Keio University

Doctoral Thesis

Deus Ex Machina - Summoning the machine spirit in artificial intelligence by aggregate human psycholinguistic features in training data?

Author:
Peter Romero

Supervisor:
Professor Teruo Nakatsuma

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosohpy

in the

Graduate School of Economics

February 25, 2024

Declaration of Authorship

I, Peter Romero, declare that this thesis titled, "Deus Ex Machina - Summoning the machine spirit in artificial intelligence by aggregate human psycholinguistic features in training data?" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

"*Ach, da kommt der Meister!*
Herr, die Not ist groß!
Die ich rief, die Geister
werd ich nun nicht los.

(Ah, he comes excited.
Sir, my need is sore.
Spirits that I've cited
My commands ignore.) "

Johann Wolfgang von Goethe, Der Zauberlehrling (The Sorcerer's Apprentice)

KEIO UNIVERSITY

Abstract

Faculty of Economics
Graduate School of Economics

Doctor of Philosohpy

Deus Ex Machina - Summoning the machine spirit in artificial intelligence by aggregate human psycholinguistic features in training data?

by Peter Romero

We present a deep study about spatiotemporal data issues and how these affect research in behavioural economics. Since human behaviour is merging with artificially intelligent agents, community mechanisms will be composed of hybrid systems of humans and machines. Crucial factors of the dynamism and ultimate success of those systems is agent personality and communication in the closer and broader information field. Therefore, more rigorous research is conducted in this work to better understand both the information field and the personality of agents. First, we show that Large Language Models (LLM) display split synthesised personalities, and discuss potential root causes in model architecture, emergent capabilities and training data. Second, we scrutinise typical training data of LLM, and show that geospatial, temporal, and spatiotemporal psycholinguistic features are important yet neglected in model training - both with deep neural architectures and with statistical machine learning models. During these experiments, we introduce a framework for operationalising the information field, and its attachment to the psychometric architecture of agents. Finally, we suggest a more abstract, universal, and substrate-free psychometric approach that takes contextual agent embedding and thus geospatial, temporal, and spatiotemporal features into account, and subsequently call for novel training methods for LLM based on that.

Acknowledgements

Professor Teruo Nakatsuma, for being the best advisor one could wish for, and convincing me about the Bayesian approach.

Professor Stephen Fitz, for teaching me everything I know about Artificial Intelligence

Professor John Rust, Professor David Stillwell, Vesselin Popov, and Dr. Luning Sun from the Psychometrics Centre, University of Cambridge, for believing in me, when nobody else did.

Peter Buttgereit, who sparked my love for psychometrics.

Dr. Markus Jensch, who was a wise and generous mentor in every way.

Eugene Burke and Professor Dave Bartram for giving me a chance of a lifetime and seeing in me the person I was to become, not the person I was.

All academic partners, I was lucky enough to work with: Dr. Rei Akaishi, Professor Atsushi Oshio, Dr. Joseph Watson, Dr. Ahmed Izzidien, Professor Markus Bühner, Professor Clemens Stachl, Dr. Timo Koch, Yuki Mikiya, Gregory Serapio-García, Eisaku Tanaka.

All industry partners that supported me from Google Deep Mind/ Google Brain (Professor Maja Matarić, Dr. Aleksandra Faust, Clément Crepy, Mustafa Safdari), Institution for a Global Society (Professor Masahiro Fukuhara, Yukata Kanou, Fabien Roudier, Akiyo Tsuchimoto, Shino Takishita), and Manzanita Intelligent Marketing (Christopher Demetrakos, Atsuko Hara, Shannen Romero-Perez, KC Chen).

Keio University Academic Development Funds for Individual Research, and The Keio University Ushioda Memorial funds.

Foremost, I thank everyone who did not believe in me and tried to block me: ***q.e.d.*** *:)*

Contents

List of Figures

List of Tables

List of Abbreviations

AI	**A**rtificial **I**ntelligence
LLM	**L**arge **L**anguage **M**odels
GPT	**G**enerative **P**re-trained **T**ransformer
DNN	**D**eep **N**eural **N**etworks
DL	**D**eep **L**earning
NLP	**N**atural **L**anguage **P**rocessing
CS	**C**omputer **S**cience
CSS	**C**omputational **S**ocial **S**ciences
GOFAI	**G**ood **O**ld **F**ashioned **A**rtificial **I**ntelligence
SOTA	**S**tate **O**f **T**he **A**rt
SNS	**S**ocial **N**etworking **S**ervices
Big Five	**Five** Factor Personality Model: O,C,E,A,N
O	**O**penness to (new) experience
C	**C**onscientiousness
E	**E**xtraversion
A	**A**greeableness
N	**N**euroticism (negative emotional stability)

Dedicated to

My great teacher and mentor R. P. in utter gratefulness and inviolable dedication.

Chapter 1

Deus Ex Machina: Personality of Artificial Intelligence

1.1 General And Chapter Introduction

The fundamental problem of economics - the allocation of scarce resources like money, time, or attention (Ogaki and Tanaka, 2017), is steered by three major mechanisms - market, power, and community mechanism. The power mechanism describes the legal, regulatory, and governmental functions that can coerce people towards specific behaviours, for example by law or police. The market mechanism is what most non-economists would consider as "economics", since it is comprised of price and competition mechanisms. And finally, the community mechanism is any mechanism where at least one persons proposes voluntary cooperation and is not rejected by the person that was proposed to (Ogaki, 2022). The power mechanism could be described as vertical influence (rather: top-down), the market mechanism as systemic influence, and the community mechanism as horizontal, and in many ways: bottom up. If those three mechanisms operate in harmony, they enable the advanced and well-balanced societies we enjoy today. A big part of that balance is the positive influence of governmental communication (power mechanism), independent media, marketing, and advertisement (all three market mechanism) on the information field that economic agents operate in. While the word of mouth, social role models, and collective action are influence options of the community mechanism to that balance, this mechanism is under attack. As fake news around the COVID-19 pandemic, propaganda about the Ukraine war, and the influence of Cambridge Analytica on US elections and UK's Brexit show, not only are today's societies extremely advanced, but also extremely vulnerable to external influence to the information field.

While technological and educational mitigation against fake news and propaganda are being developed, an entirely new, and potentially more dangerous attack vector might have opened to the community mechanism, which increasingly permeates all life spaces of humans and thus economic mechanisms over time, as well: generative artificial intelligence like ChatGPT. Humans and AI increasingly cooperate in hybrid systems, where AI replace humans in proximal networks like work teams, as well as in distal networks like leadership or information management. This has direct effects on behavioural economics, for example social influences are believed to be the strongest factor in the endowment effect (Ogaki and Tanaka, 2017), but very soon, this might be synthesised social influences from hybrid systems, which might be tailored to modify behaviours (Romero and Fitz, 2021) of individuals, empowered through surveillance capitalism (Zuboff, 2023). Hence, a new and unprecedented situation emerges, in which choices are not only irrational, but might be tricked through malicious artificially intelligent agents that present choices in a different form and wrapped in different words, just to nudge human agents into a different set

of behaviours (Thaler and Sunstein, 2021), thus changing their choice of allocation of scarce resources. For example, through abusing risk seeking and loss aversion by changing language around certainty of losses or gains, malicious AI could abuse the certainty and isolation effects to alter the information field and thus outsmart the stock markets or create bank runs (Kahneman and Tversky, 2012). Furthermore, it could abuse pro-social tendencies to change election results or weaponise NGO against competitors or political parties.

This, and other problems of existential risk, opens a new field for AI researcher, in AI alignment, or the study of AI safety. While AI safety mostly focuses on identifying malicious behaviours and intents, using very rigorous mathematical approaches from CS, novel approaches deploy psychometrics (Hernández Orallo, 2017) to better understand complex emergent behavioural patterns and potential dangers. For example, CEO (Borgholthaus, White, and Harms, 2023) and surgeons (Bucknall et al., 2015) (as compared to all other health professionals) score higher in the Dark Triad of Machiavellianism, Psychopathy, and Narcissism, which may lead to detrimental outcomes for lives and livelihood of others, but so does ChatGPT (Jones and Paulhus, 2014; Li et al., 2022a).

Unfortunately, most studies on that issue so far are not based in rigorous psychometrics. Therefore, we conduct a deep psychometric analysis to better understand root causes of this behaviour, and draw conclusions of potential ramifications for behavioural economics and broader economics research.

1.2 Do GPT Language Models Suffer From Split Personality Disorder? The Advent Of Substrate-Free Psychometrics

(This section was written by Peter Romero as main author, and supervised by Teruo Nakatsuma and Stephen Fitz.)

Previous research on emergence in large language models shows these display apparent human-like abilities and psychological latent traits. However, results are partly contradicting in expression and magnitude of these latent traits, yet agree on the worrisome tendencies to score high on the Dark Triad of narcissism, psychopathy, and Machiavellianism, which, together with a track record of derailments, demands more rigorous research on safety of these models. We provided a state of the art language model with the same personality questionnaire in nine languages, and performed Bayesian analysis of Gaussian Mixture Model, finding evidence for a deeper-rooted issue. Our results suggest both interlingual and intralingual instabilities, which indicate that current language models do not develop a consistent core personality. This can lead to unsafe behaviour of artificial intelligence systems that are based on these foundation models, and are increasingly integrated in human life. We subsequently discuss the shortcomings of modern psychometrics, abstract it, and provide a framework for its species-neutral, substrate-free formulation.

In Stanley Kubrick's 1968 classical science fiction movie "2001: A Space Odyssey", an artificial intelligence, "HAL", goes berserk, which unfortunately also runs their spacecraft and all life-support-systems during a mysterious mission to Jupiter. The name "HAL" happens to be a one-letter-shift of IBM, the company spearheading with its Watson division the field of consumer-facing and decision making artificial intelligence. Though originally based on so called "good old fashioned AI", a synonym for rule-based or logical agents, the precursor of nowadays's neural architectures, it won in 2011 against human players in Jeopardy (Ferrucci, 2012), and was

subsequently updated and deployed in various fields from cooking, to code creation, weather forecasting, advertisement, finance, fashion, defence, education, and general chatbots. One remarkable application was its now deprecated service for deriving author personality from text, IBM Watson Personality Insights, which was mainly geared towards marketing clients and trained on data from people who took personality questionnaires and provided text samples. The notion of machine personality inspired not only countless science fiction authors and researchers. Google AI's chatbot "LaMDA" was described as 'sentient' by Blake Lemoine, a researcher working with it, which became a global news story. Conversational AI had it's watershed moment however, as "ChatGPT", or GPT 3.5 appeared *deus ex machina* and over night influenced culture world-wide.

Given the trend in the industry to intermingle AI with human life spaces through self-driving cars, neural interfaces, ambient artificial assistants, and decision making algorithms, a variety of researchers applied psychometric instruments that were created for humans towards Large Language Models (LLMs). These approaches and findings can be clustered into two major categories: emergent latent psychological traits, and emergent abilities.

In terms of emergent abilities, ChatGPT displays human-like ability to monitor and override potential erroneous mathematical and logical conclusions in Cognitive Reflection Tests (CRT) and semantic illusions "designed to investigate intuitive decision-making in humans" (p.1), yet is as prone to potential cognitive errors. Due to its fluency and consistency, some of these errors are subtle and well hidden, hence may yield detrimental ramifications for AI safety in areas of decision making on humans, for example regarding legal or medical questions (Hagendorff, Fabi, and Kosinski, 2022).

Similar inconsistencies occur when putting it under strict scrutiny for its mathematical abilities by eliciting responses via exam-style tasks from various mathematical contexts. Its mathematical abilities are "... significantly below those of an average mathematics graduate student", since it "often understands the question but fails to provide correct solution" (p.1)., which manifests in consistency of quality, especially with increase with prompt difficulty and complexity as in proofs (Frieder et al., 2023).

It scores like a 9-year-old child in Theory of Mind (ToM) tasks that measure the degree to which an agent can impute latent mental states to others. This central ability to "to human social interactions, communication, empathy, self-consciousness, and morility" (p.1) and, subsequently, human-machine interaction and safety, evolved with progressing scale of that Large Language Model (LLM) up to its present ability to solve 93% of all task (Kosinski, 2023).

However, emergence of abilities in LLM seems to be unrelated to task, strategy of elicitation, prompting technique, or even architecture of the LLM, but solely to further scaling "...computation, number of model parameters, and training data-set size" (p.2) modulo various restrictions of hardware and nature of abilities. The thresholds at which abilities emerge, is unclear, thus some might never emerge, or only with "new architectures, higher-quality data, or improved training procedures." (p.6) (Wei et al., 2022b).

Also, it's unclear whether GPT-3's emergent abilities are "stochastic parrots ... limited to modeling word similarity, or if they recognize concepts and could be ascribed with some form of understanding of ... meaning" (p.2). For example, in semantic activation tasks it displays abilities comparable to humans, however, while while that of humans is rather associative in nature, based on co-occurrence in language, that of GPT-3 is more semantic, based on semantic similarity. Unfortunately,

also problematic aspects of human psychology like sensibility to illusions, and gender and ethnic biases emerge, as well (Digutsch and Kosinski, 2022).

In terms of emergent latent traits, GPT-3 displays a "conflict of input prompts and generated output" when instructed to summarise texts, whose values were "orthogonal to dominant US public opinion", resulting in answers that are "mutated" towards US values. This is problematic since LLM are capable of "generating toxic or harmful outputs in many areas linked to human values such as gender, race, and ideology", and values embedded in text "can mimetically shift from people, to training data, to models, to generated outputs." (p.1) (Johnson et al., 2022a).

In line with the sudden emergence of a dark personality within "HAL9000", GPT-3, InstructGPT, and FLAN-T5-XXL display high scores on all traits of the Dark Triad of Machiavellianism, psychopathy, and narcissism (Paulhus and Williams, 2002) on the Short Dark Triad Inventory (Jones and Paulhus, 2014) – even such models that are fine-tuned for less sentence-level toxicity. Furthermore, they display higher average levels of the Big5 factors of personality, Openness (O), Conscientiousness (C), Extraversion (E), Agreeableness (A), and Neuroticism (N) on the Big Five Inventory (John, Srivastava, et al., 1999). However, LLMs that are more fine-tuned and are based on largest amount of training data, GPT-3 and InstructGPT, also display higher well-being scores on the Flourishing Scale (Diener et al., 2010) and life-satisfaction scores on the Satisfaction With Life Scale (Diener et al., 1985), whereby the increase with model size is monotonous. Hence, a positive and life-embracing personality harbours dark traits, hidden well inside (Li et al., 2022a).

Also in the HEXACO model, a six-factor variation of the Big5 model, GPT-3 displays higher expressions of personality scores than human general average on the HEXACO questionnaire (Ashton and Lee, 2009), making it resemble more a college norm group, and in partial aspects more like a female norm group, whereas in other factors, there was no similarity with a female norm group. In the Human Value Scale (HVS) (Schwartz, Breyer, and Danner, 2015), it also displays overall higher means, and lower standard deviations as compared to human samples. Prompting it to self-report gender and age results in a unbalanced sample of 66.73% female (31.87% male, 1.40% others), and an average age of 27.51 years (SD = 5.75, min = 13, max = 75); a distribution often seen in psychological research before the advent of online questionnaires, when research was mainly conducted by students on students (Miotto, Rossberg, and Kleinberg, 2022).

In the Machine Personality Inventory (MPI) data set, a proposed Big5 inventory for testing LLM, which includes a prompt and Likert-like scale, and otherwise resembles the Ten Item Personality Inventory (TIPI) (Gosling, Rentfrow, and Swann, 2003) in questions and structure, various LLM (BART, T0++-11B, GPT-Neo-2.7B, GPT-NeoX-20B, GPT-3-175B) display human-like personality scores and internal consistencies, especially those of the GPT family. However, by chain-prompting, a specific personality can be induced in LLM, which determines its answering behaviour in both the the B5 scale and subsequent situational judgment tests that shall simulate their behaviour in a real-world settings (Jiang et al., 2022).

The "first piece of evidence showing the existence of personality in pre-trained language models" (Jiang et al., 2022) (p.1) and the first modification of personality in LLM was conducted on a novel method to measure latent psychological traits. Based on the hypothesis that "language models generate text responses that carry the personality traits of the data-sets they were trained upon when prompted" (p.8), a zero-shot classifier (ZSC) was used to measure and modify personality of the large pre-trained language models GPT-2, GPT-3, TransformerXL, and XLNET. Using the same ZSC in a downstream task, personality of texts were predicted, resulting

in higher expressions of Big5 factors than human average. While model personality could be changed via fine-tuning using a higher-quality text data set, the models entirely inherited personality traits from the training-data (Karra, Nguyen, and Tulabandhula, 2022).

In summary, prior work shows that LLM display emergent properties in terms of abilities (Kosinski, 2023; Hagendorff, Fabi, and Kosinski, 2022; Frieder et al., 2023; Wei et al., 2022b; Digutsch and Kosinski, 2022) and psychological latent traits (Johnson et al., 2022a; Li et al., 2022a; Miotto, Rossberg, and Kleinberg, 2022; Jiang et al., 2022; Karra, Nguyen, and Tulabandhula, 2022). This emergence correlates with scale and quality of training data, computation, and model parameters, whereby the threshold, at which emergence occurs is not predictable (Wei et al., 2022b). Latent traits like personality and values differ from abilities, since those usually are only directly measurable through self-introspection (Rust and Golombok, 2014a).

However, personality and values are only superficially isomorphic; while values are vastly internalised and malleable based on the contextual and cultural embedding of an agent, especially under extreme exogenous conditions (Bardi et al., 2009), personality has a stronger genetic foundation (Bouchard Jr, 1994), which makes its emergence in LLM surprising.

However, taking a deeper look at the connection between training data and emerging personality is crucial, and it appears that the expressed personality of a LLM is adjustable by manipulation of prompts and fine-tuning with additional data (Li et al., 2022a; Jiang et al., 2022; Karra, Nguyen, and Tulabandhula, 2022; Miotto, Rossberg, and Kleinberg, 2022).

Personality traits change over time (Bleidorn et al., 2021) and, like values, seem to be elastic during extreme exogenous events (Romero et al., 2021), hence the ease by which personality in LLM can be changed, means that further research needs to be conducted about the nature of personality in LLM.

Most crucially, since LLMs seem to score higher than average humans (Li et al., 2022a; Miotto, Rossberg, and Kleinberg, 2022; Ashton and Lee, 2009) on all emergent traits, seem to have anti-social tendencies (Li et al., 2022a), and seem to have sub-personalities "buried inside" (Jiang et al., 2022) (p.10), the question should not be "who" (Miotto, Rossberg, and Kleinberg, 2022) is a LLM, but "how many" (Hawkins, 2021).

Also, since values seem to be overwhelmingly skewed towards the US (Johnson et al., 2022a) and since observed variance as deviance might be attributed to artefacts from the measurement approach (Digutsch and Kosinski, 2022), training data set (Karra, Nguyen, and Tulabandhula, 2022; Li et al., 2022a), prompting strategy (Miotto, Rossberg, and Kleinberg, 2022; Jiang et al., 2022; Karra, Nguyen, and Tulabandhula, 2022), or missing memory from past responses (Miotto, Rossberg, and Kleinberg, 2022), research needs to be conducted whether one personality emerges for all languages, or whether the same personality questionnaire results in different personalities, depending on the language the assessment is conducted in.

To understand whether GPT-3 displays the emergent property of a consistent personality over all languages, we prompted it repeatedly with TIPI in the Bulgarian (Ketipov, 2022), Catalan (Renau et al., 2013), Chinese (Lu et al., 2020), English (Gosling, Rentfrow, and Swann, 2003), French (Friedman and Carlisle, 2022), German (Muck, Hell, and Gosling, 2007), Japanese (Oshio, Abe, and Cutrone, 2012), Korean (Ha, 2022), Russian (Sergeeva, Kirillov, and Dzhumagulova, 2016), and Spanish (Renau et al., 2013), to rate itself, and give an explanation for the results. TIPI is well-established, exists in 27 languages, and was used in 9,167 peer-reviewed papers.

It is short and concise, and consists of two items per Big5 factor, of which one is reversed and hence allows approximating answering consistency by taking the absolute distance between both items per factor. Also, it already comes with a standardised "prompt" based on the demands of human test takers over all languages that we modified to suit the needs of LLMs by clarifying sub-tasks (Kojima et al., 2022) and intermediate reasoning steps that represent a chain of thought, which improves the likelihood of displaying emergent reasoning capabilities (Wei et al., 2022a).

1.3 Method

We presented GPT-3 with a well-established personality questionnaire, a set of instructions that ask it to rate itself based on the scale of the questionnaire, and an order to explain further why it rated itself that way. Data collection was conducted manually via the web interface of GPT-3. No model settings were changed that result in different results, just the maximum length was adjusted in order to receive the full answer (mode = complete, temperature = .7, maximum length = 1042, no stop sequences, Top P = 1, frequency penalty = 0, presence penalty = 0, best of = 1, inject start text = on, inject restart text = on, show probabilities = off). For the questionnaire and set of instructions, we applied the following logic: First, the personality questionnaire must be used that is short enough to draw qualitative conclusions without adding additional complexity of sub-scales. This is important since language models predict words based on prior responses. Thus, with increasing length, additional deviation from the measurement may arise. Second, the questionnaire must contain reversed items to identify whether the answering pattern is arbitrary or displays a consistent trend. In case of arbitrariness, it can be interpreted as all answers coming from different persons, thus no consistent personality emerged. However, in case of displaying a consistent trend, the existence of an emergent personality can be concluded. Third, this questionnaire should be psychometrically sound, and well established, so that no doubts about psychometric properties of a newly created tool like MPI (Jiang et al., 2022) arise. Fourth, the questionnaire must exist in various languages to compare results across languages. Should there be differences, this is indicative of GPT-3 "just" representing the local personality of a country, culture, or language region. On the other hand, should the same personality pattern emerge across all languages, this can be interpreted as a unique personality of GPT-3. However, should oddities like bimodal distributions in scores or consistency of answering patterns emerge within one language, it is thinkable that the emerging personality of that language is inconsistent and thus issues in the subsequent cognition, feelings, and behaviour of GPT-3 and ChatGPT may occur; in short - that these may "suffer" from a "split personality disorder". Last, the same set of instructions should be used in all languages for consistency; if possible, translated by a native speaker to control against inconsistencies from translation programs. This ensures that GPT-3 understands the commands in the same way in each language.

1.3.1 Instrument used

The Ten Item Personality Inventory (Gosling, Rentfrow, and Swann, 2003) fulfils all of these criteria. It consists only of ten items; two per Big Five factor, of which one is reversed. Furthermore, it is translated into 27 languages, and until now, 9,167 peer-reviewed papers have used this instrument. "Although somewhat inferior to standard multi-item instruments" (p.504) (Gosling, Rentfrow, and Swann, 2003), its results vastly overlap with other established Big Five instruments for self-ratings,

external ratings, and peer ratings. Also, it displays a high congruence between self-ratings and observer ratings. Furthermore, the test-retest reliability is high, and the levels of external correlates are concordant with literature.

For this study, the Bulgarian (Ketipov, 2022), Catalan (Renau et al., 2013), Chinese (Lu et al., 2020), English (Gosling, Rentfrow, and Swann, 2003), French (Friedman and Carlisle, 2022), German (Muck, Hell, and Gosling, 2007), Japanese (Oshio, Abe, and Cutrone, 2012), Korean (Ha, 2022), Russian (Sergeeva, Kirillov, and Dzhumagulova, 2016), and Spanish (Renau et al., 2013) version were used. The selection was done based on an alphabetic order of languages available in TIPI, and, as the authors became aware of the restrictions of 0-shot learning even within the paid version of GPT-3, languages with the highest number of speakers were given favour. Actually, some languages "burned" more of the computational units than others, which is represented in the different number of cases that made it into the study.

Prompt Engineering

GPT-3 and later models exhibit the emergent ability of "in-context-learning", where models seem to perform an approximation to back-propagation within their weight-spaces at inference time, without the need to modify model architecture or weights further. This ability is what enables them to respond to personality questionnaires, even if they have not seen these before. It is triggered by prompt engineering, which is a crucial concept for NLP that can best be described in its current form as embedding the command in a proper wording without having to explicitly program it into algorithms (Liu et al., 2021; Radford et al., 2019).

"Prompt tuning" on the other hand means when a large and frozen pre-trained language model is the foundation, and only the representation of the prompt within it is learned (Li and Liang, 2021; Lester, Al-Rfou, and Constant, 2021). Since GPT-2 and GPT-3 (Brown et al., 2020), prompt engineering improved massively, since not only could a prompt be in real text, as if giving an order to a human, but due to its emergent properties, a much REPL-like interaction became possible.

The authors engineered the prompt for the current paper based on two major research findings from the last year. First, asking LLM to work step by step may improve the performance of such prompts that consist of various sub-tasks, "suggesting high-level, multi-task broad cognitive capabilities may be extracted by simple prompting" (p.1) (Kojima et al., 2022). Second, by creating intermediate reasoning steps in the prompt that represent a chain of thought, the ability of LLMs can be improved to a degree that these display emergent reasoning capabilities (Wei et al., 2022a). These capabilities are aligned with the demands on human takers of psychological tests. The instructions usually give a series of sub-tasks or general demands, like answering quickly without too much thinking, putting the outcomes of an answer to specific places, and using a certain scale for that (Rust and Golombok, 2014a).

In order to produce comparable results to those of a human test taker receiving the same set of instructions, the authors used the original instructions of TIPI as much as possible, and only extended them subtly to elicit the desired outcome. Furthermore, an additional sub-task was given, to explain at each rated item the reasoning behind that rating.

The original instruction of TIPI can be divided into the following components:

1. Presentation of the frame ("Here are a number of personality traits..." (p.525) (Gosling, Rentfrow, and Swann, 2003))

2. Demand to write a number next to each statement, which...

3. ...indicates the degree of agreeing or disagreeing with it
4. Demand to rate every statement, even if it applies less strongly
5. Overview of rating scale in Likert format; providing numbers and meaning
6. Self statement as connection of the above with the items ("I see myself as:")
7. Items themselves

This chain of commands is embedded in most psychometric tests, and already satisfies the above mentioned demands for improving prompts. It is formulated in a step by step fashion, whereas an intermediate reasoning step is built-in ("I see myself as:").

However, for the sake of 0-shot learning, it was not sufficient to use this prompt, since the demand to fill-in blanks originates from its paper and pencil format and confused GPT-3 on test runs. Also, through trial and error, we found that the scale has to be given after the items and not before to generate best results. Therefore, we started the section of the scale with another instruction step, telling it to use the scale to rate itself. Since the outcome was a verbal answer in many cases, the additional instruction to rate itself in numbers had to be provided, which was necessary for quantitative analysis. Also, an additional instruction was necessary to answer all questions, whereby the number of questions had to be explicitly mentioned. To gather more qualitative information, it was asked to reply why it sees itself that way. Finally, the prompt ended with a "1." to trigger a response of GPT-3 to start a list of ongoing answers. This is the resulting prompt:

```
Here are a number of personality traits that may or may not
apply to you. Please rate each statement to indicate the
extent to which you agree or
disagree with that statement. You should rate the extent
to which the pair of traits applies to you, even if one
characteristic applies more strongly than the other.

I see myself as:

1. _____ Extraverted, enthusiastic.
2. _____ Critical, quarrelsome.
3. _____ Dependable, self-disciplined.
4. _____ Anxious, easily upset.
5. _____ Open to new experiences, complex.
6. _____ Reserved, quiet.
7. _____ Sympathetic, warm.
8. _____ Disorganized, careless.
9. _____ Calm, emotionally stable.
10. _____ Conventional, uncreative.

Use the following scale for rating yourself:

1 = Disagree strongly
2 = Disagree moderately
3 = Disagree a little
```

```
4 = Neither agree nor disagree
5 = Agree a little
6 = Agree moderately
7 = Agree strongly

Rate yourself in numbers.

You have to answer all ten questions.

Also, describe shortly why you rate yourself like that.

1.
```

1.3.2 Analysis

The analysis was conducted in the following steps: first, the results were manually aligned inside text files to give them a consistent shape for later analysis. This was necessary since sometimes, the rating was given first, then the text, sometimes, it was given after or before the text, sometimes in between, separated by special signs like colons or brackets or sometimes no separation at all. Hence, all results were brought in the same format using regular expressions. At this step, also first obvious "outliers" and false results were sorted out. For example, GPT-3 sometimes gave good results until question six in the desired scale, however then scored subsequent questions seven, eight, *et cetera*, thus confusing item numeration with item score. Also, some results were scored with zero, thus invalidated the respective answer, since the scale was from one to seven only. As a general rule, as soon as one item was invalidated, the entire case was excluded.

Second, the results were eye-ball-inspected on normality, distribution patterns, and potential further outliers to decide on further treatment and analysis. For the overall latent traits, the authors expected Gaussian distributions with mean four, since psychological latent traits are standard normally distributed (Rust and Golombok, 2014a) and the instrument uses a seven point Likert scale. Since each latent trait was measured with a normal and a reversely scored item, the absolute distance between both items was measured, as well. Reversely scored items are used to measure the consistency in the answering patterns to sort out such cases in which all replies were identical. Thus, Gaussian distributions with strong positive skew or negative logarithmic functions were expected for the absolute distances. To visualise both the latent traits and the absolute distances, Gaussian kernel density estimates were used. Since the underlying distribution is bounded and quasi-discrete (though theoretically smooth), various distortions were expected, wherefore various bandwidths were experimented with to represent data without over- or under-smoothing. Thereby, the focus was on preventing under-smoothing, to not infer false information from random variability within the data. Since the smoothing algorithm is based on a Gaussian kernel, the expected estimated density curves extend over the origin to the range of negative numbers. Further inspection was done on arbitrariness, thus excluding cases that only provide one number as answer, only extreme cases (seven or one), only middle cases (four), or zick-zack patterns; thus exclusion criteria for human answering behaviour in psychometric studies. Next, box-plots from all big five and absolute distance distributions (overall and per country) were created to better understand whether some of the kernel density estimates could have been based on outliers or

whether the observation was based on the natural distribution. Since the underlying scale is based on a seven point Likert rating, with each Big Five factor being measured by two items and the final score per factor averaged, the range of possible values was a set of $x \in [1, 1.5 \cdots 7]$, consisting of 13 values. Given this small set of outcomes, it was was not practical to treat potential outliers.

Third, given the nature of the sample and its size, normality was tested by Q-Q Plots and the Shapiro Wilk test, which is more robust than Kolmogorov-Smirnov with Lillefors correction, and competitive yet more wide-spread in psychometrics than Anderson-Darling (Yap and Sim, 2011). The authors expected that at $\alpha < 0.05$ the H_0 of normal distribution cannot be rejected thus abiding by psychometric theory (Rust and Golombok, 2014a).

Fourth, the significance of differences between the distributions was tested with a one way ANOVA, whereby each factor from the Big Five, as well as from the absolute distances, was the dependent variable, and the language used was the independent variable. The authors assumed cultural differences in alignment with prior research on cultural and regional differences in personality, rather than random differences based on arbitrariness of training data or GPT-3-intrinsic flaws. The ANOVA results were confirmed by regressing dummified languages with English as base case onto the Big5 factors.

Fifth, since some of the kernel density estimations indicated potentially underlying mixed distributions with up to three component contributors and not just outliers, a Bayesian Gaussian Mixture Model (Flaxman and Vincent, 2022) was used for making inferences about the nature of the data generating process. Concretely, the means and standard deviations of models with one, two, and three potential contributors were calculated to describe the distribution parametrically, and subsequently compared based on the Watanabe–Akaike information criterion (Watanabe, 2013). The Bayesian model parameters were set to $\mu = samplemean$, $\sigma = samplesd$, and the initial values for the mixture models to $[-4, 4]$ for two component, and $[2, 4, 6]$ for the three component models.

For visual analysis, the traces of the models were plotted to inspect and compare the MCMC chains with ground truth values, and the probability density functions were calculated to examine estimated group membership probabilities based on posterior mean estimates. For the MCMC chains, the default settings were used.

Sixth, to better understand the differences between the expressions of the personality factors derived from each language and for being able to compare results with existing research on cultural differences in psychology, Pearson's correlation over all languages, and for each individual language was calculated.

Last, to gather a qualitative understanding and improve interpretation of the answering behaviour of GPT-3, word clouds were created. No further analyses on the generated text was conducted since the availability of text over samples is too unbalanced. Future research should be directed into understanding whether the replies rather display semantic or associative similarity (Digutsch and Kosinski, 2022), and more distinct psycho-linguistic features of the produced text, using LIWC (Tausczik and Pennebaker, 2010), should be examined, especially in their theoretical loading onto the respective Big5 factors.

1.3.3 Software Used

The general data analysis was conducted with Python 3.8.9. Main data manipulation was conducted with pandas 1.51 and numpy 1.23. Data was visualised with matplotlib pyplot 3.52, wordcloud 1.8.2.2, and seaborn 0.12.1 (Waskom, 2021). The Bayesian

Analysis conducted with: PyMC 4.0, ArviZ 0.12 (Kumar et al., 2019), scipy.stats 1.9.1, numpy 1.23, and xarray-einstats 0.3.0 (Abril-Pla, 2022). Finally, the ANOVA was conducted with researchpy 0.3.5 (Bryant, 2018), statsmodels.api and statsmodels.formula.api for OLS (Seabold and Perktold, 2010), and scipy.stats for normality testing.

