context in which psycholinguistic methods operate very much influences the outcome of the analyses. All three ways may have problems with words that have different meanings in the context of software engineering than in any other context. Examples of such terms are 'cookie,' 'cookies,' and 'CoffeeScript.' The LIWC dictionary does not take meaning into account [59] and, therefore, assigns these three words to the biologically related words category `bio`. In the context of SE, the first two are more likely to be related to website cookies rather than an edible cookie. The last is a computer language instead of a drink.
Similarly, we found occurrences of the word 'Dead' and 'Kill' to be assigned to `death`, `negemo`, and `anger`, while they related to a discussion of game-related features of Minecraft. In the case of Yarkoni, this led to higher neurotic values. We could not prevent all contextual misinterpretations with the current approach of the psycholinguistic models. However, with the preprocessing steps, we at least mitigate the risk of misinterpretation of images, email addresses and IP-addresses.

9.2.2 Internal validity

Personality change/maturation

Personality may not always remain the same over a long period. Rantanen et al. [94] found that openness, agreeableness, and conscientiousness remain stable from childhood through adulthood. They also found neuroticism and extraversion to remain more stable for men than for women [94]. However, other studies found contradicting results. Rastogi and Nagappan [95] found developers to evolve as more conscientious and extraverted and less agreeable over two consecutive years. Calefato et al. [17] found developers to become more open, agreeable, and neurotic. In our study, we did not take time into account but considered all comments available for a person. Meaning the most recent comments may be a better representation of the current personality than all older comments. As the questionnaire is our most recent caption of personality, the oldest comments may negatively influence the performance of the models. In Chapter 7, we only considered the first 100, 600, and 1200 words in the comments of a person. If personality changed for people over time, we should have used the last words to come closer to the personality scores found at the time of the questionnaire.

9.2.3 External validity

Population Validity

The SE data chosen for this study is limited to GitHub comments. Other studies on personality among software engineers relied on emails [17, 86], comments on code collaboration platforms [75], and posts on StackOverflow [13]. The GitHub comments may not be representative of all SE data due to, for example, repository guidelines and Markdown markup compared to other sources. Furthermore, as our study merely focuses on English, the approach used may not work for other languages. Although both LIWC and PI support more languages, they might not be as effective in the context of SE.

As GitHub is one of the most used platforms for code collaboration, we likely find a wide variety of personalities among the population. The larger the sample set of people, the more likely it is to generalize to the whole public. To capture an as comprehensive as possible collection of personalities, we analyzed the scores of 2,050 developers. Furthermore, we made use of a questionnaire to obtain a ground-truth for the personality scores. On this questionnaire, we received 267 responses of developers. If we compare the number of answers to earlier studies on personality, we find studies with a similar number of reactions, such as Golbeck et al. [47] with 279 responses and Carducci et al. [21] with 250 responses. Sodiya et al. [110], on the other hand, did a questionnaire with a more extensive set of people with 58 students in the first round and 1002 software engineers in the second round. Even so, Yarkoni [124] used a broader set of people; the study received 694 responses.

9.3 The Ethics of Personality Inference

Measuring personality has long been around and often there has been a concern regarding its ethical side (e.g., Messick (1965) [82]). The question about its ethics remains as of today. Although the concept has long been around, often studies forget to mention the ethics or potential negatives involved with the use of the systems they propose (e.g., [124, 47, 2, 17, 75]).

Although there are multiple use cases for personality measurement and automatic inference, one should always keep in mind the potential harm it may have on the individual. This potential harm comes in a variety of disguises: misinterpretation of information, incorrect scores, misuse of information, discrimination toward personality profiles, and probably many more implications exist. The act of personality measure is believed sometimes to dictate permanent status [82]. Hence, it undermines self-esteem and limits motivation, could decrease the diversity of talent, and the tests foster impersonal and mechanistic evaluations and decisions and the expense of individual freedom of choice of the tested [82].

To decide whether this study is ethical and how the results should or should not be interpreted and used, we identify some of the potential threats of unethical use of personality inference and the potential impact it may have. At the end of this section, we outline a recommendation on how or how not to use the results.

9.3.1 Misinterpretation of data

Even if the model is right to the last digit, what does a score X compared to a score Y mean? The scores in Figure 9.1 show Mike and Elise both have different scores but also equal scores for their traits. An openness score of 0.7 is in the 70th percentile of openness of the entire set of people. In other words, the person shows a higher Openness than more than 69% of the people. However, if we look at the example, Mike and Elise show a similar openness score. This same score does not mean they are perfectly equal, nor does it mean that they have used the same words. Mike and Elise may have identical scores due to rounding, but they are two different individuals who behave, think, and feel differently.

What does it mean to be different?

In addition to the similarities, Mike and Elise also show different scores. For example, Mike shows a conscientiousness score of 0.5 and Elise 0.6. Based on these numbers, we could conclude Elise to be more conscientious than Mike. However, this does not mean Elise is more conscientious in all scenarios, nor does it indicate how much more conscientious Elise acts. All we could conclude from these numbers is that Elise is likely more conscientious than Mike the majority of the time. Depending on the persons, this difference is noticeable in real-life examples.

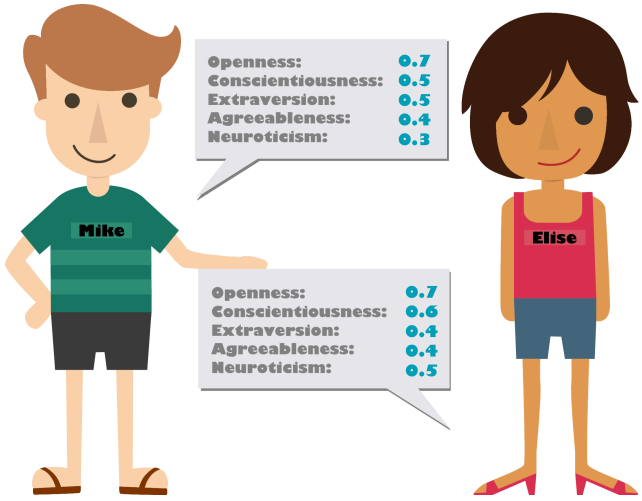


Figure 9.1: Two people, Mike and Elise, with different personality trait scores. What does this difference in personality scores mean? The cartoons are altered versions of the image of Cadinur on `https://www.cleanpng.com/`.

Wrong scores

Another essential requirement for ethical usage of personality inference is to mind the possible errors. In our example, it could be the case that Elise is much more talkative and outgoing than Mike, meaning she should receive a higher extraversion value than Mike.

Nevertheless, the model gave her a lower extraversion value. Would we base our conclusions merely on the scores Elise received, then we would draw a wrong conclusion. Multiple reasons can cause incorrect scores. One reason for the wrong scores is the interpretation of linguistics in the wrong context. For example, the word `Miss` can be the act of failing to hit or reach, feeling the absence of, but also the title used for an unmarried woman [105].

Interpretation of a word in a wrong context could lead to a deviation for the wrong personality trait. Assuming the ground-truth obtained through the questionnaire is a perfect representation of people, all methods show at least one person with an error of 0.3 or higher. Meaning there is always a person for which the concluded personality type is wrong.

9.3.2 Hiring of people

One possible use for personality inference applications would be for the hiring process of companies. One could think of a process where candidates are filtered based on their personality types. With such a filter, we can reduce a broad set of people to a few candidates with the 'desired' personality traits. Supposing the method used can picture the personality traits of the candidates perfectly, such an approach still leaves the question if it is right to do so.

Collins [28] suggested a Five-step job screening process where the inference of personality is one of the five steps. In short, the author proposes to first check for legal ground rules where one should select candidates without discrimination (e.g., gender or race). As a second step, the applicant's resumes should be checked, and reference, background, and integrity checks can be done. As a third step, the employer could do a personality analysis focusing on personality traits, as well as social dominance, bullying, and organizational citizenship behavior. The fourth step is the interview, where the employer should check for inconsistencies and ambiguities that may arise during the previous two job-screening steps. The last step could be, e.g., alcohol or drug tests, to check the finalist's integrity. Important to note from this guide is the combinations of tests that ought to be ethical. Hogan et al. [57] state that personality measurement might be able to provide valuable information on personal characteristics of the applicant, but is not an assessment of technical skills, experience, and ability to learn. The statement mainly describes why a personality screening could be a piece of the puzzle but is an incomplete puzzle without other steps and measurements.

Filtering candidates merely on personality traits might not only lead to false conclusions but also discrimination on certain personality types [114]. Personality measurement is a way to support the understanding of a person; it is not able to adequately provide a full overview of the person and should never be used as a decisive tool on its own.

9.3.3 Mental health

Closely related to the misinterpretation of data, mental health has a significant influence on the outcome of the scores and the integrity and validity of the data. Data quantifiable for research on this matter often lacks. This is partly due to the complexity of mental illness and partly due to a longstanding societal stigma making the subject all but taboo [29]. Around 45 million people around the world are believed to have a bipolar disorder [60, 122], around

20 million worldwide are believed to have schizophrenia [60, 122], and another 50 million to have dementia [122]. Between 76% and 85% of people with mental disorders receive no treatment for their disorder [120, 122] and may not even be aware of the presence of their disorder. As the number of people with mental illnesses is high, the likelihood of processing someone with a mental illness at some point in time is high.

The impact of mental disorders on the outcome of the personality analyses should not be ignored. The mental stability of a person influences the words used [30, 29]. Therefore, it also influences the outcome of personality analyses. For example, the bipolar disorder (BD), can be characterized by recurrent episodes of mania and depression [10]. BD influences the linguistic features used and can be recognized with the use of LIWC [22, 29]. Categories like `PosEmo` and `NegEmo` were found to be affected by BD. `NegEmo` also affects the outcome of Yarkoni and Golbeck.

