Universität Stuttgart

Bachelorarbeit

A methodology and initial exploration of personality traits of GitHub developers as expressed in GitHub issues

Florian Greinert

Course of Study:	Softwaretechnik
Examiner:	Prof. Dr. Stefan Wagner
Supervisor:	Dr.Daniel Graziotin
Commenced:	January 25, 2019
Completed:	July 25, 2019

Abstract

Software engineering as a collaborative task relies on the fact that people work together. This work can be influenced by personalities of the participating developers and because projects can fail because of individual differences it becomes more important to understand the human side of software engineering. This work will look at the question of GitHub issues can be used to study the personality of its authors and which, if any, constraints may apply. Using LIWC a methodology will be developed to facilitate that research with a firm grounding in personality theory and it's practical usage within the domain of software engineering using the Five Factor Model to construct a personality profile of the type of users on GitHub who open up and comment in issues.

Contents

List of Figures

List of Tables

List of Listings

List of Algorithms

1 Introduction

1.1 Motivation

Software engineering is a highly collaborative process and activity involving many developers, managers and other stakeholders at the same time. Whenever collaboration is a factor, personality of participants can affect the outcome, especially due to the nature of software development practices in which single people can hold tremendous decision making power and many developers work on definitive areas within the software itself. Cooperation and communication between developers, users, customers and all other stakeholders is essential for engineering process, which leads to clear goals for the actual development of the system. In this communication and cooperation, personality of the people involved is an important factor that can lead to failure or success. It becomes important to study and understand the human side of software engineering because of the collaborative nature, and understanding the differences and challenges personalities can provide may lead to a more positive and productive process. The understanding of the human side of software development is currently lacking in breadth and width. One of the biggest forms of collaborative development today is open source and decentralised work via GitHub, where a highly diverse group of people come together to work on projects big and small. In this environment most communication is via text in issues, bugs and other textual channels. As the choice of language use can be seen and analysed as an individual difference [PK99] here and can reflect on the personality of the writer, this style of communication can be used to determine personality traits of the participants. Specifically, the big Five Personality Traits as established by [HH95]. Thus GitHub is an important resource to establish knowledge about the personality profile of developers participating in open source development in general, and about developers using GitHub in particular. Establishing a basic personality profile for developers on GitHub may be interesting as it can be generalised in certain cases and therefore applied practically to understand and mitigate conflicts in development environments that arise due to personality of the involved parties. A general overview over personality psychology and different hypothesis, theories and models that work within the context of a persons personality must be established, and a suitable theory or model needs to be determined in order for this work to have any viable input into the field. To understand what personality, and there the self, is will be important when it comes to choosing a theoretical model for personality. This is important because there are a myriad of theories and models currently being seen as a correct way to look at personality as a phenomenon. And because of the variety of models and theories, the specific frame of this work needs to match with the chosen theory or model. Understanding personalities involved in GitHub projects can be challenging due to the decentralised and asynchronous manner in which GitHub projects move forward and the little to no real face to face interaction contributors have with each other can make it difficult to get a grasp on the personality of participating parties. GitHub issues provide us with an opportunity to linguistically analyse the interaction and to construct personality profiles and traits of those developers. With this in mind, a large enough sample size may yield meaningful information about what kind of person with what kind of personality characteristics use

GitHub for collaborative work and development. On the other hand, the people using GitHub may already be a specific sub-group of developers and there is a possibility that insights about the the developers using GitHub cannot be generalised unto all developers or developers not using GitHub. But to establish this, the nature of GitHub communication needs to be understood. Therefore, this work shall make some first steps into understanding the nature of GitHub issues, how they maybe affect the communication between developers, how they can be used to explore the personalities of those using GitHub and what constraints may apply to this type of communication. Using LIWC [PFB01] to map different usage of word and word stems to the Big Five personality traits this can be done in an automatic manner, thus making large scale analysis viable and affordable. Using GitHub's own API or a database dump, issues can be crawled through and the text in issues can be specifically extracted, prepared and then fed into LIWC. Therefore, the knowledge established by this thesis can be applied by everyone with access to GitHub and LIWC, making it possible to use this work as a basis for more detailed analysis of GitHub issues and therefore developers, or to use this to establish personality profiles for GitHub developers.

1.1.1 Related Work

NLP insights is not a new way to look at things, as previously many others have had the idea to use lexicographical models to find out what kind of people use a particular platform, or how those people change during usage over time. The idea is attractive because the myriad of platforms which are text-only which are used on a grand scale such as GitHub, Twitter, Facebook and Stack Overflow. Those all present different usage and therefore different challenges and demographics. With Facebook and Twitter having probably the most diverse cast of people using it, being almost ubiquitous in the modern media and personal landscape. Specifically interesting for software engineering of course are the platforms Stack Overflow and GitHub.
Previous research in this area included a study on the change of personality of developers in large scale distributed projects [CILV18]. Calefato et al. found out that the role, membership and contribution level have no influence on the personality traits displayed by the developers and that developers with certain traits are more likely to become contributors to an project. They used LIWC [PFB01] (linguistic Inquiry and Word Count) to map usage of different words and word stems to personality traits included in the Big Five [HH95]. Another important finding was that agreeableness and openness are important factors for successful joining a project and becoming a contributor. This paper showcases that information about the personality can be used to predict for example how teams can be structured, to use personality data as a tool to built better teams with balanced contribution to a software project.

Golbeck et al. [GRET11] analysed 279 subject's newest 2000 tweets in order to develop a methodology to extract personality attributes and formulate the personality profile of people that use twitter. The goal is to predict the personality of people in order to use that information for commercial and social purposes, such as targeted advertisement and friend recommendation. They used LIWC as a part of their evaluation of written text by the user for much of the same reason this work will use LIWC.

Another study was done to investigate the personalities of Stack Overflow users by analysing the answers and posts done. They [BHS13] found out that there is a significant correlation between the frequency of posting and the likelihood of the author. As Stack Overflow uses a voting mechanism to elevate answers to a question, Bazelli et al. also found out that users of üp"voted posts are less

likely to express significantly less negative emotions than authors of down voted posts. In this study one can see that it is not enough to simply look at the written word without context, otherwise the information about who expresses more negativity would have been lost.

To the universality of the model chosen, McCrae et al. [MC97] took the American factor structure (measured American personalities with a certain method and took it as a baseline) and compared the same methodology with different languages such as German, Portuguese, Chinese, Korean and Japanese. The results showed similar structures to the baseline one, suggesting a universal personality trait structure within this context. As the languages are in distinct language families and cultures it can be assumed that the model also chosen in this thesis at least covers a great amount of different languages and language families thus making it easier to chose this model above others when looking at GitHub issues, where different languages are used. The difference in culture which does not result in a hurdle to this kind of work is also very important to note, because the model may also be very stable across different cultures as seen in the work of McCrae et al..

Another study [Yar10] studied the personality of bloggers using both a demographic survey and the analysis via LIWC. Their findings include that bloggers do not really differ in their self-expression online and offline. Or, in their words ”The results converge with other recent findings suggesting that, contrary to popular wisdom, people do not present themselves in an idealized and overly positive way online”. This is an important discovery as it lends another layer of validity of this works proposed idea of using GitHub issues to construct personality profiles, as GitHub issues may also be relatively free of idealised self-portrayal and may betray real personality traits of those involved. Another finding worth noting is that some traits of the Big-Five Personality Model where expressed more often and easier in an online environment than others. This adds to the limitation of the model.

Due to the nature of GitHub issues, people write and respond in a semi-personal manner. Holtgraves studied how text messages, combined with relationship status can be used to determine personality profiles. [Hol11] They found out that abbreviations used in interpersonal text messaging were both a function of personality traits and the relationship status. The part about abbreviations having a function may be important when looking at GitHub issues as those ”chats”contributors have in those can vary in intimacy and professionalism and also frequently contain common and technical abbreviations. It may be important to understand if, and when, those issues become more like message chats Holtgraves analysed.

Much more works was done where lexicographical analysis of body of texts were used for different goals; evaluation of the spread of emotions via Facebook status updates, building tools to predict personality on twitter, categorising bloggers personality. The Big-Five personalities pop up constantly and have established themselves as the predominant personality model in a majority of those works for reasons this thesis will get into later in the theoretical chapter.

2 Theories of personality

epigraph

The question of personality tries to answer the question: what or who constitutes the self? And furthermore, how do we measure and quantify this? This question has been answered, with varying degree of success and rigour to the scientific theory, by many people through the centuries. Different models of personality have been developed all over the world at different points in history. This part of the thesis will establish an overview over the history of personality psychology by giving a few examples that were developed during the ages to show how the perception of personalty has changed.

2.1 Hippocrates and the four temperaments

Hippocrates, an ancient Greek philosopher often called ”Father of Medicine”, lived from 460 BC to 370 BC and described personality in four distinct temperaments or personality types. These types are derived and exist in a medical framework called Humorism in which the balance between four fluids in the body influence personality, health and well-being.

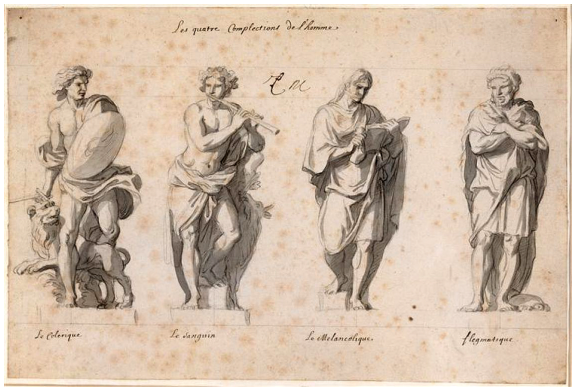

Figure 2.1: Depiction of the four types by Charles Le Brun, 17th century

The different types in 2.1 are choleric, sanguine, melancholic and phlegmatic. According to the model, human emotions, moods and behaviours are influenced by the balance between the four different fluids: black bile, yellow bile, blood and phlegm. (Be advised that the phlegm used in this model has little to no correlation to the medical term phlegm as we understand it today.) If there is an imbalance and the personality displays one set of behaviours, moods and emotions above all others it is called a sanguine, choleric, melancholic or phlegmatic personality. Within this framework, ideally, there would be a balance of all those humors to make a well-rounded personality. In this framework, a mix between the different humours will also result in a mix of those types, sometimes

showing signs of a more melancholic day even if they are sanguine most of the time. This detail, that there exists a mix of personality traits, is an important observation: the idea that a persons personality is not describable by a single adjective or trait.
Abu Ali Sina, known in the west as Avicenna, furthermore attributes

> emotional aspects, mental capacity, moral attitudes, self-awareness, movements and dreams."[Lut02] to these temperaments. Although there is little to no scientific backing to the groundwork of this hypothesis if the four fluids [Wil13], the idea of what a personality is made of can be traced back to those ideas of the past even if the framework enabling this hypothesis is pseudo-science at best. The finer points and differentiation of the four fluids and how they make up the personality is not needed, but they are intuitively understandable: Sanguine people are full of life, phlegmatic people tend to be slow and lethargic, choleric people are easy to excite and arouse and melancholic people tend to be oriented inwards and prone to brooding. This behavioural view plus the addition of Avicenna make for a compelling construct of personality that can be understood by even laypeople.

2.2 Wu Xing

Another example of this kind of scholarly hypothesis pursuit is the Chinese concept of Wu Xing (meaning moving star or planet) in which five different elements and their interaction govern attributes, interactions, time cycles and more. The concept is applied in a manner much more encompassing than for example the four temperaments. From personality, medicine to martial arts and other applications the five traditional elements are supposed to represent a fundamental cosmic truth.