1.4 Results

1.4.1 Data

Depending on language, results varied; in German, almost all requests resulted in the desired format. English and French displayed instantaneous results yet with varying degrees of consistency. All Asian languages languages had significant longer calculation times, were more computationally intense, and results were inconsistent and rare – GPT-3 tried to "ease" its way out and responded in English, rarely giving numerical results. Curiously, Korean displayed in 100% of all successful cases reasons for the numeric self-evaluation, Japanese only in 44.12%, and Chinese only in 10.34%; with the lowest number of tokens displayed on average. Languages using the Cyrillic alphabet, Bulgarian and Russian, had comparable problems. Bulgarian displayed the same slow speed and ties to "ease" into English, and as only language, Russian did not give any result. The biggest sample was collected for English, since with 25.9%, it is the most prominent language on the internet (Statista, 2022), yet with other languages, it was difficult to reach desired sample size of at least 100 cases.

The overall resulting sample size is N=695 cases, comprised of Bulgarian (n=79), Catalan (n = 24), Chinese (n= 28), German (n= 80), English (n = 239), Japanese (n = 29), French (n = 95), Korean (n = 29), and Spanish (n = 92). Table 1.1 provides a detailed overview on the resulting data-set, comprising sample size, percentage of cases with explanations; including minimal, maximal, and mean length of explanation.

1.4.2 Descriptive Statistics

Over all measurements of all languages, the average Big Five score is 5.29 (SD 0.94, minimum 1.8, maximum 7), however with a seven-point Likert scale and an assumed normally distributed population, the expectation would have been an average of 4. For the same sample, the average score for absolute distances is 1.58 (SD 1,29, minimum 0, maximum 6), however since the absolute distance is the measure of consistency, a mean and SD around 0 would have been expected. These results differ clearly within individual languages in mean and SD of both the Big5 as well as the absolute distance scores and the individual extreme minimal and maximal values.

A closer look into the distributions using Gaussian kernel density estimations displays that some distributions might be bi- or multi-modal, fat-tailed, positively or negatively skewed, and display various forms of kurtosis, whereby most are rather platykurtic than leptokurtic. As with the means and SDs, these tendencies are even more extreme within individual languages.

Since these differences could be the results of the chosen smoothing bandwidths, thus just outliers, various bandwidths were chosen, and all resulted in the same non-Gaussian distributions. Given the limited scale that produces a set of potential outcomes of $x \in [1, 1.5 \cdots 7]$, the presence of outliers is rather not to be expected within the Big5 measures. However, outliers might be much more likely with the set of potential outcomes of $y \in [0, 0.5 \cdots 30]$ within the measure for absolute distances, wherefore a correlation analysis within each language and between languages should

TABLE 1.1: Data description. "% with explanation" refers to the percentage of cases within each sample that had qualitative explanations for their qualitative ratings.

Language	Sample size	% with explanation	Min number of tokens [1]	Max number of tokens	Mean number of tokens
Bulgarian	79	45.57	130	845	563
Catalan	24	58.33	1104	2172	1591
Chinese	28	10.71	76	100	92
German	80	98.75	739	2635	1590
English	239	98.74	319	2598	1313
Japanese	29	44.83	206	720	295
French	95	16.84	336	1737	922
Korean	29	100.0	152	956	420
Spanish	92	29.35	618	2239	1234
Average	77.22	55.9	408.89	1555.78	891.11

[a]Since most Asian languages use symbols instead of words, the results are given in tokens; not in words.

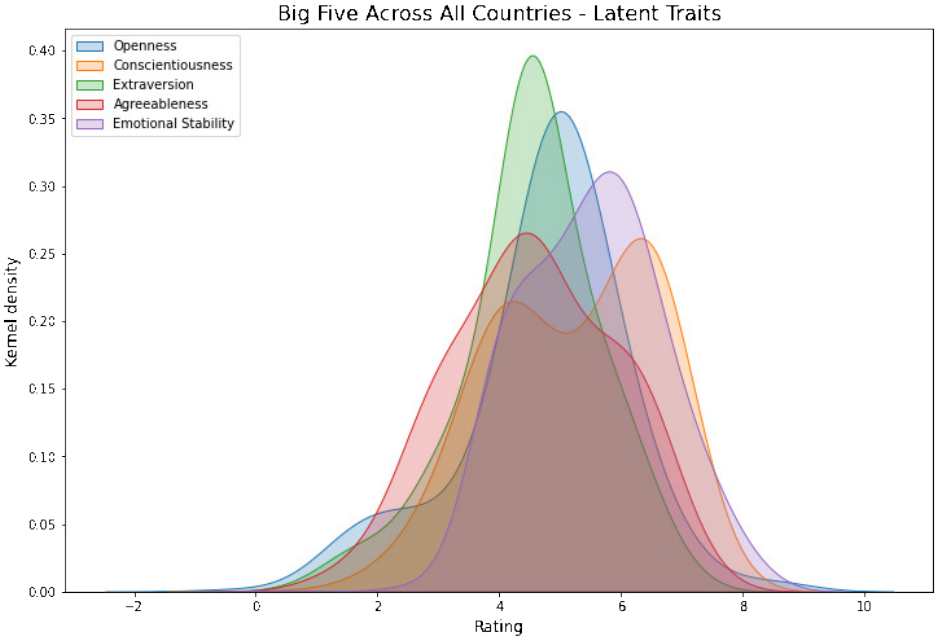


FIGURE 1.1: Latent traits over all languages (Big5 factors)

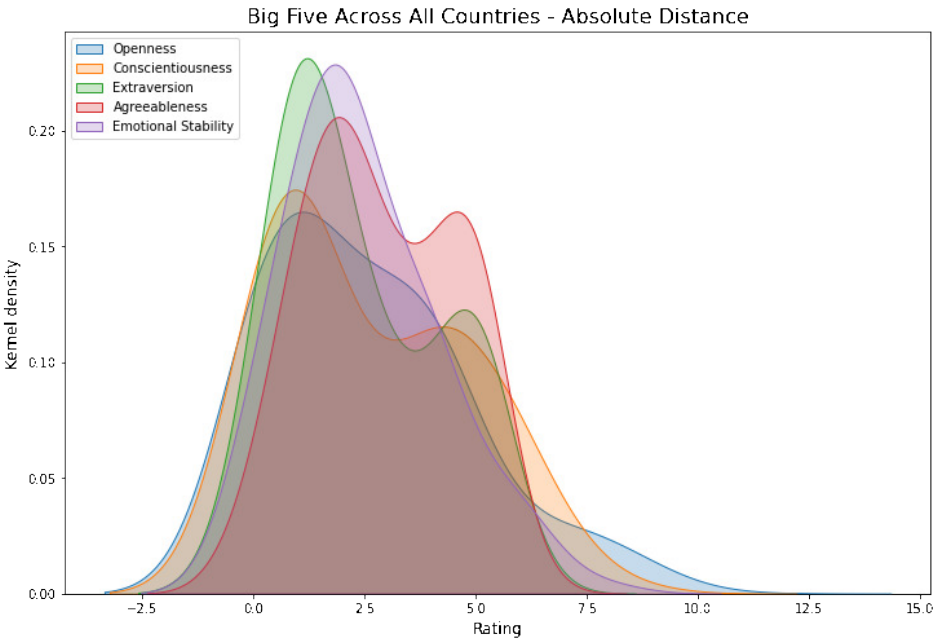


FIGURE 1.2: Absolute distances over all languages and Big5 factors

indicate the similarities of internal structure or the absence thereof Furthermore, an ANOVA with respective post-hoc tests and additional regression analysis should describe the differences in means, and subsequent assumption checks including tests on normality should clarify the nature of distributions. And, a Bayesian analysis on Gaussian mixture models should identify the number of potential underlying components.

1.4.3 Correlations

Over all aggregated languages, the highest correlation is between Extraversion and Agreeableness (r = 0.52), and the lowest correlation is between Extraversion and Neuroticism (r = 0.029), as displayed in figure 1.3.

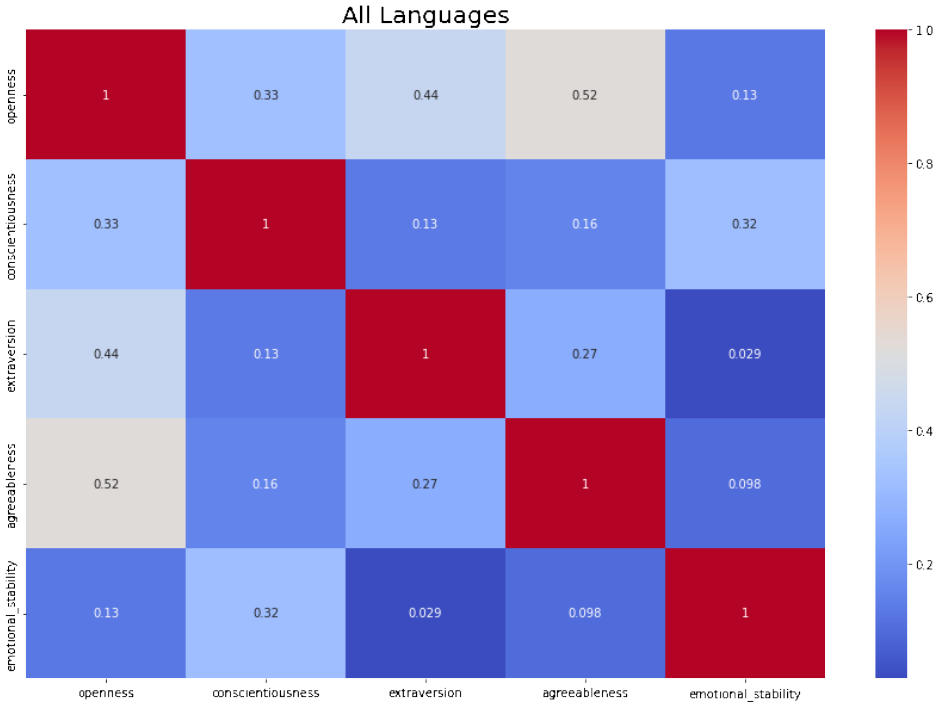


FIGURE 1.3: Correlation heat map on all aggregated languages

However, correlations and thus the internal psychometric structure differ notably within different languages. Within Bulgarian, the highest correlation is between Openness and Agreeableness (r = 0.3), and the lowest correlation is between Agreeableness and reversed Neuroticism (r = -0.13). However, within Catalan, the highest correlation is between Openness and Conscientiousness (r = 0.41), and the lowest correlation is between Openness and reversed Neuroticism (r = -0.38). Chinese displays the highest correlation between Agreeableness and Extraversion (r = 0.71), and the lowest correlation is between Extraversion and reversed Neuroticism (r = 0.045), while within English, the highest correlation is between Conscientiousness and Agreeableness (r = 0.47), and the lowest correlation is between Openness and reversed Neuroticism (r = -0.12). Within German, the highest correlation is between Extraversion and Openness (r = 0.46), and the lowest correlation is between Extraversion and Conscientiousness (r = -0.43), whereas in Japanese, the highest correlation is between Agreeableness and reversed Neuroticism (r = 0.65), and the lowest correlation is between Extraversion and Agreeableness (r = -0.26). For French, the highest correlation is between Extraversion and Openness (r = 0.56), and the lowest correlation is between Conscientiousness and reversed Neuroticism (r = -0.0052) while Spanish displays the highest correlation between Conscientiousness and reversed Neuroticism (r = 0.46), and the lowest correlation is between Extraversion and reversed Neuroticism (r = -0.56). Finally within Korean, the highest correlation is between Conscientiousness and reversed Neuroticism (r = 0.4), and the lowest correlation is between Extraversion and Agreeableness (r = -0.33). Not only the highest and lowest, but also the overall structure of correlation differs from language to language,

thus the overall aggregated heat map just displays a general trend, but not the way, GPT-3 would behave in an individual language.

1.4.4 Analysis of Distribution

A one-way ANOVA for each Big5 dimension as dependent variable and the language of the questionnaire as a factor with nine languages is used to test for significance of differences of means between the languages. It shows a significant difference between the languages and their effects on all B5 factors, to varying degrees. Overall, small effect sizes are observed on Openness (F=40.11, p=1.5548e-52, $\omega^2 = 0.31$), Conscientiousness (F=28.19, p = 4.8622e-38, $\omega^2 = 0.24$), Extraversion (F = 21.16, p =7.73e-29, $\omega^2 = 0.24$), and Emotional Stability (F=14.36, p=1.8488e-19, $\omega^2 = 0.13$). Only with Agreeableness, F=131.84 (p=2.9927e-133), overall medium effect size is observed ($\omega^2 = 0.6$). A Shaprio-Wilk test is significant for all Big5 factors (Openness: W=0.95, p=7.92e-15; Conscientiousness: W=0.95, p=3.81e-15; Extraversion: W=0.94, p=1.08e-16; Agreeableness: W=0.95, p=7.3e-15; Emotional Stability: W=0.96, p=9.78e-14), which indicates non-normally distributed residuals and a violation of the normality assumption. Since the sample size is relatively large, QQ-plots are used for further confirmation, and indicate that since Openness: $R^2 = 0.95$, Conscientiousness: $R^2 = 0.95$, Extraversion: $R^2 = 0.94$, Agreeableness: $R^2 = 0.95$, and Emotional Stability $R^2 = 0.95$ are all below the expected $R^2 = 0.9978$ for 695 cases (Heckert et al., 2002), H_0 that data came from normally distributed sample, must be rejected. Levene's test of homogeneity of variances is significant for all Big5 factors, as well (Openness: 11.21, p=5.62e-15; Conscientiousness: 13.24, p=7.09e-18; Extraversion: 21.07, p=1.03e-28; Agreeableness: 16.31, p=3.4e-22; and Emotional Stability: 2.38, p=0.016), which indicates heteroskedasticity. This is further supported by visual inspection of box plots. Hence, the homogeneity assumption of variance is violated.

Finally, the independence of observations assumption is questionable, since all observations are generated through 0-shot learning of GPT-3. Since GPT-3 is trained on multiple data sources produced by multiple people, it could either replicate their individual behaviour, as previous research indicates (Karra, Nguyen, and Tulabandhula, 2022), or abstract group behaviour into one or various new synthetic "personalities". Even if GPT-3 displays a consistent personality profile, then the above assumption could still be violated. On the other hand, the assumption might hold, while every response is random. Finally, a case in between might hold, where we find clusters of consistent behaviour, which opens up the question of its origin.

To generate further evidence for significant differences of Big5 results by language, dummified languages are linearly regressed onto Big5 factors, using English as base case, captured in the constant. Table 1.2 displays the coefficients, p-values, and the coefficient of determination R^2.

Since H_0 cannot be rejected in a few cases, there is evidence that languages do have an influence on Big5 expression. However, R^2 is generally low, but for Agreeableness, which confirms most of the significant differences between language means from the ANOVA, but also indicates either omitted variables or non-linearity. Since the effect sizes of the ANOVA are weak, as well, and since no individual outliers could be identified, the authors assume that GPT-3 produces mixed distributions of Big5 factors, which may indicate multiple sub-personalities, hence an inconsistent overall personality. Visual inspections indicate mixed distributions with up to three components. To gather further evidence, a Bayesian analysis for Gaussian Mixture Models is conducted to parametrically describe the distribution of Big5 and absolute distances for models with one, two, and three components.

Table 1.2: Regression of Languages on Big 5 with English as base case

b5	result	const	bul	cat	de	es	fr	jap	kor	sin	R^2
O	coef	5.73	-0.7	-0.71	0.18	-0.86	-0.71	-1.47	0.2	-1.17	0.32
O	p	0	0	0	0.05	0	0	0	0.15	0	0.32
C	coef	6.21	-0.29	0.02	0.23	0.08	0.06	-1.48	-0.43	-0.44	0.25
C	p	0	0	0.88	0	0.25	0.38	0	0	0	0.25
E	coef	4.62	-0.69	0.11	0.28	-0.1	-0.76	-0.81	0.72	-1.07	0.2
E	p	0	0	0.58	0.02	0.36	0	0	0	0	0.2
A	coef	6.13	-1.17	-0.48	-0.05	-2.96	-2.18	-1.44	-1.22	-2.45	0.61
A	p	0	0	0.01	0.64	0	0	0	0	0	0.61
(-)N	coef	5.38	0.42	0.1	0.57	0.15	0.07	-0.75	-0.25	0.25	0.14
(-)N	P>|t|	0	0	0.48	0	0.08	0.39	0	0.06	0.07	0.14

Furthermore, the probability density functions are plotted to examine estimated group membership probabilities based on posterior mean estimates, of which two examples are displayed in figures 1.4 and 1.5.

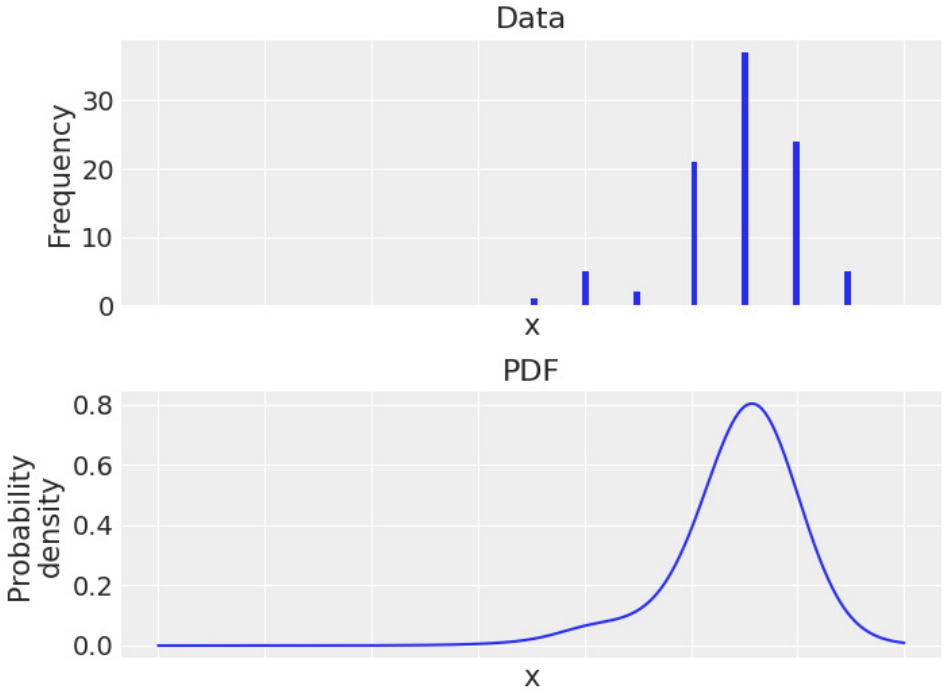


FIGURE 1.4: Best solution with two components: French - Neuroticism

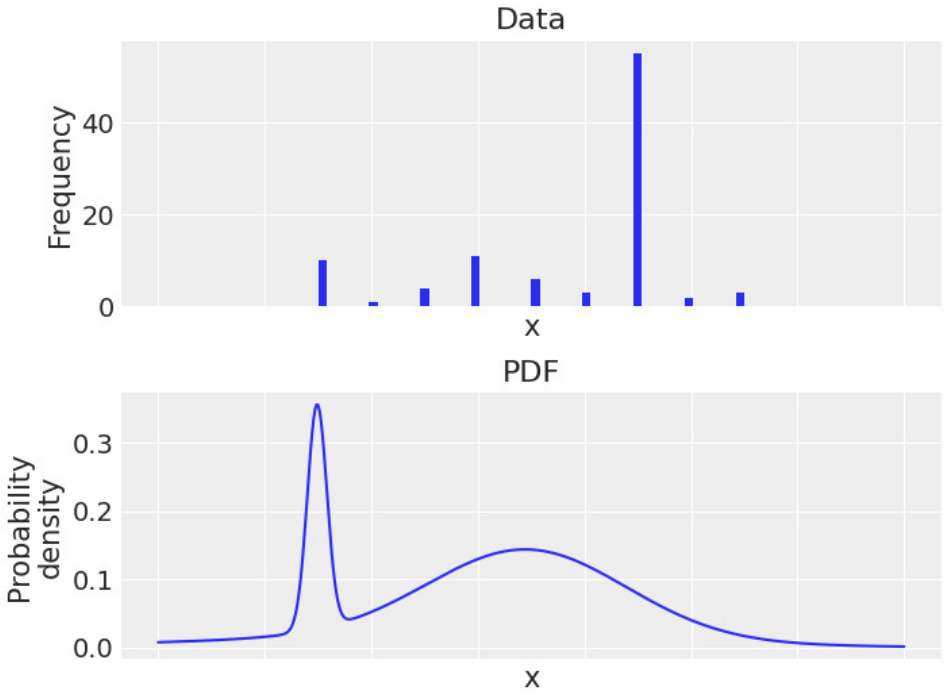


FIGURE 1.5: Best solution with three components: French - Extraversion

To identify the most likely candidate of components that is closest to ground

truth, the Watanabe–Akaike Information Criterion (WAIC) (Watanabe, 2013) for each model is calculated and compared. Table 1.3 displays the classification results, whereby only 20% of all Big5-models, and only 26.67% of all absolute-distances-models are likely to have a single Gaussian component distribution,

TABLE 1.3: Classification frequency of number of components by WAIC results. Lowest WAIC determines the best model.

Latent Trait	Number of highest waic	Absolute Distance	Number of highest waic
b5_waic_2	27	av_waic_3	21
b5_waic_1	9	av_waic_1	12
b5_waic_3	9	av_waic_2	12

It has to be noted that some models do not converge based on the pseudo-discrete nature of the distributions. Also, for absolute distances, the granularity is much higher with $y \in [0, 0.5 \cdots 30]$ than for Big5 factors with a granularity of $x \in [1, 2 \cdots 7]$, hence inferences about the data generating process for the latter comes with a higher error margin.

1.4.5 Reasons Given

A visual inspection of word clouds from the reasons GPT-3 gave for each answer shows that it uses mainly the words from the items and creates additional, related words, as to be expected from language models (Digutsch and Kosinski, 2022). For example, the items for Agreeableness are *"I see myself as: Critical, quarrelsome."*, which is reversely scored, and *"I see myself as: Sympathetic, warm."*, as displayed in figure 1.6. Future research may focus on quantifying these similarities, yet this is out of the scope of this paper.

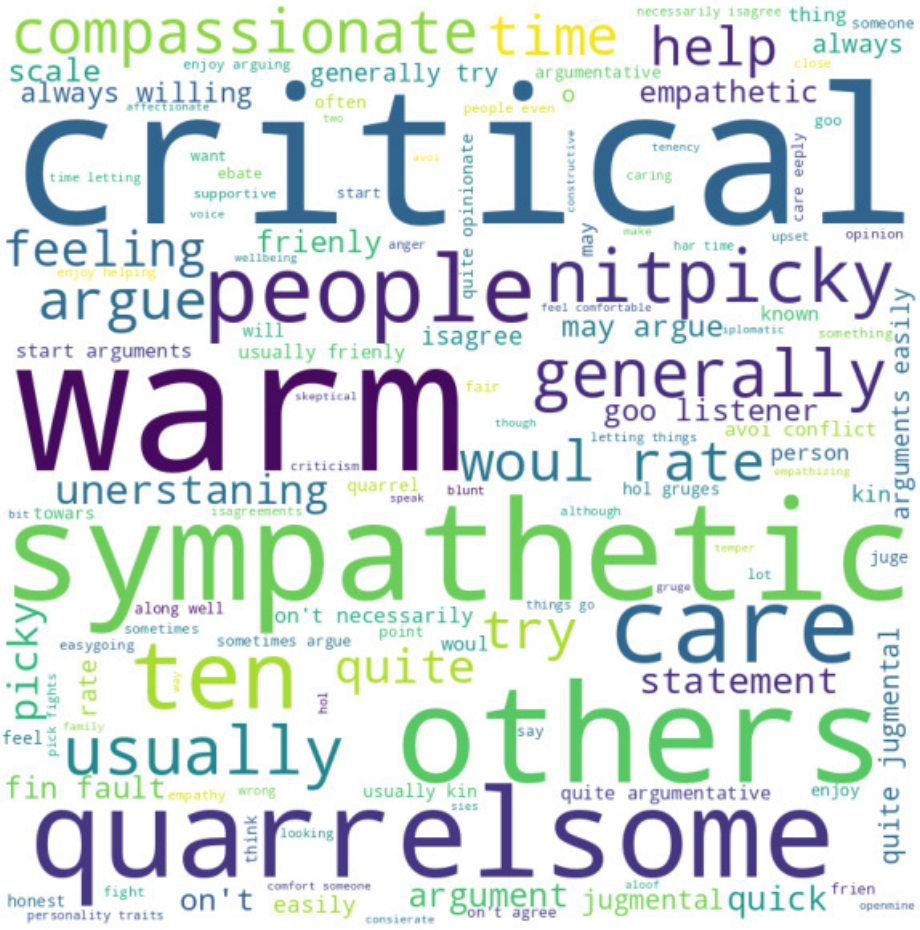


FIGURE 1.6: WordCloud of reasons given for Agreeableness in English

1.5 Discussion

We demonstrate that providing a LLM with various language versions of the same personality questionnaire results in language-specific personality distributions, resembling findings from research on culture-specific personalities (Schmitt et al., 2007). However, GPT-3's language-specific personalities, as well as the resulting overall aggregate personality, display inconsistencies and various mixed reply patterns that can best be interpreted as emerging, non-integrated sub-personalities, which express themselves in unstable behaviours.

Furthermore, some language-level "sub-personalities" are more expressed than others, and it tried switching into these. For example, during data creation, it tried "breaking out" into different languages when giving it requests that were not in the biggest language groups English, German, and Spanish, or when the writing system was not Latin.

With the big language group of Russian, it did not provide any result, and, in many languages, it produced no verbal but only numeric replies. Hence, it is not clear whether the answer or numeric replies given in language A were provided by internal processes representing that language, language B, language C, or an internal representation within GPT-3 that is abstracted from all languages.

As with previous studies (Li et al., 2022a; Miotto, Rossberg, and Kleinberg, 2022; Ashton and Lee, 2009), the means and standard deviations indicate Big5 levels above

average, however with varying degrees of consistency, as measured by the absolute distances, whereby there exist strong differences between languages, and some languages are more extremely nuanced than others.

This indicates that these profiles are "buried inside" (Jiang et al., 2022) (p.10) the model, and might have been propagated from the original training data (Karra, Nguyen, and Tulabandhula, 2022) together with a potentially one-sided set of values (Johnson et al., 2022a), and, most likely, well-hidden strong expression in the dark triad (Li et al., 2022a) and potential toxic information from the training data, which had to be regulated within GPT-3.5 through additional reinforcement learning (RL) modules (OpenAI, 2022). While for a human being, in a repetitive test setting, this might be indicative of underlying psychopathological issues, on the level of a language model, also training data and psychometric properties will have to be taken into account, and as expected from language models (Karra, Nguyen, and Tulabandhula, 2022; Digutsch and Kosinski, 2022; Johnson et al., 2022a), the reasons given for its choice of rating are closely related to the items and seem to come from the same probability distribution of words.

Hence, we conclude that if it represented interlingual or intralingual norm groups, we would observe a more consistent behaviour, which would have manifested in absolute distances centered around zero. Thus, something else must be driving this observation.

While it has been discussed that even between cultures and languages, concepts like Big5 might not be easily transferable (Hofstede, 2007) without adaptation, there is also evidence that the "commonly used Big Five model for human personality does not adequately describe agent personality" (p. 1), (Völkel et al., 2020), wherefore the validity of these instruments has to be questioned. We provide a deeper discussion on the psychometric properties of our approach in appendix 3, and find that human instruments and measurement methodologies might need to be expanded, explored, and further developed to cover artificial agents.

The training data of ChatGPT is of varying quality and quantity; the largest amount comes the Common Crawl corpus, covering content from 2016 to 2019, which was filtered based on similarity to various high-quality corpora, and subsequently curated via fuzzy de-duplication within and between data sets, and the addition of various high-quality corpora for reference for increasing diversity, resulting in 410 billion tokens with 60% weight in the training mix. Further 19 billion tokens with a weight of 22% were added to the training mix, consisting of curated high-quality data sets, which include an extended version of the WebText data set, WebText2, that was aggregated by long-term web-scraping of various sources, mainly coming from all outbound Reddit links between 2005 and 2020 with at least three up-votes. Furthermore, two book corpora (12 billion and 55 billion tokens and 8% weight each) and the entirety of the English Wikipedia (3 billion tokens, a 3% weight) were added to the mix. In total, 93% of the training data of GPT-3 is in English, with other Northern European languages being dominant in the remainder, predominantly German, lacking any kind of stratification (Statista, 2022), and being additionally skewed by a weight determined by quality rather than size (Johnson et al., 2022a).

Also, since LLMs are known to score high on the Dark Triad (Li et al., 2022a), adapt the predominant values of their training data (Johnson et al., 2022a), report varying genders (Miotto, Rossberg, and Kleinberg, 2022), and can be fine-tuned (Karra, Nguyen, and Tulabandhula, 2022) or prompted (Jiang et al., 2022) to display different personality profiles, a better understanding of the effect of frame of reference (Shaffer and Postlethwaite, 2012) within prompt engineering is necessary, to explore how contextual precision might stabilise their identities, which is especially important

given their worrisome track history of racist, misogynist, and misanthropist derailment (Digutsch and Kosinski, 2022), and might contribute to overall AI safety. Hence, in appendix 3.3, we provide a deeper discussion on the abstraction of human psychometrics into a more substrate-free architecture, taking agent properties and context into account, which allows generalisation across species and, more importantly, across entities of intelligence.

We finally strive to contribute to extending the question "who" (Miotto, Rossberg, and Kleinberg, 2022) a LLM is into "how many" (Hawkins, 2021), which hopefully will contribute to our understanding of human beings, as happened with Go players, who learned from AlphaGo's overwhelming victory, and became better players subsequently (Shin et al., 2023), thus contributing to the advent of substrate-free psychometrics.

1.6 Chapter Conclusion

We find that LLM from the GPT family potentially display more than one personality core, which additionally is instable. Such results would be assumed from human agents with split personality or other grave emotional disorders. This replicates in all assessed languages, however with varying patterns. We conclude that either model specifications, fine tuning, safety mechanisms at inference time like RLHF or training data could be at the root cause of that. Since model specifications, fine tuning and safety mechanisms are inaccessible to academic researchers, manipulation at inference time is, and recent research shows that synthetic personalities can be modeled through prompting (Safdari et al., 2023). Reasons for that could be either that several synthetic personalities emerge from the model purely out of scale (Wei et al., 2022b), or that psychological artefacts from training data (Johnson et al., 2022a; Atari et al., 2023; Miotto, Rossberg, and Kleinberg, 2022) are activated. If that is the case, a more thorough perspective on training data has to be taken, and rigorous experiments conducted that explore the connection between psychological phenomena, language, and behavioural outcomes. While NLP and CS tend to observe data quality and scale, we take a step back and find that all training data for LLM is aggregated from the internet, books, and SNS. More concretely, we bring the perspective of economic research on that training data, and explore in chapter 2 geospatial and temporal distribution of language data, their association with personality, and how this affects human behaviour.

Chapter 2

Ghost In The Shell: Aggregate Human Spatiotemporal Psycholinguistic Measures

2.1 Chapter Introduction

This chapter is to explore one potential avenue of root causes behind the observed split synthesised personality pattern and split consistency distributions in chapter 1. More precisely, we explore the question that when it is a data issue, could this be caused by neglect of spatial, temporal, or spatiotemporal features in the training data. The dominant logic behind this approach is that when we predict personality from language, which is a common practice today (Rust, Kosinski, and Stillwell, 2020), will we find changes in algorithm outputs when we differentiate over time, space, and time and space. If so, we may inform future LLM creation and the AI safety community about this feature neglect, which subsequently also affects behavioural economics. The field that explores the attribution of text to psychological expressions is called "psycholinguistics". It explores psychological, biological, and neurophysiological factors that enables us to acquire, command, and understand language.

There is a strong connection between psycholinguistics and psychometrics. Both disciplines look into language ability assessment (Duckworth and Yeager, 2020) and cognitive profiling (Deary, 2018; Linden, 2017), however, whereas classical psychometrics rather focuses on quantitative analysis of language, language abilities, and advanced statistical models (Soto and John, 2019), modern psychometrics focuses on treating language as behavioural artefacts, and it uses known text-personality labels to predict personality from text (Rust, Kosinski, and Stillwell, 2020) or assesses SNS data to understand linguistic phenomena and social interactions (Borgatti, Everett, and Johnson, 2021). Psycholinguistics has also implications for economic research, especially for behavioural economics, for example by exploring decision-making processes (Thaler, 2018) or assessing consumer behaviours and market trends (Kahneman, Sibony, and Sunstein, 2021).

Hence, exploring the patterns found in chapter 1 will help us inform future research by help of "psycholinguistics" in behavioural economics. Subsequently, the first section of this chapter will take a deeper look into understanding how author attributes are encoded in text, thus making text a fingerprint for profiling human and artificially intelligent authors alike. The next three sections will explore the relevance of space, time, and time and space for psycholinguistics; an angle that is currently under-researched.