The inference of personality results in a snapshot of a person's personality profile, depending on the time frame chosen. Therefore, a user of personality inference techniques should be aware of the potential influence of mental health in the chosen time frame, not to mention the possible backlash to the mental health a wrong judgment on personality may have on the individual (e.g., during a hiring process).

9.3.4 Privacy

For ethical personality inference, one should also adhere to the privacy of the data subject (the individual processed). In the case of this study, we improve the privacy for the data subjects by removing usernames, IP-addresses, URLs, and email addresses from messages. This study made use of IBM Personality Insights, a third-party processor. By eliminating the features above, the removed information cannot be retained by this service. As, in theory, the text provided could be used to re-identify the data subject, we adopted the `X-Watson-Learn-Opt-Out` parameter to request IBM not to retain any data from this study. Furthermore, the preprocessing steps acquired for this study obfuscate the data to mitigate the risk of re-identification partially. We urge future users of personality inference to mind the privacy of users and to adopt similar or possibly even farther going proposals in their models.

9.3.5 Why use personality inference?

Although there are negative impacts with automatic personality inference, there is also a multitude of scenarios for which the applications can be useful and ethical. We find examples for which automatic personality methods can be useful in Chapter 10.3. With any application of personality inference, one should keep in mind that there is no perfect or desired personality type, it is the combination and diversity of personalities that lead to higher team performance [20].

Takeaway

Personality inference can be wrong in multiple ways: the context, but also the score itself could be misunderstood. In no way should personality inference be used as a tool on its own to accept or reject people for job applications. The results do not cover job performance, experience, and the ability to learn [57]. Furthermore, mental health or stability may impact the outcome of the analyses, whereas wrong judgments on personality inference may torment the mental stability of the individual even more. In addition, with the use of psycholinguistic tools, one should keep in mind not to harm anyone's privacy.

Although there are negative impacts with automatic personality inference, there is a multitude of scenarios and reasons for which the applications can be useful and ethical.

The recommendations on the ethical usage of psycholinguistic models, given in this study, are not a strict guideline. However, even though the act of personality inference is seemingly possible, one should always feel obliged to honor the human behind the numbers and mind the possible negative impact the numbers may induce.

Chapter 10

Related Work

Personality in software engineering received more attention in recent years. This chapter discusses technologies related to personality inference in software engineering. We divided this chapter into related work on questionnaires for software engineering, where we look into the questionnaires used for research and the differences in academic and industrial applications. We then follow with related work for psycholinguistic models for software engineering. Finally, we cover related work on personality in software engineering to uncover for which reasons personality use can be effective for SE.

10.1 Related work on questionnaires

In the study of Sodiya et al. [110], the authors developed a questionnaire model for the assessment of personality in SE. Opposed to our research, they use a six personality assessment factor, adding cognitive ability as a sixth factor. The authors suspect SE to require this additional factor due to its technical and complex activity, requiring a high level of cognitive ability [110]. However, to the best of our knowledge, there has not been a replication study on this work. The BFI model, on the other hand, is shown to be reliable in numerous studies [44, 3, 8].

Earlier studies in the field of software engineering on personality inference can be found in the form of literature reviews. Barroso et al. [12] performed a systematic literature review for software engineering and found MBTI and Big Five popular models used to assess psychological profiles. In their study, they identify differences in findings for personality research in the academic environment compared to the industry. The authors found in academic studies there are no significant indications of personality impacting the performance of a job done, but industry studies did find significant influences of personality [12]. However, job performance lies outside the scope of this study.

Examples of models often found in industry, besides Big Five personality and MBTI [16], are DISC based on [79], HPI [56], and FSLS [42]. However, several of these commercial models have been criticized in the academic community [54], for example, for their lower formality and small theoretical or research base [96]. In particular, many of the personality tests have been associated with the 'Forer Effect' or 'Barnum effect', a psy-

chological phenomenon in which individuals give high accuracy ratings to the personality descriptions given to them by the personality model, for which the descriptions are vague enough to apply to a wide range of people [119, 54].

10.2 Related work on automated tools

Calefato et al. [18] compare existing solutions for personality inference for software engineering with literature research. In their study, they give an overview of methods and their self-reported performance. Among the methods reported are PI, Yarkoni, and Golbeck. Opposed to our approach, they do not compare the performance of models on the same SE data. All reported models tested on different SE data, which could mean the performance results are not comparable. The authors show with the use of Personality Insights, how personality can be inferred from developer emailing lists [18, 17]. Unlike our study, they only relied on the scores obtained through PI. The reasons given not to use a questionnaire are the generally low response rates and results being biased toward social desirability [18].

More studies used Personality Insights for personality inference among software engineering. For example, Paruma-Pabón et al. [86], who mined emails. In this research, the authors rely on the accuracy of PI. The authors mention the absence of a questionnaire ground-truth as a potential threat to validity [86]. We also find studies using LIWC, such as Bazelli et al. [13] who mined StackOverflow posts, and Licorish et al. [75] who mined comments on the code collaboration platform Jazz. However, none of these studies compared its performance to other methods. Furthermore, most of these studies point out their need for replication to show their reproducibility for software engineering [17, 18, 86, 75]. In our study, we partially fulfill this need.

10.3 Why is personality studied in SE?

In the previous chapter, we have shown multiple ways in which the use of automatic personality inference can be unethical and deemed wrong. However, there are also good reasons to use psycholinguistic models. Studies have shown that in software engineering (SE), personality inference could help in team composition through a combination of personalities within a team [36, 46]. Effective team formation in software engineering is difficult to achieve and it is then important to have an efficient way to assess personality from software engineers [110]. As taking questionnaires is time consuming, psycholinguistic models can help reach efficiency through automation. Other use-cases where we may still use automatic personality inference, but are not limited to, are to help explain work preferences [68], work satisfaction [1], and the effects of personality on pair programming [54, 24].

Chapter 11

Conclusions and Future Work

This chapter gives an overview of the project's contributions. After this overview, we will draw some conclusions on the results. Finally, we discuss some ideas for future work.

11.1 Contributions

In this study, we put three existing psycholinguistic methods to the test and compare their performance for software engineering data. Our contributions are four-fold:

1. We establish a baseline and compare three models on their performance to this baseline. We show three methods to perform almost equally when mean-centered, indicating the three methods may work on different scales. Furthermore, we showed possible improvements in LIWC-based methods. With transformations on LIWC category scores, we show how to reduce the negative influence of outliers on the performance.

2. We propose thirteen different preprocessing steps to clean the software engineering data, or more specifically, the GitHub communication data, and recommend the use of twelve out of thirteen steps for all three inference methods for SE data.

3. We show evidence for a required minimum recommended number of words of a hundred for the psycholinguistic tests used in this study. From our results, 600 to 1200 words show to give at least sub-optimal results while remaining efficient in resource requirements.

4. Furthermore, we show differences in the inference of openness both between fluent and non-fluent English writing and between people with and without English as their mother tongue. PI extraversion and agreeableness, Yarkoni neuroticism, and openness for all methods show differences in scores. However, the performance of the methods does not necessarily change. We found for all three methods, openness scores to be higher for non-fluent English writing than for fluent English writing. Based on earlier studies on this matter, we suspect culture and demographics to cause this difference.

11.2 Conclusions

Our work shows how psycholinguistic models could be used for automatic personality inference. Our goal was to uncover how useful psycholinguistic tests are for this purpose. For such models to be useful, they need to capture the personality of people rightfully. Earlier studies on this field lacked a comparison of multiple models to obtain a fair comparison of performance. To this end, we compared three popular personality inference models, the models proposed by Yarkoni [124], Golbeck et al. [47], and Personality Insights (PI) by IBM. Based on the resulting Big Five personality trait scores, we compared their accuracy to a personality questionnaire among software developers. From these results, we found the psycholinguistic models to introduce a peril in the reliability on an individual level. However, on a group level, all three models promise better performance and show to be more useful.

To culminate to the answering of the main research question, we have shown twelve preprocessing steps for SE communication data to improve the process for the three models. Secondly, we have shown transformations on LIWC categories to improve the overall outcome of LIWC-based methods by reducing the influence of outlier behavior. We have shown, even after all enhancing actions, a fault margin as large as 0.3 to even 0.5 is possible for all methods. Thirdly, to effectively use the psycholinguistic models, we showed the effect of the number of words on the accuracy of the models. We constructed a recommendation of at least 600 to 1200 words for useful analyses. Finally, we have shown the methods are unlikely to discriminate on certain personality types, with an exception to Personality Insights on agreeableness and the openness trait for all methods. For openness, we suspect culture or demographics to cause this effect.

Given the error margins of all three methods, ethical usage is of utmost importance. We, therefore, strongly advise users to seriously consider the adverse effects wrong usage may have on the individual analyzed.

11.3 Future work

Going further on personality inference for software engineering, we suggest three directions for future studies on personality among software engineers.

Preprocessing

In Chapter 5, the preprocessing steps suggested still leave some taint in the comments. Examples of remaining taint are emoticons and Markdown related syntax (e.g., lists, horizontal rules, emphasis, tables, and HTML formatting). The effect of these remaining elements should be investigated and, if need be, removed from the resulting data. Furthermore, some code remains in the resulting comments, albeit the removal of code blocks indicated with grave accents. A more sophisticated method for removing code from the text, such as the method by Bacchelli et al. [11], is needed. The method proposed uses island parsing to extract structured data from natural language documents. However, future studies should extend this work for more programming languages than Java only. Alternatively, the un-

natural language could be removed from the comments. Currently, comments may contain, e.g., copy-pastes of logs. A method similar to that of Jang et al. [61] could extract tables, code, and mathematical formulae from standard text.