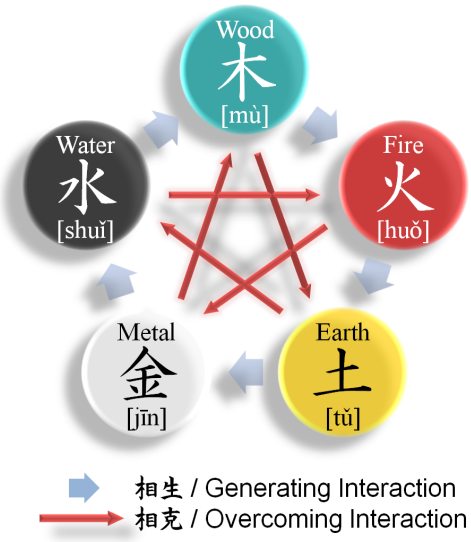

Figure 2.2: A depiction of the five elements and their interaction

The cycle of blue arrows represent nourishment and would, if applied to personality, mean that personality traits displayed by a personality that is predominantly Wood would feed into personality traits associated with Fire. For example, if Wood meant patience, strength and flexibility then displaying those traits will lead to also developing and enjoying traits that are associated with fire such as warmth, strength, endurance and joy.
The red arrows represent kind of like a rock-paper-scissors approach to personality make-up. The traits that are considered bad that are associated with fire such as rashness and restlessness can therefore be overcome and extinguished if you will by the good personality traits associated with water such as flexibility and wisdom. Personalities may be described as full of wood and like fire and as a popular shorthand for description of people it still has its uses today.
Even if the theoretical underpinnings of this hypothesis is as grounded in reality as the five fluids idea, the insight and ideas about how a personality is formed have merit. Especially the idea that certain traits temper the bad parts of a personality is a form of interaction that is intuitively understood. It also supports the idea that a personality is a complex thing, made up of a lot of moving parts that are neither static nor isolated. The interactivity between parts of personality as a form of push and pull between ones-self, and the overcoming of the weaknesses by the good parts of the personality, is seen as a virtues process. Being so all-encompassing and generalised (as seen that the cycles are applied to military strategy and martial arts in the same way), the model is also hard to understand and apply for people not familiar with the cultural background of it. Nevertheless, it is an interesting stepping stone for personality as a subject of study and is therefore mentioned, at least briefly, here. It also is important to state that the idea of personality and it's study is not a western phenomenon and has been done all over the world.

2.3 Sigmund Freud's theory of personality

One of the most famous ideas of Sigmund Freud is the hypothesis that the personality, your very self, is made up of three distinct, interactive active constructs: Id, ego and the super-ego.

Meaning it, self and above-self respectively, their function according to Freud can be summarised:
The id consists of drives that humans have no or very little control over. Instincts, desires like food, sex and thirst. Fear in the presence of danger would also fall into the control of the Id. [Che18] One might say that this represents what an organism needs, and it drives the person to obtain these things. It is present from birth and is a given as it is instinctual. It is unconscious (see the metaphor of the part of the ice berg that is submersed, but still part of the personality) The Id wants to fulfil the desires now, and indulge in it [Thu09]. Given a set of instant gratification versus delayed gratification, it will, when left unchecked, always chose the instant gratification. Newborn children display that behaviour well; they simply cannot fathom the concept of delayed gratification yet and therefore try their best to fulfil the desire now.

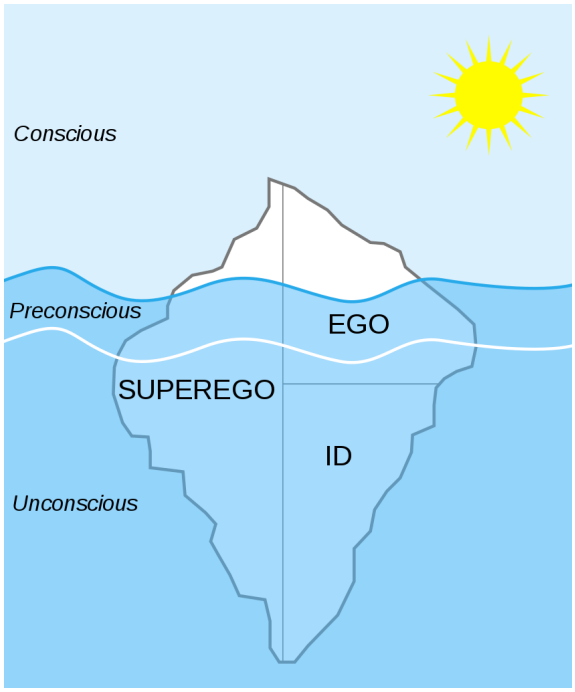

Figure 2.3: A common visualisation of the concept

”The id is inseparable from the unconscious – id wants and desires in the here and now, it doesn’t make plans for the future. Freud often claims that the unconscious (which is the same as the id) knows no time but the present, no answer but Yes.”[Thu09, p.82]

The ego, the self, is the part of the personality that is responsible to manage the cravings and impulses of the id in the immediate world around the person. [Thu09] A person that always indulges in their basic instincts is not a person that can live in a society. Or at least will have difficulties doing so. Therefore, due to social norms and practises, the ego limits and controls the impulses of the id in a manner which allows the organism to enjoy the advantages of the society.

”The ego, on the other hand, recognises time and the setbacks which go along with living in a world where one has to wait.”[Thu09, p.82]

In short, it manages the needs expressed by the id in a manner which is acceptable. The delaying of gratification is an important tool for this construct as it allows the id to indulge, but on a schedule appropriate. The ego also uses fantasy as a way to manage urges and impulses, for example instead if eating right now a person can imagine eating and therefore make the waiting until you can actually eat more bearable. [Che18]
The super-ego, or above ego, is the construct responsible for morals and cultural standards. [SGW11] It aims for ideological perfection and wants to conform to every one of rules it values. The ego uses the super ego to judge needs, impulses and shapes them with the super ego in mind into behaviour that is acceptable to both the id and the super-ego. It is not completely inaccurate to think of the ego as a judge in an eternal competition between id and super-ego. Or the fear of what happens when you break social norms, rules or anything like that.[Thu09] Freud argued that the parents have a

lasting and very important influence of this part.
The idea of opposing forces inside a personality is compelling, as is the idea of needs being moderated by societal rules and norms. The idea that outside influences directly influence the personality is novel and seems intuitive. The more ore less conscious processes that lead to a decision being influenced both by the very need for satisfaction and the appropriateness of said action that needs to be taken in order to fulfil that need. No personality stands alone, and to try to explain that furthers the understanding of personality as something complex and contextual.

2.4 The arising problems

Looking at some of those hypothesis, one feature missing in most of them is quantifiable, measurable standards. How would Freud go about telling a person how and what their Ego, Id and Superego are doing? How can they tell how wooden or metallic their personality is? Vague ideas of psychological constructs and unscientific hypothesis around personality and the self are interesting in a philosophical way, but there is nary a chance of actually using them for practical purposes. To even begin to try to describe a person not known by one of those historical models there would need to be a lot of observation, interviewing, self-reflection on the part of the observed, questionnaires and more are required. The very idea to use this knowledge for something deemed productive instead of doing it for non-productive reasons is harmful, but what this work means is that the personality traits and behaviours should be applicable to a software engineering context. This does specifically not mean that any use outside of industry and engineering is a lesser use, only that in this instance the perspective of HR, engineers and other people in a collaborative software engineering process is paramount. While interesting, this cannot be done in a economic manner and therefore would be not very helpful for this thesis. Even so, the very core concepts all of those models have in common will carry over to the more scientific models to come in the following chapters so much that a clear evolution can be seen, which makes the inclusion of those core ideas important.
A lot of hypotheses were drawn up over the ages with varying degrees of complexity. The cause, or why a person behaves the way they behave, has changed from fluids inside your body, to elements and then more or less conscious constructs of the psyche. The last seems more scientific then the rest, and it very well may be, but it is just as unquantifiable as the others presented here. Therefore, we do not need a hypothesis of personality but a model of personality. A model which is both scientific, measurable and grounded in other scientific thought. Hence, in the 20th. century, two big theories of personality emerged: trait theory and the social cognitive perspective.

2.5 Trait theory

Are people basically selfish? Some are, some aren't. Are human beings intrinsically creative? Some are, some aren't. *Personality in Adulthood, McCrae, Costa*

The common way in English to describe a personality is kind of a naive version of trait theory: we ascribe different words to people with the idea that this trait is encompassing enough to describe a wide variety of things. Aggressive, for example, has different implications what it means. Aggressive has a myriad of interpretations associated with it making it very useful as a descriptive adjective. So if Aggressive may be a trait, what exactly is a trait in this theory? According to McCrae and Costa, traits are

> ”dimensions of individual differences in tendencies to show consistent patterns of thougts[sic], feelings, and actions”[MC03, p.23]

Now the previous example of aggressive being a trait is better explained. An aggressive person has much more different set of tendencies, patterns of thoughts, feelings and actions than a person one might describe as timid. The difference now is easily seen, and moving forward, this is the definition of trait this thesis will work with. In the literature, this is also called a phenotypic definition as it will tell you how to identify a trait. This is important, because if we use such a definition as the basis for our model and theory, we are describing personalities and do not ask why they are the way they are, which is just as important as to understand why personalities display those traits to begin with. As trait is multifaceted, it is also possible to determine the degree to which a certain personality is aggressive or shy. This is helpful and matches with a naive, intuitive way of understanding personalities, as no person is completely aggressive or always shy. The nature of a model or theory that is based upon traits which are defined as above results in one very important implication:

thoughts, feelings and actions can be measured and quantified more easily than something like Freud proposed. Furthermore, the motive or reason why you have those thoughts, feelings and do those actions can be, for now, ignored or shoved aside which has practical consequences. We can shove those aside because for a perspective of how people behave and what the difference between that one personality and the others is, the reason or motive why that particular personality has those feelings or thoughts is not at all important, at least within the trait theory approach. Another point made in the trait theory is that all people display the traits identified within a spectrum. [MC03, p.25] Which means, there is a continued line from shy to bold, for example, and every personality lies at some point on this line. (Be advised that shy and bold are not real traits within this theory and only serve as an example.) When thinking about a specific type of person one might describe as aggressive, this only means that this personality is an outlier to the usual normally distributed personality traits. [MC03, p.25] Which means that personality has more of that trait which means it can be easier observed because more of the behaviour associated with that trait is displayed, more thoughts of that nature are perceived. Thus, *frequency* and *intensity* of the actions are an important data point for determination of a trait and therefore personality. Those two concept will become more important later again.
When discussing personality traits, it is prudent to think of those as tendencies, not absolute commands or predetermined outcomes in social situations. A personality that can be classified as shy would not always be so but depending on circumstances, social environment and recent experiences, may behave bold in any given situation. It is just that that personality has the tendency to be shy. [MC03, p26] It can be thought of as

average behaviour, and just like the mathematical function of average, may not say that much about any specific social interaction or number. So a shy person has the tendency to, on average, behave more reserved and closed of in social situation. Also, it is important to understand that there can be as much variety within each person as there is variety between different people. [MC03, p.27] A more poetic way to think of the distribution of trait levels between people and within people is the following:

> Summer is always hot and winter cold, but there is a wide range of temperatures within each season. [MC03, p.27]

One important aspect of those traits is consistency, which means those observable differences should be stable over a long period of time and not caused by circumstance, immediate environment (such as, for example, a trip to Saudi Arabia), flights of fancy and stress from other parts of the personality's life. One extreme example would be a person that it outspoken, gregarious and very prone to laughter person at a musical festival. That person partakes in drinking alcohol every day from breakfast to evening. Meeting that person in their normal day-to-day life then yields the observations that the person is reserved, anxious and closed off normally and changed their behaviour drastically when being confronted with music and mood altering substances such as alcohol. The behaviour in the festival environment with different circumstances than normally are the anomaly, the outlier to their traits they more often display. That behaviour should not really factor into their personality profile constructed by trait theory as it is like a numerical outlier: not important to the average and if sufficiently rare, does not indicate a tendency. [MC03, p.27] *Now this also tells us one constraint with regards to analyzing behaviour in any way shape or form and that is that one instance of observed behaviour is essentially useless when trying to construct the personality profile of the subject.* And *We do not need to understand how and why traits form in order to analyze people and construct personality profiles.* Two points that will become very important further in this thesis.