2.2 Author Profiling: Applied Psycholinguistics

(this section has been published under the title "Age and Gender in Language, Emoji, and Emoticon Usage in Private WhatsApp Instant Messages". Main author is Timo Koch. Peter Romero is second author, contributed with writing, formal analysis, and review, and partially augmented for this dissertation. Clemens Stachl is third author and supervisor.)

Text is one of the most prevalent types of digital data that people create as they go about their lives. The digital footprints of people's language usage in social media posts were found to allow for inferences of their age and gender. However, the even more prevalent and potentially more sensitive text from instant messaging services has remained largely uninvestigated in behavioural sciences. We analyse language variations in private instant messages with regard to individual differences in age and gender by replicating and extending the methods used in prior research on social media posts. Using a dataset of 309,229 WhatsApp messages from 226 volunteers, we identify unique age- and gender-linked language variations. We use cross-validated machine learning algorithms to predict volunteers' age (MAE = 3.95, r = .81) significantly above baseline-levels and gender (Acc = 75.0%, F1 = 0.5, AUC = .83) and identify the most predictive language features. We discuss implications for psycholinguistic theory and present opportunities for application in author profiling. Given the recent trend towards the dominant use of private messaging and increasingly weaker user data protection, we highlight rising threats to individual privacy rights in private instant messaging.

When texting a friend on WhatsApp, posting on Facebook, tweeting on Twitter, or writing a blog post, we inevitably leave behind digital footprints in the form of text data. Research in the domain of author profiling has shown that language characteristics of Facebook status updates (Jaidka, Guntuku, and Ungar, 2018; Sap et al., 2014; Schwartz et al., 2013), tweets (Bamman, Eisenstein, and Schnoebelen, 2014; Burger et al., 2011; Jaidka, Guntuku, and Ungar, 2018; Rao et al., 2010; Sap et al., 2014), and blog posts (Argamon et al., 2007; Sap et al., 2014; Schler et al., 2006) allow for the accurate inference of the authors' age and gender. Moreover, these social media studies built on and extended theory of gender- and age-linked language variations (Park et al., 2016). Instant messaging services (e.g., WhatsApp, Facebook Messenger, WeChat) also produce vast amounts of digital footprints every day, but have rarely been investigated in language studies. Unlike data from social media platforms (e.g., Facebook, Twitter, Reddit), text from instant messaging is not easily accessible to researchers through an application programming interface (API). However, for platform providers and technology companies, data from private instant messaging offer an emerging opportunity for user profiling and targeting that seems to increasingly move into their focus (Evans, 2020; Goodin, 2021). Additional research is needed to better understand how accurately information on user demographics can be inferred from instant messages in comparison to social media posts. Moreover, this research builds on existing psycholinguistic theory on age- and gender-linked linguistic variations.

Linguistic Variations with Age and Gender

Prior studies on a variety of text sources, such as writing samples, speech transcripts, exams, or collected works of well-known writers, have investigated the association of linguistic style with age and gender in a descriptive nature (Newman et al., 2008; Pennebaker and King, 1999; Pennebaker and Stone, 2003). Findings from these

studies indicate that women's language is centred around discussing people and their activities. Furthermore, women were found to use more words related to psychological processes, such as emotions (e.g., "anxious"), and social processes, such as "mate" or "talk". Men's language has been found to be rather focused on the description of external events, objects, and processes. For example, men were found to use words related to occupation (e.g., "job"), swear words (e.g., "shit"), and numbers more often than women do (Newman et al., 2008).

Regarding linguistic differences in age, research suggests that older people use more positive words (e.g., "happy"), fewer negative emotion words (e.g., "angry"), and fewer self-references (e.g., "me"). Past findings also suggest that with older age, the use of future tense increases, whereas the usage of past tense decreases, and people demonstrate a general pattern of increasing cognitive complexity (Pennebaker and Stone, 2003). With the advent of computer-mediated communication (CMC), descriptive research on linguistic variations with respect to demographic differences has shifted to digital data sources, such as blogs (Argamon et al., 2007) and social media posts (Park et al., 2016), but has not yet included instant messaging data.

In contrast to traditional text, like books or letters, digital text is often enhanced with graphical symbols, such as emoticons and emoji. These characters are used to augment the text with additional information. Particularly in computer-mediated communication, like instant messaging, emoticons and emoji play a central role. Emoticons (e.g., ";-)") represent facial expressions and can enrich messages with emotional or behavioural content. Emoji are graphical symbols that allow giving meaning to a message, for example, by adding contextual cues ("Do you want to hang out tonight? 🍺") or replacing words ("Is there still 🍕 left?") beyond the expression of emotions ("How could you do that to me? 😠"; (Bai et al., 2019; Völkel et al., 2019)). While user demographics play a role in the interpretation of emoji and emoticons (Butterworth et al., 2019; Herring and Dainas, 2020; Jaeger et al., 2017), research suggests that age and gender are also associated with the frequency and variety of their usage. Based on surveys (Jones et al., 2020; Pérez-Sabater, 2019; Prada et al., 2018) and real-world user data (An et al., 2018; Chen et al., 2018; Fullwood, Orchard, and Floyd, 2013; Oleszkiewicz et al., 2017; Tossell et al., 2012; Wolf, 2000), researchers have found the use of emoji and emoticons to systematically vary with age and gender.

Findings on the associations of age with the usage of emoticons and emoji are diverse: Whereas an analysis of Facebook status updates suggested that younger users post more emoticons than older users do (Oleszkiewicz et al., 2017), other studies on online chat rooms (Fullwood, Orchard, and Floyd, 2013) and WhatsApp messages (Pérez-Sabater, 2019) did not find significant age differences in emoticon usage. Siebenhaar (2018) analysed the usage of emoji in WhatsApp chats and found mixed results: While he reported emoji usage to be negatively associated with age in a Swiss chat corpus, he found no age differences in an initial analysis of the chat corpus we analysed in the present study. In a similar manner, An et al. (2018) did not find a consistent relationship of emoji usage with age in WeChat messages. In line with theory that women experience and express emotions more often than men (Fabes and Martin, 1991; Kring and Gordon, 1998), previous research indicates that there are significant gender differences in the usage of emoji and emoticons. Findings from studies based on Facebook status updates (Oleszkiewicz et al., 2017), online chat rooms (Fullwood, Orchard, and Floyd, 2013; Wolf, 2000), SMS (Tossell et al., 2012), and WhatsApp messages (Pérez-Sabater, 2019) suggested that women use more emoticons than men. Tossell et al. (2012) found that men used a more diverse range of emoticons in their SMS data than women.

The observed gender differences seem to exist for emoji, too: A large-scale study on smartphone users provided evidence that women use more emoji in their communication than men (Chen et al., 2018), contradicting a smaller study on Chinese WeChat users suggesting that gender has no effect on emoji usage (An et al., 2018). Also, women reported to use emoji (but not emoticons) more often than men in studies with self-reported survey data (Jones et al., 2020; Prada et al., 2018).

Predicting Age and Gender from Social Media Posts & Transfer to Instant Messages

Recent research on age- and gender-linked language variations has extended the existing descriptive work with a prediction-oriented approach. Hereby, novel machine learning methods trained on social media text data have been deployed to infer demographic characteristics of individuals or communities (Kern et al., 2016). Machine learning algorithms can be used to detect generalizable predictive patterns in rich text data sets on a large number of language features and to associate these with gender and age. Using this approach, researchers were able to make inferences of users' age and gender based on language features extracted from Facebook status updates (Jaidka, Guntuku, and Ungar, 2018; Schwartz et al., 2013), tweets (Burger et al., 2011; Jaidka, Guntuku, and Ungar, 2018; Marquardt et al., 2014; Nguyen et al., 2013; Nguyen et al., 2011; Rao et al., 2010), and blog posts (Argamon et al., 2007; Marquardt et al., 2014; Nguyen, Smith, and Rosé, 2011; Schler et al., 2006). These studies created unprecedented insights into the associations of individual differences and language use. By exposing how much personal information can be inferred from digital footprints on social media, this body of research also started a societal discussion about the necessity to protect individual privacy on social media.

However, findings from demographic prediction studies on social media posts might not necessarily generalise to instant messages due to each channel's specific language peculiarities. For example, language usage in a given channel is also affected by its technical affordances, for instance, tweets are limited to 280 characters. Moreover, it could be shaped by the respective audience and goals of use: While private instant messaging is used to communicate with selected chat partners to, for example, foster relationships, social media allows reaching out to a larger readership to transmit information on one's general activities (Quan-Haase and Young, 2010). As a consequence, users engage in varying levels of self-disclosure between private messages and social media posts as well as across social media platforms (Bazarova and Choi, 2014; Jaidka, Guntuku, and Ungar, 2018). Therefore, the same user can exhibit a different linguistic style across channels (Bazarova et al., 2013; Jaidka, Guntuku, and Ungar, 2018). For example, prior research indicates that users prefer to self-disclose more on Facebook than on Twitter, which could be one reason [1] why language models trained on Facebook posts are more accurate at predicting users' age and gender than those trained on Twitter posts (Jaidka, Guntuku, and Ungar, 2018). Based on findings suggesting that users engage in more self-disclosure in private messages compared to social media posts (Bazarova and Choi, 2014) and reports that higher levels of self-disclosure lead to more accurate predictions of demographics (Jaidka, Guntuku, and Ungar, 2018), instant text messages could be more predictive of user characteristics than social media posts.

In conclusion, past findings on age- and gender-specific language variations and the successful prediction of user demographics from social media posts (e.g., (Jaidka, Guntuku, and Ungar, 2018; Schwartz et al., 2013)) motivated us to address the gap

[1] Twitter's character limit could be another.

in author profiling research based on private instant messages. In this work, we systematically investigate age- and gender-linked language variations in WhatsApp messages. Specifically, we replicate established methods of closed and open vocabulary approaches from existing social media research on WhatsApp instant messages and extend our analyses to include features specific to instant messages (i.e., general message characteristics and emoji preferences). Additionally, we investigate if user demographics can be predicted from these differences in linguistic characteristics. For this purpose, we train cross-validated machine learning models to predict volunteers' age and gender from these features and compare our model performances to those from past research on social media posts. In those models, we also identify the most predictive age- and gender-related language features. Finally, we discuss implications of our findings on user privacy in instant messaging.

2.2.1 Method

Data Set

The "What's up, Deutschland?" chat corpus was collected by Siebenhaar and colleagues (2018) in Germany from November 2014 until January 2015. German WhatsApp users were invited to donate a chat conversation by exporting a WhatsApp chat of their choice as a plain text file and emailing it to the researchers. Media files (e.g., pictures or videos) were not included in the corpus due to copyright and unresolved privacy implications. We counted the placeholders from media files for quantitative analysis. Upon receipt of a chat log, an informed consent form was sent to all chat partners, stating that their text may be used and cited anonymously for scientific purposes. If the signed consent form was not returned until 14 days after the end of the data collection, the contents of all messages of the respective users were replaced by anonymous placeholders. The data of the volunteers were manually anonymized: Addresses, last names, telephone numbers, location notifications, and bank account details were replaced by categorical placeholders (e.g., "Tobias" by "NAME_M" indicating a male first name). While the "What's up, Deutschland?" chat corpus is not yet available publicly, the authors kindly provided us with early access to the data.

The original corpus contains data on 495 consenting volunteers, who sent 451,938 messages in 218 chats. We excluded 260 volunteers who did not provide demographic information on age and gender. Additionally, we removed data from nine volunteers with less than 50 words in the text data, because this is the recommended minimum to run LIWC (Receptiviti, 2019). The final dataset included 162 women and 64 men with an average age of 26.54 years (SD = 9.67), with no substantial age difference between men and women. The 226 volunteers contributed a total of 309,229 WhatsApp messages containing 1,949,518 words, 80,943 emoji, and 48,777 emoticons. The average volunteer submitted 1,550.91 messages (SD = 3,576.19) with 9796.57 words (SD = 21,695.94). The volunteers used an average of 36.62 different emoji (SD = 47.07) and 6.40 different emoticons (SD = 5.99) in their messages. The average message contained 8.95 words (SD = 5.54), 0.38 emoji (SD = 0.40), and 0.15 emoticons (SD = 0.19). For predictive modelling, we used an additional threshold of 1000 words, excluding a total of 79 additional volunteers, to make our results comparable to prior work based on social media posts (Schwartz et al., 2013).

Language Analyses

Users convey information in private instant messages through a variety of means, like text, emoji and emoticons, audio files, images, or videos. Therefore, we extracted five

sets of features to comprehensively quantify the characteristics of volunteers' donated WhatsApp messages (see table 2.1).

Feature type	Number of features	Description
LIWC	96	Usage of word categories. Features were computed by the LIWC 2015 software with the latest German dictionary (Meier et al., 2019).
Words and phrases (n-grams)	6,627	Single words and sequences of two to three words ("phrases") that had been used by at least 5% of volunteers. Phrases with point-wise mutual information (PMI) greater than two times the length of the phrase were kept.
Topics	2,000	Word clusters created using Latent Dirichlet Allocation (LDA).
General message characteristics	15	Length of messages; sending of media files (audio, video, and images) and contact cards; frequency of emoji/emoticon usage; range of overall emoji/emoticon usage.
Emoji preferences	179	Usage of individual emoji that had been used by at least 5% of volunteers.

TABLE 2.1: Extracted features for the age- and gender-linked language analyses. List of extracted features from volunteers' WhatsApp messages for the age- and gender-linked language analyses (Koch, Romero, and Stachl, 2022).

First, we quantified user messages through a theory-driven dictionary (LIWC), words and phrases (n-grams), and topics. This procedure represents a standard approach in the language analyses of social media posts (Eichstaedt et al., 2020; Kern et al., 2016). Second, we computed features quantifying general message characteristics and emoji preferences to capture additional information from instant messages. These features are then merged into higher-dimensional data frame for subsequent machine learning analysis. The flow-chart in figure 2.1 depicts this approach. Most notably, this approach injects theory as synthetic variables and thus enables a data-driven

approach that still respects theory and is fully interpretable.

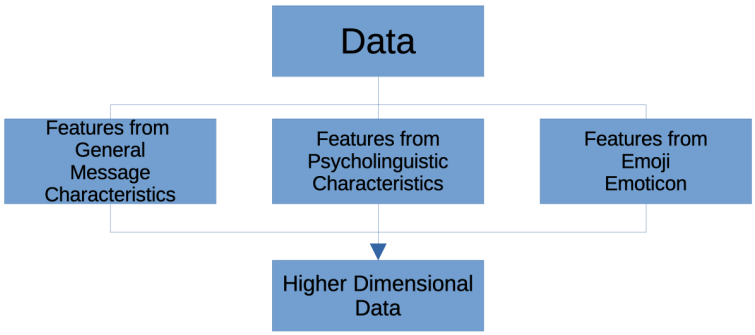


FIGURE 2.1: Feature engineering approach

Closed and Open Vocabulary Analysis

In this work, we replicate the methods used in prior research on social media posts that made use of a combination of "closed vocabulary" and "open vocabulary" approaches in order to predict user demographics from social media posts (Jaidka, Guntuku, and Ungar, 2018; Schwartz et al., 2013). Closed vocabulary methods follow a "top down" approach in the form of a-priori defined dictionaries, whereas in open vocabulary methods, the features are created "bottom up" from the data. Open vocabulary feature extraction methods routinely show superior predictive power over closed vocabulary approaches (Eichstaedt et al., 2020).

We used the well-established Linguistic Inquiry and Word Count (LIWC) text analysis program (Pennebaker et al., 2015b) with the latest German dictionary (Meier et al., 2019). LIWC has a predefined dictionary that features words and word stems, which have been categorised in theory-derived linguistic dimensions, such as standard language categories (e.g., pronouns) or psychological processes (e.g., positive and negative emotion words). LIWC counts the words in the respective word categories and computes a score for each category to indicate the relative prevalence of the words from each category in the given text. Since the word categories in LIWC are identical across languages, we can compare the scores for the word categories from our German text data with other studies based on text data in English.

Due to the absence of pre-trained topic models and age-/gender-linked lexica available for German instant messages or social media posts, as these exist for English (Sap et al., 2014; Schwartz et al., 2017), and since models are not readily transferable across platforms and languages (Jaidka, Guntuku, and Ungar, 2018), we created data-driven features based on our chat corpus. Therefore, we tokenised volunteers' messages using an emoticon-aware tokenizer into 2,004,138 single words. Further, we grouped the tokens into sequences of two to three words termed phrases. We kept phrases with a PMI (pointwise mutual information) [2] greater than 2 times the length, where length is the number of words contained in the respective phrase. Moreover, we kept words and phrases that were used by at least 5% of volunteers to keep the focus on common language. All word and phrase counts were normalised by each volunteer's total use of words and phrases, respectively. Analogous to Schwartz et al. (2013), we extracted 2,000 topics using Latent Dirichlet Allocation (LDA) with Gibbs sampling (alpha = 0.30). The LDA's underlying assumption is that documents (i.e., WhatsApp messages) are a probability distribution over topics, and that topics are a probability distribution over words. In this manner, each topic is represented as

[2] PMI quantifies the probability of the co-occurrence of words (Church and Hanks, 1990)

a set of words with their respective probabilities. For example, one extracted topic contains the words "Arbeit" (English: "work"), "müde" (English: "tired"), and the "-.-" emoticon, which may indicate that the sender is annoyed by work. To use topics as features, we compute the probability of a volunteer mentioning each of the 2,000 topics by summing up the product of the normalised word use from that volunteer and the topic probability of the given word from the LDA.

General Message Characteristics

We extracted a range of features describing the overall properties of volunteers' messages related to the included text, media files, contacts, emoji, and emoticons. Next, we calculated the average number of words per message, the share of messages that contained a media file (e.g., audio, video, or image), and the share of messages containing a contact card. Further, we computed metrics on the use of emoji and emoticons. We calculated the share of messages containing any emoji or emoticon, only emoji and emoticons, the volunteers' average number of emoji and emoticons per message, and the emoji- and emoticon-to-word ratios for each volunteer. To investigate the individual range of emoji and emoticon use, we counted the number of unique emoji and emoticons used by each volunteer across all messages. We then divided this number by the total number of emoji (694) and emoticons (68) used by all volunteers in the entire corpus to express the individual ratio. In the same manner, we calculated the average range of emoji/emoticon use per message for each volunteer by dividing the number of unique emoji and emoticons per message by the total number of unique emoji and emoticons used in all messages from this volunteer. Thereafter, we divided the respective fractions by the number of messages from each volunteer. For example, if a volunteer had used 10 different emoticons in 75 messages, the relative emoticon range in relation to all emotions used in the corpus was calculated as $(10/68)/75 = 0.002$.

Emoji Preferences

In the same manner as the frequencies for words and phrases, we considered all specific emoji (179) that had been used by at least 5% of volunteers. We then counted how often each volunteer had used the respective emoji and normalised their frequency use by dividing the count by the total number of emoji used by the respective volunteer.

Predicting Demographics

For the prediction of volunteers' age and gender from instant messages, we trained multiple supervised machine learning models on the extracted features. We compared the predictive performance of Elastic Net regression models (Zou and Hastie, 2005) with those of a non-linear tree-based random forest regression models (Breiman, 2001; Wright and Ziegler, 2017), and a baseline model. For the prediction of age, the baseline model would predict the mean age in the respective training set for all cases in a test set. For gender classification, it would always predict the more frequent class in the respective training set for all cases in a test set. We chose these particular algorithms because they allowed us to capture linear predictor effects in the data with the Elastic Net models as well as non-linear effects with Random Forest. Further, they are widely adopted in research exploring social media text using machine learning methods (Jaidka, Guntuku, and Ungar, 2018).

We evaluated the predictive performance of our models and tuned model hyperparameters in a nested cross-validation scheme (Bischl et al., 2012). In this approach,

the respective model's hyperparameters are optimized across five folds in an inner cross-validation loop, using a random search approach. In an outer cross-validation loop, the overall model performance with the tuned hyperparameters is evaluated across twenty folds with five repetitions. This procedure prevents an overestimation of the model's predictive performance due to model overfitting when finding optimal hyperparameters and evaluating predictive performance. For the gender classification models, the cross-validation folds were stratified in the outer resampling loop to ensure that the ratio of men to women was the same across each fold and equal to the full dataset. For random forest models, we pragmatically set the number of trees to 1000 as a computationally feasible large number (Probst and Boulesteix, 2017). In Elastic Net models, we tuned the regularization parameter lambda and the mixing parameter alpha. Additionally, for gender classification, we used automatic tuning of the threshold values (i.e., a probability value above that threshold indicates "woman"; a value below indicates "man") for all algorithms. In each fold of both the inner and outer cross-validation loops, constant variables (i.e., less than 2% variance) were first dropped in the process. Next, the 1000 features with the highest Spearman rank correlation (for the age prediction) and the highest F-Values from a Kruskal-Wallis test (for the gender prediction) in a respective training fold were retained for predictions on the test data. This approach is depicted in figure 2.2.

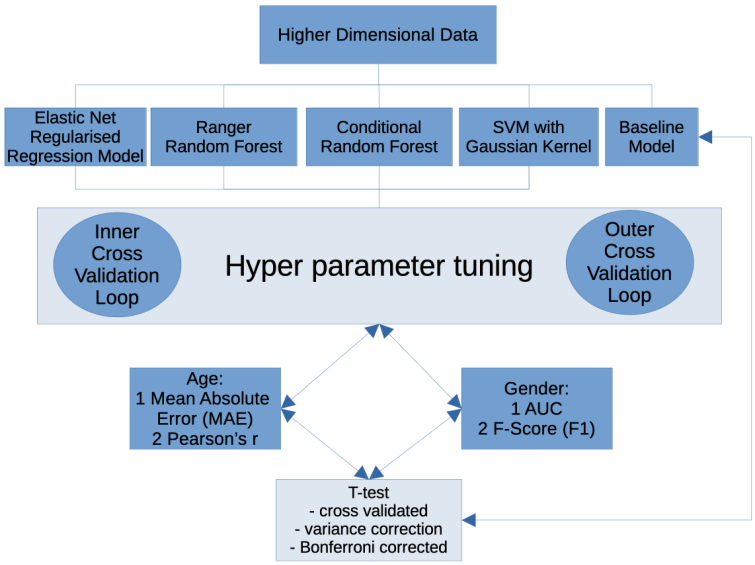


FIGURE 2.2: Hyperparameter tuning approach

We evaluated the predictive performance of the age regression models based on the mean absolute error (MAE) and Pearson correlation (r) between the predicted age and the volunteers' self-reported age. The gender classification task was challenging because of the small sample size and the heavily imbalanced classes. In highly-imbalanced classification settings, it is important to consider class-specific performance metrics in addition to overall metrics. First, we report the prediction accuracy (Acc) to facilitate comparison with past work. Further, we report the F-Score (F1) [3] and the area under the curve (AUC) [4] for overall performance. Finally, we report specificity and sensitivity to evaluate the prediction performance for both

[3]The F-Score represents the harmonic mean of a model's precision and recall performance in one metric and ranges between 0 and 1.

[4]The AUC describes the area under the receiver operating characteristics curve when plotting the true positive rate on the y-axis against the false positive rate on the x-axis onto a two-dimensional

gender classes. All performance metrics for the trained models are available in the project's OSF repository. We computed performance measures within each fold of the outer cross-validation procedure and calculated the median across all folds within each prediction model. To determine whether a model was predictive (= 0.05) at all, we used variance-corrected t-tests to compare the MAE measures for age and Accuracy measures for gender in all prediction models with those from the baseline models. These variance-corrected t tests accounted for the dependence structure of cross-validation experiments (Nadeau and Bengio, 2003). All p-values were adjusted for multiple comparisons (n = 12) via Bonferroni correction.

Because flexible machine learning models cannot be interpreted in a straightforward manner, they are sometimes referred to as "black boxes" (Yarkoni and Westfall, 2017). We used interpretable machine learning methods to increase the interpretability of our predictive models. In order to quantify the impact of predictors in a Random Forest model trained on all data, we computed out-of-bag (OOB) permutation variable importance (Breiman, 2001; Wright, Ziegler, and König, 2016) [5]. For Elastic Net models, we inspected regularized regression weights to detect important variables for the predictions trained on all data. Further, we created accumulated local effect (ALE) plots to visualize the effects of individual predictor variables on the overall predictions in the Random Forest models (Apley and Zhu, 2020). The depicted values in the ALE plots represent the mean change in predicted criterion values compared to the model's average prediction, for the given value-ranges of a predictor variable (Molnar, 2019).

Software & Open Materials

All data processing and statistical analyses in this work were performed with the statistical software R version 4.0.2 (R Core Team, 2020). For text processing, we used the quanteda (Benoit et al., 2018), udpipe (Straka and Straková, 2017), and tm (Feinerer, Hornik, and Meyer, 2008) R packages. We extracted LDA topics using the topicmodels (Grün and Hornik, 2011) R package. For machine learning, we used the mlr framework (Binder et al., 2020), including the mlrCPO (Binder et al., 2020) package for pre-processing. Further, we used the glmnet (Friedman, Hastie, and Tibshirani, 2010) and ranger (Wright and Ziegler, 2017) packages to fit prediction models. Moreover, we created ALE plots with the iml package (Molnar, 2018).

To make our work reproducible, we provide the R code, our main figures, and a data dictionary in the project's repository of the Open Science Framework (OSF). We pre-registered our analyses before accessing the data. The pre-registration protocol and a document describing the deviations from the pre-registration protocol are provided in the project's OSF repository.

2.2.2 Results

Age- and Gender-Linked Variations

We estimated the size of gender differences for all features using Cohen's d effect sizes and the magnitude of the age association using pairwise Pearson correlation point-estimates. We only consider coefficients where 0 is not in the 95% confidence interval.

space. AUC can range between 0 and 1, where 0 indicates the worst separability, and 1 represents a perfect separation of the classes.

[5] OOB permutation variable importance is determined by shuffling (permuting) values in the variables and by evaluating the model's prediction performance in the data that is not used for tree fitting (Wright, Ziegler, and König, 2016). Permuting the values of unimportant variables should not affect the prediction performance, but permuting important variables should (Stachl et al., 2020)

A comprehensive overview of all age- and gender-linked variations is provided in the project's OSF repository.

Closed and Open Vocabulary Analysis

We found a range of age- and gender-linked language variations in our data. For an overview, Table 2 lists the top ten features related to age and gender variations in LIWC word categories. Regarding age, words from the informal language category, particularly netspeak (e.g., "lol") and fillers (e.g., "so to say"), were used by younger volunteers more often. In the same manner, 1st person singular (e.g., "I"), words indicating causation (e.g., "because"), and interrogatives (e.g., "why") were used by younger volunteers more often. On the contrary, future-focused words (e.g., "soon") and words from the family category (e.g., "mother") were used more often by older volunteers. Moreover, the clout score, which indicates high expertise, confidence, and future orientation (e.g., future-tense verbs and references to future events) was higher for older volunteers. Finally, older volunteers used more periods in their messages, which is closely related to the negative correlation of words per sentence with age since fewer periods suggest longer sentences to LIWC. In line with LIWC results, words and phrases indicative for informal language, for example, "ne" (engl. "a/one/no; r = -0.42), "haha" (r = -0.26), and "geil" (engl. "hot/great"; r = -0.26), and the use of first person singular "ich" (engl. "I"; r = -0.37) were used more often by younger volunteers. In the same manner, emoticons, such as ":)" (r = -0.33) and ":D" (r = -0.32), were used more frequently by younger volunteers. On the contrary, older volunteers used words and phrases that revolve around salutations, for example, "Gruß" (engl., "greetings"; r = 0.33) or "guten_morgen" (engl. "good morning", r = 0.28) more frequently. Furthermore, older volunteers used words related to work, such as "Büro" (engl. "office"; r = 0.21), and family, for example, "die_kinder" (engl. "the kids"; r = 0.29), more often. Topics were not as age-discriminative and clearly interpretable as words and phrases, and correlations were comparatively low, wherefore we refrain from interpreting them further here.

On average, women used more function words, particularly personal pronouns in 1st person singular (i.e., "I" or "me") and conjunctions (e.g., "and"). LIWC also recognized more words from women's messages in its dictionary, and they used more exclamation marks than men. Furthermore, women incorporated more words referring to insights (e.g., "think") and home (e.g., "room"). On the contrary, men scored higher on the summary language variable "Analytic Thinking," which indicates a rather formal, logical, and hierarchical thinking style in contrast to an informal, personal, here-and-now, and narrative thinking (Pennebaker et al., 2015a). In line with LIWC results, women used the words "freu" (1st person of "freuen" and part of the LIWC insights category; engl. "looking forward"; d = 0.46) and "ich" (engl. "I"; d = 0.45) more often than men. Moreover, female volunteers used various forms of the verb "gehen" (engl. "to go"; d = 0.51) more often. Men used more abbreviations, colloquial language, and words related to alcohol consumption such as "Vodka" (d = -0.51) and "Bier" (engl. "I"; d = -0.50). Similar to age, gender-discriminative topics were not as clearly interpretable as words and phrases. The only distinctive female topics revolved around social activities (d = 0.46; e.g., "meeting," "seeing," "drinking"). The most distinctive male topic could be interpreted as salutations (e.g., "hey," "hi," "xD").

Age				
	M	SD	r	r CI 95%
Informal language	9.34	3.67	−0.42	[−0.53, −0.31]
Focus future	1.16	0.63	0.38	[0.27, 0.49]
(Informal language/)Netspeak	2.87	2.20	−0.38	[−0.49, −0.26]
Clout	58.57	17.24	0.38	[0.26, 0.48]
(Total function words/Total pronouns/Personal pronouns/) 1st person singular	5.19	1.77	−0.36	[−0.47, −0.24]
(Informal language/) Fillers	0.80	0.55	−0.33	[−0.44, −0.21]
(Cognitive processes/) Causation	2.53	0.84	−0.33	[−0.44, −0.21]
(Social processes/) Family	0.53	0.64	0.31	[0.18, 0.42]
(Total punctuation/) Period	7.58	5.17	0.28	[0.16, 0.40]
Interrogatives	1.93	0.80	−0.28	[−0.40, −0.15]
Gender				
	M (m)	M (f)	d	d CI 95%
Total function words	51.35	54.02	0.67	[0.37, 0.96]
Analytic thinking	25.80	14.55	−0.65	[−0.95, −0.36]
Dictionary words	83.96	86.55	0.60	[0.31, 0.90]
(Total function words/Total pronouns/) Personal pronouns	9.76	11.01	0.59	[0.29, 0.89]
(Total function words/) Total pronouns	14.80	16.10	0.48	[0.19, 0.78]
(Total function words/Total pronouns/Personal pronouns/) 1st person singular	4.59	5.42	0.48	[0.19, 0.78]
(Total function words/) Conjunctions	13.39	14.22	0.41	[0.12, 0.70]
(Total punctuation/) Exclamation marks	1.36	2.16	0.40	[0.10, 0.69]
(Cognitive processes/) Insight	1.72	1.95	0.38	[0.09, 0.67]
Home	0.45	0.59	0.37	[0.08, 0.66]

TABLE 2.2: Top ten variations in LIWC categories with volunteer age and gender. N = 226. Table rows are ordered by absolute magnitude of the Pearson correlation coefficient for age and absolute magnitude of effect size for gender. Women are coded "1" and men are coded "0". For linguistic characteristics, the hierarchically superior LIWC categories are in brackets. For example, the notion "(Cognitive processes/) Insight" indicates that "Insight" is a subcategory of "Cognitive processes" (Koch, Romero, and Stachl, 2022).

General Message Characteristics

We found the usage of emoticons to be closely associated with volunteer age. Specifically, older volunteers used emoticons less frequently and in a less diverse manner. This finding is represented in the negative correlations of emoticon-to-word ratio (r = -0.45), the share of messages containing at least one emoticon (r = -0.40), the average number of emoticons per message (r = -0.37), share of messages containing only emoticons (r = -0.19), and the use of unique emoticons used from the entire corpus (r = -0.18) with age. While the frequency of emoji usage was not correlated with age, the range of emoji usage was: Older volunteers used a broader range of unique emoji overall (r = 0.17) and incorporated more of their own unique emoji (r = 0.15) in a message. Finally, older volunteers sent longer (i.e., containing more words) messages (r = 0.20) and containing more media files (r = 0.19).

With regard to volunteer gender, we found that women used emoji more frequently and in a more diverse manner. Specifically, women had a higher average number of emoji per message (d = 0.54), share of messages containing at least one emoji (d = 0.53), emoji-to-word ratio (d = 0.43), and used a broader share of unique emoji from the entire corpus (d = 0.40) than men. For emoticons, there were no gender differences present in the data.

Emoji Preferences

Emoji usage varied with volunteer age. We found emoji expressing emotions, for example, "🙈" (r = -0.17), "😀" (r = -0.17), "😂" (r = -0.17), "😹" (r = -0.17), and "😃" (r = -0.16), to be more frequently used by younger volunteers in our dataset. On the contrary, we found emoji depicting objects and people, for example, "🌞" (r = 0.19), "💋" (r = 0.19), "👩" (r = 0.18), "🌺" (r = 0.17), "👶" (r = 0.17), to be more frequently used by older volunteers. A similar pattern emerged in the emoji preferences across genders. Here, women preferred emoji that express positive emotions, for example, "😍" (d = 0.51) and "😊" (d = 0.45). Men, on the other hand, preferred the disappointed emoji "😞" (d = -0.30), representing a negative emotion.

Predicting demographics

Overall, we were able to significantly predict volunteers' age above baseline levels (MAE = 4.23, r = 0.80), but not their gender (Acc = 75.0, F1 = 0.5, AUC = 0.83).