Furthermore, for the LIWC methods, outlier behavior could be reduced with transformations on category scores. Yarkoni and Golbeck may improve with a more elaborate investigation on when and which transformations to apply.

Commits, issues, and pull requests

In the current implementation of this study, we made no separation between commits, issues, and pull requests. Future work could investigate the differences in the effectiveness of all three sources. Possibly, the three sources show different behavior in terms of performance for personality inference. For example, commit messages could be written with different intentions than pull requests and, therefore, show different language habits.

Machine learning models

In this study, we compare three existing models for their performance. However, many more models exist for the inference of personality in general, but also for software engineering specifically. Machine learning models, like PI, could be more flexible than the implementations of Yarkoni and Golbeck and make it easier to become multidisciplinary [112]. ML methods could allow for a more reliable translation of personality psychology to practical applications [112]. ML models trained on software engineering data could allow for better interpretation of SE context and reach higher accuracy.

Bibliography

[1] S. Acuña, M. Gómez, J. Hannay, N. Juristo, and D. Pfahl. Are team personality and climate related to satisfaction and software quality? aggregating results from a twice replicated experiment. *Information and Software Technology*, 57:141–156, 01 2015. doi: 10.1016/j.infsof.2014.09.002.

[2] A. Alamsyah, M. F. Rachman, C. S. Hudaya, R. P. Putra, A. I. Rifkyano, and F. Nurwianti. A progress on the personality measurement model using ontology based on social media text. In *2019 International Conference on Information Management and Technology (ICIMTech)*, volume 1, pages 581–586, August 2019. doi: 10.1109/ICIMTech.2019.8843817.

[3] B. Alansari. The big five inventory (bfi): Reliability and validity of its arabic translation in non clinical sample. *European Psychiatry*, 33:S209 – S210, 2016. ISSN 0924-9338. doi: https://doi.org/10.1016/j.eurpsy.2016.01.500. URL http://www.sciencedirect.com/science/article/pii/S0924933816005046.

[4] Y. Amichai-Hamburger and G. Vinitzky. Social network use and personality. *Computers in Human Behavior*, 26(6):1289–1295, 2010. ISSN 0747-5632. doi: https://doi.org/10.1016/j.chb.2010.03.018. URL http://www.sciencedirect.com/science/article/pii/S0747563210000580.

[5] A. Angleitner and J. S. Wiggins. *Personality Assessment via Questionnaires: Current Issues in Theory and Measurement*. Springer Berlin Heidelberg, 2012. ISBN 9783642707513. URL https://books.google.nl/books?id=3q0yBwAAQBAJ.

[6] A. Ansaldo, K. Marcotte, L. Scherer, and G. Raboyeau. Language therapy and bilingual aphasia: Clinical implications of psycholinguistic and neuroimaging research. *Journal of Neurolinguistics*, 21(6):539–557, 2008. ISSN 0911-6044. doi: https://doi.org/10.1016/j.jneuroling.2008.02.001. URL http://www.sciencedirect.com/science/article/pii/S0911604408000146.

[7] P. Arnoux, A. Xu, N. Boyette, J. Mahmud, R. Akkiraju, and V. Sinha. 25 tweets to know you: A new model to predict personality with social media. *CoRR*, abs/1704.05513, 2017. URL http://arxiv.org/abs/1704.05513.

[8] B. J. Arterberry, M. P. Martens, J. M. Cadigan, and D. Rohrer. Application of generalizability theory to the big five inventory. *Personality and individual differences*, 69:98–103, October 2014. ISSN 0191-8869. doi: 10.1016/j.paid.2014.05.015. URL https://pubmed.ncbi.nlm.nih.gov/25419025.

[9] M. C. Ashton and K. Lee. The prediction of honesty-humility-related criteria by the hexaco and five-factor models of personality. *Journal of Research in Personality*, 42(5):1216–1228, 2008. ISSN 0092-6566. doi: https://doi.org/10.1016/j.jrp.2008.03.006. URL http://www.sciencedirect.com/science/article/pii/S0092656608000469.

[10] American Psychiatric Association et al. *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub, 2013.

[11] A. Bacchelli, A. Cleve, M. Lanza, and A. Mocci. Extracting structured data from natural language documents with island parsing. In *26th IEEE International Conference on Automated Software Engineering*, pages 476–479, 11 2011. doi: 10.1109/ASE.2011.6100103.

[12] A. Barroso, J. Madureira, M. Soares, and R. Nascimento. Influence of human personality in software engineering - a systematic literature review. In *19th International Conference on Enterprise Information Systems*, pages 53–62, 1 2017. doi: 10.5220/0006292000530062.

[13] B. Bazelli, A. Hindle, and E. Stroulia. On the personality traits of stackoverflow users. In *2013 IEEE International Conference on Software Maintenance*, pages 460–463, 2013.

[14] V. Benet and O. John. 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*, 75:729–50, 10 1998. doi: 10.1037//0022-3514.75.3.729.

[15] V. V. Bochkarev, A. V. Shevlyakova, and V. D. Solovyev. The average word length dynamics as an indicator of cultural changes in society. *Social Evolution & History*, 14(2):153–175, 2015.

[16] I. Briggs-Mayers, M. H. McCaulley, N. L. Quenk, and A. L. Hammer. *A guide to the development and use of the Myers-Briggs Type Indicator*. Consulting Psychologists Press, 1998.

[17] F. Calefato, G. Iaffaldano, F. Lanubile, and B. Vasilescu. 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, pages 92–101, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-5717-3. doi: 10.1145/3196369.3196372. URL http://doi.acm.org/10.1145/3196369.3196372.

[18] F. Calefato, F. Lanubile, and B. Vasilescu. A large-scale, in-depth analysis of developers' personalities in the apache ecosystem. *CoRR*, abs/1905.13062, 2019. URL http://arxiv.org/abs/1905.13062.

[19] L. Capretz. Personality types in software engineering. *International Journal of Human-Computer Studies*, 58:207–214, 2 2003. doi: 10.1016/S1071-5819(02)00137-4.

[20] L. Capretz and F. Ahmed. Why do we need personality diversity in software engineering? *ACM SIGSOFT Software Engineering Notes*, 35:1–11, 03 2010. doi: 10.1145/1734103.1734111.

[21] G. Carducci, G. Rizzo, D. Monti, E. Palumbo, and M. Morisio. Twitpersonality: Computing personality traits from tweets using word embeddings and supervised learning. *Information (Switzerland)*, 9, 5 2018. doi: 10.3390/info9050127.

[22] C. Chang, E. Saravia, and Y. Chen. 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)*, pages 374–379, 2016.

[23] Y. Chen, R. Al-Rfou, and Y. Choi. Detecting english writing styles for non native speakers. *CoRR*, abs/1704.07441, 2017. URL http://arxiv.org/abs/1704.07441.

[24] K. S. Choi, F. P. Deek, and I. Im. Exploring the underlying aspects of pair programming: The impact of personality. *Information and Software Technology*, 50 (11):1114–1126, 2008. ISSN 0950-5849. doi: https://doi.org/10.1016/j.infsof.2007.11.002. URL http://www.sciencedirect.com/science/article/pii/S0950584907001292.

[25] IBM Cloud. IBM Cloud Docs personality insight, 2019. URL https://cloud.ibm.com/docs/services/personality-insights.

[26] J. Cohen. *Statistical power analysis for the behavioral sciences*. Routledge, New York, USA, 1988. doi: https://doi.org/10.4324/9780203771587.

[27] J. Cohen. Statistical power analysis. *Current directions in psychological science*, 1 (3):98–101, 1992.

[28] D. Collins. Designing ethical organizations for spiritual growth and superior performance: an organization systems approach. *Journal of Management, Spirituality & Religion*, 7(2):95–117, 2010. doi: 10.1080/14766081003746414. URL https://doi.org/10.1080/14766081003746414.

[29] G. Coppersmith, M. Dredze, and C. Harman. Quantifying mental health signals in twitter. In *CLPsych@ACL*, 2014.

[30] G. Coppersmith, M. Dredze, C. Harman, and K. Hollingshead. From adhd to sad: Analyzing the language of mental health on twitter through self-reported diagnoses. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 1–10, 2015.

[31] P. Corr and G. Matthews. The cambridge handbook of personality psychology. pages 1–906, 01 2009. doi: 10.1017/CBO9780511596544.002.

[32] F. Coulmas. Spies and native speakers. *A festschrift for native speaker*, pages 355–367, 1981.

[33] National Research Council et al. *In the mind's eye: Enhancing human performance*. National Academies Press, 1992.

[34] Council of European Union. Council regulation (EU) no 2016/679, 2016. URL `https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng`.

[35] H. Cramér. *Mathematical methods of statistics*, volume 43. Princeton university press, 1999.

[36] 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. Team building criteria in software projects: A mix-method replicated study. *Information and Software Technology*, 55(7):1316–1340, 2013. ISSN 0950-5849. doi: https://doi.org/10.1016/j.infsof.2012.11.006. URL `http://www.sciencedirect.com/science/article/pii/S0950584912002327`.

[37] D. P. Darcy and M. Ma. Exploring individual characteristics and programming performance: Implications for programmer selection. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, pages 314a–314a, January 2005. doi: 10.1109/HICSS.2005.261.

[38] A. Davies. *The Native Speaker in Applied Linguistics*, pages 431–450. Blackwell Publishing Ltd, 01 2008. ISBN 9780470757000. doi: 10.1002/9780470757000.ch17.

[39] M. Dehghani, K. M. Johnson, J. Garten, R. Boghrati, J. Hoover, V. Balasubramanian, A. Singh, Y. Shankar, L. Pulickal, A. Rajkumar, and N. J. Parmar. Tacit: An open-source text analysis, crawling, and interpretation tool. *Behavior Research Methods*, 49(2):538–547, 2017. ISSN 1554-3528. doi: 10.3758/s13428-016-0722-4. URL `https://doi.org/10.3758/s13428-016-0722-4`.