2.5.1 Identification of traits for the theory

The naive approach to traits, via natural language, is present in most spoken languages today and therefore is easily understood. [MC03, p.25] What this means is that the way trait theory tries to conceptualise and describe traits is via natural language processes, such as they are aggressive and assertive. The idea to, at least to some part, agree to that way of describing and assessing personalities is that we as humans have done that for at least as long as language exists and therefore there can be arguably some merit to this way of doing things. Of course this way of describing is highly practical as it is understood by most humans. (They arguable is a lot of leeway on how assertive may be interpreted, but the part of "they are *trait*"wont confuse most people)

> So important are descriptions of people that thousands of terms have evolved, distinguishing minute shades of meaning that delight the poet and confound the empirical scientist. [MC03, p.29]

Which now illuminates one of the fundamental problems with trait theory: what are we actually measuring which is important to a certain context? What are the important traits within a profession, if there are any, what are the traits that are important for a significant other? And how to measure all this? It seems overwhelming and not clearly defined. As seen in the next chapter, Carl Jung described a scale on the basis of Extraversion versus Introversion, with other personality traits aligning on that axis. More on that later. This is one of the most popular scales invented and is used to this day, with other addition, in the model called the Myers–Briggs Type Indicator. Working within the Introversion-Extraversion axis, Guilford and R. B. Guilford found that different traits were existent along that axis, supporting its validity and use. Guilford went on to developed a personality model that had 10 traits in total called the Guilford-Zimmerman Temperament Survey.[MC03, p.30] Another personality researcher, Hans Eysenck, concluded via factor analysis that Extraversion was one of the two fundamental personality dimensions found in many different models and hypothesis. The other being something he identified as emotional instability or maladjustment and called Neuroticism because apparently people who were seen and diagnosed as neurotics displayed those traits.[MC03, p.30]

So far we have two scales with four different end points identified: Extraversion vs Introversion and Maladjustment vs no Maladjustment. As is seen later, these two axes will reappear in the model used today in the most popular model of personality derived from trait theory. Thus lending legitimacy to the underlying theoretical model used by that theory.

This approach greatly decreases the amount of knowledge needed to understand a certain verdict or personality profile. Using other models, such as the ones introduced earlier in this thesis, are not as easily understood out of the box so to speak. Freud's id, ego and superego are important but not easily understood and not are not supposed to be. They are not meant to be used by laymen. One could argue that they are also not meant to for a practical framework that is not psychology or psychoanalysis. The natural language approach therefore already has several advantages that are plain to see. MacCrae, in his book Personality in Adulthood, further states that due to the nature of personality and it's importance to culture, play, tradition and all parts of live, surely lead to natural languages around the world being able to describe the full spectrum of personality differences.[MC03, p.34] Time and interaction over long periods of human history made sure that, at least as practically as possible, that most traits are accounted for. While that is not a scientific accurate method of establishing his argument, it is intuitive and hard to argue against, so for this work we assume he is correct.

Now this leads to another problem: if natural languages are indeed fully able to describe all traits a human might posses, why are there so many words that have the same meaning or are synonyms? As previously stated, slight nuances and shades of differences are indeed something most natural language have when it comes to descriptions. Why this is the case, this work cannot begin to comprehend, but what can be said is that in order for the trait theory to be useful in a practical context there need to be some form of pruning all those words down into a manageable dictionary for people to work with otherwise one might get lost in minutiae [MC03, p.34] like thinking really hard about the difference between considerate and thoughtful for a given personality.

After establishing a dictionary that can be worked with, R. McCrae and Costa Jr. settled on a three factor personality model with two previously mentioned factors: Neuroticism and Extraversion. With a third they called Openness to Experiences. [MC03, p.34] Names for Neuroticism and Extraversion where both taken from Eysenck's research because the traits they both identified where incredibly similar for all intends and purposes. During their research McCrae and Costa saw limitations of their own devised system as some traits that are arguable real were not present in their model, one example being self-discipline. However, one can never be sure what is lacking in a given model, hypothesis or theory.
Then finally, R. Goldberg constructed a five-factor model of personality by using the natural languages again to start from the beginning. [MC03, p.35] This model had a lot of similarities with McCrae's and Costa's model and they found out that indeed, their three factors are represented in this new five factors model. Both their initial theory and the new one where strengthened by their similarity and now a considerable argument about the five factor model's completeness could be made. And indeed, with scales and measuring methodologies being made for all the five traits, a comprehensive model of personality was published under the name of NEO-PI, the NEO Personality Inventory with five traits present: Neuroticism, Openness, Agreeableness, Extraversion and Conscientiousness. [MC03, p.35] This model will be explained in more detail later in the next chapter which looks at two different direct models, both evolving from the hypothesis and theories mentioned before.

3 The need for a model of personality

Theoretically, in order to analyse and differentiate different personality types, one would need to have some kind of basis to determine what a personality is. The fundamental theory this thesis is operating under was introduced in the previous chapter as Trait Theory. The idea if personality and history associated with the search for a way to describe it is old. When asked about people, one might respond with a myriad of adjectives and sets of behaviour that person displays such as angry, solemn, funny, quiet, boisterous, annoying and many more. The question if there is one defining character trait of a person is hard to answer, as people and therefore their personalities are complex, conditional and contextual. If a person is always rude to you, that does not necessarily mean that person is rude to all and every time. Therefore to categorise that person as a rude person would be wrong. In fact, rudeness might only be a very small part of its personality. The difference between what you perceive, a rude person, and the other person is, a person with a different idea what would be considered rude, could be the point of conflict as well. Furthermore, rude or friendly is more of a description of behaviour, which can be part of a personality, but is not really a personality trait itself. One might argue that the reasons that lead to a person being friendly are more in line of what one might think of when talking about personality and personality traits. Those are some of the reasons why personality is not an easy topic to breach, and why a model and the understanding about the model is needed for this kind of work.
Adjectives to describe a person are also limited in the sense that not all types of internal processes, internal or external behaviour, thoughts and values can be described with a simple adjective that is easy and intuitively understandable. What does funny or melancholic mean? How universal is the understanding of those two words? Also, neither of those words can really be used to describe a person for the purpose of software engineering. Is a funny engineer better suited for some position than a melancholic one? Is melancholic or funny even an attribute that is important in the context of working together? Are there attributes that are much more relevant to the software engineering process? Ideally, the chosen model would use adjectives and personality traits that are easily applicable for this framework of software engineering. The aforementioned Trait Theory operates under the assumption of a Five Factor Model with all behaviour, private, personal and professional modelled.
There needs to be a system or model that could accurately describe a persons personality, or traits that form the personality, which is robust and holistic. Holistic in a sense that most behaviours could be explained by those traits of personality. Furthermore, the traits named by the model should be meaningful in the context of software engineering because this thesis seeks to explore that particular field. There needs to be an easy way to get the personality of developers on GitHub that can scale. Without proper processing the model is useless, so we need to find a model that can be processed for human consumption and consultation. That means one might need to visually present

it in a way that can be useful for human resources, project leaders or other people that want to use the information for practical purposes.
Two popular models will be explained in the following part, beginning with the popular MBTI (Meyers-Briggs Type Indicator) that many people use online for self-description. The second, FFM or Big-Five, will be explained in much more detail as it is a natural outgrowth of Trait Theory and will be the basis for personality analysis in this thesis.

3.1 Meyers-Briggs Type Indicator

The Meyers-Briggs Type Indicator, or MBTI, is a model of personality developed before and during the second world war and is based in the writing of Carl Gustav Jung. His idea of personality is based on the dichotomous relationship of two different poles of cognitive behaviour of personality: rational functions such as thinking and feeling, and irrational functions like sensation and intuition. The difference between the two can be seen as the difference between what you do consciously and perceivable, and what is experienced. To make the point clearer, one might imagine a wrong feeling in the stomach in response to a situation that may seem dangerous. This feeling of danger is not a conscious process the person follows as they assess the situation, think about ramification and then decide to feel the danger. They simply have the intuition to feel the said danger.
The other hand of the functions are the rational ones, those you can explain and perceive more clearly and may even know where those feelings and thoughts come from. The MBTI relies in self-reporting questionnaires primarily because it constructs personality type by inferring it from observed and reported behaviour. The MBTI does have valid claims as a personality model, but it cannot hold the promise to accurately encompass all of a persons personality in its 16 types. And its usage and utility, for example a corporate setting to determine work related performance indicators, can be questioned. [Pit05] And as this study uses the premise of using the data for software engineering purposes, this leads to a devaluation of the models worth for this thesis.

As seen in 3.1 each personality is constructed out of two halves that are exclusive to each other: thinking vs feeling and intuition vs sensing. A thinking person can then be either intuitive or a sensuously, so to speak. A feeling person can also have either. Within the possible combination a personality attributes then introverted or extroverted to all their types, resulting in someone being for example, an extraverted (or extroverted) sensing introverted thinking type, resulting in a complete Myers-Briggs Type of ESTP. The possible combinations and types can be seen in 3.1.

It also provides a list of famous people who supposed to be one of the types from that model. As it also indicates what that person values (for example, intellectual vs emotional pursuits) it can be very hard for people to self-report correctly because they of course want to be the personality type that values the things they value, a problem that is called socially desirable reporting. Therefore, a bias can show itself. Though, in a setting which low stakes such as self reporting of personality, this is not a big problem. But should be kept in mind. [PV07] Another problem that arises from a model using this way of reporting, is the scale of effort needed to apply it to communities that are

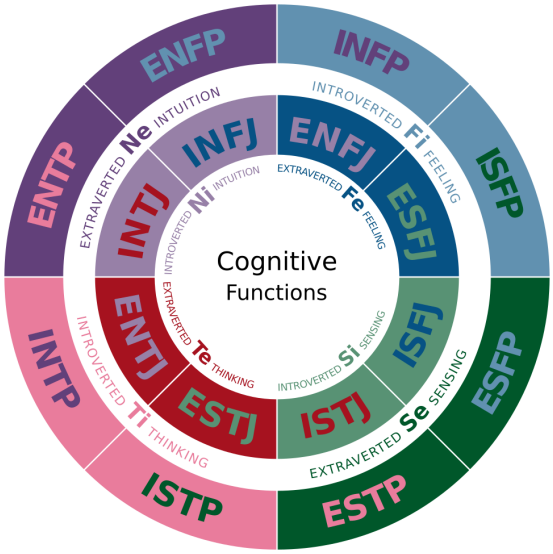

Figure 3.1: Myers-Briggs Type Indicator types and functions

large such as GitHub, Stack Overflow or the likes. To use this model in the type of work this thesis wants to accomplish would have been hard and relied on people to respond. Such a study comes with it's own set of problems and challenges and cannot be reliably automated within GitHub, which is a high barrier for using this model. Therefore, another personality model was used instead. Therefore, this part will not go into further detail about how this model works.

3.2 Big-Five

The predominant personality model many related works presented use is the so called Big Five-model or Big Five personality traits, short FFM. The usage of this model [MJ92] ranges from academic to application for human resources [HH95] [BM91]. It is popular and does not have to rely on self-reporting, which in turn is more reliable than self-reporting using approaches such as MBTI, which is much more popular among people for self-description. [Pit05]
It can be based on the words people are using to determine which type of personality traits one might display. [PK99] This in turns means that using lexicographical analysis it is possible to determine the personality traits when the only thing we have from that person is written texts, blog posts, status updates from example Facebook or GitHub issue texts. To put it simply: people who display a certain personality dimension more often use different words more frequently than people with another dominant personality dimension. Dominant means, displays more words associated with that dimension. [PK99] It simply says people using this kind of words more often display more traits associated with openness for example. This way, a hexagonal traits map can be built which means the characteristics can be shown in a more precise and nuanced matter.