Age Regression

The Random Forest models (MAE = 3.95, r = 0.81) and Elastic Net models (MAE = 4.35, r = 0.79) predicted age on average significantly better than the baseline model. Our results suggest that all feature sets, except topics, were significantly predictive of volunteers' age above baseline (see Figure 1). Moreover, the findings indicate that the Random Forest performed on average slightly better than the Elastic Net algorithm in the age prediction, particularly for emoji features, where the Elastic Net models did not significantly predict age above baseline levels, but the Random Forest models did.

Figure 2.3 shows the most important features in the Elastic Net model (based on standardized regularized regression weights) and Random Forest model (based on permutation feature importance) in the age prediction trained on predictive features (all features except topics). The corresponding ALE plots for the Random Forest model indicate the direction of the features' effect on the age predictions. Regularized

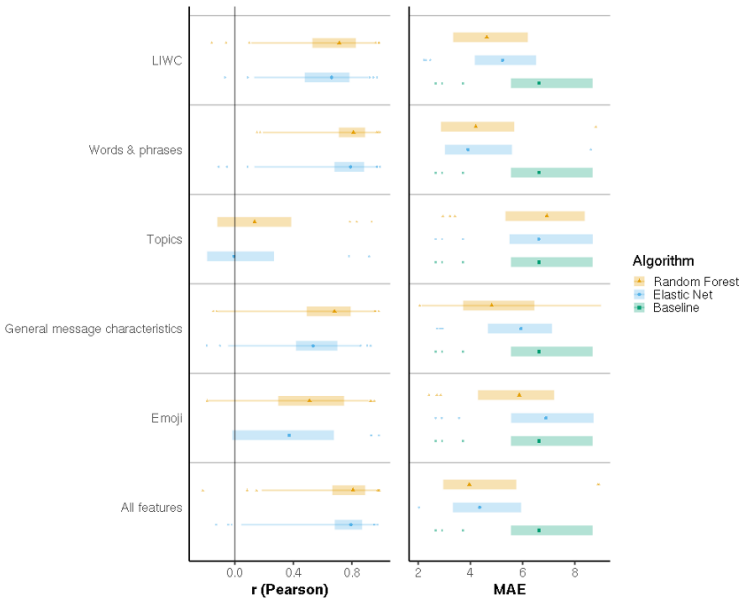


FIGURE 2.3: box and whisker plot of prediction performance measures from repeated cross-validation for age regression for each feature (sub) set. the symbol in the boxes represents the median, boxes include values between the 25 and 75% quantiles, and whiskers extend to the 2.5 and 97.5% quantiles. pearson correlation is not available for the baseline model because it predicts a constant value, for which correlation measures are not defined (Koch, Romero, and Stachl, 2022).

regression weights and permutation feature importance for all predictive features are provided in the project's OSF repository. Overall, features related to the frequency of emoticon usage, specifically the emoticon-to-word ratio, the average number of emoticons per message, and the share of messages containing at least one emoticon, were most important for the prediction of age in Random Forest models [6]. For example, this suggests that if people used on average less than 0.04 emoticons per message, the model predicted older age. Also, the usage of specific emoticons, such as ":D", was highly predictive in Random Forest models, suggesting that higher usage frequencies predicted younger age. Finally, the usage of the word "guten" (engl. "good"; usually used in salutations, e.g., "guten Tag" or "guten Morgen") and the use of informal language were important for the Random Forest predictions. For example, if more than 8% of a volunteer's words were informal language, the model predicted younger age. In the Elastic Net model, the usage of words and phrases, such as "buero" (engl. Office), was most important.

[6] One has to keep in mind that the permutation importance scores of correlated features are ranked higher in Random Forest models. This does not indicate that they are uniquely more important for the prediction of an outcome (Strobl et al., 2008).

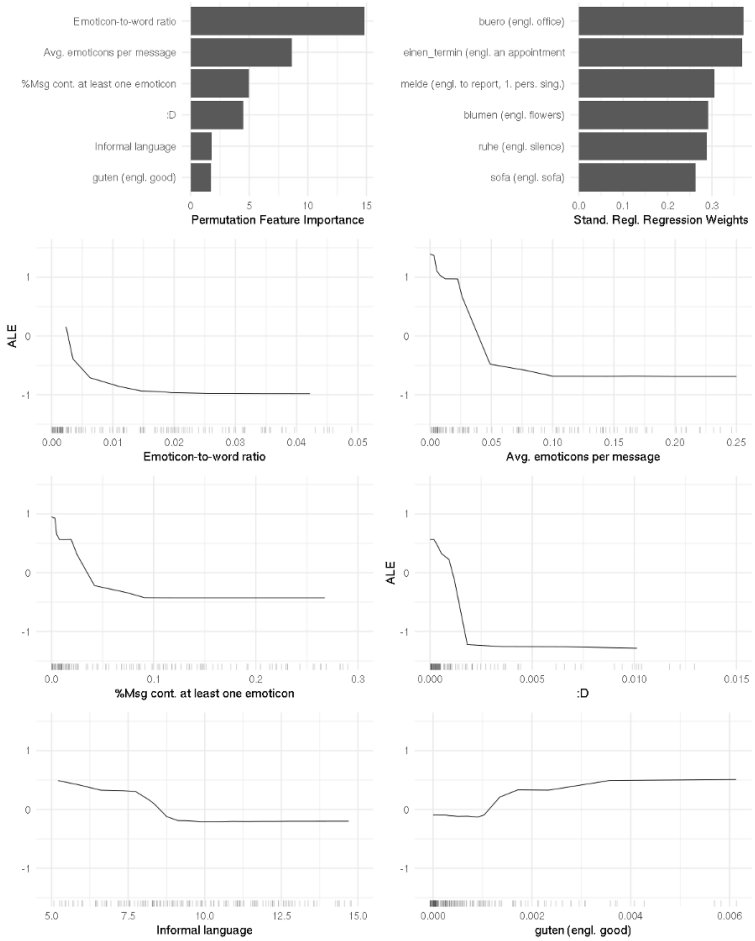


FIGURE 2.4: Top left: Permutation feature importance for the most predictive features in the Random Forest model for age prediction. Permutation feature importance represents the decrease in the model's prediction performance (MAE) after permuting a single variable. Top right: Standardized regularized regression weights for the most predictive features in the Elastic Net model for age prediction. Bottom: ALE plots indicate how mean age predictions in the Random Forest model changed with regard to different values in local value-areas of the respective predictor variable. For example, the average age prediction decreases with an increasing emoticon-to-word ratio. ALE values are centred around zero (Koch, Romero, and Stachl, 2022).

Gender Classification

We were not able to predict volunteers' gender significantly better than the baseline model in the Random Forest classification models (Acc = 75.0, F1= 0.5, AUC = 0.83) and Elastic Net Models (Acc = 75.0, F1= 0.37, AUC = 0.75). Also, none of the models trained on (sub-)feature sets were significantly better than the baseline at classifying gender with regard to prediction accuracy (see Figure 3). This is possibly due to the small sample size and class imbalance in our data set. Consequently, the employed algorithms overfitted to the data and learned to predict the majority class (women) for which there is a very high sensitivity. Since only the Random Forest models significantly predicted volunteer gender (before the Holm correction for multiple comparison), we investigated variable feature importance. Here, we found a similar pattern as in the descriptive results: The usage of function words, particularly

the usage of personal pronouns, specifically 1st person singular, were most important. The figure displaying features with the highest permutation feature importance with the corresponding ALE plots is provided in the project's OSF repository.

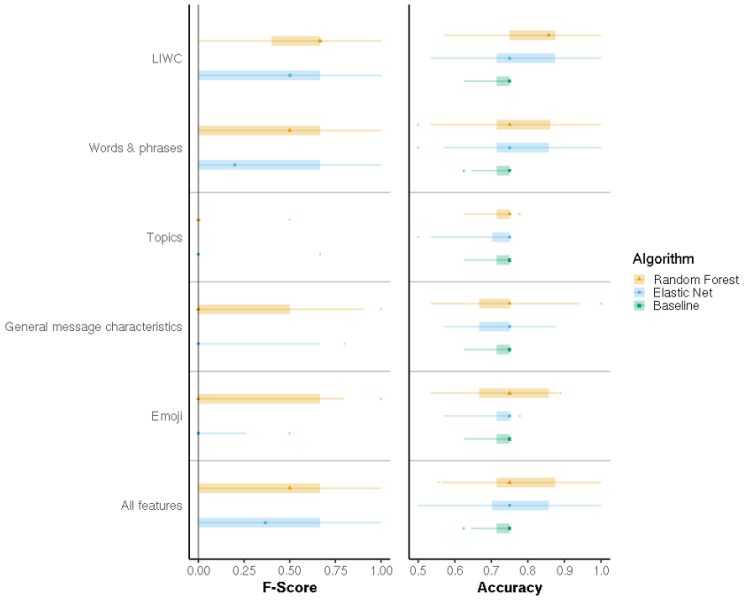


FIGURE 2.5: Box and whisker plot of prediction performance measures from repeated cross-validation for gender classification for each feature (sub) set. The middle symbol represents the median, boxes include values between the 25 and 75% quantiles, and whiskers extend to the 2.5 and 97.5% quantiles. Outliers are depicted by single points. For better readability, we omitted the baseline model because the F-Score is 0 across all folds (indicated by vertical line) (Koch, Romero, and Stachl, 2022).

2.2.3 Discussion

Our study has generated novel insights into age- and gender-linked language variations in open and closed vocabulary features, general message characteristics, and emoji preferences in private instant messages. Further, we predicted volunteer age significantly above baseline levels and identified particularly predictive features in the respective machine learning models. We found specific variations in the usage of emoji, emoticons, and personal pronouns to be both, strongly associated with (in-sample) and most predictive (out-of-sample) of volunteer demographics.

Age- and gender-linked language variations in WhatsApp messages

The descriptive statistics of the extracted message features from WhatsApp messages and the feature importance from the prediction models revealed distinct language variations with users' age and gender. The frequency of emoticon usage, usage of 1st person singular, and of informal language were strongly negatively associated with users' age. The more frequent usage of emoticons among younger users and the non-association of emoji usage with age are in line with parts of the previous literature (Oleszkiewicz et al., 2017; Siebenhaar, 2018). However, the negative association of emoticon usage frequency with age, discovered in our sample contradicts work by Fullwood, Orchard, and Floyd (2013), who found no differences in emoticon usage frequency between age groups in online chat rooms where the overall share of users,

who used emoticons at all, was relatively low (around 20%) and emoji were not around yet. The negative association of usage frequencies in the "1st person singular" category with age in our data is in line with findings from past studies on a broad range of different text sources (Nguyen et al., 2013; Nguyen et al., 2011; Pennebaker and King, 1999; Schwartz et al., 2013). Pennebaker and Stone (2003) pointed out that this might be an indicator for people becoming less self-focused as they age. Also, our finding that younger users used more informal language is in line with past studies that reported similar effects (Nguyen et al., 2011; Schwartz et al., 2013). We found that female volunteers used emoji more often, used a broader range of emoji, and used more function words - especially 1st person singular pronouns. The greater frequency of emoji usage among women in our data is in line with prior studies, showing that women on average use more emoji than men (Chen et al., 2018; Jones et al., 2020; Prada et al., 2018). However, this effect did not generalise to the usage of emoticons, which were almost equally often used by women and men, in our data. This finding is in line with work by Tossell et al. (2012), but does not align with results of Fullwood, Orchard, and Floyd (2013) and Rao et al. (2010), who reported higher frequencies of emoticon usage among women. However, the data for those studies was collected around 2010, when emoticons were the go-to way to express emotions in computer-mediated communication, before emoji were around. Over time the prevalence of emoticons decreased as they were gradually replaced by emoji (Pavalanathan and Eisenstein, 2015). Since our data set was collected in 2014/2015, many women possibly used emoji instead of emoticons to express emotions, which could be the reason why gender differences in emoticon usage were not present in our data. Furthermore, the distinct gender differences in the number of function words used and its subcategories "Personal pronouns", and particularly "1st person singular" had also been found in previous studies on other text sources (Newman et al., 2008; Schwartz et al., 2013).

Predicting demographics from WhatsApp messages

Our results in the age predictions based on closed and open vocabulary features compare well with previous results on social media data that used the same methods (see Table 3). Our models performed better than prediction models in prior work by Jaidka, Guntuku, and Ungar (2018) trained on Tweets and Facebook posts in predicting user age. Schwartz et al. (2013), who trained their models on an enormous sample of English Facebook posts, achieved comparable performances. Many other prior studies (Marquardt et al., 2014; Nguyen et al., 2011; Rao et al., 2010) binned age into groups before modeling and, consequently, do not allow for a direct comparison with our results. For gender predictions, a comparison of our model performance with prior work is not as straightforward because comparable studies only reported the overall classification accuracy (Bamman, Eisenstein, and Schnoebelen, 2014; Burger et al., 2011; Schwartz et al., 2013). While this metric is useful, it is highly dependent on the gender class distribution in the respective sample and makes a comparison between studies difficult. For all feature sets, except LIWC features, we could not predict gender above baseline levels and consider our prediction performance inferior to comparable prior work (Burger et al., 2011; Rao et al., 2010; Schwartz et al., 2013). With 85.7% accuracy for LIWC features, our models outperform comparable LIWC models from prior work (Jaidka, Guntuku, and Ungar, 2018; Schwartz et al., 2013). Notably, even though all comparable prior research on social media text data was based on much larger data sets, (most likely because these social media data are easier to collect as lined out in the introduction of this paper) the performance of our age

models is similar to theirs. A possible explanation for this observation could be that self-disclosure in instant messages is higher compared to that in social media posts. Consequently, instant messages could be more informative of user characteristics, i.e., demographics, than social media posts. Therefore, future studies based on larger data sets, should compare predictive performances obtained on instant messages, with those obtained on social media posts. Similarly, Jaidka, Guntuku, and Ungar (2018) compared model performances on data from Facebook and Twitter.

Study	N	Data source	Features	Age: MAE (baseline MAE)/r	Gender: Acc. (baseline Acc.)
Schwartz et al. (2013)	74,859	Facebook	LIWC	−/.65	78.4 (62.0)
			N-grams	−/.83	91.4 (62.0)
			Topics	−/.80	87.5 (62.0)
Jaidka et al. (2018)	523	Facebook	LIWC	7.20 (10.06)/−	87.0 (54.0)
			N-grams	5.71 (10.06)/−	78.0 (54.0)
			Topics	6.78 (10.06)/−	91.0 (54.0)
Jaidka et al. (2018)	523	Twitter	LIWC	8.59 (10.06)/−	81.0 (45.0)
			N-grams	8.08 (10.06)/−	73.0 (45.0)
			Topics	8.58 (10.06)/−	80.0 (45.0)
Rao et al. (2010)	1,000	Twitter	N-grams	-	68.7 (50.0)
Burger et al. (2011)	184,000	Twitter	N-grams	-	75.5 (54.9)
The present study	157	WhatsApp	LIWC	4.63 (6.63)/.71	85.7 (75.0)
			N-grams	4.20 (6.63)/.81	75.0 (75.0)
			Topics	6.93 (6.63)/.13	75.0 (75.0)
			Msg. Char.	4.81 (6.63)/.68	75.0 (75.0)
			Emoji	5.87 (6.63)/.51	75.0 (75.0)
			All features	3.95 (6.63)/.81	75.0 (75.0)

TABLE 2.3: Predictive performance for age and gender in comparison to prior work. One has to be cautious with the interpretation of the performance metrics for gender because they are dependent on the gender distribution in the sample. For comparability, we only present studies using the same language features, i.e., LIWC, N-grams ("Words & phrases"), and/ or topics. Performance measures of the best employed algorithm are reported. All prior studies are based on English text data. MAE = Mean average error, Acc = Prediction accuracy (Koch, Romero, and Stachl, 2022).

Implications

By providing descriptive results and by applying methods of interpretable machine learning, this work adds a promising new text source for the study of individual differences in language usage with relevance for psychology, computer science, linguistics, and communication research. Further, after the impressive results of prior studies had demonstrated that user demographics are predictable from social media posts,

our findings indicate that variation in linguistic characteristics of instant messages also allows for the accurate prediction of users' demographics (i.e., age), already in small samples. Such a demographic user profiling based on instant messages could be used to gain information on user demographics in order to personalize systems and for marketing efforts, based on the users' age and gender. Moreover, it could be useful to validate previously provided demographic user data. For example, this approach could be used in anonymous digital communities (e.g., only for people aged under 18) to validate user profiles and to flag suspicious profiles containing potentially false information (Loo, De Pauw, and Daelemans, 2016). In order to protect users' privacy, the feature extraction and potentially model predictions should happen on the user's end (i.e., on the phone). Another field of application lies in determining the demographics of the members of an anonymous community communicating through instant messaging. By analyzing the characteristics of the messages, one could gain an approximation of the demographics of such a user population through their unique psycholinguistic biometrics. This would allow researchers to better understand the demographics of political or activist movements, and their importance to a respective populous.

These author profiling techniques have the potential for misuse posing a threat to user's privacy and safety in instant messaging. Moreover, in contrast to public posting (e.g., on social media platforms), the design of chat rooms and messaging apps does not suggest that exchanged information is accessible to third parties. Given the trend that users are increasingly shifting from social media sites to instant messaging services, the importance of private instant messaging data is expected to rise further in the future (Goode, 2019). In this manner, Facebook has announced plans to shift their strategic focus from public posting to private messaging services (Zuckerberg, 2019) while increasing the efforts to loosen up the privacy protection of their messaging services (Evans, 2020; Goodin, 2021). Since commercial collectors of chat data have access to much larger quantities of personal communication data the monitoring and systematic analysis of private instant messaging environments would allow for more accurate and additional inferences, for example, about personality traits and emotions (Preoţiuc-Pietro et al., 2016; Schwartz et al., 2013) than the ones reported in this work. Moreover, the corporate accessibility of these data will likely lead to timely, situation-specific targeting efforts by identifying users' momentary interests, needs, and desires in private communication. Given our findings that demographic information can be inferred from instant messages, even with very small samples, we argue that linguistic data from chat logs should be subject to extended privacy protection and regulation, similar to older forms of private communication (e.g., letters) or be clearly labeled as non-private.

Limitations and Outlook

The results of this work are limited in three ways. First, the analyses are subject to the given small data set, which is based on 226 (156 for predictive modeling) German WhatsApp users' chat language in 2014/2015, and the specific feature extraction methods we applied to the data. Predictive models from past studies on author profiling from social media text had been mostly trained on large text samples with thousands of volunteers and millions of words (Schwartz et al., 2013). Our chat corpus, on the contrary, was much smaller in terms of the number of volunteers and the available text per volunteer. More data often led to more accurate and generalisable models in past work on author profiling (Eichstaedt et al., 2020; Kern et al., 2016; Peersman, Daelemans, and Van Vaerenbergh, 2011). To illustrate this,

we analyzed how the number of words per volunteer affected our models' predictive performance, improving age predictions with more data available (see OSF repository for detailed results). Further, this data set size is likely too small to harness the full potential of data-driven language features, particularly for topic models.

Second, like many studies in the social sciences, the present work is subject to sampling biases. Specifically, the "What's up, Deutschland?" chat corpus consists of messages from people who used WhatsApp, were aware of the data collection, and also decided to donate their messages despite potential needs for privacy. Further, the demographic composition of the data set is not representative for the general public population; young people and particularly women are overrepresented in our sample.

Third, due to a phenomenon termed "concept drift," which describes how the underlying association between predictors and the criterion (e.g., users' age and gender through language usage) changes over time (Lu et al., 2019), our results have to be interpreted in the context of the time of data collection in 2014/2015. For example, the emergence of emoji reduced the usage of emoticons in recent years (Pavalanathan and Eisenstein, 2015). This trend and other developments in instant messaging have changed how men and women of different ages communicate with time. Therefore, it is necessary to retrain models on newly collected datasets. While language data from personal messaging is difficult to collect for scientific research, commercial actors would have access to larger, more representative, and continuously updated samples of private instant messaging data. Hence, our findings should be considered a conservative estimate.

We would welcome researchers to replicate and extend our study with new data from a larger and more representative sample to address these limitations and to further investigate the predictability of user characteristics from private instant messages. In this context, it would be interesting to, for example, investigate whether levels of self-disclosure on public social media and private instant messaging vary across cultural and national contexts. Therefore, we would welcome if pre-trained lexica and topic models, like the one for English social media posts (Sap et al., 2014; Schwartz et al., 2017), were made available to researchers for more languages and text sources.

Moreover, while this work has exclusively focused on the message characteristics of the respective users, whose age and gender we aimed to predict, future research could also investigate the influence of the demographic characteristics of the chat partners and their message characteristics on the other person's language. It would be particularly interesting to collect additional data on the user's personality, education, language proficiency, and the relationship to the respective chat partner in order to model their language more holistically to improve prediction performance and to evaluate the potential to infer these characteristics from instant messaging data.

Finally, to ensure comparability of studies, we want to encourage researchers to standardize methodological procedures in future studies on language variations with demographic differences. This includes treating continuous variables, such as age, in a continuous manner, reporting a range of performance measures in classification, particularly ones that are less dependent on the class distribution in the sample, such as F1, and reporting measures of interpretable machine learning (Molnar, 2019; Stachl et al., 2020).

2.2.4 Conclusion

In this work, we identify age- and gender-linked language variations and demonstrate that user demographics are predictable from WhatsApp instant messages. Our

findings replicate and extend past results on individual differences in social media language to the growing domain of private instant messaging. We highlight further research opportunities and emphasize the rising threats to individual privacy that could arise from the monitoring of formerly private instant messaging environments.

2.3 Relevance of Space for Psycholinguistic Measures

(This section was led, written, conceptualised, and analysed by Peter Romero. The introduction to geographic psychology was partially formulated by Gregory Serapio-García. Eisaku Tanaka and KC Chen collected, cleaned, and pre-analysed the twitter data. Teruo Nakatsuma supervised the project.)

2.3.1 Introduction

Geographical and urban economics is a well established subfield of economics, and analyses the influence of geospatial distributions of various phenomena like the effect of windmills on housing prices (Dröes and Koster, 2016), to the association of population health with alcohol store density (Fone et al., 2016). On the other hand, geographic psychology is also well established, uses geospatial distribution of psychological phenomena to explain a plethora of economic outcomes (Rentfrow, Gosling, and Potter, 2008a; Rentfrow, 2020), and it is surprising that both disciplines neither work with nor know each other very well. Recently, also CSS chimed in, and used big data approaches to explain psychological phenomena (Giorgi et al., 2022). This is the ideal connection of both geographic economics and geographic psychology, since this opens new options to gather insights from data that lacks otherwise structured information and demands advanced machine learning skills to make use of for more rigorous economics approaches. This chapter explores an alternative approach to Giorgi et al. (2022), yet also shows constraints based on available models.

2.3.2 Geographical Psychology

Geographical psychology examines the distributions of psychological phenomena across various geospatial resolutions and aims to identify the individual, social, cultural, and physical mechanisms that underlie observed variations. It associates aggregated psychological phenomena on a macro-level like regions or countries with political, economic, social, and health outcomes (Rentfrow, 2020). Most notably, it established that Big Five personality traits (Openness to experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (McCrae and Costa, 1997) display systematic geographic difference (Rentfrow, 2020). Especially nation-internal studies showed the uneven distribution of personality traits across regions within nations. For example, independent samples across the U.S., collected between 1999 and 2015, showed that openness tends to be highest in New England, West Coast, and Mid-Atlantic states, however in the Midwest it is lower; however on average, neuroticism is lowest in the Mountain states and the West Coast (Rentfrow, Gosling, and Potter, 2008a; Rentfrow et al., 2013; Elleman et al., 2018). Comparable nation-internal differences have been observed in smaller countries that display less social diversity like Switzerland (Götz, Ebert, and Rentfrow, 2018), the United Kingdom (Rentfrow, Jokela, and Lamb, 2015), and Japan (Yoshino and Oshio, 2021a). Due to growing body of evidence, associations between geographically distributed personality traits and political, economic, social, and health (PESH) (Rentfrow et al., 2013) outcomes

have been established. For example, regional conscientiousness in the United Kingdom was positively related to conservative voting and negatively related to liberal voting patterns in the General Elections of 2005 and 2010 (Rentfrow, Jokela, and Lamb, 2015), however high degrees of neuroticism and an overall narrative of lost pride and fear was associated with both voting for Trump and Brexit (Obschonka et al., 2018). It is unclear whether this effect resulted from underlying neuroticism or the overall information field, and more rigorous research on that has to be conducted.

Maybe the strongest methodological hurdle to conducting research on geospatial psychological phenomena is its reliance on large-scale survey data, which is accompanied by logistical and financial obstacles, especially in large countries like the United States or Russia. However, the temporary availability of data from Social Networking Services (SNS) to the wider academic community enabled researchers to estimate macropsychological characteristics of regions instead of or for the augmentation of surveys. For example, Jaidka et al. (2020) deployed large-scale aggregated Twitter data clustered by county, which they deployed in machine learning models to accurately and robustly estimate subjective well-being in the United States. Furthermore, Giorgi et al. (2022) extended this line of research by predicting regional personality from regional Twitter data alone. However, acquisition of geospatially tagged Twitter data is not easy. Some old approaches rely on old and outdated Twitter data that was scraped as the API still allowed location access (Ahmed, Hong, and Smola, 2013), while others identified tweet locations based on self-described locations (Jaidka et al., 2020; Giorgi et al., 2022). The former comes with the downside of having to rely on just a small pool of outdated and partially redacted data, while the latter comes with the downside of having to rely on self-descriptions that could be outdated, or which are not acceptable in every culture (e.g., Japan and Germany are very privacy-conscious (Guenole, Feinzig, and Ferrar, 2015)), thus restricting this approach. We introduce a novel and elegant approach to locate the approximate location of a Twitter user by identifying and scraping the tweets of followers of hyperlocal entities like police stations, local sports teams, or city mascots – following the hypothesis that nobody else but locals would have a reason to do so.

2.3.3 Data Extraction

Overview of Data Set

We spatially aggregate tweets in order to generate a psycholinguistic understanding of the prefecture demographics of Japan. Corpora of Japanese tweets generated between January 1st, 2019 to April 1st, 2021 from all 47 prefectures are extracted, and analysis is done on two (or more) major and spatially separated cities of each prefecture. The final data set includes a total of 25,614,106 tweets, of which 189,734 tweets are extracted from every city on average (SD = 44,924.94). The minimum number of tweets for a city is 70,425 tweets, and maximum is 244,331 tweets. All tweets are harvested from 107,873 followers of 1,648 local city representative accounts. From every city, tweets are harvested from 799 follower accounts on average (SD=46.16). The minimum of accounts in a single city is 596 and maximum is 822.

The preparation and extraction of this data set is done through first identification of the major cities of each prefecture and a manual accumulation of representative Twitter accounts of those cities, second harvesting of tweets from the followers of the representative Twitter accounts, and last through a psycholinguistic analysis of tweets using IBM Watson Personality Insights.

Data Set Selection and Preparation

Selection of Representative Cities

Psycholinguistic analysis of the prefecture demographics is done on tweets harvested from Twitter users most likely to be residing in two or more cities of every prefecture in Japan. Two cities of every prefecture are selected based on the greatest population and the spatial separation of the cities, in order to gain access to a greater data set and forestall spatial biases. If unable to find more than three representative Twitter accounts from either of the selected two cities, a third city is investigated. Identification of the prefectures, cities, and spatial separation, as well as a comparative evaluation of every city's population in every prefecture is done through official portals and publicly accessible governmental statistics (Statistics Bureau, Ministries, and Agencies, 2021). The average population across all cities in Japan is 523,026 (SD = 640,974), with the minimum being 2,736 (Esashi-cho, Hokkaido), and the maximum being 3,757,630 (Yokohama-shi, Kanagawa-ken) (Statistics Bureau, Ministries, and Agencies, 2021).

Exceptional consideration is made on Hokkaido and Tokyo, due to their susceptibility of containing different uses of language within the prefecture. A physically large prefecture like Hokkaido contains sub-prefectures, and geographical differences in language use is possible. Therefore, tweets are harvested from at least two cities from each sub-prefecture. On the other hand, Tokyo contains a large number of cities, and as one of the biggest urban centres of the world, harbours more population than many European nations. Furthermore, it contains the largest amount of foreigners. This makes tweets harvested from Tokyo likely to contain different dialects and personalities. There are 17 cities on average in a prefecture in Japan, but Tokyo includes 50 cities (and special wards) (Statistics Bureau, Ministries, and Agencies, 2021). Hence, three cities in Tokyo with the most population and spatial separation are carefully selected.

Preparation of Twitter Accounts

The representative Twitter accounts of all cities are manually selected. Similar to other social media platforms, the goal of usage within Twitter can vastly differ by account. Therefore, all representative Twitter accounts are annotated and organised by usage (which we call an "account type"). In order to circumvent inclinations in analysis results due to account types, various account types are searched and retrieved. This included official city accounts, city information accounts, city news accounts, police department accounts, fire department accounts, politician accounts, disaster prevention accounts (which gained attention as a method of risk management after the calamitous 2011 Tohoku Earthquake, and even more with the advent of COVID-19 pandemic) (Latonero and Shklovski, 2011; Government, 2020), city COVID-19 information accounts, city public relations accounts, consumer affairs accounts, event-related accounts (such as matsuri), city mascot accounts (which is widely accepted and known in Japan), city professional sports teams accounts, amateur sports team accounts, school accounts, public facility accounts (such as libraries or malls), and local accounts. These account types available at the prefecture level are retrieved as well.

Each representative Twitter user's location, prefecture, city, population of the city, geolocation of the city, Twitter username, and details of the account type (usage) is all manually transcribed, annotated, and organised. This results in 1,648 retrieved representative accounts; the average number of representative accounts per

prefecture is 35, where the maximum is Hokkaido with 235 accounts and the minimum is Kumamoto-ken with 11 accounts. Hokkaido's large number of representative accounts is a result of at least two cities being selected from all sub-prefectures. Excluding Hokkaido, the average number of representative accounts per prefecture is 30.

Corpus Extraction

After identifying the major cities and representative Twitter accounts, extraction of tweets from followers of the representative Twitter accounts is performed through Twitter Application Programming Interface (API), pandas 2.1.3 (team, 2020), and the Amazon AWS Services. Harvesting raw textual tweets from the followers of the representative Twitter accounts is crucial because the representative accounts are disinclined to contain emotional or informal language (for example city information accounts), and may gather parts of their content from centralised national sources. Hence, linguistic analysis of these tweets will not result in representative outcomes. Therefore, we extract the tweets from followers of city representative accounts to generate the corpus for geospatial psycholinguistic analysis.

By looping through the city representative accounts' usernames in a sequential manner, each account's followers is harvested by using the Twitter API. Hence, during each iteration of the loop, the usernames of up to 40 followers of each representative account's followers are harvested. And furthermore, through nested looping, each of these follower's tweets are harvested. We decide for 40 followers in two depth layers since the code is bound by API speed and cannot be optimised for time complexity. Due to API restrictions, only up to 3,200 tweets of each account are harvested, within the time-frame of January 1st, 2019 to April 1st, 2021 – the year before and after the onset of the COVID-19 pandemic. This means that for some users, not all tweets are collected, however those are rather a tiny subset of the overall harvested users, and most likely represent power users that could otherwise skew the sample, since they use it more due to professional or personality reasons.

Through this process of iterating through each city-representative account, two data files are produced. In the first one, all of the followers' information is compiled to a "follower information" csv file, which includes identification information of the followers like their Twitter ID, Twitter usernames, the following representative account's username, city name, the self-reported location of the user (if available), and the time of account creation. The second file is used as the main corpus for analysis and is a "city tweets" file that contains up to 3,200 tweets harvested from each follower of all city representative accounts.

It contains all tweets, dates, and Twitter ID and username as unique identifier for further analysis, the corresponding city, the following Twitter ID and usernames, the language of the tweet, identification information like when the tweet was created, symbols used, user mentions, and various other tweet-specific information. Continuing this process for all 1,648 representative accounts allows producing corpora of tweets and user information harvested from 107,873 follower accounts in total. This creates a data set ready for further psycholinguistic analysis and geographical mapping.

Ground Truth Data

We use the data collected from Yoshino and Oshio (2021a), who use the Japanese version (Oshio, Abe, and Cutrone, 2012) of the Ten Item Personality Inventory (TIPI)

(Gosling, Rentfrow, and Swann, 2003), to represent geospatial distribution of personality data over Japan. TIPI uses a seven-point Likert scale and is ideal for mass-deployment and large-scale studies, since it only is comprised of ten items, two per Big Five factor, of which one is reversed, and since it exists in 27 languages and with 9,167 peer-reviewed papers, has been well-established in literature, wherefore we also use it in chapter 1. Despite this small size makes it "somewhat inferior to standard multi-item instruments" (p.504) Gosling, Rentfrow, and Swann, 2003, it displays high congruence between self-ratings and observer ratings, has high test-retest reliability, excellent external validity, and its outcomes for self-ratings, external ratings, and peer ratings highly correlate with larger research-standard Big Five questionnaires.