[40] S. J. Dollinger and F. T. L. Leong. Volunteer bias and the five-factor model. *The Journal of Psychology*, 127(1):29–36, 1993. doi: 10.1080/00223980.1993.9915540. URL `https://doi.org/10.1080/00223980.1993.9915540`.

[41] K. S. Faught, D. Whitten, and K. W. Green Jr. Doing survey research on the internet: Yes, timing does matter. *Journal of Computer Information Systems*, 44(3):26–34, 2004. doi: 10.1080/08874417.2004.11647579. URL https://www.tandfonline.com/doi/abs/10.1080/08874417.2004.11647579.

[42] R. Felder. Learning and teaching styles in engineering education. *Journal of Engineering Education -Washington-*, 78:674–681, 01 1988.

[43] A. Field, J. Miles, and Z. Field. *Discovering Statistics Using R*. SAGE Publications, 2012. ISBN 9781446258460. URL https://books.google.nl/books?id=wd2K2zC3swIC.

[44] A. Fossati, S. Borroni, D. Marchione, and C. Maffei. 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*, 27:50–58, 01 2011. doi: 10.1027/1015-5759/a000043.

[45] C. L. Gagné. Psycholinguistic perspectives, June 2017. URL https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199695720.001.0001/oxfordhb-9780199695720-e-13.

[46] A. Gilal, J. Jaafar, M. Omar, S. Basri, and A. Izzatdin. Balancing the personality of programmer: Software development team composition. *Malaysian Journal of Computer Science*, 29, 03 2016. doi: 10.22452/mjcs.vol29no2.5.

[47] J. Golbeck, C. Robles, M. Edmondson, and K. Turner. 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*, pages 149–156, October 2011. doi: 10.1109/PASSAT/SocialCom.2011.33.

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

[49] L. Gou, M. Zhou, and H. Yang. Knowme and shareme: Understanding automatically discovered personality traits from social media and user sharing preferences. *Conference on Human Factors in Computing Systems - Proceedings*, 4 2014. doi: 10.1145/2556288.2557398.

[50] G. Gousios. The ghtorrent dataset and tool suite. In *Proceedings of the 10th Working Conference on Mining Software Repositories*, MSR '13, pages 233–236, Piscataway, NJ, USA, 2013. IEEE Press. ISBN 978-1-4673-2936-1. URL http://dl.acm.org/citation.cfm?id=2487085.2487132.

[51] F. Grosjean. *Studying bilinguals*. Oxford University Press, 2008.

[52] R. HalicioGlu, H. K. Akin, and Y. Fedai. *International Advanced Researches & Engineering Congress 2017 Proceeding Book*. International Workshops (at IAREC'17), 2017.

[53] R. V. Hamilton. A psycholinguistic analysis of some interpretive processes of three basic personality types. *The Journal of Social Psychology*, 46(2):153–177, 1957. doi: 10.1080/00224545.1957.9714317. URL https://doi.org/10.1080/00224545.1957.9714317.

[54] J. E. Hannay, E. Arisholm, H. Engvik, and D. I. K. Sjoberg. The effects of personality on pair programming. *IEEE Transactions on Software Engineering*, 36(1):61–80, 03 2010. doi: 10.1109/TSE.2009.41.

[55] Ong Hee. Validity and reliability of the big five personality traits scale in malaysia. *International Journal of Innovation and Applied Studies*, 04 2014.

[56] R. Hogan. *Hogan personality inventory*, volume 2. Hogan Assessment Systems Tulsa, OK, 1995.

[57] R. Hogan, J. Hogan, and B. Roberts. Personality measurement and employment decisions: Questions and answers. *American Psychologist*, 51:469–477, 05 1996. doi: 10.1037/0003-066X.51.5.469.

[58] A. Howard, M. Choi, and Wu T. K. An alternative to personality questionnaires: Introducing personality tasks. 2019. URL https://questpartnership.co.uk/an-alternative-to-personality-questionnaires-introducing-personality-tasks/.

[59] M. Ireland and M. Mehl. Natural language use as a marker of personality. *Oxford handbook of language and social psychology*, pages 201–218, 01 2014.

[60] S. L. James, D. Abate, K. H. Abate, S. M. Abay, C. Abbafati, N. Abbasi, H. Abbastabar, F. Abd-Allah, J. Abdela, A. Abdelalim, I. Abdollahpour, R. S. Abdulkader, Z. Abebe, S. F. Abera, O. Z. Abil, H. N. Abraha, L. J. Abu-Raddad, N. M. E. Abu-Rmeileh, M. M. K. Accrombessi, D. Acharya, P. Acharya, ..., and C. J. L. Murray. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: a systematic analysis for the global burden of disease study 2017. *The Lancet*, 392(10159): 1789–1858, November 2018. ISSN 0140-6736. doi: 10.1016/S0140-6736(18)32279-7. URL https://doi.org/10.1016/S0140-6736(18)32279-7.

[61] M. Jang, J. D. Choi, and J. Allan. Improving document clustering by removing unnatural language. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 122–130, 2017.

[62] H. Jodai. An introduction to psycholinguistics. *Online Submission*, 2011.

[63] O. John, L. Naumann, and C. Soto. *Paradigm shift to the integrative big five trait taxonomy: History, measurement, and conceptual issues*, pages 114–158. Guilford Press, 01 2008.

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

[65] O. P. John, E. M. Donahue, and Kentle R. L. The "big five" inventory - versions 4a and 54. *Journal of Personality and Social Psychology*, 1991. doi: 10.1037/t07550-000.

[66] P. J. Kajonius. Cross-cultural personality differences between east asia and northern europe in ipip-neo. *International Journal of Personality Psychology*, 3(1):1–7, 2017.

[67] P. Kezwer. The extroverted vs. the introverted personality and second language learning. *TESL Canada Journal*, 5(1):45–58, October 1987. doi: 10.18806/tesl.v5i1.514. URL https://teslcanadajournal.ca/index.php/tesl/article/view/514.

[68] M. V. Kosti, R. Feldt, and L. Angelis. Personality, emotional intelligence and work preferences in software engineering: An empirical study. *Information and Software Technology*, 56(8):973–990, 2014. ISSN 0950-5849. doi: https://doi.org/10.1016/j.infsof.2014.03.004. URL http://www.sciencedirect.com/science/article/pii/S0950584914000639.

[69] W. H. Kruskal and W. A. Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260):583–621, 1952. doi: 10.1080/01621459.1952.10483441. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1952.10483441.

[70] S. Laviosa and M. G. Davies. *The Routledge handbook of translation and education*. Taylor & Francis, 2020.

[71] J. Lee. The native speaker: An achievable model? *Asian EFL Journal*, 7:152–163, 06 2005.

[72] J. Lee, M. Zhou, and X. Liu. Detection of non-native sentences using machine-translated training data. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pages 93–96, 2007.

[73] K. Lemhöfer, T. Dijkstra, H. Schriefers, H. Baayen, J. Grainger, and P. Zwitserlood. Native language influences on word recognition in a second language: A megastudy. *Journal of experimental psychology. Learning, memory, and cognition*, 34:12–31, 02 2008. doi: 10.1037/0278-7393.34.1.12.

[74] J. C. Li. Effect size measures in a two-independent-samples case with nonnormal and nonhomogeneous data. *Behavior Research Methods*, 48(4):1560–1574, 2016. ISSN 1554-3528. doi: 10.3758/s13428-015-0667-z. URL https://doi.org/10.3758/s13428-015-0667-z.

[75] S. A. Licorish and S. G. MacDonell. Personality profiles of global software developers. In *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*, EASE '14, pages 45:1–45:10, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2476-2. doi: 10.1145/2601248.2601265. URL http://doi.acm.org/10.1145/2601248.2601265.

[76] P. M. Lightbown and N. Spada. Do they know what they're doing? l2 learners' awareness of l1 influence. *Language Awareness*, 9(4):198–217, 2000. doi: 10.1080/09658410008667146. URL https://doi.org/10.1080/09658410008667146.

[77] F. Mairesse, M. Walker, M. Mehl, and R. Moore. Using linguistic cues for the automatic recognition of personality in conversation and text. *J. Artif. Intell. Res. (JAIR)*, 30:457–500, 9 2007. doi: 10.1613/jair.2349.

[78] A. S. Mak and C. Tran. Big five personality and cultural relocation factors in vietnamese australian students' intercultural social self-efficacy. *International Journal of Intercultural Relations*, 25(2):181–201, 2001. ISSN 0147-1767. doi: https://doi.org/10.1016/S0147-1767(00)00050-X. URL http://www.sciencedirect.com/science/article/pii/S014717670000050X.

[79] W. M. Marston. *Emotions of Normal People*. Brace & Co, Harcourt, 1928.

[80] T. McArthur, J. Lam-McArthur, and L. Fontaine. *Oxford companion to the English language*. Oxford University Press, 2018.

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

[82] S. Messick. Personality measurement and the ethics of assessment. *The American psychologist*, 20:136–42, 03 1965. doi: 10.1037/h0021712.

[83] D. Mizza. 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. doi: 10.15640/jehd.v3n3a8.

[84] M. Mostafa, T. Crick, A. C. Calderon, and G. Oatley. Incorporating emotion and personality-based analysis in user-centered modelling. In Max Bramer and Miltos Petridis, editors, *Research and Development in Intelligent Systems XXXIII*, pages 383–389, Cham, 2016. Springer International Publishing. ISBN 978-3-319-47175-4.

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

[86] O. H. P. Pabón, F. A. González, J. Aponte, J. E. Camargo, and F. Restrepo-Calle. 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)*, pages 8–14, May 2016. doi: 10.1109/CHASE.2016.010.