Therefore this model is a good pick to use for the type of analysis this thesis is about because the feasibility of using GitHub issues for this kind of analysis is much easier when using a widely used model that can be applied without as much interaction between the person analyzed and the person analyzing. A summary and explanation of the personality traits and dimensions will follow.

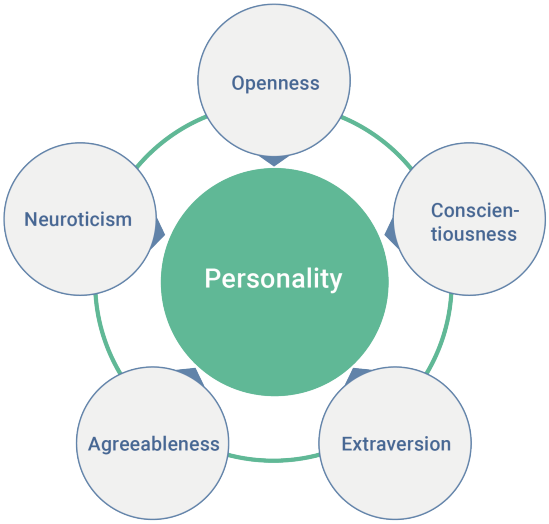

Figure 3.2: The Big Five personality dimensions

The five personality dimensions as seen in 3.2 are not single personality traits as is, but mean a whole lot of different behaviours bundles into one of those dimensions as I will call them. One might think of those dimensions as fields with different behaviour, preferences and traits that can be distinguished from other fields that display different sets of behaviour, preferences and traits. A person has all these fields in different levels, thus establishing their personality as a whole. Through this work, field and dimension are therefore be be considered interchangeable.

3.2.1 Traits

Previously in 2.5, five personality traits were mentioned and it was explained how they came to be. It should be again noted that those traits are not descriptions of deterministic behaviour but indicators that show tendencies over long periods of time. A personality that scores high on Openness might not be open on 12 different encounters in a row but over the span of a year will be. That means a singular data points or observations will not be enough to statistically make an assumption about the personality in question. Also as previously stated, within the trait there are different levels of the trait in question present being called low scorers and high scorers from here on out.

Openness

Generally speaking, those kind of personalities will be more creative, will seek out new experiences, make novel associations between even distant subjects. They are open to self-examination and have fewer problems with new roles, tasks and such assigned to them. Individuals with a high Openness to new experiences as described by the FFM are not indifferent with doing new and novel things, they are constantly seeking new things out and this behaviour is active and not passive. [MC85].

Conscientiousness

Colloquially also called diligence. Individuals who score high on this field may display behaviour that makes them reliable to others. They take obligations to others seriously, display desire to complete tasks well. They could be described as organized and efficient.
A lot of behaviour in this field is desirable for workers, managers and students. [HPPL07] For example, a problem with procrastination can be observed more often by people with low score in this field. [DS02]

Extraversion and introversion

Extraversion and introversion are two poles of behaviour when it comes to interaction with the outside world. Generally speaking, people who enjoy human interaction in and of itself, present themselves as sociable and comfortable in group settings tend to be extroverts. While personalities who act more reserved, enjoy solitary activities more and are more reluctant to join into large group activities can be described as introverts. Extraversion and introversion are not to be confused with not being shy and being shy. Both poles simply describe what a given person might find more rewarding when it comes to activities. These two poles are very popular to sketch a personality in colloquial speech but which descriptive power goes much further than the description used in the FFM. Using the FFM extroverts and introverts are not holistic personality types but indicate habits, traits and behaviour.

Agreeableness

Agreeableness influences the individuals willingness for cooperation and to be selfless in group settings. It helps with identification with others. [JBH97] Personalities that score high in this field display traits such as trusting, altruism, compliance, modesty and sympathy more often than individuals who score low. [Mat12] Generally speaking, an agreeable person fits easier into groups, causes less friction between them and the group, fits easier into hierarchies and tends to try to preserve the peace of the group while being honest and straight-forward.

Additional definition of traits		
Trait	Low scorer	High scorer
Agreeableness	Critical, skeptical, shows condescending behaviour, tries to push limits, expresses hostility directly	Sympathetic, considerate, warm, compassionate, arouses liking, behaves in a giving way
Conscientiousness	Eroticizes situations, unable to delay gratification, self-indulgent, engages in fantasy, daydreams	Behaves ethically, dependable, responsible, productive, has high aspiration level
Extraversion	Emotionally bland, avoids closer relationships, over control of impulses, submissive	Talkative, gregarious, socially poised, behaves assertively
Openness	Favors conservative values, judges in conventional terms, uncomfortable with complexities, moralistic	Values intellectual matters, rebellious, nonconforming, unusual thought processes, introspective
Neuroticism	Calm, relaxed, satisfied with self, clear-cut personality, prides self on objectivity	Thin-skinned, basically anxious, irritable, guilt-prone

Table 3.1: California Q-Set Items Defining the Five Factors by McCrae, Costa and Busch (1986). Consult the next part before interpreting this table as some of those behaviours can invoke unnecessary value judgment in readers not familiar with the ideas presented there. Data sources from [MC03, p.56]

Neuroticism

The idea of a personality like this goes back to Galen of Pergamon, who categorised personality types by the mixes of four bodily fluids/humours. He called individuals whom the FFM model would categorise with a high neuroticism score melancholics. His characterisation of those people being fearful, anxious or sad is not that different from FFM even if his reasons for giving them these categories were completely incorrect. Personalities which score high on neuroticism experience feelings like anxiety, fear, anger, envy, jealousy, guilt, loneliness and worry more often. It is also much more likely for them to feel a depressed mood. [Tho08] Generally speaking they tend to view situations as more threatening than they might be, worry more and will be self-conscious.

3.2.2 Judgmental view of traits

Before going deeper into the Five Factors Model and how to measure it, there must be some words about judging and rating the different traits present. These traits must be viewed outside the good-bad dichotomy or they lose a lot of value. The five traits describe behaviour that must not be organized with those two terms, so be advised to view them separately and as distinct patterns of behaviour, thoughts, emotions and values. The prime example, right down to the name, is Neuroticism and it's description. People can easily fall into the trap of looking at it as bad the moment they understand what it means. However, as they are completely distinct from one another, a personality scoring high in Neuroticism score does not lose points in Openness, Agreeableness and the other traits. They are not more likely to be closed off introverts, which is seen as less desirable as being an open extrovert, than a personality which scores low in the category. To ascribe good or bad adjectives to the traits would be a serious mistake, as would be viewing them in a non-distinct way. Especially the non-distinct view of the whole trait structure would hamper its usefulness for describing personalities completely.
Therefore, it is important not to interpret additional characteristics unto the traits to keep the model clear and working. To illustrate, some more examples: Introverted people are not hostile, only reserved. They are not, by definition, bad at being in social outings, only do not want to be there that much. Just because a person does not want to be there or prefers different kind of social settings does not make them a worse personality that is depressed or always gloomy. It also does not indicate some form of anxiety or other psychological irregularities. A low score on the Openness to new Experiences scale are also not grumpy malcontents every time they see a new vegetable on their plate, that would be an unfavorable reading of that trait indeed. As they are less likely to show violent emotions [MC03, p.51] which is indeed not something bad either, even if they would prefer to experience new connections, relationships and excitements in a context they are familiar with smaller differences than to things they already know. To finish this warning, be not mistaken to think that good and bad are universal. So even if one might be tempted to ascribe good and bad to those traits, what good and bad is varies from culture to culture, person to person and country to country easily. To be on the safe side, no judgment should be done at all if possible. McCrae and Costa concluded as well that well-adjusted and extraverted people rate themselves as most happy and satisfied people. [MC03, p.51] but using happiness and satisfaction as the be and en all of a personalities goodness is problematic. Sure is that those traits pop up all the time, and people of all kinds of positions, social standings and occupation display a wide variety of those. Meaning that they are not being bred out, they kind of stand the test if time as well as any other thing humans are and do. Over the course of human history, really undesirable traits had the chance to die off yet they did not. Traits and their consequences in tendencies, emotions, thoughts and behaviour are as diverse as people, and all kinds of people are necessary for a society to function. Not all people can be loners, but loners that work well alone are still extremely important for certain tasks. Conservative people (not the political alignment) who stick to good solutions in the past are just as needed as those that will try out new things, so the value of low and

Alternative systems classification of traits in the FFM					
System	N	E	O	A	C
GZTS	Emotional Stability* Objectivity* Friendliness* Personal Relations* Masculinity*	General Activity Restraint* Ascendance Sociability	Thoughtfulness	Friendliness	Restraint
EPQ	Neuroticism	Extraversion		Psychoticism*	Psychoticism*
MBTI		Extraversion	Intuition	Feeling	Judgment
CPI	Realization*	Internality*			Norm-favoring

Table 3.2: N: Neuroticism E: Extraversion O: Openness A: Agreeableness C: Conscientiousness GZTS: Guilford-Zimmerman Temperament Survey EPQ: Eysenck Personality Questionnaire both mentioned in the previous chapter when discussing the development of Trait Theory. MBTI is the other personality model discussed before the FFM above in 3.1. CPI: California Psychological Inventory, a model and personality inventory not discussed in this paper.
: marks traits that whose opposites are measured. For example, Friendliness measures the opposite of friendliness. Data sourced from [MC03, p.56]

high scores on those scales can be seen. Robert Hogan argues that advantages gained by individuals also account for differences in trait and trait levels so evolutionary those are indeed all things needed for a societies survival.[MC03, p.51]

3.2.3 Validity of the FFM and relationship with other models

Looking at this table, we see that the FFM is indeed quite encompassing. It also is much less verbose than some other models. As to the comprehensiveness, according to McCrae and Costa, they tested their version of the FFM against the presumed more granular California Q-Set Items Definition in table 3.1 shown above. Due to the fact that it came out before the FFM as described was completed. [MC03, p.52]. They did this the following way: 403 people used the California Q-Set Item Set Definition for self-description, and afterwards compared them to their own Five Factor Model. They found out that a lot of those were identical to their own factors. Thus boosting both models validity. The table above shows a detailed comparison with other models and what their versions of the five factors are called.

3.2.4 Quantifying a personality

Fortunately, as it is tendencies that we are measuring, one needs to look for behavioural patterns, or in this case word usage, over time. Personalities can also, as mentioned before, vary to the degree in which they do show the trait via word usage. Therefore a single personality may showcase a range of scores when it comes to the different traits.[MC03, p.38] Going into this problem there are three ways that measuring a personality was done: Self-reporting, observer-rating and observation. To explain: describing ones-self, being described and being observed. For this thesis and the framework of GitHub issues, the first two will be completely unpractical. Specifically with the amount of effort required to inquire the developer themselves and finding people who can even describe the developers in question. As this would be a logistical nightmare, the third option naturally is the only one with a chance of success. Also the only option where no input from a person, observed or observed-adjacent, is required. The validity of such an observation can be called into question, however, that natural language use is tied to personality is a clear and focal point of this thesis and will not be argued. [PK99] Furthermore, related work [BHS13] was shown that there is no need for input from the author of corpus of text to be analysed for it to be meaningful. This supporting evidence that the approach detailed in the next section is a good way to tackle this thesis's problems and questions. Thus, there will be strictly an analysis of the usefulness of only using the GitHub issue's text from the users to construct a personality profile with not further interaction needed from the author. This results in a much easier method which can be automated very easily. It also completely eliminates self-bias in reporting and there will be little to no administrative overhead. Furthermore, by using a process already present in a lot of related works (either as the sole tool or as a set of tools used) this work will not break completely new ground with it's tools and theory, but concentrates on the validity it's singular tool usage in a chosen domain, that is GitHub issues. The historical development and ultimate end point (as of now) for the chosen underlying theory of personality, Trait Theory, in the Five Factor Model (FFM) have been explained and shown, resulting in a convincing argument for the given model and it's traits in this work. The forgone chapter is the theoretical underpinning of the method chosen in the following chapter, and therefore serves as a comprehensive reminder that both the model (and where it came from) and tool chosen are a credible

way to tackle the question of how we may analyse GitHub issues for personality and personality profiles for the purpose of finding a way to use that information to further support a software engineering process and support developers, teams and communities to better their methods, processes and ultimately their outcome as a team, company and group.