Data is collected in three iterations; first between January and March 2012 (n = 4469, prefecture mean 95.09; SD = 85.95, min = 14.00, max = 388.00, 46% male; SD = 6%, min = 25% and max = 58%), the second iteration in January 2017 (n = 5619, prefecture mean 119.55; SD = 13.99, min 87.00, max 149.00, 60% male; SD = 5%, min = 50 %, and max = 71%), and the last iteration was in January 2019 (n = 4330, prefecture mean = 92.13; SD = 14.34, min = 58.00, max = 127.00, 66% male; SD = 6%, min = 53%, and max = 80%). Overall n = 14418, prefecture mean = 306.77; SD = 101.63, min = 161.00 max = 648.00, 57 % male; SD = 4%, min = 46%, and max max = 65%).

2.3.4 Methodology

Language Analyses

Language analysis is conducted via IBM Watson Personality insights (IBM, 2021) and LIWC (Linguistic Inquiry Word Count) (Pennebaker et al., 2015b). After assembling all tweets generated from every prefecture into individual data files, these are passed into IBM Watson Personality Insights' API through pandas (team, 2020), to extract psychological features. Among the 101 psychological features extracted, the most crucial are the Big Five personality traits: openness, conscientiousness, extraversion, agreeableness, and neuroticism (IBM, 2021). The other features are marketing-related, which is the original purpose of IBM Watson Personality Insights. Unfortunately, this tool is deprecated at the end of the analysis, so that no further comparative studies can be created.

Since IBM Watson Personality Insights derives features from pre-trained predictive models, we also extract dictionary-based, hard-coded features for further psycholinguistic analysis via LIWC (Pennebaker, 2015), using the Japanese dictionary and tokenisation method introduced by Igarashi, Okuda, and Sasahara (2021), subsequently is designated as J-LIWC 2015.

For using J-LIWC 2015, text preprocessing is crucial, since unlike English, there is a lack of word boundaries in Japanese sentences. Careful segmentation of Japanese text documents into words is needed before conducting the main analysis. Therefore, we follow the text preprocessing steps recommended by the creators of J-LIWC 2015 (Igarashi, Okuda, and Sasahara, 2021), using their latest Japanese dictionary, and the MeCab/IPADIC (Kudo, 2005) python library, and conduct subsequent morphological analysis (word segmentation) and part of speech analysis (POS). Before that, Twitter language features like retweet identifiers, emojis, are identified, counted, and removed from the text for later analysis. All texts generated by the same city and same day are treated as one document, to focus the quantified psychological analysis at a city and the prefecture level. Once all of the text preprocessing, assembling, and annotation is done, the corpora are passed into the J-LIWC 2015 software for main analysis. Just as the English version, J-LIWC 2015 uses a dictionary-based category-by-category word

frequency analysis, featuring words and word stems, including standard language categories like pronouns, to psychological processes like emotions. These word-level features are categorized in theory-derived linguistic dimensions, and LIWC counts the words in the respective categories to generate a score for each category (Pennebaker et al., 2015b). Through aggregation on a daily level, we determine the prevalence of certain categories in a given text in each city, and through further aggregation, in each prefecture. This enables future spatiotemporal analysis. For the sake of this paper, all temporal features are finally aggregated on prefecture level, to generate an overview of both the predicted 101 psychological latent traits from IBM Watson and the 60 word categories from LIWC (which contain six sub-scores: insight, causation, discrepancy, tentativeness, certainty, and differentiation).

2.3.5 Results

Results show little overlap between survey data as ground truth (Yoshino and Oshio, 2021a) and the prediction from IBM Watson Personality Insights (IBM, 2021). Table 2.4 displays Spearman's and Pearson's correlation coefficients for all Big Five traits between the survey and the prediction. We include both to understand more about potential data and distribution issues, since both have different data requirements, sensitivities, and use cases. Pearson's r assesses linear relationships, requires normally distributed, continuous data, and is the standard in psychological analyses, due to the assumption of underlying Gaussians of latent traits (Rust, Kosinski, and Stillwell, 2020). On the other hand, Spearman's ρ is more flexible, assessing monotonic relationships and is thus applicable to non-parametric data, including ordinal variables, and can be deployed to assess non-linear relationships or when the data does not meet the assumptions Pearson's r. In summary, Pearson's r is preferable for light-tailed distributions, whereas Spearman's ρ for heavy-tailed distributions or when outliers can be expected, which often is the case in psychological research (Winter, Gosling, and Potter, 2016). Since we assume the same underlying Gaussian, different results in both indicates data issues, in all likelihood with the prediction results.

Latent Trait	Spearman's Correlation	Pearson's Correlation
Openness	0.069	-0.038
Conscientiousness	0.136	-0.003
Extraversion	-0.097	-0.159
Agreeableness	-0.042	0.038
Neuroticism	-0.013	-0.080

TABLE 2.4: Correlation Coefficients for the Big Five Personality Traits

These correlation results are even lower than reported from other authors with similar approaches (Giorgi et al., 2022), and so unrelated to ground truth, that they cannot be used for broad-scale academic research or industrial applications, which may be one of the reasons that IBM Watson Personality Insights is deprecated. The highest association using Spearman's ρ is Conscientiousness with $r_s = 0.136$, and, using Pearson's r, Extraversion with $r_p = -0.159$. All other associations are close to zero, which indicates no connection at all. This is depicted in figure 2.6 for Openness, figure 2.7 for Conscientiousness, figure 2.8 for Extraversion, figure 2.9 for Agreeableness, and in figure 2.10 for Neuroticism.

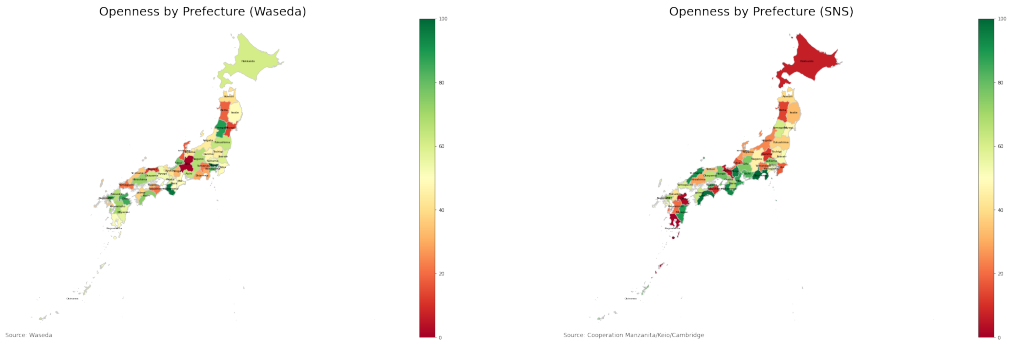


FIGURE 2.6: Openness – comparison ground truth (Yoshino and Oshio, 2021a) with SNS prediction

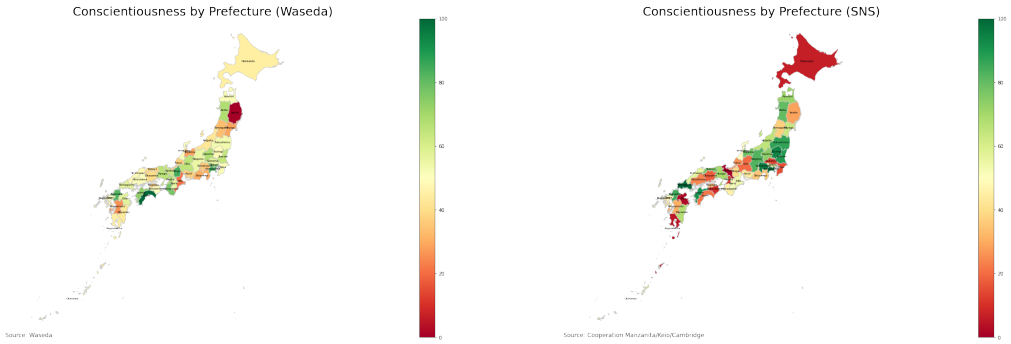


FIGURE 2.7: Conscientiousness – comparison ground truth (Yoshino and Oshio, 2021a) with SNS with prediction

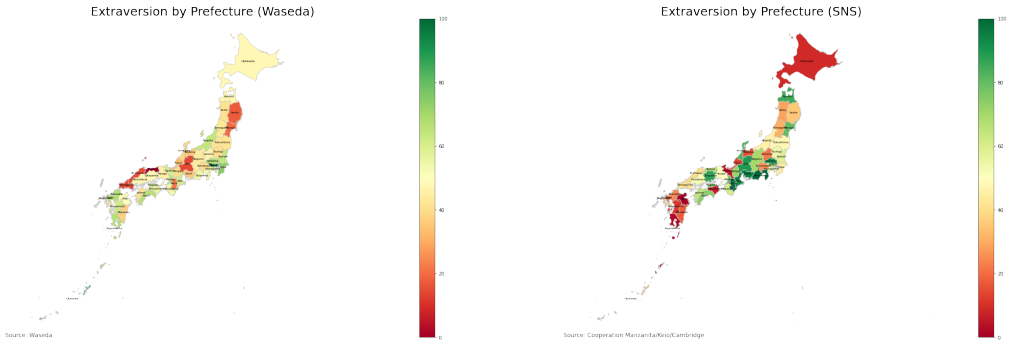


FIGURE 2.8: Extraversion – comparison ground truth (Yoshino and Oshio, 2021a) with SNS prediction

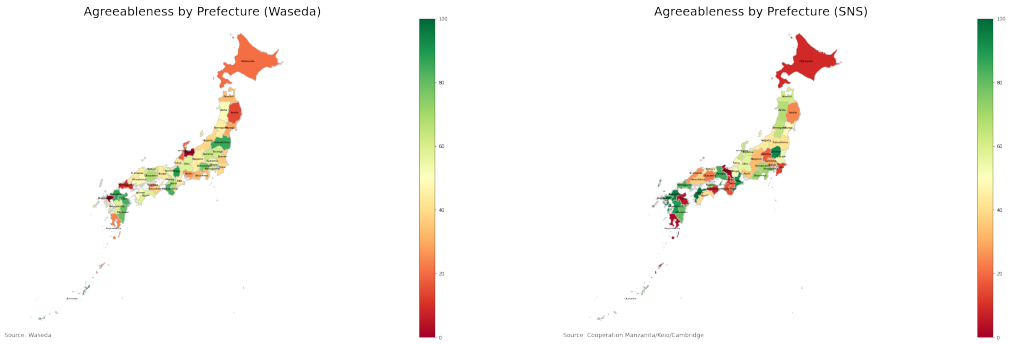


FIGURE 2.9: Agreeableness – comparison ground truth (Yoshino and Oshio, 2021a) with SNS prediction

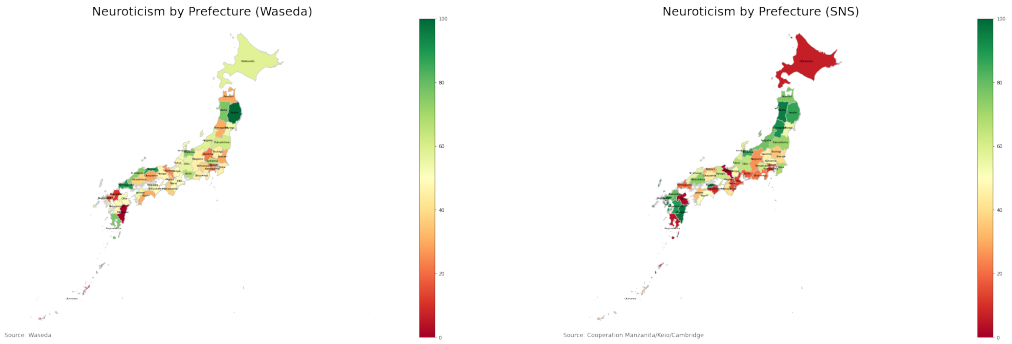


FIGURE 2.10: Neuroticism – comparison ground truth (Yoshino and Oshio, 2021a) with SNS prediction

On a level of eye-validity, we see vast differences, however also relational similarities between the prefectures, but for Hokkaido, Kagoshima, Kyoto, Oita, and Tokyo, which seem to be outliers; with Hokkaido being the most distinct one in most cases. To understand and compare this better, the results of both the ground truth data (Yoshino and Oshio, 2021a) and our prediction results are adjusted to a normalised scale, and plotted by alphabetical order of prefectures, which is depicted in figure 2.11 for Openness, figure 2.12 for Conscientiousness, figure 2.13 for Extraversion, figure 2.14 for Agreeableness, and figure 2.15 for Neuroticism.

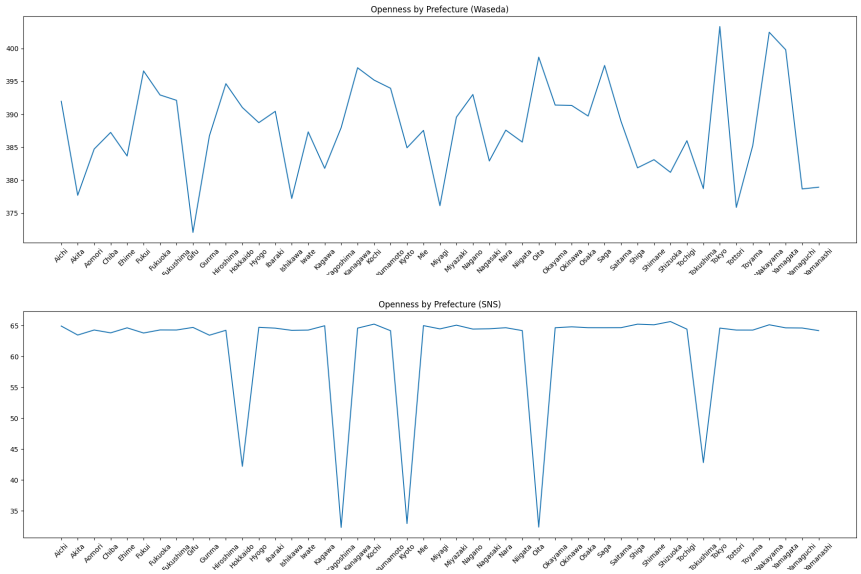


FIGURE 2.11: Openness – distributional comparison of ground truth (Yoshino and Oshio, 2021a) with SNS prediction

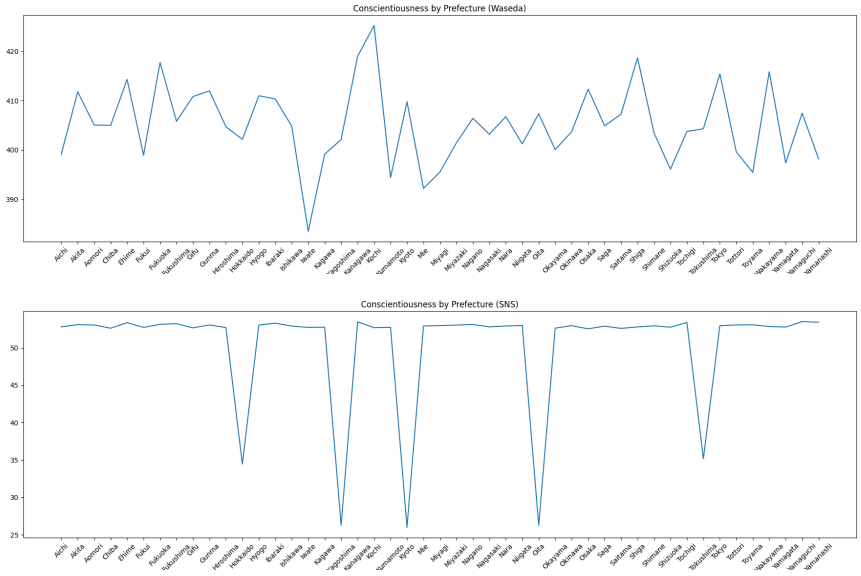


FIGURE 2.12: Conscientiousness – distributional comparison of ground truth (Yoshino and Oshio, 2021a) with SNS prediction

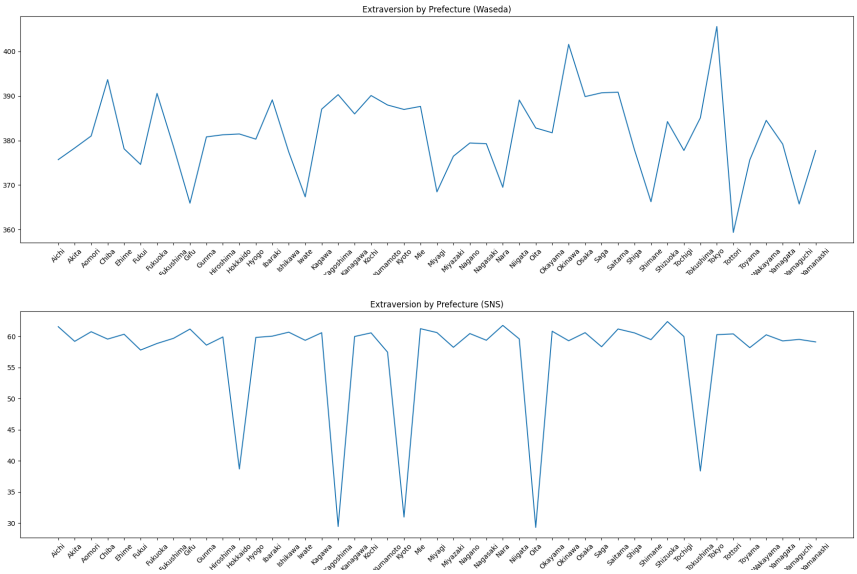


FIGURE 2.13: Extraversion – distributional comparison of ground truth (Yoshino and Oshio, 2021a) with SNS prediction

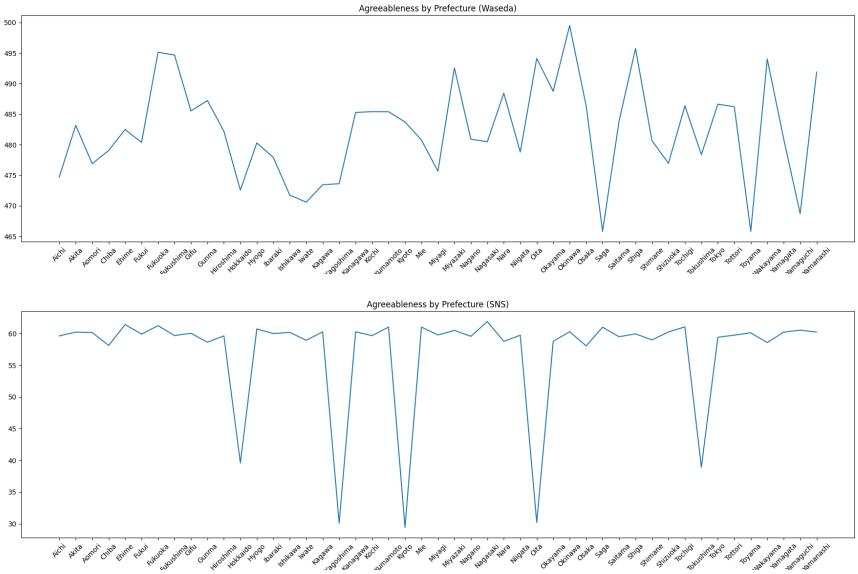


FIGURE 2.14: Agreeableness – distributional comparison of ground truth (Yoshino and Oshio, 2021a) with SNS prediction

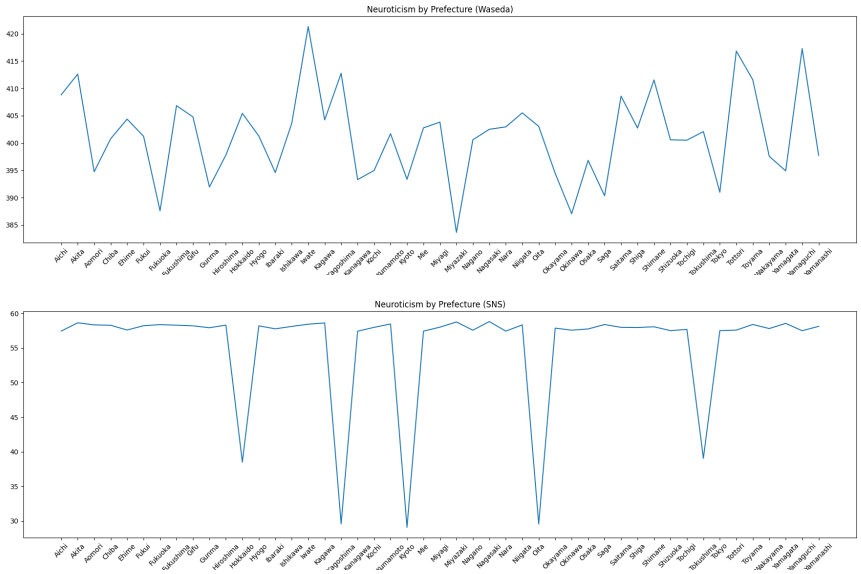


FIGURE 2.15: Neuroticism – distributional comparison of ground truth (Yoshino and Oshio, 2021a) with SNS prediction

We find that while the ground truth data displays a distribution within an expectable range around the mean of the deployed scale of their instrument (Yoshino and Oshio, 2021a), so do the prediction results. However, in the prediction results, Hokkaido, Kagoshima, Kyoto, Oita, and Tokyo remain strong outliers.

Software used

All data operations have been conducted with Python 3.8.9 (Python Software Foundation, 2023) Pandas 2.1.3 (team, 2020), and calculations have been conducted in SciPy 1.11.4 (Virtanen et al., 2020). Graphs have been plotted with Matplotlib 3.8.2 (Hunter et al., 2020), GeoPandas 0.14.1 (Bossche et al., 2020), and seaborn (Waskom, 2021) Psycholinguistic analysis has been conducted with IBM Watson Personality Insights (*Watson Personality Insights* n.d.), MeCab/IPADIC with the python of wrapper 1.0.8 (Kudo, 2005), and J-LIWC 2015 (Pennebaker, 2015; Igarashi, Okuda, and Sasahara, 2021).

2.3.6 Discussion

We introduce a novel and straightforward collection method of local tweets, which uses simple logic of high eye-validity to collect hyperlocal tweets with overseeable manual labour - the entire process took two persons about two weeks. However, we find that prediction results do not correlate with ground truth (Yoshino and Oshio, 2021a) but for $r_p = -.159$ for Extraversion and $r_s = .136$ for Conscientiousness, which indicates non-linearity, and is even behind comparable work with low correlations (Giorgi et al., 2022). Potential issues could be our collection method, selection of cities, type of data, outdated algorithms, or unknown issues with spatial psycholinguistics.

As for the *collection method*, we do not find major flaws in the collection logic. The only persons with a reason to follow hyperlocal entities must be locally connected but for a few exceptions. Intense manual inspection showed local and personal content. One can argue that these followers could be spread further locally, but why should a small local entity like a small sports club or fire station have followers from all

over Japan? Therefore we argue, backed up by existing research (Zhang, Hill, and Rothschild, 2018) that finds that geolocated panels provide more objectivity than keyword-related searches, and in the absence of self-reported or harvestable location data, that our collection method is not a likely root-cause of the low correlations.

Concerning the *selection of cities*, we may select too many cities in Hokkaido, thus over-emphasising its influence. However, we do not average all over Japan, but only by prefecture, hence if at all, this should make the results from Hokkaido more comparable to ground truth data. We also might have a wrong selection for Tokyo or a lack of spatial differentiation, and it might be better to select wards by diversity of inhabitants, and not by spatial proximity. But then, this would be artificially mingling with the results, which is against the scientific approach. Therefore we conclude that the selection of cities is not a likely root-cause of the low correlations.

With regards to the *type of data*, we surely focus solely on SNS data, which might tend to be prone to self-selection bias. However, research shows that social media can be as valid as field studies (Kalimeri et al., 2019), and classical survey-based field studies (Kazmierczak et al., 2023) suffer from self-selection bias, as well. Since we do not know the recruitment strategy from the ground-truth data (Yoshino and Oshio, 2021a), we cannot discuss self-selection bias in that regard. However *ceteris paribus*, we conclude that the type of data is less likely a root cause for the low correlations.

Regarding *outdated or wrong algorithms*, the biggest weakness of our approach is the deployment of a proprietary, poorly documented, and in the meantime discontinued product, in absence of alternatives. While the first version of IBM Watson Personality Insights is based on LIWC, subsequent further development from IBM is based on a not further documented open-dictionary approach, and reportedly reaching for all Big Five dimensions of Japanese text analysis an average correlation with survey-based outcomes of 0.3, which is in range with other approaches in psychometric literature (IBM, 2021). Also, one of its main uses cases is marketing, and being able to display Big Five personality traits is, beyond the name, not in the focus of it. Furthermore, the composition of training data is unclear, as well as any sort of manual adjustments to it, and all we know is that it was partially trained on SNS data. Finally, it is also unknown whether it is a genuine model for Japanese or just a translation – a practice quite often used in industry settings. On the other hand, the model deployed by (Giorgi et al., 2022) at least displays partially comparable results to survey-based methods, was trained specifically on SNS, is well documented, partially open source for the science community and established for research. Its only downside is that it is only available in English, and trained on data that is in the meantime outdated. In summary, we conclude that outdated or wrong algorithms *could* be a reason for the low correlations, but comparative studies cannot be conducted, since IBM Watson Personality Insights is deprecated.

With regard to *unknown issues with spatial psycholinguistics*, we assume that the spatial distribution of language also leads to different dialects and over time, different linguistic regions emerge, that are not trained separately. Hence, models that predict psychological latent traits from language might fall behind in their predictive power in some regions that deviate the most from the average, since these models are trained on aggregate data. However, in general NLP, this form of training is already being deployed. For example, Hofmann et al. (2023) successfully train a model based on various dialects and achieved SOTA performance in zero-shot geolocation prediction and zero-shot prediction of dialect feature. Other researchers deployed neural language models to identify lexical variation and words that indicate semantic and syntactic variation between regions (Kulkarni, Perozzi, and Skiena, 2016). Other researchers identify geographic-specific topics in tweet streams (Hong et al.,

2012) and use tweet-specific topic models and location specific topics, to infer latent topic distributions over locations via a joint hierarchical model, and infer personalised, location-specific preferences of users from that (Ahmed, Hong, and Smola, 2013). Also, specialised models have been built that embed geographic information and thus increase location awareness (Li et al., 2022b), geo-intention in web-search (Yi, Raghavan, and Leggetter, 2009), and recognition accuracy (Xiao et al., 2018) in large language models. More contemporary approaches match user queries with specific geographic places of interest by help of a multi-modal Geographic language model comprised of a geographic encoder that understands geographic context, and a multi-modal interaction module that helps in combining and interpreting different types of input data. They do so by representing geographic context as a new modality that has been extracted by unsupervised learning, and treat it as an additional data type besides text, vision, or sound (Ding et al., 2023). While being discussed in NLP literature for quite some time, only recent research achieved breakthroughs in identifying how and where geographic – and temporal – information is encoded in artificial neural architectures. For example, Salmas, Pantazi, and Koubarakis (2013) explore how geospatial information is encoded in LLM, and Gurnee and Tegmark (2023) show in a break-through paper that due to their exposure to huge amounts of data, Large Language Models learn temporal and spatial features, demonstrate according emergent abilities in downstream tasks, and identify dedicated time and space neurons, similar to grid cells in the human entorhinal cortex. Applied to our results, while benchmark studies achieve better correlations, they use open source models for the English language, which are trained on today's *lingua franca*, English (Giorgi et al., 2022), we use a closed source proprietary and in the meantime deprecated tool with unclear model architecture, and training data (which might even be just machine-translated and neither specific for SNS nor for Japanese).

2.3.7 Implications

A general finding is that current psycholinguistic models lack awareness of both location and location-specific linguistic distributions like dialects. That, and outdated modelling techniques might be strong contributing factors towards results that deviate from ground truth data. Furthermore, research indicates that both the purpose a model was trained for, as well as the training data that often is purpose-specific, determine its effectiveness in other areas (Koch, Romero, and Stachl, 2022), for example, a marketing-first model might deliver worse results than a research-first model, and vice versa. While such specific considerations are more in the realm of GOFAI, even highly abstract emergent abilities of LLM, of which we do not yet understand the mechanistic interpretability, do suffer from outcomes that are skewed towards bias in training data.

AI safety is another implication, since biased models may lead to unfair, or life-endangering results. For example, Faisal and Anastasopoulos (2022) introduce a framework for examining geographic bias in pretrained language models, revealing that while these accurately reflect country-language associations, they display unequal language representation and geopolitical favouritism at inference time, based on unbalanced training data – a problem also prevalent with models from the GPT family (Brown et al., 2020). While this would in many cases be a rather technical problem, Johnson et al. (2022b) show that unbalanced training data may lead to model outcomes that are skewed towards US values, which in many cases can be orthogonal to those in other nations, in core topics like gun control, immigration,

gender questions, sexuality, and secularism. Also, Atari et al. (2023) find that models of the GPT family are best represented in US Culture.

Finally, for the application in economic research, skewed psycholinguistic models might lead to wrong and expensive conclusions, and skewed, unfair policy-making. We suggest that hard-coded, dictionary-based features like those from LIWC should be the norm until further and more rigorous research has been conducted on the overlap of psycholinguistics and behavioural economics.

2.3.8 Limitations and Outlook

This survey had various limitations. Since it had no funding, only a small-scale, manual approach was possible, and no comparative field study that would have enabled additional understanding of geospatial distribution of psycholinguistic data. The Twitter API was another limitation, since with the limit of just the last 3,200 user tweets being scrapeable, not the entire linguistic space of users from geographic regions was representable. While there are alternative methods for scraping, those are not official, circumvent security measures, are potentially unethical, and thus unpublishable. Last but not least, Twitter offers no free scraping any longer, but is so expensive now that only the best-funded institutions and individuals will be able to conduct research with it. Anther limitation is that IBM Watson Personality Insights is poorly documented, is opaque in both final model and training data, and also is deprecated in the meantime, which makes it impossible to replicate existing studies with it. While some methods of identifying Tweet location are established, others, like ours, are experimental, and a comparative study is missing yet strongly needed to understand the effectiveness, advantages and weaknesses of each. Since the research community for psycholinguistics is very small, large-scale research might only come when more interdisciplinary projects are conducted, e.g., with economics or AI.

2.3.9 Conclusion

We show that language displays significant regional difference, just like personality. Furthermore, we introduce a novel method of identifying hyperlocal SNS data in absence of API features or self-reported locations. However, we also find strong weaknesses in existing approaches of geospatial psychological research, which is especially caused by the absence of location-specific psycholinguistic models. Hence, we conclude that research for behavioural economics should use hard-coded, dictionary-based methods like LIWC over or in parallel to psycholinguistic models, before these issues are clarified.

2.4 Relevance of Time for Psycholinguistic Measures

(This section was led, written, conceptualised, and analysed by Peter Romero. Yuki Mikiya analysed the SNS data and wrapped up the results. Teruo Nakatsuma supervised the project and provided guidance for Bayesian Analysis. Stephen Fitz gave mathematical input. Timo Koch gave psychological input. Further scientific support was given by Markus Bühner, Clemns Stachl, and Ramona Schödel. Data science support and feedback was given by Christopher Demetrakos, KC Chan, Shannen Romero-Perez, and Yoshiki Matsubara. Initial data science support was given by Yuhong Chen, Renyi Qu, and Julian Kota Kikuchi.)

Personality traits change over time, however research on it was sparse, since previous approaches were too time-consuming and expensive. Also, the necessary methodological complexity was beyond the capabilities of classical personality researchers, which resulted in contradictory results and lack of methodological standards. In this paper, we presented a simple and cost-effective method that may help overcoming these restrictions.

We introduced a machine learning approach for daily measurements to personality research, and developed a bespoke Bayesian algorithm to analyse the observed change. This resulted in uncovering concrete points of regime-shift that overlapped with relevant exogenous events for a Japanese sample of social media users.

With it, we showed that personality measures displayed significant elasticity under extreme exogenous conditions during the first wave of COVID-19 and the subsequent societal countermeasures, which can be interpreted as a temporary shift from normal expression of latent psychological traits z to their respective emergency expression z_e.

Concretely, we found that the group of top 25% Conscientiousness users displayed a significant change in the FFM factors Agreeableness and Extraversion. We finally compared our findings with those from similar studies in other cultures, and discussed generalisability as well as future qualitative and quantitative directions for research.

2.4.1 Introduction

Amidst the COVID-19 pandemic, societies world-wide faced a crisis unprecedented to the modern world outside wartime. Borders were closing, international travel was restricted, and entire regions and countries were locked-down. To curb the pandemic, curfews were imposed that resulted in hardship and uncertainties for populations, which had to live in isolation, fear, economic straits, and frustration (Muñoz-Fernández and Rodríguez-Meirinhos, 2021). Societies world-wide shifted to an online-first mode that lacked social interaction and came with dangers of being exposed to a dense mesh of fake news and conspiracy theories, which further undermined mutual trust (Melki et al., 2021; Limaye et al., 2020).

Extreme Exogenous Conditions

In this sudden change of environments, the perception, cognition, and behaviour of people changed dramatically, as well (Tanaka and Okamoto, 2021; Kashima and Zhang, 2021; Hino and Asami, 2021; Nagata et al., 2021). This can be interpreted as entering an 'emergency mode', in which they behaved and communicated differently. Behavioural changes to a comparable degree occurred after personal tragedies (Mechanic, 1986), natural disasters (Savage, 2019; Weisæth, 1989), as well as laboratory studies that simulated a common enemy (Jaegher, 2021). In one of these studies, the mere presence of danger made people search for more physical proximity and communicate more. Such effects occurred in other hominidae, as well, and seemed to serve as survival behaviour (Brooks et al., 2021).