[87] T. Pei-Lee, C. Y. Chen, W. C. Chin, and Y. Y. Siew. Do the big five personality factors affect knowledge sharing behaviour? a study of malaysian universities. *Malaysian Journal of Library & Information Science*, 16(1):47–62, 2017. URL `https://mjlis.um.edu.my/article/view/6682`.

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

[89] J. Pennebaker, M. Francis, and R. Booth. *Linguistic inquiry and word count (LIWC)*, 1 1999.

[90] J. Pennebaker, C. Chung, M. Ireland, A. Gonzales, and R. Booth. *The Development and Psychometric Properties of LIWC2007*. LIWC.net, Austin, Texas, 1 2007.

[91] V. Pieterse, M. Leeu, and M. Eekelen. How personality diversity influences team performance in student software engineering teams. In *2018 Conference on Information Communications Technology and Society (ICTAS)*, pages 1–6, 3 2018. doi: 10.1109/ICTAS.2018.8368749.

[92] T. Poling, D. Woehr, L. Arciniega, and C. A. Gorman. The impact of personality and value diversity on team performance. 1 2004.

[93] B. Raad and M. Perugini. *Big Five factor assessment: Introduction*, pages 1–26. Hogrefe & Huber Publishers, Gottingen, Germany, 1 2002. ISBN 0-88937-242-X.

[94] J. Rantanen, R. Metsäpelto, T. Feldt, L. Pulkkinen, and K. Kokko. Long-term stability in the big five personality traits in adulthood. *Scandinavian journal of psychology*, 48:511–8, 1 2008. doi: 10.1111/j.1467-9450.2007.00609.x.

[95] A. Rastogi and N. Nagappan. On the personality traits of github contributors. In *2016 IEEE 27th International Symposium on Software Reliability Engineering (ISSRE)*, pages 77–86, 2016.

[96] J. H. Reynierse, D. Ackerman, A. A. Fink, and J. B. Harker. The effects of personality and management role on perceived values in business settings. *International Journal of Value-Based Management*, 13(1):1–13, 2000.

[97] J. Rolland. *The Five-Factor Model of Personality Across Cultures*, pages 7–28. 12 2002. ISBN HB: 0-306-47354-2 PN: 0-306-47355-0. doi: 10.1007/978-1-4615-0763-5_2.

[98] M. J. Rosen and D. Carlson. *Just My Type: Understanding Personality Profiles*. Nonfiction - Young Adult. Lerner Publishing Group, 2016. ISBN 9781467795791. URL https://books.google.nl/books?id=lsOLCwAAQBAJ.

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

[100] C. Ross, E. S. Orr, M. Sisic, J. M. Arseneault, M. G. Simmering, and R. R. Orr. Personality and motivations associated with facebook use. *Computers in Human Behavior*, 25(2):578–586, 2009. ISSN 0747-5632. doi: https://doi.org/10.1016/j.chb.2008.12.024. URL http://www.sciencedirect.com/science/article/pii/S0747563208002355.

[101] S. Sagadevan, N. H. A. H. Malim, and M. H. Husin. Sentiment valences for automatic personality detection of online social networks users using three factor model. *Procedia Computer Science*, 72:201–208, 2015. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2015.12.122. URL http://www.sciencedirect.com/science/article/pii/S1877050915035838.

[102] N. Salleh, E. Mendes, J. Grundy, and G. S. J. Burch. An empirical study of the effects of personality in pair programming using the five-factor model. In *2009 3rd International Symposium on Empirical Software Engineering and Measurement*, pages 214–225, October 2009. doi: 10.1109/ESEM.2009.5315997.

[103] D. Schmitt, J. Allik, R. McCrae, V. Benet, J. Veríssimo, and U. Reips. The geographic distribution of big five personality traits. *Journal of Cross-Cultural Psychology*, 38: 173–212, 03 2007. doi: 10.1177/0022022106297299.

[104] M. Schoonvelde, G. Schumacher, and B. Bakker. 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*, 7:124–143, 2019. doi: 10.5964/jspp.v7i1.964.

[105] O. C. Schultheiss. Are implicit motives revealed in mere words?: Testing the marker-word hypothesis with computer-based text analysis. *Frontiers in psychology*, 4:748–748, October 2013. ISSN 1664-1078. doi: 10.3389/fpsyg.2013.00748. URL https://pubmed.ncbi.nlm.nih.gov/24137149.

[106] H. Schwartz, J. Eichstaedt, M. Kern, L. Dziurzynski, S. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. Seligman, and L. Ungar. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8:e73791, 09 2013. doi: 10.1371/journal.pone.0073791.

[107] S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples)†. *Biometrika*, 52(3-4):591–611, 12 1965. ISSN 0006-3444. doi: 10.1093/biomet/52.3-4.591. URL https://doi.org/10.1093/biomet/52.3-4.591.

[108] J. Shen, O. Brdiczka, and J. Liu. Understanding email writers: Personality prediction from email messages. In *User Modeling, Adaptation, and Personalization: 21th International Conference*, volume 7899, pages 318–330, 6 2013. doi: 10.1007/978-3-642-38844-6_29.

[109] E. Smith, R. Loftin, E. Murphy-Hill, C. Bird, and T. Zimmermann. Improving developer participation rates in surveys. In *2013 6th International Workshop on Cooperative and Human Aspects of Software Engineering (CHASE)*, pages 89–92, May 2013. doi: 10.1109/CHASE.2013.6614738.

[110] A. S. Sodiya, H. Longe, A. Onashoga, A. Oludele, and L. Omotosho. An improved assessment of personality traits in software engineering. In *InSITE 2007: Informing Science + IT Education Conference*, 1 2007. doi: 10.28945/3164.

[111] C. Soto. *Big Five personality traits*, pages 240–241. SAGE Publications, Thousand Oaks, California, US, 1 2018.

[112] C. Stachl, F. Pargent, S. Hilbert, G. Harari, R. Schoedel, S. Vaid, S. Gosling, and M. Buehner. Personality research and assessment in the era of machine learning. 08 2019. doi: 10.31234/osf.io/efnj8.

[113] H. H. Stern, E. E. Tarone, H. H. Stern, G. Yule, and H. Stern. *Fundamental concepts of language teaching: Historical and interdisciplinary perspectives on applied linguistic research*. Oxford university press, 1983.

[114] E. F. Stone-Romero. Personality-based stigmas and unfair discrimination in work organizations. In *Discrimination at work: The psychological and organizational bases*, pages 255–280. Lawrence Erlbaum Associates, 01 2005. ISBN 0-8058-5207-7.

[115] Student. The probable error of a mean. *Biometrika*, 6(1):1–25, 1908. ISSN 00063444. URL http://www.jstor.org/stable/2331554.

[116] Y. Tausczik and J. Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29:24–54, 3 2010. doi: 10.1177/0261927X09351676.

[117] A. Thorne and H. G Gough. *Portraits of type: An MBTI research compendium*. Davies-Black Publishing, 1991.

[118] twitter_developers. Counting characters - twitter developers, n.d. URL https://developer.twitter.com/en/docs/basics/counting-characters.

[119] K. D. Vohs. Barnum effect, 2016. URL https://www.britannica.com/science/Barnum-Effect.

[120] P. S. Wang, S. Aguilar-Gaxiola, J. Alonso, M. C. Angermeyer, G. Borges, E. J. Bromet, R. Bruffaerts, G. de Girolamo, R. de Graaf, O. Gureje, J. M. Haro, E. G. Karam, R. C. Kessler, V. Kovess, M. C. Lane, S. Lee, D. Levinson, Y. Ono, M. Petukhova, J. Posada-Villa, S. Seedat, and J. E. Wells. Use of mental health services for anxiety, mood, and substance disorders in 17 countries in the who world mental health surveys. *Lancet (London, England)*, 370(9590):841–850, September 2007. ISSN 1474-547X. doi: 10.1016/S0140-6736(07)61414-7. URL https://pubmed.ncbi.nlm.nih.gov/17826169.

[121] S. S. Wang and M. A. Stefanone. Showing off? human mobility and the interplay of traits, self-disclosure, and facebook check-ins. *Social Science Computer Review*, 31(4):437–457, 2013. doi: 10.1177/0894439313481424. URL https://doi.org/10.1177/0894439313481424.

[122] WHO. Mental disorders, November 2019. URL https://www.who.int/news-room/fact-sheets/detail/mental-disorders.

[123] F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945. ISSN 00994987. URL http://www.jstor.org/stable/3001968.

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

Appendix A

BFI Questionnaire

In this chapter, we first show the questionnaire questions sent to the participants. The actual questionnaire was taken through Qualtrics[1] under the university domain. Second, we show the scoring model used to infer personality scores through the BFI method.

A.1 Questionnaire questions

Hello! We are researchers from the Delft University of Technology, the Netherlands exploring unconventional solutions to infer developer personality. The questionnaire comprises of 44 questions on personality and 5 demographic questions and will take around 7 minutes of your time.

Please try to answer truthfully according to your interpretation of the questions. There are no 'desired' answers to the questions.

All data processed for this study are protected under the European General Data Protection Regulation (GDPR). Therefore, we ask you to read the following carefully.

1. I voluntarily consent to be a participant in this study and understand that I can refuse to answer questions. I can withdraw from the study within one month from my submission without having to give a reason.

2. I understand that taking part in the study involves the completion of a survey that indicates my personality.

3. I understand that the information I provide will be used for the completion of a scientific study.

4. I understand that personal information collected about me that can identify me, such as [e.g., my email address], will not be shared beyond the study team.