4 Methods

This chapter will deal with methods used to analyse and ultimately provide some discussion about GitHub issues as a source for personality research. The way to study GitHub issues is either through a tool to pull issue text, names, authors from repositories live or through a dump of the database, which is magnitudes more practical as the immediacy a live tool provides is not needed in this kind of work. Therefore, a database dump was downloaded and used to look at the structure of GitHub issues and the surrounding constructs such as comment, authors and repositories. This chapter is meant to explain the methods used to facilitate an initial exploration of GitHub issues in the context of Natural Language Processing. This chapter is therefore sectioned into the different domains of this problem: LIWC, the tool used for analysis. GitHub issues, the subject being analysed. The database, from which the issues are sourced and the correlation between LIWC and the FFM. And finally the method to pull it all together into a workable work flow by both manual input into LIWC and a simple tool that processes the data for each interaction.
Each of those sections will deal with problems within that domain and how it influences the results. Later in the discussion, different solutions and approaches will be presented in order to deal with those problems.

4.1 LIWC

Linguistic Inquiry and Word Count [PBJB15] is a program developed for analysis of text for different properties in categories. Those categories range from pure analytical statistics such as word count, words per sentences to word groups that deal with social, cognitive and biological processes. When analysing a text it will show you the total percentage of all words in that corpus which are from a certain category. The valid input ranges from .txt-files, .rtf to .pdf-files with computer readable text. It is quite user-friendly and easy to use while also providing meaningful analysis. Their own manual states that most abbreviations are not a problem, as are misspellings. However, that only rings true when we have large amount of text to input for that particular unit of analysis. Many of the related works used this analysis as-is, with their own dictionaries (which analyses different words for different categories specified by the person constructing the dictionary) and in tandem with other lexicographical techniques. In of it self, it is a program which tags words with certain categories that do not really mean anything by themselves; they are simply descriptors of the words context so to speak. To ascribe meaning beyond that is the challenge of the user.
By their own admission [PBJB15] common internet notations can be a problem if they are a large percentage of the corpus. Those range from email-addresses to URLs, Hashtags and Twitter handles. However in the Manual provided for LIWC it is said that

> ”Note that many types of “net speak” that is used as shorthand interpersonal communication (e.g., “lol”, “4ever”) are captured by the LIWC2015 dictionary and do not need to be altered in your text files.”[PBJB15, p.16]

so this eases some of the apprehensions one might have with using LIWC for a task such as analysing GitHub issues that are, by their nature of a decentralised online platform that can be quite informal, rife with both abbreviations and internet notations. But there are words and abbreviations (mostly those that contain symbols not commonly used in words) that still provide a problem:

Typed entry	Change to
w/	with
b/	between
&	and
‘cause	because
and/or	and- or
‘an or ‘n	and
mos	months
sec	second
@	at

Table 4.1: Some of abbreviations in net jargon that still provide a challenge to the LIWC system and should, if possible, be changed to another word before analysis.[PBJB15, p.16]

As seen it 4.1 those entries will inadvertently influence the result in a negative way, which means it will be less accurate. For this work, those entries will not be changed and thus will be a drag on the accuracy. Apart from that, LIWC and it's function analyses common words so some misspelled scientific names or incorrectly used technical words and jargon aren't that important. Remember that the personal word usage is an indicator of personality but it is not that important if, at some point, they are misspelled as we do not try (and indeed, cannot do that without a huge loss of validity) to construct personalities from a single comment, word or sentence. We need a rolling average of the personality traits and therefor need more input than a few sentences. The more text available, the better and the more complete correct the LIWC analysis will be as it is closer to a true average of the users word usage. Looking at how it works, how easy it is usable and the related work section builds the picture of a tool that can be used for the purpose of analysing GitHub issues. It's wide spread use is a good indicator for its popularity and it is not hard to see why.
To be brief, LIWC will be used on the GitHub issue texts extracted from the database to analyse the written communication of specific users; those that opened the issue and are communicating (if applicable) with users that are also present within the issue. The text will be used as raw input for the LIWC program and the output is used for further refinement into the FFM. There will be a section later to deal with this and how this is possible.
For example, this is the way the process works: This comment4.1 is to be analysed with the LIWC programme, therefore this needs to be put into a form it can ready. The raw text is then put into a text-file and read into the LIWC program. (Documentation is found in the manual if required.) The result is this output:

The input were also all the answers the issue reporter provided in the issue itself, along with more posts from the issue reporter. Therefore, the text does feature those particular words in the given categories with the shown regularity. So what does this mean? For now, not a whole lot. The statistics are there, as are the categories, but the interpretation is completely up to the user.

So reading this, as of now there is not a whole lot of use for us in the context of actually interpreting the result of such an analysis. But the process is clear, and is used just as described. How to solve the problem of interpretation is shown in a later section.

4.2 GitHub Issues

Literally the one of the biggest issue in this thesis is the issue of issues. The GitHub issue is a specific kind of pseudo-forum post made by contributors, users and developers in a repository. This issue is meant to facilitate the communication to effectively fix bugs and help development with feedback from stakeholders. An issue usually consist of a text post made by the issue author with a description of a problem, bug, inconsistency or defect. Other users can then comment on the issue, commonly to affirm that they also have the problem or to answer questions for the author. They may expand the

mvirkkunen commented 2 hours ago

In a program that just has `fn main()` and a busy loop I can access resources very concisely like this:

```
led_pin.set_high();
```

In RTFM 0.5.0 the same thing looks like this:

```
c.resources.LED_PIN.set_high();
```

This is pretty verbose and my immediate reaction is that I don't like it. The `c.resources.` prefix gets repetitive really fast, and I don't like that all my resource names are SCREAMING. In practice in my current project I just created a variable for every resource at the start of every function, but this feels like pointless boilerplate code I shouldn't need to write:

```
fn some_task(c: some_stack::Context) {
    let mut led_pin = c.resources.LED_PIN;
    // ....
}
```

I understand the design decisions for each part of the current syntax are probably as follows:

1. `c`: Magic variables are not immediately obvious. An actual argument is.
2. `resources`: It creates a namespace to avoid conflicts between resources and other things like `spawn`.
3. `LED_PIN`: The SCREAMING is because they're all essentially `static` variables.

All of this *makes sense*, but the end result is just *not nice* - at least in my opinion. Personally I would prefer to just have resources appear magically as lower-case local variables, because I know to avoid name conflicts.

I don't have a perfect or even a good solution to propose but I didn't see any discussion about the increasing verbosity of accessing resources, so I thought I'd open an issue for discussion.

Figure 4.1: GitHub issue with a good amount of text and quoted code snippets.

original text with examples, screen shots and more.
Take, for example, the issue 4.1 quoted above. What we see here is a typical example of a issue with code snippets and text. (Not shown here are the answers and comments made by the author after the issue was posted. But they follow the same rules as opening posts.)
Containing large amount of code snippets is in and of itself not a problem, as are most technical and scientific blocks of information, however we lack in substance of text because code snippets often are used instead of an explanation of something. For good reason do we use code to illustrate and show a problem, however for this venture a descriptive, natural language explanation would of course be better. If the issue has additional posts by the author this issue is mitigated because those can be more similar

Category	Score	**Category**	Score
pronoun	12,84	ppron	6,31
i	4,73	we	0,00
you	0,90	shehe	0,00
they	0,68	article	6,76
negate	2,70	number	0,90
affect	5,63	posemo	2,48
negemo	3,15	anx	0,68
anger	1,58	sad	0,23
social	4,95	family	0,00
friend	0,00	female	0,00
male	0,00	cogproc	21,85
insight	2,93	cause	3,83
discrep	4,50	tentat	5,63
certain	1,80	see	0,68
hear	0,68	feel	0,23
body	0,00	sexual	0,00
ingest	0,00	a chieve	0,13
focuspast	2,93	focuspresent	12,84
focusfuture	0,45	motion	2,70
space	5,18	time	3,38
work	4,95	leisure	0,00
home	0,00	relig	0,00
death	0,00	swear	0,00
netspeak	0,00	assent	0,00
Total Word Count	444		

Table 4.2: The score is the %-value of the given words in that category of all the words in the text analysed. The categories are specifically chosen for practical reasons and cover both correlation tables in the section about correlation.

to a conversation, or better yet, textual messages between different people thus resulting in much more personal use of words which are better for analysing personality traits. That is the reason author posts in the issue are included as well.
The reason all other posts are more or less ignored is that we somehow have to specify a unit of analysis so to speak. That means that there must be a specific personalty we want to construct a profile for, that we want to analyse. And doing that on the basis of author or who opened the issue is a practical way to limit the scope of whose text is to be extracted. Therefore, the author and all text of the author inside the issue was selected as the scope of extraction.

Now we have of course still the problem of scale as that one issue is not enough to make any assumption about the author due to reasons explained in the chapter about Trait Theory.

4.3 Database

The data set utilised by this thesis was the database dump by the GHTorrent project. [] Featuring a lot of data (sans sensitive data of course) one might need to query information about projects, users, statistics, issues and probably more. Due to the nature of GitHub and it's activity, they of course cannot have all of the data at any given time and do not claim to even have all of it when it comes to dates already passed. It is however, the most practical way to analyse and work with data from GitHub. Live data is also usable of course, but the easy in which one might extract from a database is unparalleled.
A MongoDB dump of the issues from 2015 was downloaded, extracted and converted to a (huge) .json file for further processing. This was deemed much more economical than running a MongoDB locally because the queries were much slower than manual processing of basically text file with hundreds of thousands of .json objects encoded in it. The file was sporting 51GB of raw issue data with all relevant information in it: user and body of the messages along with meta data, URLs and more. But only the user who posted it and the message body is interesting for this work.
The MongoDB dump is in a special .bson file which can be converted using functions provided by a MongoDB installation to convert it into "human readable .json files", which are the ones used to extract the data here. The advantages over classical database usage is the immediacy of all the data points that are needed and also that the tool developed is already working with natural language texts, therefore a connection or translation is not needed at all. Simply operating with the big file is enough to gather information.
The aforementioned unit of analysis is also present in those files as every issue can be identified by the person that started it. There is a second .bson file with all issue comments separately present that supplements the other data into a more complete set of information: the body (or opening post) of the issue and the comments made by the opened of the issue in the issue itself. However, due to the nature of these issues and the analysis method used individual issues are not actually distinguished. The unit of analysis therefore is the opening texts and comments made in issues by a particular user/personality identified by their user name (or as it is called in the data base itself: owner). We need this amount of text simply to get an average going for the analysis. For details why so much text in needed please advise the theoretical part of this work.