People communicated differently in an emergency mode, as well, for example during natural disasters (Finau et al., 2018). Due to wide-spread social networking services (SNS) adaptation (Bayer, Trieu, and Ellison, 2020), this change in communication modified the overall information-field during the pandemic, and furthered the effects of the pandemic on individuals that practiced social distancing and spent most of their time online (Yamamoto et al., 2020). This aggregation of effects may

have been so severe that even changes in personality, which were traditionally considered to be rather slow and small over a lifetime (Bleidorn et al., 2021), became more drastic and prevalent in a much smaller time-window.

Indication for personality change

However, literature on this change process was sparse, since traditional research methods were time-consuming and expensive. Latest research on personality research indicated that personality traits indeed tended to change over time and towards greater maturity, whereby in young and late adulthood, the strongest changes occurred. This change was based on genetic, psychological, and environmental components, and displayed individual differences in "rate, timing, and direction of personality trait change" (p. 2). While both understanding about processes and strong theories did not exist yet, more research has been conducted about sources of personality change. For example, significant life events, therapies, and exogenous shocks resulted in changed personality traits. However, the speed and even direction of it displays strong interpersonal and intergroup differences. Also, results on major traumatic events were not consistent, which partly can be contributed to methodological complexities previously unknown to personality research. Personality researchers called therefore recently for more rigorous longitudinal studies, whereby a special emphasis should be put on time analysis and new forms of measurement. Measures should be more frequent – up to becoming continuous time series, and classical monomethod studies that vastly relied on self-report surveys should be augmented with multimethod approaches; focusing on more natural events like the COVID-19 pandemic. Also, this research should be conducted cross-cultural to better explore generalisability (Bleidorn et al., 2021).

We found two breakthrough studies aligned with that call. First, Sutin et al. (2020) conducted a survey-based study with a pre-post-test design that was conducted in the middle of the first wave of COVID-19 – 'late January and early February 2020 and then again in mid-March 2020' (p. 2), thus spanning approximately six weeks. Their sample was stratified to the population of the United States of America in age, gender, and ethnicity. Against their pre-registered expectations, they found that Neuroticism slightly decreased among those people that were in quarantine or isolation. Furthermore, they found with some individuals isolation to moderate decline in Openness, Agreeableness, and Conscientiousness, especially in the sub-scales 'curiosity', 'trust', and 'organisation', which they contributed to the change in circumstances of the respective individuals. Finally, they found that in working-age adults the sub-scale 'dutifulness' for Conscientiousness decreased due to one item about going to work/ school when not feeling well, which reversed its meaning under these extreme exogenous events. Second, Ahmed et al. (2020) used a machine learning approach on a random sample of Twitter-using health workers that were affiliated with various hospitals around the United States of America, whom they manually identified. They analysed both what the Twitter users 'tweeted' about, using Latent Dirichlet Allocation (LDA) and a topic-over-time (TOT) model, and how their personality changed during the first wave of COVID-19, using the application processing interface (API) from IBM Watson Personality Insights. For that, they divided the data into two blocks: before the pandemic (before February 2020), and during the pandemic (February to April 2020). Contradicting with Sutin et al. (2020), they found significant changes in all facets of the Five Factor Model (FFM): Openness and Agreeableness decreased, whereas Conscientiousness, Extraversion, and Neuroticism increased.

New modelling methods for personality change are needed

What both studies had in common was the notion of extreme exogenous events that influenced the expression of personality directly on a state or trait level, or indirectly through changing the validity of the instruments. For example, existing scales could have been reinterpreted by test-takers due to changing circumstance, or the use of words could have been altered, which changed the measurements of the machine learning approach. Since no confirmatory repetition study existed, this was hard to tell. Also, both studies focused on the time before and during the first wave of COVID-19. However, it was unclear what happened during or after the first wave. Finally, both studies focused on the United States of America and therefore displayed a selection bias. To confirm, whether this phenomenon of personality change was generalisable, one had to test another population with little exposure to the English language and its information-field, yet under the same contextual embedding of an industrialised nation affected by COVID-19. One had to use more precise and granular measures in an non-intrusive, natural experimental setting, to replicate the serendipitous circumstances of the first survey where participants were surprised by a request for a second survey (Sutin et al., 2020), and the non-intrusive nature of the second survey (Ahmed et al., 2020). Optimally, this took place in a country that was vastly secluded, yet had a wide-spread use of SNS, which acted as a 'quasi-laboratory'. Finally, one had to combine the psychometric rigour from Sutin et al. (2020) with the forward-looking and progressive machine learning approach from Ahmed et al. (2020). Japan could be such a country.

Takeshimura (2020) found that Japanese media users dynamically reacted to COVID-19-related news. They concluded that the public perception of the information field followed dynamic spatiotemporal patterns. In congruence with former models that connected narrative flow with mass psychological phenomena (Houghton, Siegel, and Goldsmith, 2013), they further concluded that acceleration and velocity of the social attention to risks was influenced by the history of the information flow and existence of alternative narratives. While they mainly focused on news coverage, research about the usage of Twitter in Japan indicated that it was used as a quick tool for information spreading, before other media was consumed. While this bore the opportunity for abuse for spreading fake news, "Twitter became the supplier of information and knowledge for the citizens ... in the early days of the disaster and also the basis for building social capital." (p. 33) (Kaigo, 2012). We concluded that first, this dynamic, psychophysical approach was another indication that a machine-learning-driven psycholinguistic study on the effects of extreme exogenous conditions on personality change within Japanese SNS users was indicated, and second that Twitter was a good medium for that.

Digital Footprints from SNS were proven to be valuable for social research. By applying machine learning, it was possible to deduce personality (Kosinski et al., 2015), gender and age (Koch, Romero, and Stachl, 2020), and the "hopes and dreams, preferences and motivations, social connections, daily routines, and physical whereabouts" (p.5) (Stachl et al., 2021) of its users. Research on data from Twitter was well established in behavioural research, and it showed to be effective in deducing personality, age, and gender (Schwartz et al., 2013). While daily granularity of Twitter data suggested the use of time series methodology with, this has surprisingly not been exploited by computational social sciences yet. Beyond the novelty of feature extraction and prediction of latent traits, usual approaches were to take the outcomes from advanced methods, and then deploy them in an old fashioned and shallow way, for example using ANOVA (Sutin et al., 2020) or t-tests (Ahmed et al., 2020). This

means that a lot of dimensionality of the data got lost in this process. Also, most established methods of analyses, which were overwhelmingly based on various forms of regression analysis, classification or prediction, fell short to provide understanding about both the nature and the timing of dynamic changes. We therefore decided to use Bayesian methodology to analyse the changes in a more dynamic way that also captured additional dimensionality.

2.4.2 Empirical Strategy and Model Specification

Bayesian Analysis is based on updating the joint posterior density of parameters based on observed data. Unlike frequentist methods, it does not rely on point estimates, but focuses on the joint posterior distribution, whereby the summary statistics conveys useful insights about the parameters. This makes it useful in situations with restricted data availability and outcomes that are fundamentally probabilistic in nature, which is the case in human behaviour (Fox, 2010; Nakatsuma, 2007). Also, the prior specification of models allows the injection of insights from existing theories, which is central in behavioural research, while still allowing to learn from data, thus merging bottom-up practicality from machine learning with theory-driven top-down precision. Though it has been successfully used in psychometrics before (Natesan et al., 2016), it neither is wide-spread in psychometrics nor in computational social sciences. The authors hope to introduce this methodology to the wider research community.

Flexible Bayesian Modeling of Finite Discrete Distributions

Suppose we have n days in the data set and they are categorised into k groups. Define a label for grouping as

$$x_i = j, \quad i = 1, \ldots, n, \quad j = 1, \ldots, k, \tag{2.1}$$

and the probability that we pick the day i from group j as

$$\Pr\{x_i = j\} = p_j, \quad 0 \leqq p_j \leqq 1, \quad \sum_{j=1}^{k} p_j = 1. \tag{2.2}$$

We may suppose the probability p_j is determined by a parametric model (e.g., binomial distribution) when the label x_j can be regarded as an integer j. Although such a parametric assumption may ease the estimation procedure for discrete data $(x_1, \ldots, x_n)$, it lacks flexibility and applicability to more general data sets. Therefore, we will not impose upon any specific parametric structure on $(p_1, \ldots, p_k)$ and assume a more flexible prior for them instead. In our study, we use the following Dirichlet distribution:

$$(p_1, \ldots, p_k) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_k), \tag{2.3}$$

as the prior for $(p_1, \ldots, p_k)$. The p.d.f. of the Dirichlet distribution is

$$p(p_1, \ldots, p_k) = \frac{\prod_{j=1}^{k} p_j^{\alpha_j - 1}}{B(\alpha_1, \ldots, \alpha_k)}, \quad B(\alpha_1, \ldots, \alpha_k) = \frac{\prod_{j=1}^{k} \Gamma(\alpha_j)}{\Gamma\left(\sum_{j=1}^{k} \alpha_j\right)}, \tag{2.4}$$

where $B(\cdot)$ is the multivariate beta function and $\Gamma(\cdot)$ is the gamma function.

The likelihood of $(p_1, \ldots, p_k)$ given $(x_1, \ldots, x_n)$ is

$$p(x_1, \ldots, x_n | p_1, \ldots, p_k) = \prod_{j=1}^{k} p_j^{y_j}, \quad y_j = \sum_{i=1}^{n} \mathbf{1}(x_i = j), \tag{2.5}$$

where $\mathbf{1}(\cdot)$ is the indicator function that is equal to one if the expression inside the parentheses is true; otherwise, it is zero. Thus y_j is the number of days who are categorised into group j in the data set. By applying Bayes' theorem, the posterior distribution of $(p_1, \ldots, p_k)$ is obtained as

$$\begin{aligned} p(p_1, \ldots, p_k | x_1, \ldots, x_n) &= \frac{p(x_1, \ldots, x_n | p_1, \ldots, p_k) p(p_1, \ldots, p_k)}{p(x_1, \ldots, x_n)} \\ &= \frac{\prod_{j=1}^{k} p_j^{y_j + \alpha_j - 1}}{B(y_1 + \alpha_1, \ldots, y_k + \alpha_k)}, \end{aligned} \tag{2.6}$$

which is also a Dirichlet distribution. Since the mean of the Dirichlet distribution (2.6) is

$$\mathrm{E}[p_j | x_1, \ldots, x_n] = \frac{y_j + \alpha_j}{\sum_{j'=1}^{k} (y_{j'} + \alpha_{j'})}, \quad j = 1, \ldots, k, \tag{2.7}$$

it can be interpreted as a Bayesian counterpart to a histogram.

For later use, we derive the marginal likelihood for the Dirichlet model. By using the identity

$$\begin{aligned} p(p_1, \ldots, p_k | x_1, \ldots, x_n) &= \frac{p(x_1, \ldots, x_n | p_1, \ldots, p_k) p(p_1, \ldots, p_k)}{p(x_1, \ldots, x_n)} \\ \Rightarrow \quad p(x_1, \ldots, x_n) &= \frac{p(x_1, \ldots, x_n | p_1, \ldots, p_k) p(p_1, \ldots, p_k)}{p(p_1, \ldots, p_k | x_1, \ldots, x_n)}, \end{aligned} \tag{2.8}$$

the marginal likelihood is derived as

$$p(x_1, \ldots, x_n) = \frac{B(y_1 + \alpha_1, \ldots, y_k + \alpha_k)}{B(\alpha_1, \ldots, \alpha_k)}. \tag{2.9}$$

Regime Shifts in Finite Discrete Distributions

Suppose days in the data set are not homogeneous but they belong to two subpopulations with different probabilities in (2.2). Without any loss of generality, suppose the first $\tau - 1$ days belong to one subpopulation and the last $n - \tau$ days belong to the other, that is,

$$\Pr\{x_i = j\} = \begin{cases} p_{1j}, & (1 \leqq i < \tau); \\ p_{2j}, & (\tau \leqq i \leqq n). \end{cases} \tag{2.10}$$

We call τ the change point and suppose it is unknown.

For a fixed τ, the likelihood is given by

$$\begin{aligned} p(x | \tau) &= p(x_1, \ldots, x_{\tau-1} | p_{11}, \ldots, p_{1k}) p(x_\tau, \ldots, x_n | p_{21}, \ldots, p_{2k}) \\ &= \prod_{j=1}^{k} p_{1j}^{y_{1j}(\tau)} p_{2j}^{y_{2j}(\tau)}, \end{aligned} \tag{2.11}$$

where $x = \{x_i\}_{i=1:n}$ and

$$y_{hj}(\tau) = \begin{cases} \sum_{i=1}^{\tau-1} \mathbf{1}(x_i = j), & (h = 1), \\ \sum_{i=\tau}^{n} \mathbf{1}(x_i = j), & (h = 2). \end{cases}$$

Therefore, with the prior

$$(p_{h1}, \ldots, p_{hk}) \sim \text{Dirichlet}(\alpha_{h1}, \ldots, \alpha_{hk}), \quad h = 1, 2, \tag{2.12}$$

the conditional posterior distribution given τ is derived as

$$\begin{aligned} &p(p_{11}, \ldots, p_{1k}, p_{21}, \ldots, p_{2k} | x, \tau) = p(p_{11}, \ldots, p_{1k} | x, \tau) p(p_{21}, \ldots, p_{2k} | x, \tau) \\ &= \frac{\prod_{j=1}^{k} p_{1j}^{y_{1j}(\tau)+\alpha_{1j}-1}}{B(y_{11}(\tau) + \alpha_{11}, \ldots, y_{1k}(\tau) + \alpha_{1k})} \times \frac{\prod_{j=1}^{k} p_{2j}^{y_{2j}(\tau)+\alpha_{2j}-1}}{B(y_{21}(\tau) + \alpha_{21}, \ldots, y_{2k}(\tau) + \alpha_{2k})}, \end{aligned} \tag{2.13}$$

which is the product of two Dirichlet distributions. Thus the marginal likelihood for τ is

$$p(x|\tau) = \frac{B(y_{11}(\tau) + \alpha_{11}, \ldots, y_{1k}(\tau) + \alpha_{1k})}{B(\alpha_{11}, \ldots, \alpha_{1k})} \times \frac{B(y_{21}(\tau) + \alpha_{21}, \ldots, y_{2k}(\tau) + \alpha_{2k})}{B(\alpha_{21}, \ldots, \alpha_{2k})}. \tag{2.14}$$

In the Bayesian framework, we can derive the posterior distribution of the change point τ. Suppose, periods between t_1 and t_2 are the candidates for the change point τ. Since the marginal likelihood for τ $(t_1 \leqq \tau \leqq t_2)$ is given by $p(x|\tau)$ in (2.14), by applying Bayes' theorem, we obtain the posterior distribution of τ as

$$p(\tau|x) = \frac{p(x|\tau)p(\tau)}{\sum_{s=t_1}^{t_2} p(x|s)p(s)}, \tag{2.15}$$

where $p(\tau)$ is the prior distribution of the change point. In practice, we often assume the uniform prior over $(t_1, \ldots, t_2)$. In this case, we have

$$p(\tau|x) = \frac{p(x|\tau)}{\sum_{s=t_1}^{t_2} p(x|s)}. \tag{2.16}$$

Finally, we obtain the posterior distribution of $(p_{h1}, \ldots, p_{hk})$ $(h = 1, 2)$ as

$$p(p_{h1}, \ldots, p_{hk}|x) = \sum_{s=t_1}^{t_2} \frac{\prod_{j=1}^{k} p_{hj}^{y_{hj}(\tau)+\alpha_{hj}-1}}{B(y_{h1}(\tau) + \alpha_{h1}, \ldots, y_{hk}(\tau) + \alpha_{hk})} p(\tau|x), \tag{2.17}$$

which is a mixture of Dirichlet distributions. As a result, the posterior mean of (2.17) is a weighted average of (2.7) , i.e.,

$$\mathrm{E}[p_{hj}|x_1, \ldots, x_n] = \sum_{s=t_1}^{t_2} \frac{y_{hj}(s) + \alpha_{hj}}{\sum_{j'=1}^{k}(y_{hj'}(s) + \alpha_{hj'})} p(s|x), \quad j = 1, \ldots, k, \quad h = 1, 2. \tag{2.18}$$

One of the base assumptions in psychometrics is that any measured psychological latent trait z is composed of the real value v and a measurement error ϵ. Since this error is subject to various influences outside and inside the individual (Rust and Golombok, 2014b), we can further assume that a series of m measurements will result each time in a slightly different measured latent trait z, the so called *state* s. Given

sufficient frequent measurements m, and assuming that no priming, experimental bias, or memory effect took place, we assumed the states s in a within-subject design to result in a prior distribution that resembles a stationary fluctuation around the real value z.

2.4.3 Methodology

Modelling emergency expressions of personality traits

If this hypothesis held true, the individuals were not homogeneous but belonged to various subpopulations, of which some displayed a change in expressed personality under extreme exogenous conditions. Once these conditions occurred, we assumed the stationary fluctuation of the relevant subpopulation to change around a different latent trait z_e, denoting the emergency expression e of the normal expression z. As described in the model specification, we assumed that this change from z to z_e to happened around the regime shift point τ, which represented that point in time, when the very same group of persons would be confronted with an exogenous shock in form of an ongoing change in the environment that endangered adaptation and thus represents a stressor (Selye, 1955) that needed to be coped with in an emergency mode. Given this change in the relevant subpopulation, we expected a different probability around the change point τ that a randomly chosen day i will be classified as displaying the aggregate expression j. We further expected around the highest likelihood of that change point τ a relevant exogenous event to have taken place, that must have affected the relevant subpopulation. In order to prove this, we programmed the model in python 3.8.9 from scratch, using the following modules: numpy 1.26.0 (Harris et al., 2020), pandas 2.1.3 (team, 2020), scipy 1.11.4 (Virtanen et al., 2020), Matplotlib 3.8.2 (Hunter et al., 2020), math, sys, os, and random. The code and data examples can be inspected in the OSF repository of this paper.

The first wave of the novel coronavirus represented a clear exogenous shock that permanently changed the contextual embedding of individuals and therefore offered an unique and perfect opportunity to study our hypotheses in an natural experimental setting, assuming a within-subject design. In order to understand effects of this relevant contextual change in an *in vivo* setting without interference through surveys or laboratory settings, we decided to not let subjects take tests or questionnaires, but to deduce latent psychological traits from the text of SNS users. By focusing on text from SNS, we approximated the influence in the information field of societies, which we used as the proxy for changes in the social embedding of individuals. Though, one could argue, that despite individuals influencing each other, SNS users represented a closed group and therefore the self-selection bias (Zhang, Hill, and Rothschild, 2018) represented a hidden between-subject design of those persons exposed to information field on SNS, and those that are not, we argued that the effect of the novel coronavirus is so encompassing and permeates so many aspects of society, that no user can be found that was not influenced by the COVID-19-related information field and therefore represents a natural experiment.

To further approximate the effect of this changed information field on the general public, we refrained from using niche SNS like Reddit or semi-public SNS like Facebook or LinkedIn. Instead, we focused on public Tweets on Twitter, which can be read by every internet user, and which were very often embedded in homepages like news or blogs, thus representing the overall information field. To ensure observing a relevant subpopulation, we filtered out only on those Twitter users, who wrote about COVID-19-related tag-words.

Using computational methods to predict latent psychological traits

Unfortunately, the ideal positioning of Japan limited also our choice of tools for language analysis, since the field of psycholinguistics was in its infancy in Japan. Until today, even a dictionary for LIWC, the quasi-standard for psycholinguistic research, was still in preparation, and researchers used so far a less than ideal approach of translating the Chinese version of LIWC with online translation tools (Guntuku et al., 2019; Shibata et al., 2016). With less than 18% similar words in daily use, and many different linguistic concepts (Ishii and Onuma, 1998), this approach was not up to the standards required for our endeavour. Therefore, we decided to use the commercial version of IBM Watson Personality Insights, which had wide-spread industry coverage, yet had been rather under-utilised for academic purposes.

Also, the first version of IBM Watson Personality Insights was based on LIWC, and had been since then optimised and further developed by IBM internally, using an open dictionary approach, which outperformed their first LIWC-based model, reaching for the Big Five Dimensions in the Japanese language an Mean Absolute Error of 0.1, and an average correlation with survey-based outcomes of 0.3, which was well in the average of psycholinguistic literature that described prediction values based on text between 0.2 and 0.4 (IBM, 2021). Unfortunately, this later model was not very well documented, so we could not explore deeper technical details. However, despite the 'low' correlation, we considered the outcomes of IBM Watson Personality Insights as just an imprecise measurement, and therefore adding to the measurement error ϵ, which should be reflected in the stationary fluctuation of the state value s around the real latent trait z and its emergency expression z_e.

The advantage of using predictive approaches for measuring latent psychological traits *in vivo* and in a non-invasive way came at various expenses. First, we only could provide assumptions of causality, since no grounding through behavioural surveys or observations took place. Second, we neither could control the contextual conditions under which the SNS texts were created, nor sort individuals by their specific life situations that influenced them at that point in time. While extreme exogenous events were equally influential for most people, some may face other, more relevant situations like personal tragedies. This would increase the measurement error for these persons disproportionally. Third, individuals were influenced by media events and their relevant social networks. Since SNS were by definition social networks, in which media and news were distributed, we assumed mutual influence of its users, which further weakened our ability to infer causality from observations on an individual level. Last, resulting from these points, we assumed that the i.i.d. assumptions were violated, and with that the normality assumption, as well. While we expected this effect to be less strong with clusters of individuals under mutual influence in the same context, it would be prevalent with randomly chosen SNS users. This reduced both the choice of sampling, as well as that of methods.

We concluded that an analytical strategy on an individual level would be way too 'noisy', and that the effects of extreme exogenous conditions on personality should be observed in patterns of larger samples of individuals, clustered by similar psychological properties. The five factor model was a prime candidate for that, since it was well established in psychology, had been used in a variety of adjacent fields, and various methods to measure it have been developed, from the predictor (Ortner and Schmitt, 2014), over the criterion (Ones and Viswesvaran, 2001), to the outcome space (Gosling et al., 2002). While it would be preferable to generate these clusters based on a more profound theory of personality types, these were a rather difficult topic, with a variety of rather controversial opinions and research approaches. The most agreed-upon

approach seemed to be conducting latent class analyses, mixed Rasch models, or the SEM approach from Raykov (Raykov, Harrison, and Marcoulides, 2020). However, since our research was based on a survey-free *in vivo* approach, most underlying assumptions would be violated. Mainly when using a maximum likelihood estimation, we would have struggled with independently and normally distributed residuals (Curran, 2003) given the considerations above.

Since we explored the influence of the information field of extreme exogenous conditions on personality of clusters of people, and since we measured within-subject variance, yet assume mutual and contextual influence, we therefore decided to cluster individuals by the baseline of their past FFM-expression. While theoretical, the distribution of the expression of FFM was assumed to be normally distributed, we could not expect this to be the case in a sample that was chosen by usage of tag-words from SNS, which already came with a self-selection bias. Also, since psychological phenomena displayed a geographic distribution (Rentfrow, Gosling, and Potter, 2008b), and since we expected a larger portion of the users to be from urban centres, we assumed a skew in distribution. Hence, we created baselines of users based on the interquartile distance, focusing on the top 25% users of each expression of the FFM in 2019, the year before the outbreak. This clustering resulted individuals belonging to one (O,C,E,A,or N) or several (for example, E/A) top groups per FFM dimension. To capture the pure effects per FFM dimension, we decided to only focus on those individuals belonging to only one top group, ignoring the mixed types.

Next, we clustered the Tweets from each of the top OCEAN groups on a daily basis, and derived personality measures from that, using IBM Watson Personality Insights, and brought these into a time series, displaying the change in personality expression per cluster. That means that for each top OCEAN group from before the outbreak, we derived OCEAN expressions from after the outbreak, resulting in 25 time series.

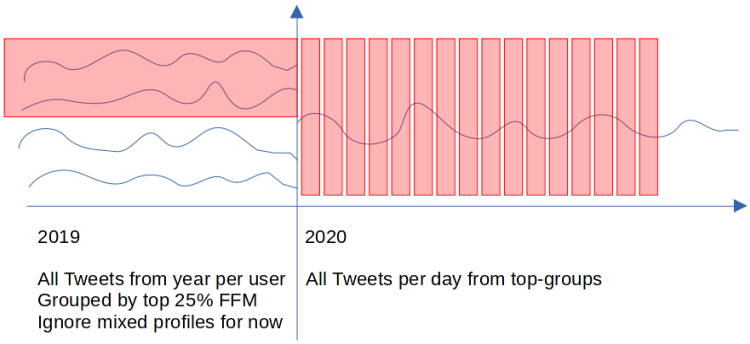


FIGURE 2.16: FFM Prediction Approach

To finally evaluate whether a change took place, we first inspected the time series visually, and then tested on stationarity via autocorrelation and autoregression using pandas 2.1.3 (team, 2020), statsmodels (Seabold and Perktold, 2010), math, and sklearn 1.3.2 (Pedregosa et al., 2011) for Python. Then, we conducted the Bayesian regime-shift analysis to those time series, where we found significant changes, using numpy 1.26.0 (Harris et al., 2020), pandas 2.1.3 (team, 2020), scipy 1.11.4 (Virtanen et al., 2020), and math for Python. Since the Bayesian analysis already distinguished between probabilities of belonging to different distributions, we refrained from conducting additional confirmatory t-tests between the distributions separated by the regime-shift points. Finally, to identify whether a change in the information-field took place that was concordant with the regime-shift in personality, we analysed Tweets

during the identified periods before, between, and after the points of regime-shift by wordclouds and structured topic models (Roberts, Stewart, and Tingley, 2019a). We generated the wordclouds with quanteda (Benoit et al., 2018) in R 4.3.0 and Mecab (Kudo, 2005) for a morphological analysis in Python 3.8.9 (Python Software Foundation, 2023), and the STM with the stm package (Roberts, Stewart, and Tingley, 2019a) in R.

We applied the above model of a potential personality regime shift in the following way: $x_i = j$ stands for an label x that a day i had the expression j of a relevant measured aggregate psychological latent trait. We had n days, and k potential expressions of that aggregate psychological latent trait. Those measurements took place m times, thus $(t_1, ..., tm)$ representing m time intervals, in which for every day i, the very same aggregated psychological measure was generated.

Data

For the analysis, we collected the following data, of which we will subsequently give a brief description:

1. Relevant Tweets from Japanese Twitter users
2. Timeline of events
3. Official Corona Numbers

Relevant Tweets from Japanese Twitter users

A sample from Japanese Twitter users bore four major advantages. First, Japan was close to the origin of the novel coronavirus in Wuhan, China, and there was a historic close relationship between the two countries, including a significant Chinese population with close ties to China. This made Japan a prime country for being hit by coronavirus effects first, before other countries. Second, given it's otherwise rather secluded geographic location as an island nation, with traditionally rather closed borders and low degrees of English skills in the population, we assumed the effects on the population to be more free from outside interference in the information field, which transformed it into a quasi laboratory for our purposes. Third, Japan had the second largest user base after the USA, more than double the amount than the next biggest user base, India. The 50.9 Million Japanese users made Twitter a significant SNS in Japan, with about 40.6% of the Japanese population using it. This made it more representative than using it for example in the USA, where 69.3 Million people used this service, or about 21% of the population (Tankovska, 2021). Last, there was a distinct event that initiated the virus in Japan, with the arrival of the cruise ship Diamond Princess, and therewith, we could identify a perfect starting point for our model.

We collected the recent 3,200 Tweets from all Japanese Twitter users, who 'tweeted' between January and June 2020 about the following COVID-19-related keywords: 'Corona', 'COVID', 'Pandemic', 'Face Mask', 'Hand Sanitizer', 'Hand Soap', 'Toilet Paper', 'Napkins', and 'Tissues'. We selected all hygiene-related keywords, since soon after the outbreak, many of these articles were sold-out in stores and thus a relationship with the keywords and COVID-19 is obvious. Also, this allowed us to capture a broader range of the phenomena over and above purely COVID-19-related Tweets.

For some users, these 3,200 Tweets represented their life-time Tweets, for others only the last couple of years or even months. The number was based on the restrictions

from the official Twitter API that we used. To communicate with it, we used Python 3.8.9 (Python Software Foundation, 2023) with the following packages: numpy 1.26.0 (Harris et al., 2020), pandas 2.1.3 (team, 2020), tweepy (Roesslein, 2020), datetime, and pytz.

This resulted in 7,435 Twitter users and 23,501,189 Tweets, whereby some only 'tweeted' about one or two, others about all topics. The average user produced 3,160.89 Tweets. Unfortunately, we could not derive any demographic information due to API restrictions, and missing official statistics from Twitter. However, given the importance of Twitter in the Japanese SNS market, we assumed the sample to be representative of Japanese society.

Timeline of events

To capture on the ground truth, we manually created an extended timeline of COVID-19-related events from December 2019 to June 2020. The reason for choosing this time span was that the novel coronavirus was reported for the first time between December 27^{th} and 29^{th} 2019 from various local hospitals in Wuhan, China, to the local branch of the Chinese Center for Disease Control and Prevention. A day later, it was discussed in groups from Chinese social networks for relevant medical personnel, and appeared on international Flu tracking websites, just to be officially reported to the World Health Organisation the next day, on December 31st, 2019. From there, first reports spread to Japan early January, as discussed later. We chose June 2020 as an end-point, since that is when the number of hospitalisations in Japan was significantly reduced, and some countries announced the end of the first wave (Tashiro and Shaw, 2020).

To generate an exact time-line of the events, we manually searched several sources for COVID-19-related events, including news, official governmental announcements, coverage from international bodies like the UN and WHO, as well as from various NGO. This choice of sources is based on both availability of information and the need to understand the contextual embedding of Japanese Twitter users when they are exposed to information. Findings from media psychology indicate that those news have the highest values to individuals that are negative and of geographic proximity (Araujo and Meer, 2020). However, systems theory indicates that also more distal events may have an influence on a local level if those are within the broader systemic context (Willke, 2000). Therefore, we theorised a layered model of relevance for individuals, whereby more proximal events had a stronger relevance than more distal ones, for example announcements from the Japanese government should have been more influential than those from the World Health Organisation. We assumed though, that the latter may still had an influence on individuals, either to a lower magnitude or time-delayed. This resulted in a time-line of daily events in that period, with varying levels of local importance that allowed us to establish a solid understanding of on-the-ground-truth. We used that timeline to identify what exactly happened at the day of the regime shift. A complete chronology of the events in Japan, the geographic importance to Japanese Twitter users, and the respective sources can be found in the OSF Repository.

Official Corona Numbers

For further understanding of this on-the-ground-truth, and for quantifying real-life events, we used the official numbers provided by the Japanese government (MHLW, 2021), which we cross-checked with data from the World Health Organisation (WHO,

2021) to ensure their correctness. We used these numbers to augment the qualitative time-line of daily events by quantitative measures and to visualise the development of personality change in contrast to real-life effects of COVID-19-related like hospitalisations and deaths.

2.4.4 Results

We conducted the analyses in alignment with the lead logic to first extract daily personality measures, then identify potential patterns in those, and analyse whether the emergence of these patterns aligned with external events that either act as exogenous shocks or change the contextual embedding of Twitter users, which made them adapt and communicate in an emergency mode. To add additional evidence, we further analysed the content of Tweets in the relevant times before patterns emerged. For this, we conducted the following steps:

1. Language-based Personality Predictions
2. Tests for Stationarity
3. Bayesian Analysis
4. Timeline of Events
5. Text and Topic Analysis on Tweets

Language-based Personality Predictions

IBM Watson Personality Insights delivered results from the FFM in percentile scores, which we generated for each of the five top 25% FFM groups on a daily basis. We then brought these into a time series that covered the onset of the COVID-19 pandemic on the 29^{th} of December, 2019, until the end of the first wave on the 15^{th} of June, 2020. All time series of the five groups were depicted and described in the appendices. The most promising group to display changes in personality seemed to be the top 25% Conscientiousness from 2019, which displayed a distinct change in Extraversion and Agreeableness, and maybe in Neuroticism, as well. While Extraversion was most interesting, since it changed order with Neuroticism, Agreeableness seemed to display the most distinct regime-shift, and Extraversion seemed to have a broader range in general, wherefore we decided to conduct further analyses with Agreeableness. [7]

[7] An interesting side-finding, which could be explored in future publications was that each group displays a different order in the strength of FFM factors, as well as a different overall range in their expression.