[1] https://www.qualtrics.com/

5. I understand that this study will process my public GitHub data obtained through the GitHub API and I understand that any of this data is processed through the IBM Watson Personality Insights service.
6. I understand that my data obtained from this questionnaire will be combined with personality scores inferred from my public comments on GitHub.
7. I understand that I must be at least 16 years old to participate in this study.

For more questions about the study please do not hesitate to contact us: Frenk van Mil at f.c.j.vanmil@student.tudelft.nl

I acknowledge that I have read the above and agree with the processing of my data provided through this questionnaire.

I agree □

Personality questions

In this section, the personality questions are asked. In the end, these questions should give an indication of your personality.

	I am someone who..	$--$	$-$	$-/+$	$+$	$++$
1.	is talkative	☐	☐	☐	☐	☐
2.	tends to find fault with others	☐	☐	☐	☐	☐
3.	does a thorough job	☐	☐	☐	☐	☐
4.	is depressed, blue	☐	☐	☐	☐	☐
5.	is original, comes up with new ideas	☐	☐	☐	☐	☐
6.	is reserved	☐	☐	☐	☐	☐
7.	is helpful and unselfish with others	☐	☐	☐	☐	☐
8.	can be somewhat careless	☐	☐	☐	☐	☐
9.	is relaxed, handles stress well	☐	☐	☐	☐	☐
10.	is curious about many different things	☐	☐	☐	☐	☐
11.	is full of energy	☐	☐	☐	☐	☐
12.	starts quarrels with others	☐	☐	☐	☐	☐
13.	is a reliable worker	☐	☐	☐	☐	☐
14.	can be tense	☐	☐	☐	☐	☐
15.	is ingenious, a deep thinker	☐	☐	☐	☐	☐
16.	generates a lot of enthusiasm	☐	☐	☐	☐	☐
17.	has a forgiving nature	☐	☐	☐	☐	☐
18.	tends to be disorganized	☐	☐	☐	☐	☐
19.	worries a lot	☐	☐	☐	☐	☐
20.	has an active imagination	☐	☐	☐	☐	☐
21.	tends to be quiet	☐	☐	☐	☐	☐
22.	is generally trusting	☐	☐	☐	☐	☐
23.	tends to be lazy	☐	☐	☐	☐	☐
24.	is emotionally stable, not easily upset	☐	☐	☐	☐	☐
25.	is inventive	☐	☐	☐	☐	☐
26.	has an assertive personality	☐	☐	☐	☐	☐
27.	can be cold and aloof	☐	☐	☐	☐	☐
28.	perseveres until the task is finished	☐	☐	☐	☐	☐
29.	can be moody	☐	☐	☐	☐	☐
30.	values artistic, aesthetic experiences	☐	☐	☐	☐	☐
31.	is sometimes shy, inhibited	☐	☐	☐	☐	☐
32.	is considerate and kind to almost everyone	☐	☐	☐	☐	☐
33.	does things efficiently	☐	☐	☐	☐	☐
34.	remains calm in tense situations	☐	☐	☐	☐	☐
35.	prefers work that is routine	☐	☐	☐	☐	☐
36.	is outgoing, sociable	☐	☐	☐	☐	☐
37.	is sometimes rude to others	☐	☐	☐	☐	☐
38.	makes plans and follows through with them	☐	☐	☐	☐	☐
39.	gets nervous easily	☐	☐	☐	☐	☐
40.	likes to reflect, play with ideas	☐	☐	☐	☐	☐
41.	has few artistic interests	☐	☐	☐	☐	☐
42.	likes to cooperate with others	☐	☐	☐	☐	☐
43.	is easily distracted	☐	☐	☐	☐	☐
44.	is sophisticated in art, music, or literature	☐	☐	☐	☐	☐

General questions

1. English is my mother tongue
   `English is the language I learned as a child. Growing up in a bilingual home could mean you have more than one mother tongue language`
   Yes: ☐ No: ☐ Maybe: ☐
2. I am fluent in written English
   `The answer to this question may depend on your interpretation of 'fluency'. There is no right or wrong to this question.`
   Yes: ☐ No: ☐ Maybe: ☐
3. In what country did you spend most of your youth?
   [ ]
4. Would you like to receive a response from this study?
   `I wish to receive a response from this study`
   Yes: ☐ No: ☐

Thank you for your participation, your contribution is very much appreciated!

A.2 BFI scoring model

In this section, we show the scoring model for the BFI questionnaire. We calculate the Big Five personality scores based on the 44 questions of BFI [63, 65, 14] shown in the questionnaire above. The scores can be calculated according to the scoring method defined for BFI [63, 65, 14]:

```
Averaging the following question items for each personality trait
(with R the reversed score).

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

Reversed scores calculation: 6 - <score>
```

The numbers after a trait indicate the question number. For example, 'Openness: 5, 10, 15, (...)' takes the scores given to questions 5, 10, 15, etc. All scores are on a scale from one to five. 'One' indicates a strong disagreement, whereas a 'five' indicates a strong agreement. All questions indicated with an 'R' should be taken as a reverse. For example, '8R' means $6 - score$, where *score* is the score given to question 8.

Appendix B

Glossary

In this appendix we give an overview of frequently used terms and abbreviations.

BFI: The abbreviation for 'Big Five Inventory'. A questionnaire framework for the inference of personality [63, 65, 14].

BFP: The abbreviation for 'Big Five Personality'.

Golbeck: Term used for this study to indicate the implementation based on the proposed psycholinguistic model of the study of Golbeck et al. [47].

LIWC: Linguistic Inquiry and Word Count. A text analysis program that assigns category scores based on the words used in text.

MAE: The abbreviation for Mean Absolute Error. A term often used to indicate accuracy. Unlike the RMSE, the MAE weighs all data points equally. The MAE is calculated with the following formula:

$$MAE_t = \frac{1}{n}\sum_{u=1}^{n} |y_{u,t} - \hat{y}_{u,t}|$$

Where n is the number of users compared, u the user compared, $y_{u,t}$ the predicted personality trait score for user u and trait t, and $\hat{y}_{u,t}$ the observed personality trait score for user u and trait t.

Mix-Max normalization A normalization process for which all scores are put to a scale between zero and one. The Min-Max normalization uses the formula

$$X_{normalized} = \frac{X_{raw} - x_{min}}{X_{max} - X_{min}}$$

For this study, $X_{normalized}$ indicates all normalized personality scores, X_{raw} represents all raw personality score, and X_{min} and X_{max} the minimum and maximum raw personality score observed, respectively.

PI: Personality Insights. A tool by IBM which can be used to extract Big Five personality traits from text.

RMSE: Abbreviation for Root Mean Squared Error. A term often used to indicate accuracy. The RMSE is essentially the square root taken from the MAE. Unlike the MAE, the RMSE weighs outliers more heavily. The RMSE is calculated with the following formula:

$$RMSE_t = \sqrt{\frac{1}{n}\sum_{u=1}^{n}(y_{u,t} - \hat{y}_{u,t})^2}$$

Where n is the number of users compared, u the user compared, $y_{u,t}$ the predicted personality trait score for user u and trait t, and $\hat{y}_{u,t}$ the observed personality trait score for user u and trait t.

SE: Abbreviation for Software Engineering.

Trait: In this study, the word 'trait' is used as a short replacement for 'Personality Trait'. A personality trait is either one of five Big Five personality traits [111]—openness, conscientiousness, extraversion, agreeableness, and neuroticism.

Yarkoni: Term used for this study to indicate the implementation based on the proposed psycholinguistic model of the study of Yarkoni [124].

Appendix C

Additional tables

In this chapter, additional tables missing in the main chapters of the document can be found.

C.1 Additional tables chapter 7

Table C.1: 95% Confidence interval of the absolute error for each method and trait for the different number of words fed to the analyses. The columns indicate the number of words given to the analyses (i.e, 100 words, 600 words, 1200 words, and 3000 or more words). The columns from left to right: the trait and method for which the confidence interval is given, the 95% confidence interval of the 100 word dataset, the 95% confidence interval of the 600 word dataset, the 95% confidence interval of the 1200 word dataset, and the 95% confidence interval of the 3000 word dataset

Trait / Word count	**100**	**600**	**1200**	**3000**
PI Conscientiousness	0.17	0.15	0.16	0.15
PI Extraversion	0.15	0.15	0.17	0.16
PI Agreeableness	0.16	0.15	0.15	0.16
PI Neuroticism	0.12	0.10	0.10	0.10
Yarkoni Openness	0.12	0.11	0.11	0.10
Yarkoni Conscientiousness	0.12	0.15	0.14	0.13
Yarkoni Extraversion	0.16	0.12	0.13	0.12
Yarkoni Agreeableness	0.13	0.10	0.12	0.09
Yarkoni Neuroticism	0.12	0.15	0.15	0.16
Golbeck Openness	0.08	0.07	0.07	0.08
Golbeck Conscientiousness	0.11	0.07	0.07	0.07
Golbeck Extraversion	0.12	0.13	0.12	0.12
Golbeck Agreeableness	0.10	0.11	0.11	0.11
Golbeck Neuroticism	0.19	0.17	0.18	0.16

C.2 Additional tables chapter 8

In this section, we show the additional tables for the comparison of English proficiency missing in the main chapters of this document.

C.2.1 Comparison of Fluency with Maybe

In this section, we show some additional tables when 'Maybe' remains for the comparison of fluency for English proficiency.

Wilcoxon summed rank test

Table C.2: Wilcoxon summed rank test - Fluency [Maybe vs No]. The first column indicates the method and the trait for which the differences in means are checked. The V column depicts the V-statistic. The column p.value shows the p-value found for the method which is assumed significant if below 0.05. The column r shows the r effect size with the effect column showing the corresponding effect size interpretation.