4.4 Correlation to FFM

The interpretation of the output from LIWC is an issue which will be addressed here. The problem lies in the translation from LIWC categories (which are by themselves useless for this thesis) to the aforementioned FFM traits which are easier to understand and more importantly, grounded in personality psychology. Those can be used with caution while LIWC categories cannot be used for much in this context. While translating those categories to FFM is way out of scope for this thesis, they are precedential works where the exact step was also needed and was done.
For one of the work [GRET11] see the related work section. The other[Yar10] is also

interesting and was done by the NIH (The National Institute of Health), the primary agency responsible for bio-medical and public health research in the United States. Both works found significant correlation between the LIWC categories and FFM traits by utilising both analysis of text and other methods. See table 4.3 for the values used by the twitter study. [GRET11] To explain the correlation, remember that LIWC has the percentage values of words of a given category in relation to all analysed words in the text. Meaning that if a certain category has for example 14% of the words and due to the correlation they have each a value of 0.06 the total sum for that correlation (14% of the words multiplied with the correlation value) is added to the trait score. There are some values that are quite high in the table, however it should not be forgotten that they are also a lot of negative values to balance that out.
In the study of the NIH[Yar10] the correlation between FFM traits and LIWC categories is further explored and tested. Previous studies[FF08][HP09][PK99] have been confirmed [Yar10] and correlation has been shown again in that study. Specifically, the NIH study replicated 50% of correlations mentioned by Pennebaker King in their 1999 study and replicated 15 out of a total of 24 correlations Hirsh Peterson in their 2009 study mentioned. So far, the study confirms and replicates so much that pure chance is not an option to explain those phenomena. However, be advised that the study solely analysed blogger texts.
Knowledge gained by previous studies are in accord with the findings in that study, thereby strengthening its argument about the correlation being correct. The works of Costa McCrae about Extraversion and Neuroticism are consistent.[Yar10] Words pertaining to negative emotions (Anxiety, fear, Sadness, Anger and the whole category of Negative Emotions in 4.2) correlate positively to Neuroticism.[CM80][LK91] Extraversion was found to positively correlate with use of the words pertaining to interpersonal interaction and positive emotions. (This includes the whole category of Positive Emotions). [LD01][PDF90] Also included in that are the categories Social Processes, Friends, Sexuality and 2nd Person References. The words correlated to Agreeableness are of the 1st Personal Plural References, Family, Friends and the whole category of Positive Emotions. Negatively correlated are the words of the Negative Emotions category and Swear words.
Even the correlation is well established by this point, some constraints and unusual findings still are present, for examples and details please refer to the work itself as this work does not have the scope to handle those.
As can be seen in the given tables 4.3 and 4.44.5 are not the same and have some difference. Therefore it was decided to user the NIH study table as it is closer to standard LIWC categories and it also uses does not use text features such as exclamation marks, commas and the like in its analysis. This was deiced to be prudent because issue texts are full of such features without them being part of the sentences themselves (code snippets, examples, net speak). However, the tool has both tables present and it can be changed quickly to allow either of them to be used or even a combination of both. Finally, the NIH and Twitter study are both used as a basis for this method because issues are a mix of both short bursts of information and blogger posts about specific subjects, making both correlation tables a viable option.

Language Feature	Extro.	Agree.	Consc.	Neuro.	Open.
”You”	0.068	0.364*	0.252*	-0.212	-0.020
Articles	-0.039	-0.139	-0.071	-0.154	0.396*
Auxiliary Verbs	0.033	0.042	-0.284*	0.017	0.045
Future Tense	0.227	-0.100	-0.286*	0.118	0.142
Negations	-0.020	0.048	-0.374*	0.081	0.040
Quantifiers	-0.002	-0.057	-0.089	-0.051	0.238*
Social Processes	0.262*	0.156	0.168	-0.141	0.084
Family	0.338*	0.020	-0.126	0.096	0.215
Humans	0.204	-0.011	0.055	-0.113	0.251*
Negative Emotions	0.054	-0.111	-0.268*	0.120	0.010
Sadness	0.154	-0.203	-0.253*	0.230	-0.111
Cognitive Mechanism	-0.008	-0.089	-0.244*	0.025	0.140
Causation	0.224	-0.258*	-0.155	-0.004	0.264*
Discrepancy	0.227	-0.055	-0.292*	0.187	0.103
Certainty	0.112	-0.117	-0.069	-0.074	0.347*
Perceptual Processes					
Hearing	0.042	-0.041	0.014	0.335*	-0.084
Feeling	0.097	-0.127	-0.236*	0.244*	0.005
Biological Processes	-0.066	0.206	0.005	0.057	-0.239*
Body	0.031	0.083	-0.079	0.122	-0.299*
Health	-0.277*	0.164	0.059	-0.012	-0.004
Ingestion	-0.105	0.247*	0.013	-0.058	-0.202
Work	0.231	-0.096	0.330*	-0.125	0.426*
Achievement	-0.005	-0.240*	-0.198	-0.070	0.008
Money	-0.063	-0.259*	0.099	-0.074	0.222
Religion	-0.152	-0.151	-0.025	0.383*	-0.073
Death	-0.001	0.064	-0.332*	-0.054	0.120
Fillers	0.099	-0.186	-0.272*	0.080	0.120
Punctuation					
Commas	0.148	0.080	-0.24*	0.155	0.170
Colons	-0.216	-0.153	0.322*	-0.015	-0.142
Question Marks	0.263*	-0.050	0.024	0.153	-0.114
Exclamation Marks	-0.021	-0.025	0.260*	0.317*	-0.295*
Parentheses	-0.254*	-0.048	-0.084	0.133	-0.302*

Table 4.3: The feature correlation between LIWC categories and FFM traits.[GRET11]
*: Values marked with * are cited significant as they posses a $p < 0.05$.
The values are percentiles, meaning the the word ”You”for example will ädd”0.068 (or 6.8%) to the trait score of Extraversion. (Note: LIWC will give out the percentage of all the ”You”words used in the analysed text.)

Language Feature	Extro.	Agree.	Consc.	Neuro.	Open.
Total pronouns	0.06	0.06	-0.21***	0.11**	-0.02
First person sing.	0.12**	0.01	-0.16***	0.05	0
First person plural	-0.07	0.11**	-0.1*	0.18***	0.03
First person	0.1*	0.03	-0.19***	0.08*	0.02
Second person	-0.15***	0.16***	-0.12**	0.08	0
Third person	0.02	0.04	-0.06	0.08	-0.08
Negations	0.11**	-0.05	-0.13**	-0.03	-0.17***
Assent	0.05	0.07	-0.11**	0.02	-0.09*
Articles	-0.11**	-0.04	0.2***	0.03	0.09*
Prepositions	-0.04	-0.04	0.17***	0.07	0.06
Numbers	-0.07	-0.12**	-0.08*	0.11*	0.04
Affect	0.07	0.09*	-0.12**	0.06	-0.06
Positive Emotions	-0.02	0.1*	-0.15***	0.18***	0.04
Positive Feelings	0.01	0.11*	-0.11**	0.14**	-0.02
Optimism	-0.08*	0.05	0	0.15***	0.16***
Negative Emotions	0.16***	0.04	0	-0.15***	-0.18***
Anxiety	0.17***	-0.03	-0.02	-0.03	-0.05
Anger	0.13**	0.03	0.03	-0.23***	-0.19***
Sadness	0.1*	0.02	-0.03	0.01	-0.11*
Cognitive Processes	0.13**	-0.06	-0.09*	-0.05	-0.11**
Causation	0.11**	-0.09*	-0.02	-0.11**	-0.12**
Insight	0.08	0	-0.08	0.01	-0.05
Discrepancy	0.13**	-0.07	-0.12**	-0.04	-0.13**
Inhibition	0.09*	-0.13**	-0.07	-0.08	-0.05
Tentative	0.12**	-0.11*	-0.06	-0.07	-0.1*
Certainty	0.13**	0.1*	-0.06	0.05	-0.1*
Sensory Processes	0.05	0.09*	-0.11**	0.05	-0.1*
Seeing	-0.01	0.03	-0.04	0.09*	0.01
Hearing	0.02	0.12**	-0.08*	0.01	-0.12**
Feeling	0.1*	0.06	-0.01	0.1*	-0.05
Social Processes	-0.06	0.15***	-0.14***	0.13**	-0.04
Communication	0	0.13**	-0.06	0.02	-0.07
Other references	-0.08*	0.15***	-0.14***	0.15***	-0.02
Friends	-0.08*	0.15***	-0.01	0.11**	0.06
Family	-0.07	0.09*	-0.17***	0.19***	0.05
Humans	-0.05	0.13**	-0.09*	0.07	-0.12**

Table 4.4: Part 1 of the correlation tables by the NIH study. *p < .05; **p < .01; ***p < .001.

Language Feature	Extro.	Agree.	Consc.	Neuro.	Open.
Time	0.01	-0.02	-0.22***	0.12**	0.09*
Past Tense Vb.	0.03	-0.01	-0.16***	0.1*	0
Present Tense Vb.	0.06	-0.01	-0.16***	0	-0.06
Future Tense Vb.	-0.02	-0.06	-0.08	-0.01	-0.01
Space	-0.09*	0.02	-0.11**	0.16***	0.04
Up	-0.1*	0.09*	-0.15***	0.11**	0.09*
Down	-0.04	-0.02	-0.11**	0.11**	0.06
Inclusive	-0.02	0.09*	0.11**	0.18***	0.07
Exclusive	0.1*	-0.06	0	-0.07	-0.16***
Motion	-0.02	0.02	-0.22***	0.14***	0.04
Occupation	0.05	-0.12**	0.01	-0.04	0.06
School	0.06	-0.07	0.02	-0.01	-0.04
Job/Work	0.07	-0.08*	0.04	-0.07	0.07
Achievement	0.01	-0.09*	-0.05	0.05	0.14***
Leisure	-0.05	0.08*	-0.17***	0.15***	0.06
Home	0	0.03	-0.2***	0.19***	0.05
Sports	-0.01	0.05	-0.14***	0.06	0
TV/Movies	-0.02	0.05	0.05	-0.05	-0.06
Music	-0.02	0.13**	0.04	0.08*	-0.11**
Money/Finance	0.04	-0.04	-0.04	-0.11**	-0.08
Metaphysical	-0.01	0.08	0.07	-0.01	-0.08
Religion	-0.03	0.11**	0.05	0.06	-0.04
Death	0.03	0.01	0.15***	-0.13**	-0.12**
Physical States	0.03	0.14***	-0.09*	0.09*	-0.05
Body States	0.02	0.1*	-0.04	0.09*	-0.07
Sexuality	0.03	0.17***	0	0.08*	-0.06
Eating/drinking	-0.01	0.08	-0.15***	0.03	-0.04
Sleep	0.1*	0.02	-0.14***	0.11**	-0.03
Grooming	0.05	-0.01	-0.2***	0.07	-0.05
Swear words	0.11**	0.06	0.06	-0.21***	-0.14**

Table 4.5: Part 2 of the correlation tables by the NIH study. *p < .05; **p < .01; ***p < .001.

4.5 Work flow

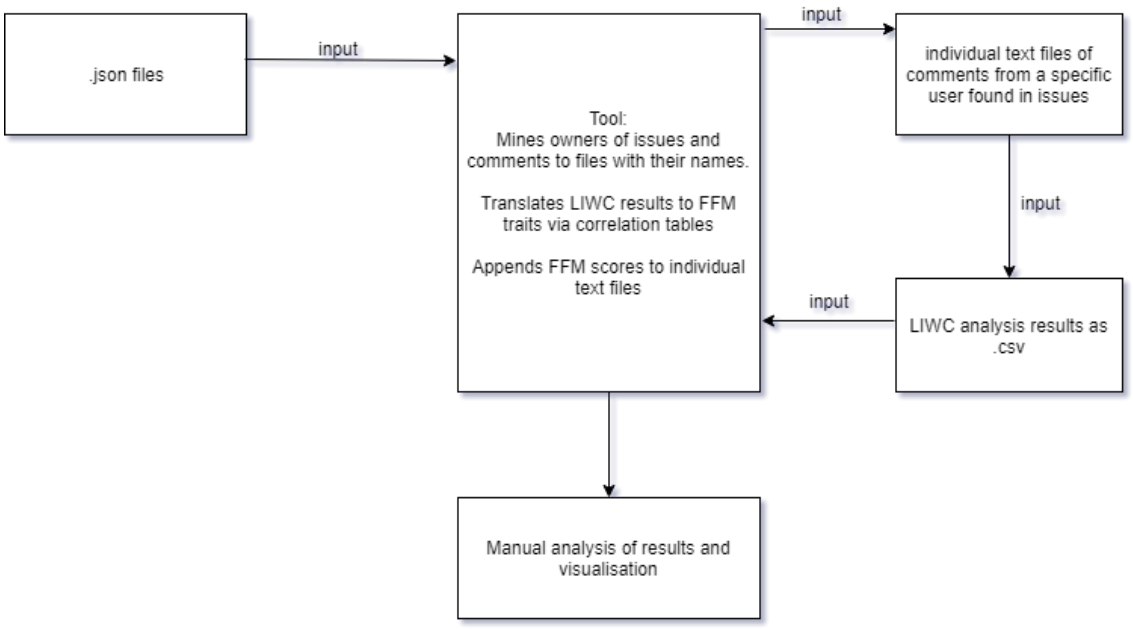

Figure 4.2: Operational work flow from .json file to translated file with FFM traits.