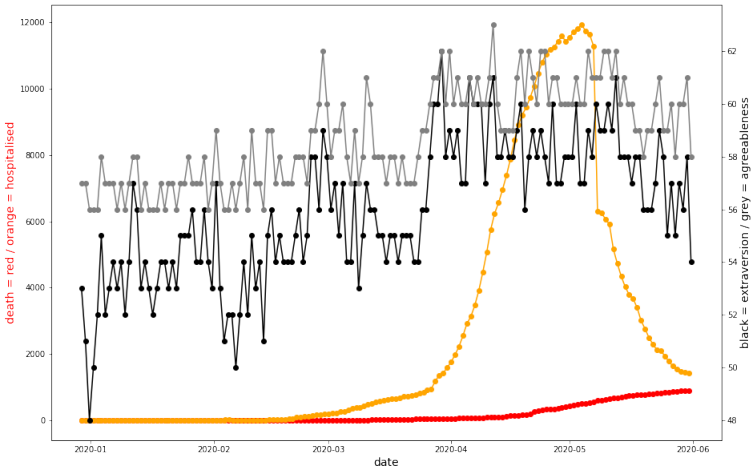


FIGURE 2.17: Extraversion and Agreeableness vs Hospitalisations and Deaths

Figure 2.17 shows an overlay of both the time series of Agreeableness and Extraversion of the top 25%Conscientiousness group from 2019 with the time series of the official COVID-19-related deaths and hospitalisations. It shows that these peak during the regime-shift, which seemed to start around the first inflection point of Hospitalisations. The FFM percentiles were plotted on the ordinate and the date on the abscissa. The authors concluded that there is enough evidence to conduct further analyses as planned.

Tests for Stationarity

We tested for stationarity by autocorrelation and autoregression. As expected, the autorcorrelation function (ACF) plot for Neuroticism indicated stationarity. It degraded with increasing lag quickly to zero and fluctuated around it. However, aligned with prior expectations, the ACF plot for Extraversion indicated non-stationarity and thus indicated that a regime-shift took place. Also, the ACF plot for Agreeableness indicated non-stationarity and thus indicated that a regime-shift took place, as well. In summary, Extraversion and Agreeableness displayed non-stationarity and were candidates for further analyses. These results were further confirmed with an autoregression model, creating a seven day forecast with a lag of 29 days. Extraversion (RMSE: 2.515) performed worse than Agreeableness (RMSE: 1.343). Therefore, aligned with visual inspection, the Bayesian Analysis to determine the point of regime-shift was conducted using the Agreeableness time-series from the 2019 top 25% Conscientiousness group.

Bayesian Analysis

To determine the point of regime-shift, the model from section 2.4.2 was implemented in Python 3.8.9 (Python Software Foundation, 2023) on the Agreeableness time series displayed in figure 2.18 in grey. The resulting updated posterior distribution represented the probability for a regime-shift for each day that could have been caused by the COVID-19 pandemic, and was calculated using Equation 2.16. It is plotted in figure 2.18.

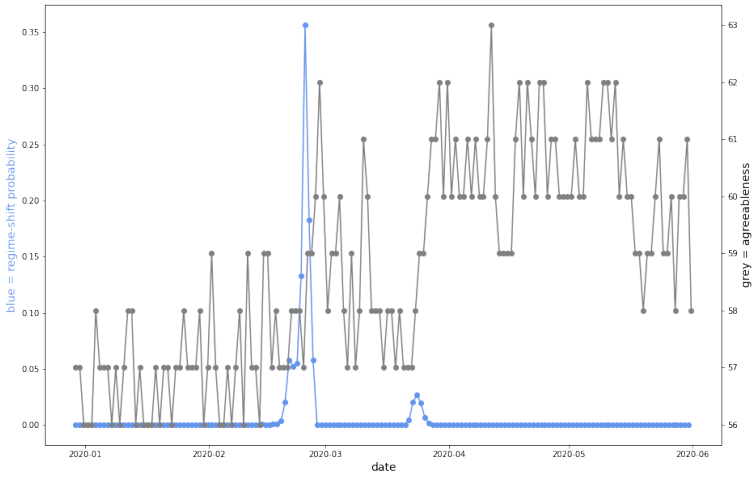


FIGURE 2.18: Time series of Agreeableness vs Probability of Regime-Shift

Counter-intuitively, and against prior assumptions, this posterior plot indicated two (plus one) plausible candidates of regime-shift in the data, whereby the first point was the most likely, and was preceded by a smaller burst. The first distinct point of regime-shift took place at day 58, which was the 24^{th} of February 24, 2020, and it represented a probability of 36% that the regime-shift took place at that day. It was preceded by a smaller burst on day 54, which was the 21^{st} of February, 2020, and it represented a probability of 6% that the regime-shift took place at that day. The second spike was at day 86, which was the 25^{th} of March, 2020, and it represented a 3% probability that the regime-shift took place at that day. However, since the posterior distribution represented the probability of a regime-shift, the first distinct point was most important for interpreting the outcomes. Given the probabilistic nature of human behaviour though, we treated each point as a potential regime-shift in the interpretation.

Results from Timeline Of Events

For this, we created a timeline of daily COVID-19-related events to understand, which exogenous event took place around the regime-shift points of personality change. In line with various theories on relevance of geospatial embedding, we included such from international bodies like WHO, more distal areas like the EU and US, more proximal areas like China, and from Japan. The full timeline can be viewed in our OSF repository.

The 'real' point of regime-shift then occurred on 24^{th} of February, 2020. At that day, an official expert meeting for the novel coronavirus took place in Japan, in officials from the Ministry of Health, Labour and Welfare discussed ways to contain group infections, secure hospital beds and help people acquiring better understanding about the virus. A day later, that same ministry established a Cluster Response Section in accordance to the Basic Policies for Novel coronavirus Disease Control, which was decided upon the same day by the newly formed Headquarters for Novel coronavirus Disease Control of the Japanese government and announced by prime minister Abe Shinzo to the public. Finally, he announced just two days later, on 27^{th} of February, 2020, that all schools in Japan would be told to close until April to respond to the outbreak.

The burst before the 'real' point of regime-shift occurred on 21^{st} of February, 2020. At that day, the organisers of the Nagoya Women's Marathon cancelled participation for the public and only allowed elite athletes to participate amid fear of increased infection in crowds. Also, the Paralympics Boccia tournament was postponed to protect athletes. A couple of days earlier, on the 17^{th} of February, 2020, the Ministry of Health, Labour and Welfare presented for the first time guidelines on seeking medical help in case of COVID-19 infection. This was interpreted as a 'pre-activation' of the public awareness of real-life ramifications of COVID-19 in mass events, which then was 'activated' for the first time on the 21^{st} of February, 2020, where the said decisions where made. We further assumed a pre-existing 'pre-activation' of public awareness of everything related mass-sports-events, since Japan was preparing for the 2020 Olympics that were planned to take place in Tokyo.

The second point of regime-shift took place on March 25^{th}, when the governor of the Tokyo Metropolitan Government Koike Yuriko urged residents to stay at home during weekends, just a day after prime minister Abe Shinzo and the International Olympic Committee (IOC) announced the Tokyo 2020 Olympics to be postponed to 2021 amid the uncertainties of the pandemic.

The two points of regime-shift had in common that these occurred around official announcements from the government, surrounded by announcements from the Olympics, concerning mass-events. Since Japan had been preparing for the Olympics, and since these were heavily advertised in Japan, involving many parts of the population, the relevance was greater for the population than other events. Also, Japan was known for its conservative culture and high trust in the government (Kim and Voorhees, 2011), so again, official announcements from the government were relevant for Japanese people. In line with the theory outlined earlier, negative and spatially proximal news were more relevant in public perception.

It is therefore a good explanation that in the uncertain and fear-loaded atmosphere during the onset of the COVID-19 pandemic, the official announcements from the IOC and the prime minister led to the perceived personality changes. Especially the high conscientiousness group might have been receptive to messages from authorities, since rule-obedience is highly correlated with (Bègue et al., 2015). Furthermore, the displayed significant increase of agreeableness and extraversion was aligned with findings that humans cooperate more strongly in situations of adversary and danger (Dawans et al., 2012-06-01), for which it was helpful to be more nice to others (agreeableness) and to reach out to them in the first place (extraversion). The slight increase in neuroticism, that was also reported by other studies (Sutin et al., 2020) was not significant, which makes sense, since in emergency situations, it was relevant to have higher degrees of emotional stability as an emergency expression z_e of ones latent traits z.

Results from Text and Topic Analysis On Tweets

To understand what Japanese Twitter users were talking about in the respective time slots between the regime-shifts, and to better understand whether COVID-19 or other, more seasonal topics dominated, we conducted quantitative content analyses of the Tweets from the top 25% group of high Conscientiousness, on which the analysis on regime-shift points was based.

First, we divided the paper into six periods, which are listed in table 2.5. All Tweets on and before 31^{st} of October, 2019, formed a baseline for overall understanding of what the top 25% Conscientiousness group was talking about. To better understand the influence of the pre-holiday and holiday season on Tweets, we created

the next period on and between 1^{st} of November, 2019, and 28^{th} of December,2019 – the day before the outbreak. For the time from the official outbreak towards the first burst before the actual point of regime shift, we created a time period on and between the 29^{th} of December, 2019, and 20^{th} of February, 2020. This time span described the most interesting part of psychological change, where fear of the unknown was more present than actual, scientifically-proven facts. The time on and between this first burst and the actual point of regime shift, was on and between the 21^{st} of February, 2020, and 24^{th} of February 2020. Though being a small time-span, we included it to understand what users 'tweeted' about between both official events described before, when COVID-19 got increasingly more influence on their lives. We further included it to test, whether both events are separable or whether we should consider them as one big event of regime shift. To better understand what users spoke about between the actual first and second point of regime shift, we included the time on and between 25^{th} of February, 2020, and 23^{rd} of March, 2020. And finally, to understand what users 'tweeted' about after the second point of regime-shift, we took the time on 24^{th} of March, 2020, and thereafter.

Period	**Description**	**Number of tweets**
- - 31.10.2019	baseline in general (included to get an overall understanding of what people are talking about)	277,348
01.11.2019-28.12.2019	pre-holiday season (included to understand influence of period shortly before outbreak)	111,392
29.12.2019-20.02.2020	outbreak until first regime shift (included to understand short burst before actual first point)	128,300
21.02.2020-24.02.2020	short burst to first regime shift (included to understand whether short burst is different to actual first point of regime shift)	9,753
25.02.2020-23.03.2020	first regime shift to second regime shift (included to understand what happened before second point of regime shift)	75,889
24.03.2020-	after second regime shift (included to understand what happened after second point of regime shift)	240,853

Table 2.5: time span and the number of tweets for each period

Second, we pre-processed that text data from each period by removing standard stop words that were listed in stopwords-iso Japanese, Twitter-specific stop words ("retweet", "RT", "tweet", "post", "follow", and "https"), numbers, and words that are related a retweet contests ("apply", "campaign", "retweet contests", "present", and "win a retweet contest") that took place at the time of data collection. We removed Hiragana because it was mainly used for postpositional particles in Japanese texts and therefore did not provide additional meaning for analysing topics (Catalinac and Watanabe, 2019). The pre-processing was conducted partly manually, partly with the quanteda subpackage for stopwords (Benoit et al., 2018) – using R, and the International Components for Unicode (ICU) dictionary for tokenisation. For the sake of an English publication, we translated all text with the DeepL, using its API for R. Furthermore, we eliminated information about word order in the resulting corpora since the wordcloud and the Structured Topic Model (STM) we used in the next steps were based on bag-of-words design.

Third, we analysed what Twitter users 'tweeted' about in each period using wordclouds, which was based on simple calculation of word frequencies from a random sample of 4,000 words in each corpus to control against sample size bias and reduce computational complexity. We restricted those wordclouds to the top 150 words, to make them more readable. Since wordclouds represented text content in this shallow way, the before-mentioned pre-processing helped extracting more meaning from them. For generating the wordclounds themselves, we used the quanteda textplots package in R. We then manually analysed and interpreted the content of each wordcloud. Those wordclouds are depicted in our OSF repository. People 'tweeted' increasingly about COVID-19-related topics, and the restrictions for daily life, that came with it. Words indicative of lifestyle, fun, and music faded, whereas survival and emergency-related words emerged, indicating that the population entered an emergency mode. As discussed before, this maybe was indicative that the most conscientious people focused increasingly on cooperation against an outer threat, which demands extraversion and agreeableness. In line with that hypothesis, it was indicative that People used more often Katakana terms such as 'COVID-19', 'masks', and 'news'. It is used for both technical terms, but also for non-Japanese words, indicating that the virus was associated with something hostile that came from abroad. This led the authors to the conclusion that the points of regime-shift indicated clear shifts in both narrative and displayed personality, based on extreme exogenous conditions.

Fourth, we analysed the pre-processed Tweets, merging-together all periods into one clean corpus for creating a STM, which is the more progressed version of a Latent Dirichlet Allocation (LDA). The goal behind this step was to better understand what topics were discussed and how these topics developed over time, capturing higher dimensionality of text than wordclouds can do. For this, we used the stm package in R (Roberts, Stewart, and Tingley, 2019b). Previous studies (Blei and Lafferty, 2009; Griffiths and Steyvers, 2004) pointed out that the results of STM analyses were variable when the number of topics was changed. Therefore, we used the searchK algorithm from the stm package to determine the optimum number of topics, which was 28 for the subsequent analysis. The detailed outcomes of this analysis are depicted in our OSF repository.

Overlapping with the results from the wordclouds, lifestyle topics dominated, one economic/ everyday life issues topic was constantly present, and so was one COVID-19-related topic. This mainly covered the pandemic itself, the restrictions it brought to the daily lives of people, as well as the resilience those evoked, forcing them to switch to a survival or endurance-mode. Interestingly, no topics covered the Olympics, nor holiday-related aspects, nor the government.

To further understand how the proportion of topics changed over time, we plotted the expected topic proportion of two select topics over the six periods from table 2.5: the COVID-19-related topic, and the topic that covers economic/ everyday concerns. Also, while the COVID-19-related topic was represented in the wordclouds as well, the topic covering commercial/ everyday concerns only became visible in the stm, which indicates that it is representative of 'baseline' of everyday life, and therefore is a good comparison to the emerging topic from the pandemic.

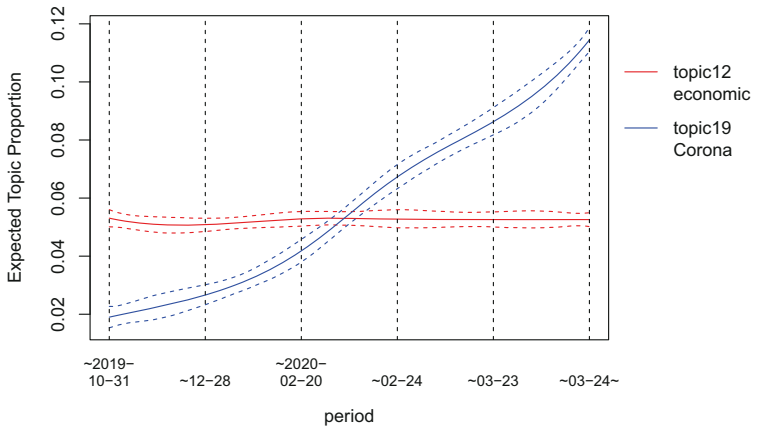


FIGURE 2.19: Change in topic estimation by STM

Figure 2.19 confirms that the proportion of the COVID-19-related topic 19 increased steadily over all six periods, whereas the 'baseline' economic topic 12 remained constant – no matter whether the government requested the citizens to stay at home and shorten the business hours, phases of states of emergency, or 'free' periods in between. We assumed that this was because COVID-19 was such an important issue that permeated all life spaces, and brought so many restrictions and risks, that it was hard to avoid. In summary, as compared to 'baseline' topics, COVID-19-related topics became over time more all-encompassing, intangible, yet permeating all areas of life.

Results from IBM Watson

Top 25 % Openness from 2019:

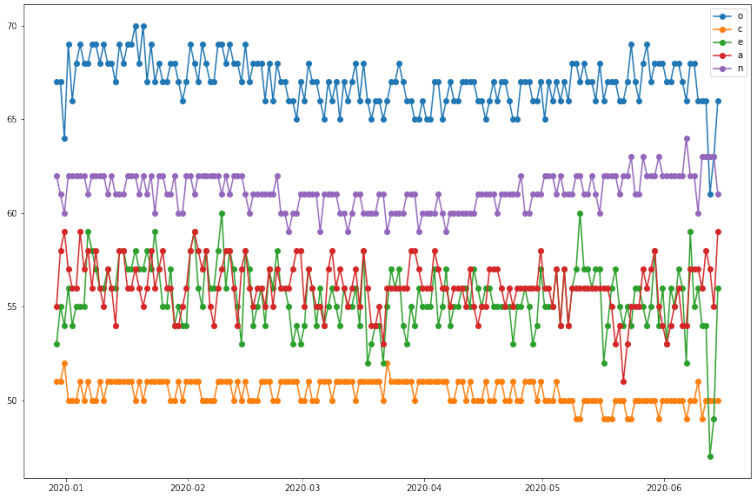


FIGURE 2.20: Top 25 % Openness from 2019

In the top 25% Openness group from 2019, most FFM expressions are basically flat and seem to follow a stationary fluctuation. The most distinct trend is a lower level of Neuroticism and Openness between March and mid-May, yet this general trend lacks sharp peaks. Extraversion and Agreeableness display a more stable trend, yet also a broader range. Conscientiousness scores lowest around the 50^{th} percentile, Extraversion and Agreeableness cluster between the 53^{rd} and 59^{th} percentile, Neuroticism fluctuates around the 63^{rd}, and Openness between the 66^{th} and 69^{th} percentile.

Top 25 % Conscientiousness from 2019:

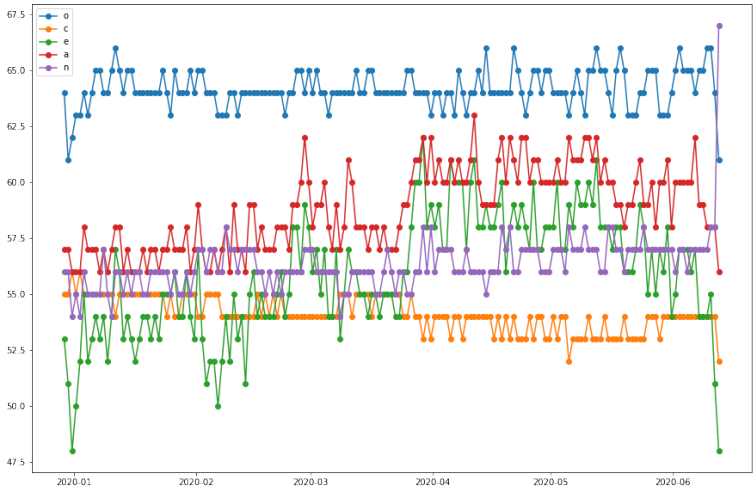


FIGURE 2.21: Top 25 % Conscientiousness from 2019

In the top 25% Conscientiousness group from 2019, clear spikes are visible in Extraversion and Agreeableness in late February and late March; both displaying a clear regime-shift. Neuroticism seems to be slightly elevated, as well. Extraversion and Agreeableness display a broader range, while Openness and Conscientiousness

display a broader range mainly towards May and June but else seem to follow a stable trend. Openness Clusters fluctuates around the 63^{rd} percentile, whereas Conscientiousness around the 54^{th} percentile. Extraversion and Agreeableness display a clear curve, however seem to be clustered together. At first, they fluctuate between the 51^{st} and 56^{th} percentile, then between the 57^{th} and 63^{rd} percentile, displaying a clear regime-shift. Furthermore, Neuroticism slowly moves up from fluctuating between the 55^{th} and 57^{th} at the beginning towards the 56^{th} to 58^{th} percentile at the time of regime-shift. It appears that Extraversion and Agreeableness form a clear regime shift, whereas Neuroticism just a change in general trend. Furthermore, it is important that the only curves that cross their tendencies are Neuroticism and Extraversion, from before the regime-shift Neuroticism scoring higher, to after the regime-shift, Extraversion scoring lower. This is could be indicative towards an underlying phenomenon, however, given the rather strong overall fluctuation of Extraversion, could also be a measurement error.

Top 25 % Extraversion from 2019:

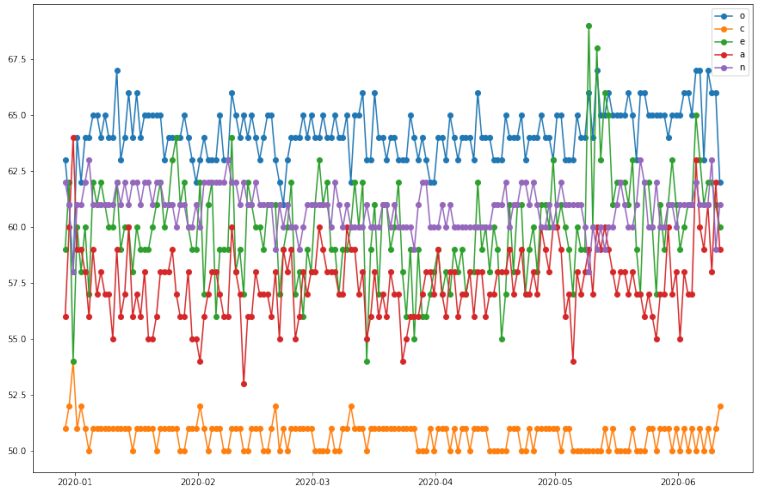


FIGURE 2.22: Top 25 % Extraversion from 2019

In the top 25% Extraversion group from 2019 displays an overall flat tendency, following a stationary fluctuation, however Extraversion displays a very broad range that is broken by some spikes, which seem to not indicate a trend change, though. Extraversion and Agreeableness display the largest range, Openness and Neuroticism medium range, and Conscientiousness the smallest range. Also, Conscientiousness scores much lower, slightly around the 50^{th} percentile whereas all other FFM factors appear clustered mainly between the 55^{th} and 65^{th} percentile. Despite being a logical conclusion, it is unfortunately not clear from the IBM Watson Personality Insights Technical Manual, whether the underlying distribution is normally distributed. Therefore, the meaning between this observable percentile distance is rather unclear.

Top 25 % Agreeableness from 2019:

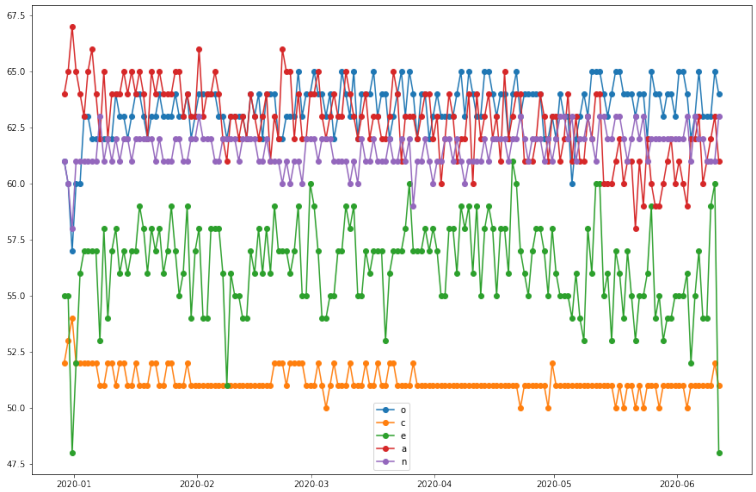


FIGURE 2.23: Top 25 % Agreeableness from 2019

In the top 25% Agreeableness group from 2019 displays an overall flat tendency, following a stationary fluctuation, with a very slight downward-trend in Agreeableness. Conscientiousness is the only FFM factor with a rather small to constant range. Agreeableness, Openness, and Neuroticism cluster together between 60^{th} and 65^{th} percentile, whereas Extraversion fluctuates around the 57^{th}, and Conscientiousness around the 51^{st} percentile.

Top 25 % Neuroticism from 2019:

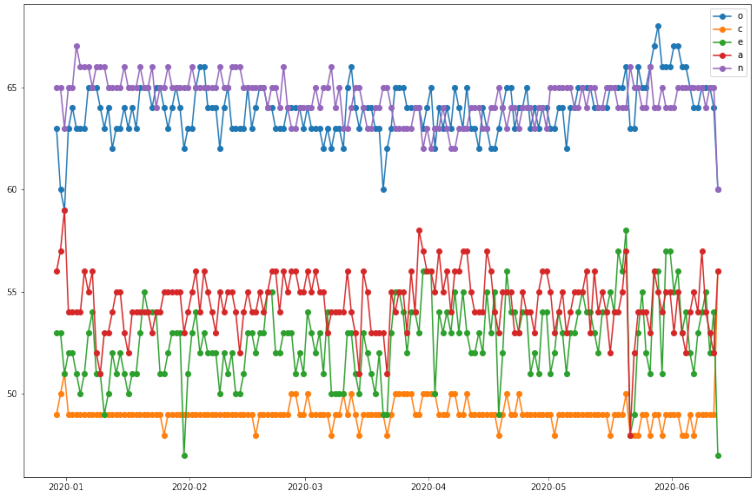


FIGURE 2.24: Top 25 % Neuroticism from 2019

Finally, in the top 25% Neuroticism group from 2019 displays an overall flat tendency, following a stationary fluctuation. Extraversion and Agreeableness display the broadest, Openness and Neuroticism a medium, and Conscientiousness the smallest, almost constant range. While Conscientiousness scores lowest, it still clusters with Extraversion and Agreeableness between the 49^{th} and 57^{th} percentile, whereas Openness and Neuroticism cluster together between the 63^{rd} and 66^{th} percentile.

Autocorrelation and Autoregression

As figure 2.25 shows, the autorcorrelation function (ACF) plot for Neuroticism indicates stationarity, as expected. It degrades with increasing lag quickly to zero and fluctuates around it.

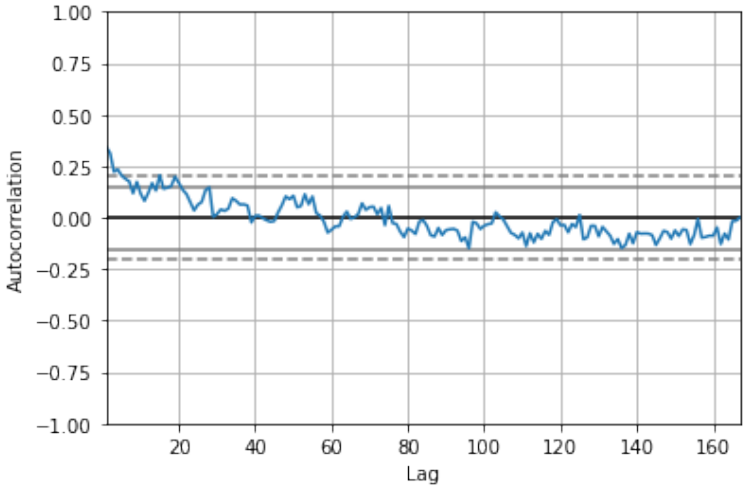


FIGURE 2.25: Autocorrelation Neuroticism

As 2.26 shows, the ACF plot for Extraversion indicates non-stationarity and thus indicates that a regime-shift took place.

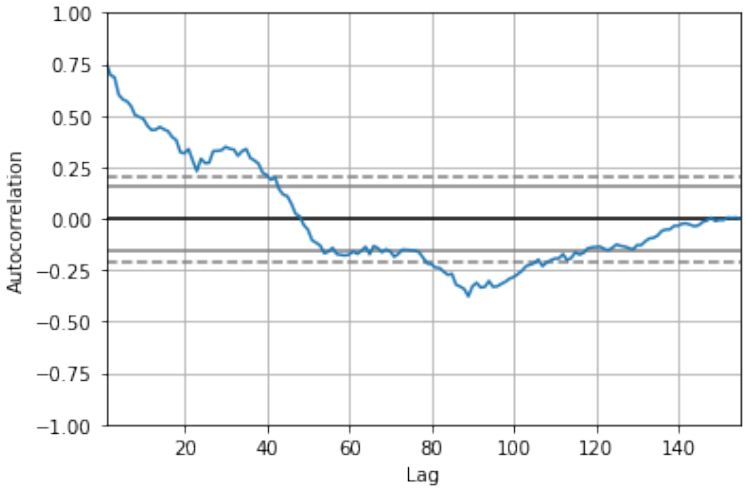


FIGURE 2.26: Autocorrelation Extraversion

As figure 2.27 shows, the ACF plot for Agreeableness indicates non-stationarity and thus indicates that a regime-shift took place.

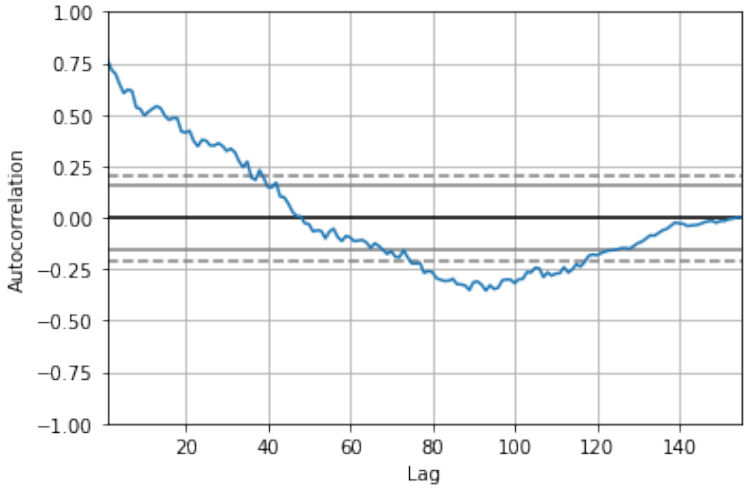


FIGURE 2.27: Autocorrelation Agreeableness

Wordclouds over periods

Figure 2.28 shows displays the events during the periods shown in table 2.5.

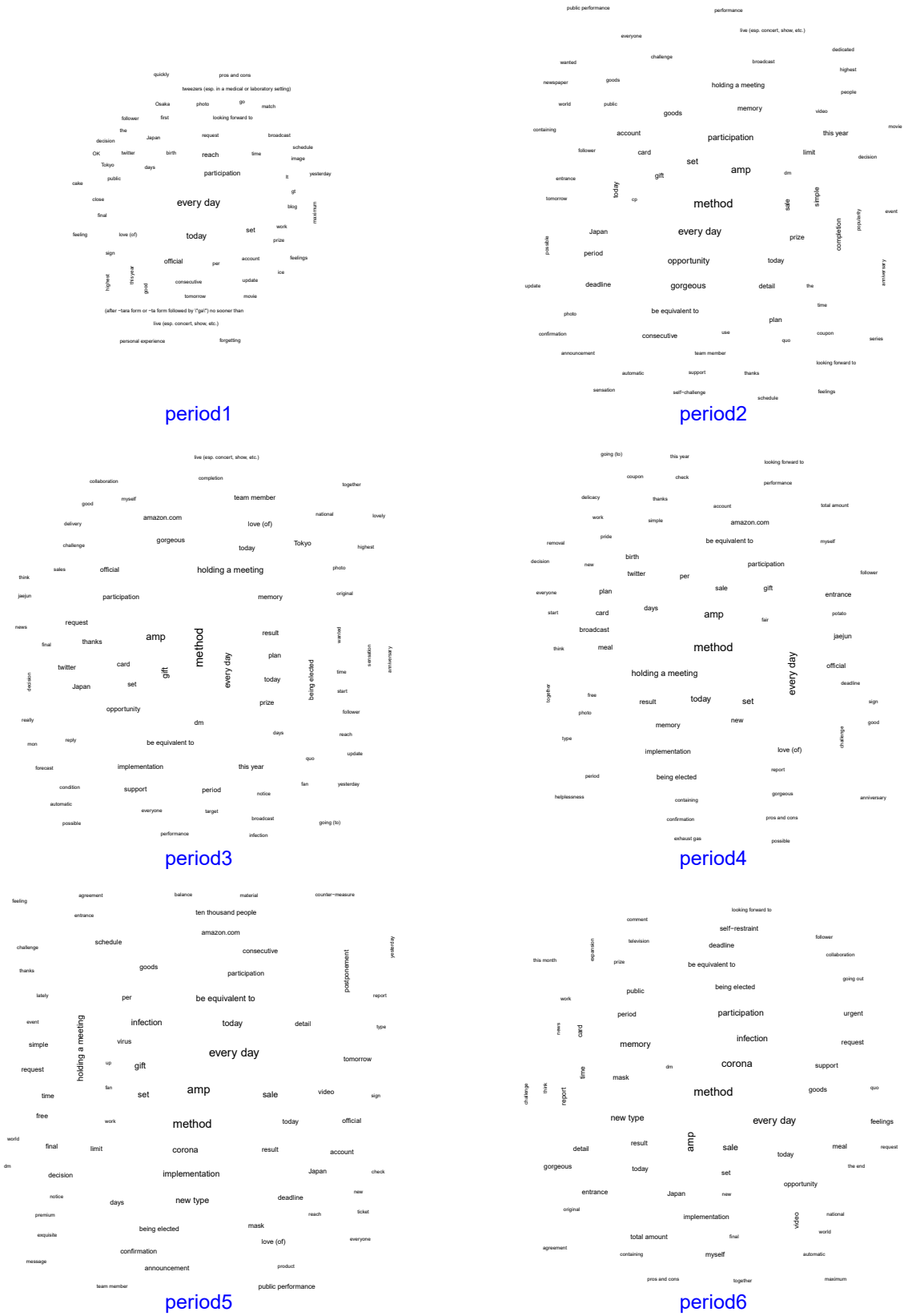


FIGURE 2.28: Variation of the top 150 words over the six periods

In the first period (- to 31.10.2019), the word that was most often used, was 'daily', followed by 'method'. Otherwise, as expected over such a long period, a mix

of words was in use, like 'account', 'coupon', 'win', 'photo', 'to tweet', 'fun', 'product', 'photo', 'follower', and Kim 'Jae-joong', a musician from South Korea, who is famous in Japan. In the second period (01.11.2019 to 28.12.2019), the word that was most often used, was 'method', followed by 'daily'. As expected, season-related words were mostly in use, like 'Christmas', 'gift', 'account', 'Amazon', and 'Nitori', a Japanese chain of IKEA-like stores. Also, a lottery must have taken place, indicated by 'ticket', 'chance', and 'game'. The word 'summer' appeared, and 'Jae-joong' was present, again. In the third period (29.12.2019 to 20.02.2020), the word that was most often used, was 'method', followed by 'daily'. Again, season-related words were used like 'anniversary', 'new-year', 'new-year's present', 'present', and 'Amazon'. The lottery was present, again: 'ticket', 'fun', 'result', 'coupon', 'original', 'card', and 'follower'. Furthermore, people 'tweeted' about 'Tokyo', 'Japan', 'check', and 'start'. Also, 'Jae-joong' was present, as well. In the fourth period (21.02.2020 to 24.02.2020), the word that was most often used, was 'method', followed by 'daily' and 'follow'. However, now 'Corona', 'Virus', and 'test' came up frequently; as did 'proud', 'world', 'involvement', 'every', 'Japan', 'Tokyo', and 'present'. Finally, 'Jae-joong' was there again, with his 'New' 'Single'. In the fifth period (25.02.2020 to 23.03.2020), the word that was most often used, was 'method', followed by 'today', 'amp', and 'Corona'. Further, 'Virus', 'Mask', 'Life', 'Restriction', 'Humans', 'Tokyo', and 'Japan' were 'tweeted' about, clearly indicating the prevalence of the pandemic and its ramifications to daily in the information-field. Also, 'Delayed', 'Plans', and 'Event' occurred more often, indicating the postponement of the Tokyo 2020 Olympics. A small competition was 'tweeted' about, where won could 'win' if one opened an 'account'. People spoke about 'Luxury', and consistently about 'Jae-joong', again. An interesting finding in that period is that Katakana was used more often than in other periods. Katakana is a Japanese writing-system, which is used for technical terms and foreign words. In the sixth period (24.03.2020 to), the word that was most often used, was 'method', followed by 'Corona' and 'today'. It was followed by other pandemic-related words like 'Mask', 'Virus', 'Infection', and 'Body'. The next important group of words were related to the 'counter-measures' from the 'government' of 'Japan' and 'Tokyo' that issued a 'declaration' of a state of 'emergency'. Related to that, people 'tweeted' more about the 'restriction' of life, mutual 'support' and 'cooperation', 'time-frames' of the pandemic, 'planning' of their lives, and about 'meals'. Some other topics were the 'news', 'new models', 'goods', and 'publications'. Remarkably, 'Jae-joong' is missing from the most often tweets. His popularity plummeted after outrage from his fans over an unfelicitous April fool's day controversy, where he posted on his Instagram that he tested positive for COVID-19 due to his 'careless' lifestyle that disregards governmental recommendations.