Column	V	p.value	p<0.05	r	effect
PI Openness	180	0.68	FALSE	-0.09	negligible
PI Agreeableness	170	0.49	FALSE	-0.14	small
Yarkoni Agreeableness	198.5	0.95	FALSE	-0.01	negligible
Golbeck Conscientiousness	108.5	0.02	TRUE	-0.5	medium
Golbeck Agreeableness	119	0.04	TRUE	-0.44	medium
Golbeck Neuroticism	211.5	0.66	FALSE	-0.09	negligible

Table C.3: Wilcoxon summed rank test - Fluency [Maybe vs No]. The first column indicates the method and the trait for which the differences in means are checked. The V column depicts the V-statistic. The column p.value shows the p-value found for the method which is assumed significant if below 0.05. The column r shows the r effect size with the effect column showing the corresponding effect size interpretation.

Column	V	p.value	p<0.05	r	effect
PI Openness	2063	0.15	FALSE	-0.3	small
PI Agreeableness	3076.5	0.08	FALSE	-0.36	medium
Yarkoni Agreeableness	2555.5	0.91	FALSE	-0.02	negligible
Golbeck Conscientiousness	2010.5	0.11	FALSE	-0.33	medium
Golbeck Agreeableness	1883	0.05	TRUE	-0.42	medium
Golbeck Neuroticism	2483	0.91	FALSE	-0.02	negligible

Table C.4: Wilcoxon summed rank test - Fluency [Maybe vs No]. The first column indicates the method and the trait for which the differences in means are checked. The V column depicts the V-statistic. The column p.value shows the p-value found for the method which is assumed significant if below 0.05. The column r shows the r effect size with the effect column showing the corresponding effect size interpretation.

Column	V	p.value	p<0.05	r	effect
PI Openness	1721	0.61	FALSE	-0.13	small
PI Agreeableness	2464	0.03	TRUE	-0.54	large
Yarkoni Agreeableness	1896	0.9	FALSE	-0.03	negligible
Golbeck Conscientiousness	2336	0.08	FALSE	-0.43	medium
Golbeck Agreeableness	2096	0.39	FALSE	-0.21	small
Golbeck Neuroticism	1623	0.38	FALSE	-0.21	small

Unpaired Student t-test

Table C.5: Student t-test - Fluency [Maybe vs No]. The first column indicates the method and the trait for which the differences in means are checked. The V column depicts the V-statistic. The column p.value shows the p-value found for the method which is assumed significant if below 0.05. The column d shows the Cohen's d effect size with the magnitude column showing the corresponding effect size interpretation.

Column	V	p.value	p<0.05	d	magnitude
PI Conscientiousness	1.59	0.12	FALSE	0.49	medium
PI Extraversion	-0.49	0.63	FALSE	-0.17	small
PI Neuroticism	-0.62	0.54	FALSE	-0.2	small
Yarkoni Openness	-0.58	0.57	FALSE	-0.2	small
Yarkoni Conscientiousness	0.76	0.45	FALSE	0.24	small
Yarkoni Extraversion	-0.09	0.93	FALSE	-0.03	negligible
Yarkoni Neuroticism	-0.89	0.38	FALSE	-0.28	small
Golbeck Openness	2.75	0.01	TRUE	0.88	large
Golbeck Extraversion	1.07	0.29	FALSE	0.34	medium

Table C.6: Student t-test - Fluency [Maybe vs No]. The first column indicates the method and the trait for which the differences in means are checked. The V column depicts the V-statistic. The column p.value shows the p-value found for the method which is assumed significant if below 0.05. The column d shows the Cohen's d effect size with the magnitude column showing the corresponding effect size interpretation.

Column	V	p.value	p<0.05	d	magnitude
PI Conscientiousness	2.35	0.03	TRUE	0.51	large
PI Extraversion	0.8	0.43	FALSE	0.14	small
PI Neuroticism	-1.7	0.1	FALSE	-0.39	medium
Yarkoni Openness	-3.52	0	TRUE	-0.79	large
Yarkoni Conscientiousness	-0.23	0.82	FALSE	-0.05	negligible
Yarkoni Extraversion	0.13	0.9	FALSE	0.03	negligible
Yarkoni Neuroticism	1.52	0.14	FALSE	0.39	medium
Golbeck Openness	-0.8	0.43	FALSE	-0.18	small
Golbeck Extraversion	1.34	0.19	FALSE	0.24	small

Table C.7: Student t-test - Fluency [Maybe vs No]. The first column indicates the method and the trait for which the differences in means are checked. The V column depicts the V-statistic. The column p.value shows the p-value found for the method which is assumed significant if below 0.05. The column d shows the Cohen's d effect size with the magnitude column showing the corresponding effect size interpretation.

Column	V	p.value	p<0.05	d	magnitude
PI Conscientiousness	0.32	0.75	FALSE	0.07	negligible
PI Extraversion	0.97	0.34	FALSE	0.31	medium
PI Neuroticism	-0.59	0.56	FALSE	-0.17	small
Yarkoni Openness	-1.58	0.13	FALSE	-0.54	large
Yarkoni Conscientiousness	-1.27	0.22	FALSE	-0.3	small
Yarkoni Extraversion	0.23	0.82	FALSE	0.06	negligible
Yarkoni Neuroticism	2.57	0.02	TRUE	0.73	large
Golbeck Openness	-4.23	0	TRUE	-1.09	large
Golbeck Extraversion	-0.09	0.93	FALSE	-0.02	negligible

One-way ANOVA & Kruskal-Wallis test

To compare the influence of fluency, we apply the One-way ANOVA [43] on normally distributed traits and the Kruskal-Wallis [69] test on non-normal traits. Normality is tested using the Shapiro-Wilk test [107]. From the results (see tables C.8 and C.9), we can observe PI agreeableness, Yarkoni openness and neuroticism, and Golbeck openness and conscientiousness to have significant differences in means caused by different classifications of fluency.

Table C.8: One-way ANOVA test on the association between personality and Fluency. The first column shows the method and trait name. Df depicts the degrees of freedom. The F.value is the F-statistic. The p.value the p-value found followed by the significance for $p < 0.05$. The eta.sq column is the Eta-squared effect size. The effect columns shows the effect size interpretation.

Column	Df	F.value	p.value	p<0.05	eta.sq	effect
PI Conscientiousness	2	2.77	0.06	FALSE	0.02	
PI Extraversion	2	0.91	0.4	FALSE	0.01	
PI Neuroticism	2	1.69	0.19	FALSE	0.01	
Yarkoni Openness	2	7.91	0	TRUE	0.06	small
Yarkoni Conscientiousness	2	0.71	0.49	FALSE	0.01	
Yarkoni Extraversion	2	0.04	0.96	FALSE	0	
Yarkoni Neuroticism	2	5.21	0.01	TRUE	0.04	small
Golbeck Openness	2	9.46	0	TRUE	0.07	medium
Golbeck Extraversion	2	0.63	0.54	FALSE	0	

Table C.9: Kruskal-Wallis test on personality affect by Fluency

Column	Df	Chisq	p.value	p<0.05
PI Openness	2	2.18	0.34	FALSE
PI Agreeableness	2	7.44	0.02	TRUE
Yarkoni Agreeableness	2	0.03	0.99	FALSE
Golbeck Conscientiousness	2	6.13	0.05	TRUE
Golbeck Agreeableness	2	5.09	0.08	FALSE
Golbeck Neuroticism	2	0.75	0.69	FALSE

Comparison to the ground-truth

Table C.10 shows the RMSE (Root Mean Squared Error) values and MAE (Mean Absolute Error) values of all three fluency groups (Yes, No, and Maybe).

Table C.10: Comparison of RMSE and MAE values of Fluency for all three fluency groups (Yes, No, and Maybe). The first three columns show the MAE (Mean Absolute Error) values of the Maybe, No, and Yes groups. The next three columns show the RMSE (Root Mean Squared Error) values of these groups. For both the MAE and RMSE, the lower the better.

	MAE Maybe	MAE No	MAE Yes	RMSE Maybe	RMSE No	RMSE Yes
PI O.	0.12	0.16	0.14	0.15	0.19	0.18
PI C.	0.16	0.09	0.18	0.19	0.11	0.23
PI E.	0.16	0.18	0.20	0.20	0.22	0.25
PI A.	0.27	0.23	0.32	0.30	0.27	0.35
PI N.	0.14	0.13	0.18	0.17	0.17	0.22
Yarkoni O.	0.14	0.15	0.13	0.17	0.18	0.15
Yarkoni C.	0.20	0.12	0.17	0.23	0.14	0.21
Yarkoni E.	0.15	0.18	0.19	0.20	0.22	0.23
Yarkoni A.	0.18	0.13	0.18	0.21	0.16	0.22
Yarkoni N.	0.19	0.18	0.18	0.23	0.23	0.22
Golbeck O.	0.09	0.13	0.11	0.11	0.15	0.13
Golbeck C.	0.18	0.23	0.15	0.24	0.25	0.19
Golbeck E.	0.14	0.18	0.16	0.17	0.23	0.19
Golbeck A.	0.13	0.16	0.13	0.15	0.19	0.16
Golbeck N.	0.34	0.35	0.32	0.40	0.39	0.37

C.2.2 Comparison of Mother tongue

In this section, we show some additional tables when 'Maybe' remains for the comparison of mother tongue for English proficiency.

Wilcoxon summed rank test

Table C.11: Wilcoxon summed rank test - Mother tongue [Maybe vs No]. The first column indicates the method and the trait for which the differences in means are checked. The V column depicts the V-statistic. The column p.value shows the p-value found for the method which is assumed significant if below 0.05. The column r shows the r effect size with the effect column showing the corresponding effect size interpretation.