Starting with a .json file created by MongoDB's bsondump method which converts a MongoDB data base dump into a human readable .json file for further consumption. It is versatile because MongoDB dumps can be converted to files easily readable by programming languages. Processing .json files is a tested and popular method of data entry and as such there are a lot of tools around to do that. Second as far as natural language processing goes a .json file is very close to actual language because it can be understood very quickly even without much knowledge about the subject the file is about.

There are two .json files: one with all the issue opening text and another one with all comments made by the user in other issues. Both these combined texts represent a unit of analysis for a certain personality. The tool itself is a Python script processing the .json file line by line and separating the comment texts into .txt files of their respecting authors. It removes certain words and letters such as `\n` or `\r` which have no value for the analysis and only bloat the word count. They are also not present in the actual issue text one would find when reading the issue online. They are there for formatting purposes only and therefore serve no purpose here. After collecting these files they are each individually loaded into LIWC to analyse them with the categories described in the the 4.4 section. These categories show correlation and are therefore selected in the native LIWC categories options menu. The output file generated by LIWC is a .cvs file with the scores of each file according to the selected categories and as a .csv file it can be easily read by the most rudimentary of scripts and programs. Hence reading the file into the tool and translating the LIWC categories to

FFM trait scores by way of multiplying the score from the correlation table with the amount of words in the LIWC category it is associated with. These scores are then taken and analysed manually.
As seen in 4.2 with regards to manual work, the only thing being done outside of the tool is the dumping of the MongoDB database and the input into LIWC. The output file generated by LIWC itself does not need to be changed in any way or even looked at. As long as it is a .csv file and configured in the default way the tool will read it and process it further.
The output from the tool is thus: a .csv file with each row representing a personality's FFM score calculated normalised on a 0 to 1 scale sorted alphabetically by personalities user name. Their name can be obscured if needed and are not taken into account in the final analysis.
Finally, the tool also runs some rudimentary statistical procedures to calculate the average, mean, percentiles ect. of all the gathered texts to give some information about them.

5 Results

Using the database dump to extract large amounts of a users issue texts and comments is simple albeit time consuming task as the files used to source them have millions of lines. However, the operation simply takes some time to finish. The first and then the second step of creating the files and adding the issue opening texts then adding other comments made in issues by the same personality take roughly the same amount of time. Writing and reading files can be expensive however Python was a good choice for that endeavour.

The amount of personalities analysed was n=455. The amount was only limited by how long one would want to let the tool run uninterrupted. The amount of text available ranges from as little as 6 lines to over 38k lines of written comments and issue texts. As the amount varies, some constraints of the validity would be implied however this data is interpreted. Being too restrictive with what users to analyse would also be bad because GitHub users do differ in their verbosity and users with less comments than other still are part of the user base. Just be advised that their results may be less accurate in their depiction average personality scores.

median	min	max	range
213	6	38428	38422
25 percentile	50 percentile	75 percentile	90 percentile
76	213	869	2294

Table 5.1: Basic stats about the analysed files of each personality and their texts. These are about lines or sentences.

These 5.1 are the basic statistics about the analysed texts. As mentioned, the range is pretty large resulting in the average not saying much. Therefore, the percentiles give a more accurate understanding of the distribution of the length of the texts and comments. As shown in 5.1 most of all texts extracted are around between 213 and 869 with only very little getting to over 2000.

About the statistics about the individuals files and the containing words 5.2, there are quite the large amount of words analysed on a personality per personality basis albeit the range is again very high, the same problems as above arise. Even then, a minimum of 3120 words is not too bad. However, the problem is following: How many words are actually usable when it comes to the analysis? In table 5.3 we can see that the amount of data that is useful for this analysis makes this still worthwhile, even if there are outliers within that percentage. The outliers come from either issues that contain

median	min	max	range
9922	155	2972908	2972753
25 percentile	50 percentile	75 percentile	90 percentile
3120	9922	42480	2294

Table 5.2: Basic stats about the analysed files of each personality and their texts. These are about lines or sentences.

only code snippets, log content or issues that have Asian characters in them. LIWC does not naively support those and the only other standard dictionary is German. There are however different dictionaries available on the internet if one is so inclined.

median	min	max	range
57.39	3.76	79.97	76.21
25 percentile	50 percentile	75 percentile	90 percentile
47.625	57.39	63.805	68.172

Table 5.3: Basic stats about the analysed files of each personality and their texts. These are about lines or sentences.

Now looking at the result of the tools calculations, the different tallied scores of the FFM traits by the score value they had in each category of the LIWC categories. The total score of every word that contributed to a trait was divided by all LIWC categories that had correlation to that very trait.

5.1 FFM Trait Scores

5.1.1 Neuroticism

Neuroticism			
Min	Max	Range	Median
-32.58	4340	4373	21.28
25 percentile	50 percentile	75 percentile	90 percentile
6.38	21.28	89.20	252.80

Table 5.4: Neuroticism FFM trait scores

As seen in 5.4 the analysed 455 personalities displayed an mean score of 21.28 for Neuroticism as calculated by the NHI table and the words of their correlating category from LIWC. There are posts (seen by the minimum) that would indicate that the personalities are less neurotic, but every percentile speaks another story: words that

indicate neurotic though patterns and behaviour are found in the majority of posts. 21.28 is the median score they have accumulated by the NIH table and their respective LIWC categories. Meaning that when one looks at all the categories the personality used words from, most had a score of around 21.28 from that particular trait, namely Neuroticism.

5.1.2 Extraversion

Extraversion			
Min	Max	Range	Median
-10615.88	3.94	10619.82	-9.25
25 percentile	50 percentile	75 percentile	90 percentile
-39.12	-9.25	-2.80	-1.18

Table 5.5: Extraversion FFM trait scores

The Extraversion table is a lot more negative, and very to none positive correlation could be found.
As seen in 5.5 the personalities displayed a negative score as the mean value and indeed most percentiles confirm this: most issue texts and comments show that the words used would indicate a personality is less extraverted. This would mean that most users do not show any average behaviour (implicated through their word usage) that would classify them as extraverted.

5.1.3 Openness

Openness			
Min	Max	Range	Median
-21409.17	-0.62	21408.54	-61.76
25 percentile	50 percentile	75 percentile	90 percentile
-297.26	-61.76	-18.93	-9.71

Table 5.6: Openness FFM trait scores

Openness fare little different than Extraversion, being even more negatively correlating than the previous trait.
The table 5.6 shows that in this instance there have been personality whose issue texts and comments come out positively correlating with Openness. Of course there could have been individual posts or comments that fit the criteria, but as personalities are the unit of analysis this is the result nonetheless.

5.1.4 Agreeableness

Openness			
Min	Max	Range	Median
0.46	17429	17428	42
25 percentile	50 percentile	75 percentile	90 percentile
12.79	42.20	197.87	545.64

Table 5.7: Agreeableness FFM trait scores

Agreeableness correlations can be widely found in the comments and issue texts. Surprisingly, given the previous scores and positive Neuroticism scores. However, as discussed in the next part, this might be actually not that difficult to explain.

5.1.5 Conscientiousness

Openness			
Min	Max	Range	Median
-2422.44	37.17	2459.61	-10.31
25 percentile	50 percentile	75 percentile	90 percentile
-82.00	-3.11	-50.92	-2.16

Table 5.8: Conscientiousness FFM trait scores

And finally conscientiousness. It can also be said that most analysed comments and texts do not show many signs of word usage suggesting conscientious behaviour on the side of the author.

5.2 Interpretation and constraints

5.2.1 The journey

Gathering the results was not straight forward as it seems as the correlation tables and the resulting translation to scores was not that simple.

5.2.2 The interpretation

This is the picture this thesis has drawn about the natural language texts posted by the users in issues and comments in issues:

The words they use would suggest that they are neurotic (a lot of positive correlation to Neuroticism), introverted (negative correlating scores for Extraversion) ,closed off to new things and rather would stick to routine(lots of texts that correlate negatively to Openness), considerate and compliant (high Agreeableness scores) personalities that are not very disciplined and probably do not plan ahead very much. (negative correlation to Conscientiousness) Can this be explained?
Issues are highly diverse and do not really follow a consistent template, often resulting in posts that are a error or the likes with a short and concise sentence boiling down to "this does not work please fix it". Bots are also not infrequent, resulting in something like "Broken xyz". (Think of a Jenkins that writes a message into issues when the build fails) They do not need to adhere to social etiquette or niceties usually and posts inside issues can be frequently ignored by the opening user. Language barriers also make some sentences seem more harsh than they perhaps are meant to be. Insufficient language skills result in texts that can throw the calculation into balance and perhaps they are interpreted more negatively than they should be.
Hyperbole also are often employed on the internet to make a point but as there was no sentiment analysis in this thesis irony, overreactions and the likes are taken as they are and therefore influence the score.
One of the most important aspects is that issues are not your standard natural language text and should not be seen as such. Issues by their very definition are more negatively tinted for the sole reason that they are opened when something is broken, lacking or missing something to work like the opened wishes. That very mindset might very much influence the word usage in those instances. Imagine, if you will, writing a letter to the financial authority of your choice because they overtaxed you. That letter could not be reasonable taken as a normal interaction between you and another person and therefore the analysis of word usage would result probably a different image than a letter to your grandmother. The same constrain applies here.
Furthermore the amount of net speak, abbreviations that are not in the dictionary, code snippets,log file entries and more have certainly influenced the result drastically. Parts of those not analysable texts are analysed by LIWC and therefore by the tool developed even though they might not represent any correlation whatsoever. Also constructs like lists, tables and more are also represented in those comments read into the tool and also further muddle the waters.
It should not be taken completely as gospel that the people of GitHub are like these results show. However, online decentralised collaboration is not as easy and the amount of distance might result in some people showcasing much more behaviour that is commonly associated with behaviour that can be interpreted as reserved, unwillingness to step out your comfort zone. That most comments tend to negatively correlate with Conscientiousness might be explainable that the issue texts and comments usually spring from an unwanted occurrence such as a bug or a missing feature. Their very nature might not lend itself to a lot of conscientious word usage. That is only a speculation however. The high correlation with Agreeableness is also surprising, given the other results. As compliance and informal, familiar speech is also a part of that the internet might support this kind of word usage. On the internet, everyone acts much more familiar with one another than they would in real life thus resulting in a higher score. In issues they complain as if their car broke to their significant other. The compliance argument for the high score of Agreeableness might come from the asking tone many

issues/feature requests written in issues, feature. That would be an explanation worth exploring.
All in all, the results are not that surprising given the circumstance of issues and their comments and the constraints of the issue format given in this chapter. Remember even if they have lore scores in some traits, it just means they are less so and they do not lack Openness, they are just low on that score and therefore display different behaviour.

6 Discussion

To explore the possibility to get a grip on developers personalities via GitHub in general and issues in particular, this thesis set out to understand the personality theory and methodology necessary to support that endeavour. The constraints that have shown themselves during the interpretation of the texts mined are severe and might even be crippling. There needs to be an understanding how much the framework of working off of issues influences the though processes behind the issues as this framework is not neutral. It is not neutral because for an issue to be opened, there needs to be a bug or a missing/lacking feature. This results in the text being perhaps biased from the very beginning and therefore cannot be simply read as is. Furthermore, other than the framework of the mind these users had when writing these texts, the texts themselves are not common natural language texts are operate meant to be read in a very specific context: that of GitHub issues. Log file entries, code snippets, short concise sentences and the like all further colour the results in a way that is not addressed in this work. The general idea is sound however, if the constraints are massive. However, they are avoidable perhaps with a lot of filtering done prior to analysis with LIWC. That was not done and if there is another work that would try to use GitHub issues as a base it would need to certainly do that. And also understand how the previous mindset influences the traits displayed in the word usage.
The results do suggest however a general trend of GitHub issue text to display behaviour that indicate a little reserved personality that is agreeable and neurotic. (Going by the scores) That is perhaps the kind of person that is even inclined to open an issue. Issues therefore might be already biased because the option of actually opening an issue is not seen as an option by all people but by an subset of developers acting on GitHub. This self selecting of users using this feature might make a generalisation upon the greater GitHub community and then on developers in particular difficult. The related work section for example has a similar observation: that those users which are voted positively more often than others are those that have a high amount of Agreeableness score. They are selected on part by that and therefore cannot be generalised unto developers, not even unto users present in Stack Overflow.
However, even if there is a self selecting happening with the personalities who open up GitHub issues, those kind of people are still part of the general software engineering population and with this an interesting question is how they fit into that group?
The base line thoughts and validity of this kind of analysis however was well documented in this thesis and the usage of LIWC was a good idea that gave results that could be used for further processing. Both LIWC and it's connection to trait theory has been explained in the theoretical part and the practical part connected them successfully. Even if the validity of the actual results are muddied by constraints, the basic procedure was a success. If the results are not on par, the methodology at least proved to be versatile, not too complicated and practical.

GitHub issues are a wellspring of natural language texts that can be used to identify traits in developers specifically who open up those issues, even if the circumstances surrounding their conception is meta knowledge that needs to be addressed before a more in-depth analysis that is more precise can be accomplished.
Works that build off of this thesis need to solve the problem of the issues being full of text that influence the LIWC result even if the authors word usage under normal circumstances would not warrant that influence on traits. Another problem that must be solved is an efficient way of filtering out net speak, constructs and abbreviations used by the GitHub community in their posts and texts. Perhaps a GitHub dictionary specifically for LIWC would be prudent and very welcome in this kind of research as surely most researchers who use natural language texts found in the internet have stumbled upon that very same problem.
Lastly, the implications of a specific type of personality that opens up GitHub issues may be like the personalities who open up support tickets in other platforms, and that link might be very much worht exploring.

7 Conclusion

GitHub as a platform features a highly diverse range of projects to work on, try out and explore. Many developers seek out contribution to one ore more projects, joining in on a collaborative, remote and decentralised way of working on developing software. To deepen the understanding of those developers was set out as a goal and GitHub issue texts and comments, a very specific kind of natural language text featured in projects big and small, was used as a resource for it. Hopefully being able to generalise some insight unto the greater GitHub community and software engineers in particular. The personality that resulted in that analysis was found to be neurotic and agreeable while scoring low on Agreeableness, Openness and Conscientiousness. However constraints were discovered that hampers the ability to say for certain that the personality shown through NLP in those issues can be summarised as a good average of personality of a developers. This is because the very nature of issues and their surrounding circumstances may promote a specific kind of communication with a specific kind of tone. Thus, the resulting personality profile, while valid in the sense that those personalities constructed did open up and comment in issues, may showcase a general trend to feature certain traits more often than a more free form of self expression otherwise would suggest. Which means that those traits displayed in issues are part of the personality as a whole, but the whole range of personality traits and their according word usage may not be found in those texts.
And even if the general GitHub population is already not the same as the general software engineering populace, the insight gathered on the part of personality that is displayed with and within issues is perhaps valuable for understanding a personality in it's entirety.Regarding and understanding the mindset with which the personality approaches the creation of a certain natural language text is more important when the texts are only created when either not satisfied with a piece of software, when it lacks certain functions or when enraged because of a bug. Therefore, if has been found that the traits displayed by the GitHub users that create those issues and comment in issues are still valuable to understand as the process of issue tracking is important from a developing and communication perspective. The parts of a developers personality discovered in issue texts are still part of the whole personality and even if it is perhaps unwarranted to say that the whole range of personality traits are revealed through issue texts and comments, they are needed to round off a complete personality inventory when looking at GitHub developers as a population. This work has therefore discovered an integral part of how to look at personality in GitHub and the constraints that apply to that kind of research on that platform.
Personality analysis as a tool to streamline and improve software engineering practises is not simple and needs many sources for even one accurate personality profile, however as GitHub features more natural language texts than issues and comments, the potential

is still there for a holistic approach that would also feature issue texts and comments heavily. It only needs to be able to factor in the special circumstances and constraints found during this thesis.

Bibliography

[] URL: http://ghtorrent.org (visited on 06/01/2019) (cit. on p. 44).

[BHS13] B. Bazelli, A. Hindle, E. Stroulia. "On the personality traits of stackoverflow users". In: *2013 IEEE international conference on software maintenance*. IEEE. 2013, pp. 460–463 (cit. on pp. 16, 37).

[BM91] M. R. Barrick, M. K. Mount. "The big five personality dimensions and job performance: a meta-analysis". In: *Personnel psychology* 44.1 (1991), pp. 1–26 (cit. on p. 31).

[Che18] K. Cherry. *What Are the Id, Ego, and Superego?* 2018. URL: https://www.verywellmind.com/the-id-ego-and-superego-2795951 (visited on 06/01/2019) (cit. on pp. 21, 22).

[CILV18] F. Calefato, G. Iaffaldano, F. Lanubile, B. Vasilescu. "On developers' personality in large-scale distributed projects: the case of the apache ecosystem". In: *2018 IEEE/ACM 13th International Conference on Global Software Engineering (ICGSE)*. IEEE. 2018, pp. 87–96 (cit. on p. 16).

[CM80] P. T. Costa, R. R. McCrae. "Influence of extraversion and neuroticism on subjective well-being: happy and unhappy people." In: *Journal of personality and social psychology* 38.4 (1980), p. 668 (cit. on p. 45).

[DS02] S. Dewitte, H. C. Schouwenburg. "Procrastination, temptations, and incentives: The struggle between the present and the future in procrastinators and the punctual". In: *European Journal of personality* 16.6 (2002), pp. 469–489 (cit. on p. 33).

[FF08] L. A. Fast, D. C. Funder. "Personality as manifest in word use: correlations with self-report, acquaintance report, and behavior." In: *Journal of personality and social psychology* 94.2 (2008), p. 334 (cit. on p. 45).

[GRET11] J. Golbeck, C. Robles, M. Edmondson, 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*. IEEE. 2011, pp. 149–156 (cit. on pp. 16, 44–46).

[HH95] P. J. Howard, J. M. Howard. "The Big Five Quickstart: An Introduction to the Five-Factor Model of Personality for Human Resource Professionals." In: (1995) (cit. on pp. 15, 16, 31).

[Hol11] T. Holtgraves. "Text messaging, personality, and the social context". In: *Journal of research in personality* 45.1 (2011), pp. 92–99 (cit. on p. 17).

[HP09] J. B. Hirsh, J. B. Peterson. "Personality and language use in self-narratives". In: *Journal of research in personality* 43.3 (2009), pp. 524–527 (cit. on p. 45).

[HPPL07] D. M. Higgins, J. B. Peterson, R. O. Pihl, A. G. Lee. “Prefrontal cognitive ability, intelligence, Big Five personality, and the prediction of advanced academic and workplace performance.” In: *Journal of personality and social psychology* 93.2 (2007), p. 298 (cit. on p. 33).

[JBH97] J. A. Johnson, S. R. Briggs, R. Hogan. *Handbook of personality psychology*. Elsevier, 1997 (cit. on p. 33).

[LD01] R. E. Lucas, E. Diener. “Understanding extraverts’ enjoyment of social situations: The importance of pleasantness.” In: *Journal of personality and social psychology* 81.2 (2001), p. 343 (cit. on p. 45).

[LK91] R. J. Larsen, T. Ketelaar. “Personality and susceptibility to positive and negative emotional states.” In: *Journal of personality and social psychology* 61.1 (1991), p. 132 (cit. on p. 45).

[Lut02] P. L. Lutz. *The rise of experimental biology: an illustrated history*. Springer Science & Business Media, 2002 (cit. on p. 20).

[Mat12] J. L. Matsumoto D. *Culture and Psychology*. Wadsworth-Cengage Learning, 2012 (cit. on p. 33).

[MC03] R. R. McCrae, P. T. Costa. *Personality in adulthood: A five-factor theory perspective*. Guilford Press, 2003 (cit. on pp. 24–27, 34–37).

[MC85] R. R. McCrae, P. T. Costa Jr. “Openness to experience”. In: *Perspectives in personality* 1 (1985), pp. 145–172 (cit. on p. 33).

[MC97] R. R. McCrae, P. T. Costa Jr. “Personality trait structure as a human universal.” In: *American psychologist* 52.5 (1997), p. 509 (cit. on p. 17).

[MJ92] R. R. McCrae, O. P. John. “An introduction to the five-factor model and its applications”. In: *Journal of personality* 60.2 (1992), pp. 175–215 (cit. on p. 31).

[PBJB15] J. W. Pennebaker, R. L. Boyd, K. Jordan, K. Blackburn. *The development and psychometric properties of LIWC2015*. Tech. rep. 2015 (cit. on p. 40).

[PDF90] W. Pavot, E. Diener, F. Fujita. “Extraversion and happiness”. In: *Personality and individual differences* 11.12 (1990), pp. 1299–1306 (cit. on p. 45).

[PFB01] J. W. Pennebaker, M. E. Francis, R. J. Booth. “Linguistic inquiry and word count: LIWC 2001”. In: *Mahway: Lawrence Erlbaum Associates* 71.2001 (2001), p. 2001 (cit. on p. 16).

[Pit05] D. J. Pittenger. “Cautionary comments regarding the Myers-Briggs Type Indicator.” In: *Consulting Psychology Journal: Practice and Research* 57.3 (2005), p. 210 (cit. on pp. 30, 31).

[PK99] J. W. Pennebaker, L. A. King. “Linguistic styles: Language use as an individual difference.” In: *Journal of personality and social psychology* 77.6 (1999), p. 1296 (cit. on pp. 15, 31, 37, 45).

[PV07] D. L. Paulhus, S. Vazire. “The self-report method”. In: *Handbook of research methods in personality psychology* 1 (2007), pp. 224–239 (cit. on p. 30).

[SGW11] D. Schacter, D. T. Gilbert, D. M. Wegner. *Psychology (2nd Edition)*. New York: Worth, 2011. URL: http://www.amazon.com/Psychology-Daniel-L-Schacter/dp/1429237198/ref=sr_1_1?s=books%5C&ie=UTF8%5C&qid=1313937150%5C&sr=1-1 (cit. on p. 22).

[Tho08] E. R. Thompson. "Development and validation of an international English big-five mini-markers". In: *Personality and individual differences* 45.6 (2008), pp. 542–548 (cit. on p. 34).

[Thu09] P. Thurschwell. *Sigmund Freud*. Routledge, 2009 (cit. on pp. 21, 22).

[Wil13] W. F. Williams. *Encyclopedia of pseudoscience: From alien abductions to zone therapy*. Routledge, 2013 (cit. on p. 20).

[Yar10] T. Yarkoni. "Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers". In: *Journal of research in personality* 44.3 (2010), pp. 363–373 (cit. on pp. 17, 44, 45).

All links were last followed on June 21, 2019.

Declaration

I hereby declare that the work presented in this thesis is entirely my own and that I did not use any other sources and references than the listed ones. I have marked all direct or indirect statements from other sources contained therein as quotations. Neither this work nor significant parts of it were part of another examination procedure. I have not published this work in whole or in part before. The electronic copy is consistent with all submitted copies.

---

place, date, signature