Column	V	p.value	p<0.05	r	effect
PI Openness	459.5	0.33	FALSE	-0.4	medium
PI Agreeableness	649.5	0.73	FALSE	-0.14	small
Yarkoni Agreeableness	418.5	0.21	FALSE	-0.51	large
Golbeck Conscientiousness	491	0.45	FALSE	-0.31	medium
Golbeck Agreeableness	461.5	0.34	FALSE	-0.39	medium
Golbeck Neuroticism	694.5	0.51	FALSE	-0.27	small

Table C.12: Wilcoxon summed rank test - Mother tongue [Maybe vs No]. The first column indicates the method and the trait for which the differences in means are checked. The V column depicts the V-statistic. The column p.value shows the p-value found for the method which is assumed significant if below 0.05. The column r shows the r effect size with the effect column showing the corresponding effect size interpretation.

Column	V	p.value	p<0.05	r	effect
PI Openness	116	0.29	FALSE	-0.44	medium
PI Agreeableness	175	0.7	FALSE	-0.16	small
Yarkoni Agreeableness	105.5	0.18	FALSE	-0.54	large
Golbeck Conscientiousness	118	0.31	FALSE	-0.42	medium
Golbeck Agreeableness	116	0.29	FALSE	-0.44	medium
Golbeck Neuroticism	141.5	0.67	FALSE	-0.17	small

Table C.13: Wilcoxon summed rank test - Mother tongue [Maybe vs No]. The first column indicates the method and the trait for which the differences in means are checked. The V column depicts the V-statistic. The column p.value shows the p-value found for the method which is assumed significant if below 0.05. The column r shows the r effect size with the effect column showing the corresponding effect size interpretation.

Column	V	p.value	p<0.05	r	effect
PI Openness	5615	0.51	FALSE	-0.05	negligible
PI Agreeableness	5548	0.6	FALSE	-0.04	negligible
Yarkoni Agreeableness	4993	0.52	FALSE	-0.05	negligible
Golbeck Conscientiousness	4980.5	0.5	FALSE	-0.05	negligible
Golbeck Agreeableness	4972.5	0.49	FALSE	-0.05	negligible
Golbeck Neuroticism	4448.5	0.07	FALSE	-0.13	small

Unpaired Student t-test

Table C.14: Student t-test - Mother tongue [Maybe vs No]. The first column indicates the method and the trait for which the differences in means are checked. The V column depicts the V-statistic. The column p.value shows the p-value found for the method which is assumed significant if below 0.05. The column d shows the Cohen's d effect size with the magnitude column showing the corresponding effect size interpretation.

Column	V	p.value	p<0.05	d	magnitude
PI Conscientiousness	-1.12	0.31	FALSE	-0.63	large
PI Extraversion	-0.62	0.56	FALSE	-0.27	small
PI Neuroticism	1.37	0.22	FALSE	0.44	medium
Yarkoni Openness	0.33	0.75	FALSE	0.1	small
Yarkoni Conscientiousness	0.48	0.65	FALSE	0.14	small
Yarkoni Extraversion	-0.59	0.58	FALSE	-0.16	small
Yarkoni Neuroticism	0.73	0.49	FALSE	0.21	small
Golbeck Openness	0.39	0.71	FALSE	0.17	small
Golbeck Extraversion	-0.18	0.87	FALSE	-0.07	negligible

Table C.15: Student t-test - Mother tongue [Maybe vs No]. The first column indicates the method and the trait for which the differences in means are checked. The V column depicts the V-statistic. The column p.value shows the p-value found for the method which is assumed significant if below 0.05. The column d shows the Cohen's d effect size with the magnitude column showing the corresponding effect size interpretation.

Column	V	p.value	p<0.05	d	magnitude
PI Conscientiousness	-0.62	0.56	FALSE	-0.27	small
PI Extraversion	-0.49	0.64	FALSE	-0.23	small
PI Neuroticism	1.45	0.19	FALSE	0.46	medium
Yarkoni Openness	0.07	0.95	FALSE	0.03	negligible
Yarkoni Conscientiousness	0.45	0.66	FALSE	0.14	small
Yarkoni Extraversion	-0.89	0.4	FALSE	-0.31	medium
Yarkoni Neuroticism	1.2	0.27	FALSE	0.42	medium
Golbeck Openness	-0.51	0.63	FALSE	-0.32	medium
Golbeck Extraversion	-0.32	0.76	FALSE	-0.14	small

Table C.16: Student t-test - Mother tongue [Maybe vs No]. The first column indicates the method and the trait for which the differences in means are checked. The V column depicts the V-statistic. The column p.value shows the p-value found for the method which is assumed significant if below 0.05. The column d shows the Cohen's d effect size with the magnitude column showing the corresponding effect size interpretation.

Column	V	p.value	p<0.05	d	magnitude
PI Conscientiousness	1.39	0.17	FALSE	0.25	small
PI Extraversion	0.35	0.73	FALSE	0.05	negligible
PI Neuroticism	0.39	0.7	FALSE	0.06	negligible
Yarkoni Openness	-0.61	0.55	FALSE	-0.08	negligible
Yarkoni Conscientiousness	0.02	0.98	FALSE	0	negligible
Yarkoni Extraversion	-0.71	0.48	FALSE	-0.1	small
Yarkoni Neuroticism	1.1	0.28	FALSE	0.16	small
Golbeck Openness	-3.5	0	TRUE	-0.43	medium
Golbeck Extraversion	-0.4	0.69	FALSE	-0.06	negligible

One-way ANOVA & Kruskal-Wallis test

To check if there are differences in the values given to the question about their mother tongue, we apply the One-way ANOVA [43] on normally distributed traits and the Kruskal-Wallis [69] test on non-normal traits. Normality is tested using the Shapiro-Wilk test [107]. From both the ANOVA and the Kruskal-Wallis test, only Golbeck openness was found to have a significant difference in mean scores caused by the mother tongue property (see table C.17). However, this difference was found to be small in accordance with Cohen's interpretation of effect size [26].

Table C.17: One-way ANOVA test on personality affected by Mother tongue showing all significant traits. The first column shows the method and trait name. Df depicts the degrees of freedom. The F.value is the F-statistic. The p.value the p-value found followed by the significance for $p < 0.05$. The eta.sq column is the Eta-squared effect size. The effect columns shows the effect size interpretation.

Column	Df	F.value	p.value	p<0.05	eta.sq	effect
PI Openness	2	0.07	0.93	FALSE	0	
PI Conscientiousness	2	2.15	0.12	FALSE	0.02	
PI Extraversion	2	0.26	0.77	FALSE	0	
PI Agreeableness	2	0.27	0.76	FALSE	0	
PI Neuroticism	2	0.67	0.52	FALSE	0.01	
Yarkoni Openness	2	0.16	0.85	FALSE	0	
Yarkoni Conscientiousness	2	0.06	0.94	FALSE	0	
Yarkoni Extraversion	2	0.32	0.73	FALSE	0	
Yarkoni Agreeableness	2	0.8	0.45	FALSE	0.01	
Yarkoni Neuroticism	2	0.7	0.5	FALSE	0.01	
Golbeck Openness	2	3.92	0.02	TRUE	0.03	small
Golbeck Conscientiousness	2	0.32	0.73	FALSE	0	
Golbeck Extraversion	2	0.1	0.91	FALSE	0	
Golbeck Agreeableness	2	0.76	0.47	FALSE	0.01	
Golbeck Neuroticism	2	0.92	0.4	FALSE	0.01	

Table C.18: Kruskal-Wallis test on personality affect by Mother tongue showing all significant traits. The first column indicates the method and trait name. The second column (Df) indicates the degrees of freedom. Chisq depicts the Chi-squared value. The p.value is the p-value found followed by the significance if $p < 0.05$.

Column	Df	Chisq	p.value	p<0.05
PI Openness	2	1.44	0.49	FALSE
PI Agreeableness	2	0.4	0.82	FALSE
Yarkoni Agreeableness	2	2.1	0.35	FALSE
Golbeck Conscientiousness	2	1.16	0.56	FALSE
Golbeck Agreeableness	2	1.48	0.48	FALSE
Golbeck Neuroticism	2	3.54	0.17	FALSE

Comparison against ground-truth

Table C.19 shows the RMSE (Root Mean Squared Error) values and MAE (Mean Absolute Error) values of all three fluency groups (Yes, No, and Maybe).

Table C.19: Comparison of RMSE and MAE values of Mother tongue for all three groups. The first three columns show the MAE (Mean Absolute Error) values of the Maybe, No, and Yes groups. The next three columns show the RMSE (Root Mean Squared Error) values of these groups. For both the MAE and RMSE, the lower the better.

	MAE Maybe	MAE No	MAE Yes	RMSE Maybe	RMSE No	RMSE Yes
PI O.	0.10	0.14	0.17	0.13	0.17	0.20
PI C.	0.24	0.17	0.20	0.30	0.21	0.24
PI E.	0.22	0.19	0.19	0.26	0.24	0.23
PI A.	0.32	0.31	0.31	0.34	0.34	0.34
PI N.	0.26	0.16	0.19	0.28	0.21	0.22
Yarkoni O.	0.06	0.13	0.14	0.08	0.16	0.16
Yarkoni C.	0.20	0.17	0.17	0.21	0.21	0.20
Yarkoni E.	0.22	0.18	0.18	0.27	0.23	0.23
Yarkoni A.	0.22	0.18	0.17	0.23	0.21	0.21
Yarkoni N.	0.23	0.18	0.17	0.25	0.22	0.21
Golbeck O.	0.09	0.11	0.13	0.10	0.13	0.15
Golbeck C.	0.12	0.16	0.14	0.16	0.20	0.19
Golbeck E.	0.17	0.16	0.17	0.19	0.19	0.21
Golbeck A.	0.12	0.13	0.14	0.12	0.16	0.17
Golbeck N.	0.26	0.33	0.31	0.32	0.38	0.37