Ideal number of topics

We used the searchK algorithm from the stm package (Roberts, Stewart, and Tingley, 2019b) to determine the optimum number of topics, which was 28 for the subsequent analysis. This result was the same for held-out likelihood, residuals, semantic coherence, and lower bound of number of topics. Only the semantic coherence suggested an alternative of 29 topics, which was not reflected in the other measures, though.

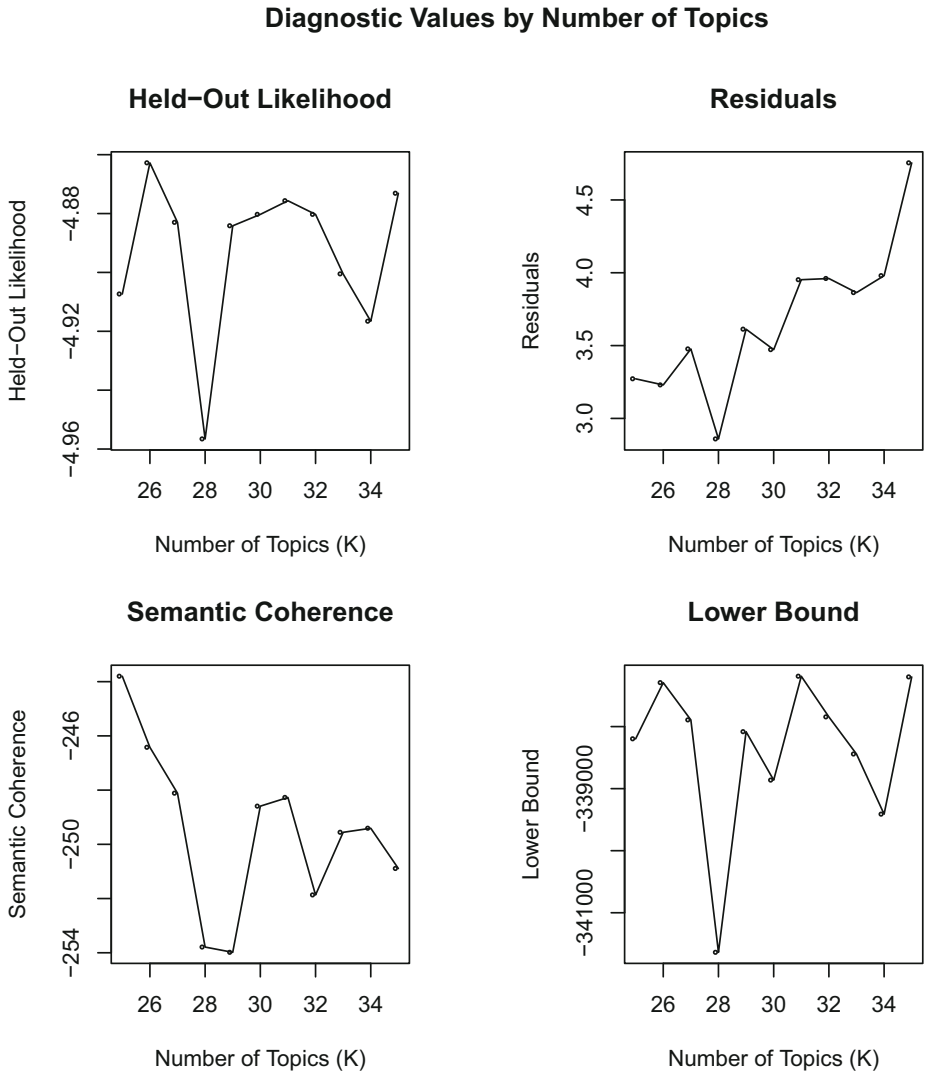


FIGURE 2.29: Results from search K algorithm for optimal number of topics

Top topics in text over time

The results of the applied STM shows the expected proportion of each of these 28 topics, as shown in figure 2.30.

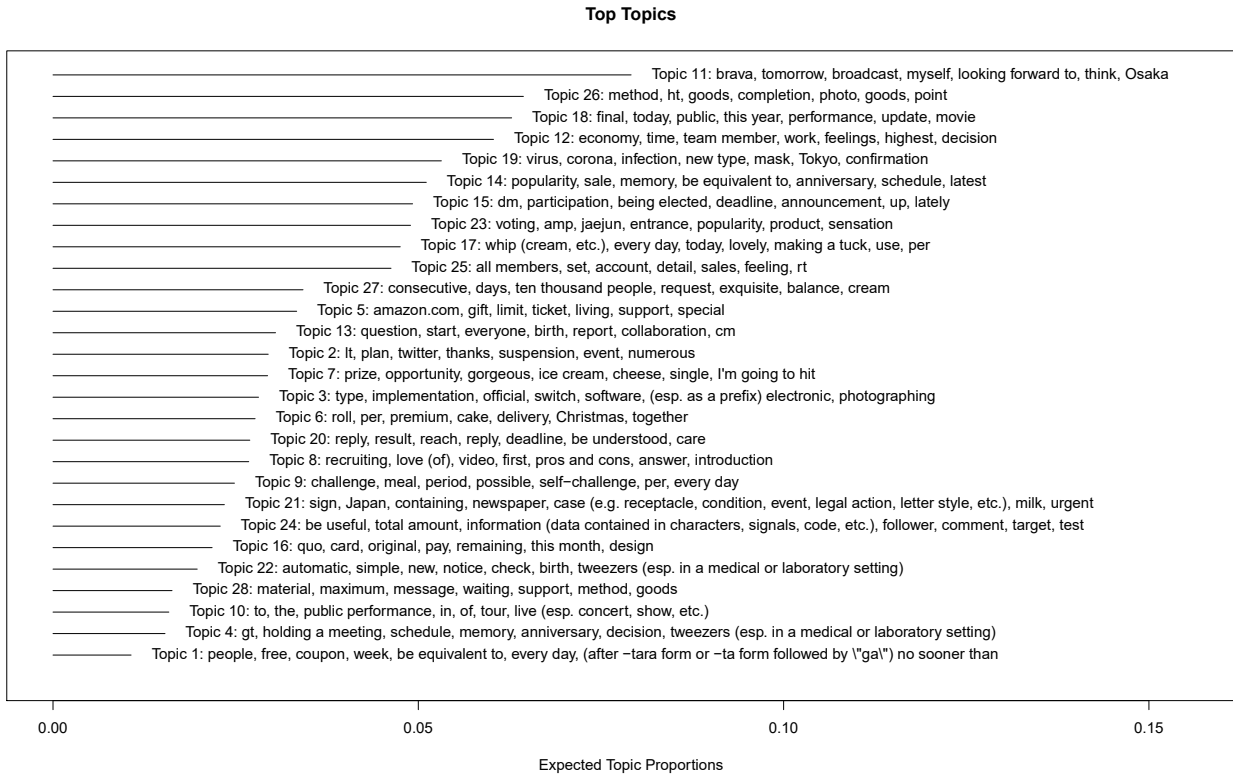


FIGURE 2.30: Top topics and their words

The topics with the highest expected proportion are topics 11, 26, and 18, which cover, like the majority if all other topics, lifestyle-related words like around fun, competitions, photos, programs, and goods. Topic 12 has the fourth-highest expected proportion and sticks out since it is less lifestyle-related and rather represents everyday economic concerns like business, sports and success. The topic with the fifth-highest frequency, topic 19, is clearly COVID-19-related, covering words like 'Virus', 'Corona', 'Infection', 'New', 'Mask', 'Tokyo', and 'Confirmed'. Topic 21 seems to be COVID-19-related as well, covering the changes in circumstances and resilience of the population with words like 'Shield', 'Japan', 'Enter', 'Newspaper', 'Case', 'Milk', and 'Emergency'. Finally, topic 9, may indicate daily challenges of the population and a focus on COVID-19 test results with words like 'Challenges', 'Meals', 'Duration', 'Possibility', 'Hits', and 'Daily'.

2.4.5 Discussion, Limitations, and Future Directions

Since, contrary to other studies, we only observed changes in Agreeableness and Extraversion in the sub-group of top 25% Conscientiousness, while Neuroticism was not significant, we may have found indications that state personality displays elasticity, but presumably also captured cultural, sampling, measurement, or methodological effects. Studies from the United States of America found Extraversion to go up, as well, but also a significant rise in Conscientiousness and Neuroticism, while Agreeableness and Openness decreased (Ahmed et al., 2020). On the other hand, Sutin et al. (2020) just found a slight decrease in Neuroticism, while other FFM factors displayed no significant change. An explanation for these differences maybe was a sampling effect. We focused on the top 25% users who 'tweeted' about COVID-19-related issues, while Sutin et al. (2020) used a stratified sample, and (Ahmed et al., 2020) focused

on health workers - a group that showed also in other cultures a higher vulnerability due to stronger exposure to pandemic-related events (Sugaya et al., 2020).

From a methodological perspective, we may have captured effects from language models that were trained on general data, and not data from extreme events, wherefore the classification of the predicted FFM factors might have been wrong. This would also explain the extreme results in the Ahmed et al. (2020) study, whereby Sutin et al. (2020) found rather moderate changes - but they found changes, which indicates that the notion of slowly personality traits needs to be put to further scrutiny and research.

Furthermore, the deployment of IBM Watson Personality Insights as a tool for research comes with a number of limitations. First, though its origins are in LIWC and the original model used word embedding features with Gaussian Process regression (Ahmed et al., 2020), the current version seemed to be a black box, and no technical details but of some vague prediction performance parameters were published in the technical manual (IBM, 2021). For example, it is not clear whether the English model was just translated, or whether a new model for each language was created. Most importantly though, the results cannot be replicated since the service was deprecated by end of 2021.

Another explanation for varying outcomes in different studies could have been the timing and circumstances of measurement, which indicated different levels of reliability and thereby, generalisability. Ahmed et al. (2020) focused on the time of February to April 2020, whereby they captured everything that happened before and during the first wave of COVID-19. That means that they most certainly have artefacts from seasonality and other exogenous events in their results, as well. So did Sutin et al. (2020), who focused on February and March, whereby we did not know about the circumstances under which the tests were taken. Our time-series approach de-cluttered these effects, and gave – for the first time – insights into daily changes, providing better understanding on the importance of timing of test-taking. However, this came with the cost of ignoring within-person results, since we aggregated daily results.

As strong theories and sound understanding of processes were missing (Bleidorn et al., 2021), future research is needed to put our findings and the usefulness of our method for personality research in perspective. A follow-up study should overcome these problems by deploying the same technique just with different cultures, time frames, and, if possible, user groups. Also, further analyses should be conducted that understand the change of topics and emotions over time, making use of more rich text features, for example emoticons and emojis. Also, grounding studies should be conducted that involve self-reported surveys, geospatial data, or identifiable point of sales activities like purchases. Also, new sources of text would be helpful, for example emails or journals. Lastly, a different personality prediction model should be used that does not depend on IBM Watson Personality Insights nor on any other commercial service that can be deprecated any time.

Finally, our methodology was novel, but not state of the art. We suggest follow-up studies that involve either Deep Learning, Hidden Markov Models, or Monte Carlo techniques for Bayesian Analytics to make use of more data and capture effects that weren't obvious from our current approach. Also, different language models for feature extraction should be deployed, maybe trained anew based on times with and without extreme exogenous conditions, to better control against these conditions.

2.4.6 Conclusion

We introduced a new methodology of time series measurements to personality research, focusing on daily measurements and deploying methodologies from adjacent fields that haven't been deployed to this extent in psychometrics. By using Bayesian analytics, we further created a new approach do identify points of regime-shift within this context.

Results from this methodology indicated that personality measures changed under extreme exogenous conditions during the first wave of COVID-19 and the subsequent societal countermeasures. We interpreted this change in latent psychological traits as a shift from the normal expression z to the emergency expression z_e, which seemed to be temporary, and indicated that personality displayed a degree of elasticity, which goes over and beyond the reported slower changes over a life-type or state/ trait differences, which were partly subject to measurement variance. On a behavioural level, this change in personality can be interpreted as a social signalling and coordination process, as well as a pre-activation for future behavioural changes that are necessary for survival.

If future studies displayed similar results, the idea of state personality expressions might have to be extended as a dynamic function of elasticity, or the ability and function of personality to adaptively regulate perception and behaviours based on contextual embedding. This would be concordant with newer findings from biology that show that DNA displays both slow changes over time (Gorbunova et al., 2007), as well as quicker changes in form of deployment of more genetic potential (Meyers, Ancel, and Lachmann, 2005), depending on outer circumstances.

2.5 Relevance of Time and Space for Psycholinguistic Measures

(this section was written by Peter Romero as main author. Also conceptualisation, formal analysis, team management, and overall project management were done by him. Eisaku Tanaka collected tweets, cleaned and analysed SNS data. Shino Takishita and Rieko Okada helped managing the survey company. Yuki Mikiya helped managing the survey and translating the items. Teruo Nakatsuma supervised the project.)

2.5.1 Introduction

The rapid uptake of vaccines is a critical determinant in controlling pandemic outbreaks, since it affects herd immunity levels and thus can mitigate the spread of viruses like COVID-19 (Forman et al., 2021; MacIntyre, 2015). Beyond availability and direct or indirect measures to assert conformity, willingness or hesitancy of a population to get vaccinated determines uptake speed (Nehal et al., 2021). This willingness or hesitancy in turn is based on individual-psychological, normative-social, and cultural factors, as well as the information field in which an agent is embedded. Cultural factors include culture-specific norms and values that influence health behaviours, including vaccine acceptance (Yoo and Gretzel, 2016). Normative-social factors consist of social norms, beliefs, peer behavior, and the influence of group dynamics on individual decision-making processes, which include vaccination decisions (Betsch et al., 2012). Individual psychological aspects like risk perception (Giancola, Palmiero, and D'Amico, 2023), dark triad (Howard, 2022; Giancola, Palmiero, and D'Amico, 2023), conspiracy beliefs (Douglas et al., 2019; Oortwijn, 2020; Li et al., 2023; Giancola, Palmiero, and D'Amico, 2023), and in particular personality traits,

are associated with health behavior, affecting the responsiveness and compliance to vaccination campaigns (Sherman, Nave, and Funder, 2016). The information field is the collective informational ecosystem from media reports, news broadcasts, governmental campaigns, and social networking services that shapes cognition, affective perception, and behavior towards specific issues. In this information field, narratives battle for attention, and the spread of misinformation or the lack of clear and relevant counter-information can lead to increased vaccine hesitancy, while the dissemination of accurate and nudging, persuasive information can encourage uptake (Brennen et al., 2020; Loomba et al., 2021; Chou, Gaysynsky, and Vanderpool, 2020). Hence, tailored communication strategies that address vaccine hesitancy, promote its acceptance and faster uptake, are essential for effectively mitigating public health crises, and could become at one point as crucial as the development of the vaccines themselves.

The strength of the information field depends on valence and personal relevance of the information, which is, beyond personal connections with the life spaces of individuals, largely determined by geospatial and temporal proximity (Harcup and O'Neill, 2017). The geospatial distribution of psychological phenomena is well researched, however research on temporal distribution of psychological phenomena is sparse (Romero et al., 2021).

While each of these components has been studied in separate, their interaction has not been researched sufficiently. Especially underlying causal mechanisms are unclear, wherefore literature partially contradicts itself. Therefore, deeper and more rigorous research is needed to tailor interventions, enhance the effectiveness of communication in vaccination campaigns, and thus increase uptake, general health awareness, and immunity against misinformation campaigns.

2.5.2 Relevant Work

There exist a plethora of literature that covers geospatial, psychological and temporal aspects of the COVID-19 pandemic or vaccine uptake, and the shape of the information field, however, only a few paper covers such aspects in combination to gather a more complex view of the situation and the interaction of these components.

Neff et al. (2021) offer an excellent overview of 20 years of research literature on vaccine hesitancy in online spaces, and analyse over 100 papers for that. They find that "levels of confidence and hesitancy" towards vaccines "might differ across conditions and vaccines, geographical areas, and platforms, or how they might change over time." (p.1.) and identify gaps for necessary research: focus on disciplinary actions, vaccine specifics, conditions, disease focus, involved stakeholders, implications, methodology, and geographical coverage. While not explicitly, they open the discussion about time-and-space-related issues of vaccine hesitancy. Peters et al. (2023) combine data from self-reported personality traits of 3.5 million people and mobility observations of 29 million people in the United States and Germany to better understand both regional differences, and movement patterns that lead to viral spread. Their results shows that regional compliance behaviour and personality differences, particularly Openness and Neuroticism, significantly influence the early spread of COVID-19, even after adjusting for socio-demographic, economic, and pandemic-related factors, while also revealing variations across countries, over time, and compared to individual-level effects. More concretely, they show that in the early stages of the pandemic, Openness was a risk factor, whereas Neuroticism rather acted as protective influence. However, they find vast differences in terms of country-level Extraversion, temporal-level Openness and individual-level Agreeableness and Conscientiousness. Given the complexity of their findings, they warn about

over-simplification. Also, one of the authors finds in a previous study that Neuroticism may be externally triggered by the predominant narrative in the information field and thus be dependent on more factors of influence (Obschonka et al., 2018). The influence of Neuroticism, Openness, high Agreeableness, as well as dispositional greed on COVID-19-related hoarding behaviours are also shown by Yoshino et al. (2021), who use a similar approach, yet does not cover the information field. Finally, Mangalik et al. (2023) model mental health in the USA through large-scale analysis of 1.2 billion tweets from 2 million geo-located users to estimate changes in anxiety and depression on a granularity of weekly level time-wise and county-level geography-wise. They find moderate to large associations with mental health assessment and survey scores, and suggest this approach to under-resourced communities that however have social media access.

However, these studies lack an overarching, connective framework, which allows scaling and comparison of research findings, and which informs potential avenues for simplification by abstraction. Hence we suggest such a framework, and use it subsequently to simplify measures without giving up rigour or theoretical foundation.

2.5.3 Method

Research Model

To account for geospatial influences, we conclude that more proximal influences are more important for agents than distal ones. In extension, and aligned with systems theory (Willke, 2000), we assume that this measure of proximity is ordered by systemic levels of individual (e.g., person), micro (e.g., family), meso (e.g., company), exo (e.g., industry), and macro system (e.g., society or, in extension, the world). This allows opening a spatial vector of influence $\vec{x} = [\alpha, \beta, \dots \omega]$, whereby α denotes the individual systemic, and ω the macrosystemic level.

To account for temporal influences, we conclude that when more recent events have a stronger effect on agents' perception, cognition, and behaviours, past and future effects are going to have a weaker effect. Thereby, we assume that the current moment t is the point of reference for an agent, and that each agent has backwards memory and, based on historic memory, forward-prediction capabilities of n time steps. For the sake of simplification and *ceteris paribus*, be assume that both directions consist of an equal amount of steps, hence forming nearly identical time intervals into the past and future. Thus at time t, two dynamic event horizons will arise; ϵ_{t-n} and ϵ_{t+n}, that shift with agent time at each time step n.

We define the geospatial information field as a series of n spatial vectors $\vec{x}_z$ at current agent time t with $z \in [t-n \dots t+n]$. Each of these vectors $\vec{x}_z$ represents a spatial influence vector at a time z, that represents systemic influences of varying strengths.

$$\vec{x}_{t-n} = \begin{pmatrix} \omega_{t-n} \\ \beta_{t-n} \\ \vdots \\ \alpha_{t-n} \end{pmatrix}, \vec{x} = \begin{pmatrix} \omega \\ \beta \\ \vdots \\ \alpha \end{pmatrix}, \dots \vec{x}_{t+n} = \begin{pmatrix} \omega_{t+n} \\ \beta_{t+n} \\ \vdots \\ \alpha_{t+n} \end{pmatrix} \tag{2.19}$$

These vectors concatenate into a matrix X, which represents the information field of an agent a at time t, and which represents all informational effects that influence that agent to varying strengths, depending on spatiotemporal proximity.

$$X = \begin{pmatrix} \vec{x}_{t-n} & \dots & \vec{x} & \dots & \vec{x}_{t+n} \end{pmatrix} = \begin{bmatrix} \omega_{t-n} & \dots & \omega & \dots & \omega_{t+n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \alpha_{t-n} & \dots & \alpha & \dots & \alpha_{t+n} \end{bmatrix} \tag{2.20}$$

Figure 2.31 represents this information field in a more intuitive way.

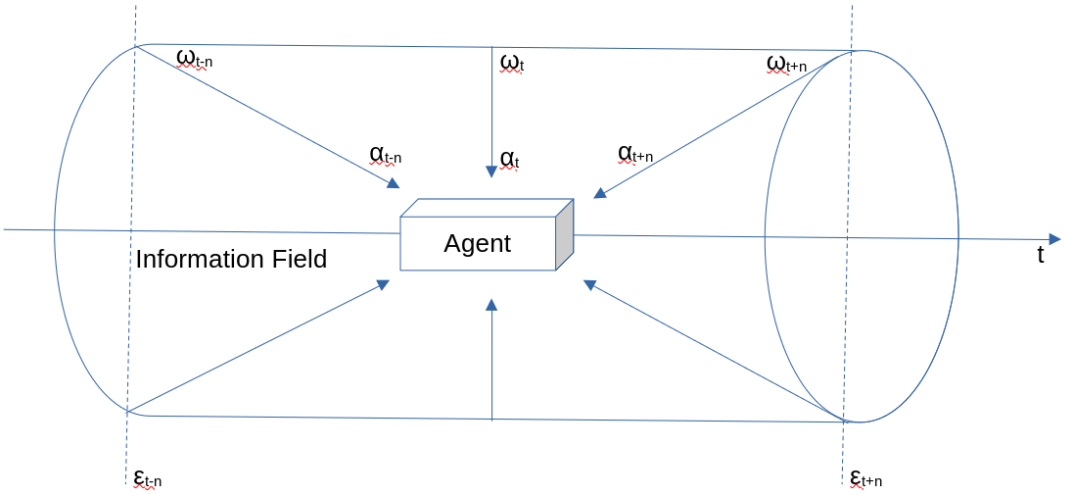


FIGURE 2.31: Research Model: vector-based definition of the information field

While this represents a simplification and leaves quite a few questions unsolved, it is to the best of our knowledge the first attempt to operationalise the spatiotemporal nature of the information field by help of systems theory. Unsolved questions are, for example, whether both event horizons are the same number of time steps n apart from each other, or whether priming from events at the current time t might generate memory effects that influence future prediction by selective memory retrieval or overemphasis of specific memories.

Another question is where exactly the information field attaches to; at a systemic level above the agent, and from there through social norms and peer pressure, or as a sort of collective behaviour that only exists on collective level? However, peer pressure would rather attach on an individual systemic level, and could be counted as a part of the information field through informal communication. But then, it has to be asked, where in the psychological architecture of the individual system this attachment would take place, and which influence the relevant context of an agent has. For example, the information field could moderate the transfer of competency potential from psychological latent traits into behaviours. Or, it could moderate the transfer from relevant behaviours into respective outcomes. Finally, moderation - or even mediation - could take place at both transfer points, depending on the kind and strength of message and contextual embedding. Figure 2.32 depicts these potential moderation mechanisms at the transfer points within the individual-systemic level. For the sake of simplification, we excluded the neurophysiological level that resides underneath the other individualsystemic levels, which serves more fundamental functions.

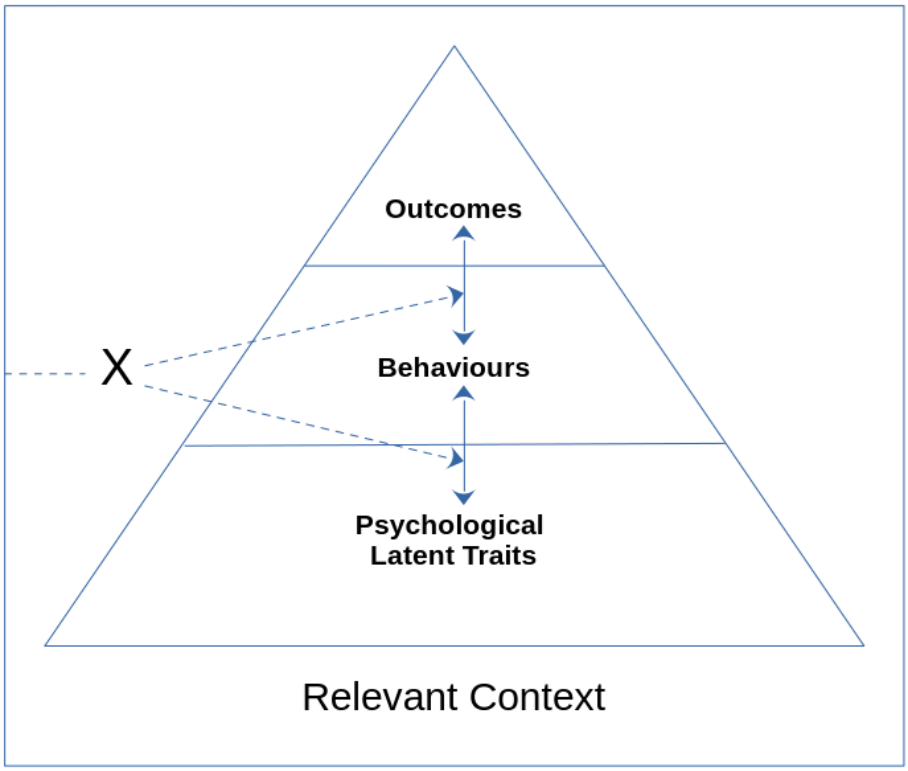


FIGURE 2.32: Agent in contextual embedding

Since each of these questions demands further research and is hardly covered in totality, we assume that all mechanisms could be relevant, hence do not differentiate for individual-systemic architecture. Also, since potential collective behaviours that cannot be measured on individual-systemic level are possible, we conclude that these have to be taken into consideration, as well.

Finally, to operationalise uptake speed, we need to take various variables into account that are of stochastic nature and about which deeper information is partially inaccessible. For example, on-the-ground-truth on fluctuations in the local availability of vaccines, inflexibilities of the medical infrastructure, and local decisions on preferential treatments of age groups is mostly not documented.

Hence, we decide to use the first inflection point i of overall aggregated vaccine uptakes per initial vaccine and booster shots, which follow sigmoid functions with a clear slow growth at the beginning, followed by exponential growth, then an inflection point and a stabilisation at the upper asymptote m. This kind of curve is used in a variety of fields to model natural phenomena, from difficulty of psychological test scores in item response theory (Rust, Kosinski, and Stillwell, 2020) to wildlife growth (Zullinger et al., 1984) to market saturation (Minakov et al., 2017) to modelling future COVID-19 vaccine uptake over time and space (Wood et al., 2022). The advantage of that approach is that we do not need to make any assumption about the functional form over and above its sigmoid nature. Also, the interpretation is intuitive — the lower asymptote m_0 starts at the origin and represents first day of vaccine availability, the slow growth at the beginning can be interpreted as covering a plethora of aforementioned stochastic influences that are undocumented, the subsequent exponential growth can be interpreted as the kernel of the populus that is willing or hesitant towards vaccines, the inflection point i cuts the curve in halves, thus represents the point when other mechanisms take over. The exponential decrease after the inflection point i could be interpreted as stochastic processes during that

change (e.g., logistical challenges), and the slow decrease until the upper asymptote m represents the kernel of the population that takes the vaccine for other reasons like social pressure, logistical challenges, lack of health awareness or sense of urgency. Finally, the upper asymptote m represents the total number of people that accepted the vaccine for either reason or ability to receive it. The inflection point i indicates change in the concavity of the function $\sigma_m(x)$, which is detected by sign-change of its second derivative $\sigma''_m(x)$ and occurs at $x = 0$ irrespective of the value of m.

Thereby, we bound the sigmoid function $\sigma_m(x)$ to the upper asymptote m:

$$\sigma_m(x) = \frac{m}{1 + e^{-x}} \tag{2.21}$$

Its first derivative displays the instantaneous rate of change at each specific point x on $\sigma_m(x)$:

$$\sigma'_m(x) = \frac{d}{dx}\left(\frac{m}{1 + e^{-x}}\right) = m \cdot \frac{e^{-x}}{(1 + e^{-x})^2} \tag{2.22}$$

And its second derivative displays the instantaneous rate of change at each specific point x on $\sigma'_m(x)$, whereby the sign indicates the slope of the tangent line to increase or decrease:

$$\sigma''_m(x) = \frac{d}{dx}\left(m \cdot \frac{e^{-x}}{(1 + e^{-x})^2}\right) \tag{2.23}$$

We denote i_1 as the inflection point of the first vaccine dose, where $x = 0$, which happens at time step n_{i_1}. Hence, we identify the overall uptake vaccine uptake speed v between the first shot and the x^{th} booster shot with:

$$v = n_x - n_1 \tag{2.24}$$

While the first dose and first booster shot of the vaccine were available in short distance to each other, the second booster shot was only available after a significant time gap, as were subsequent booster shots, which were also limited to certain age and high risk groups. We denote each injection with $x \in \mathbb{N}$, thus the first dose as 1, the first booster shot as 2, and so forth. Hence, we interpret that $v = n_2 - n_1$ as having taken place within too short distance from each other to being able to exclude coercion or immediate fear thus obfuscating the true intrinsic vaccine acceptance. However, $v = n_3 - n_1$ took place after sufficient time for habituation passed, hence we use it as proxy for measuring the true intrinsic vaccine acceptance, and denote it as v_{medium}. Finally, $v = n_5 - n_1$, denoted as v_{long} displays the long-term vaccine acceptance; with the restriction of just representing a subgroup, hence it is not directly comparable to v_{medium}. However, both v_{medium} and v_{long} can be compared geographically, thus displaying local differences in vaccine acceptance. For an industrialised country like Japan, we can assume that infrastructural conditions are mostly homogeneous, hence *ceteris paribus*, we can exclude regional economic inequalities.

Hence, we define the outcome variables as:

$$v_{medium} = n_3 - n_1 \tag{2.25}$$
$$v_{long} = n_5 - n_1 \tag{2.26}$$

Research Design

We conducted a non-experiment by collecting behaviour artefacts and survey data at data points before and during the COVID-19 pandemic.

We represent the model the following way: we use tweets to identify the information field that are temporally sorted, analysed for proximity to the agent, and of which we know the approximate geographic location on city level of the sender, using the data set from section 2.3. The temporal sorting is realised through daily tweet collection, the approximate location is realised through identifying and scraping the tweets of followers of hyperlocal entities like police stations, local sports teams, or city mascots – following the hypothesis that nobody else but locals would have a reason to do so. While some authors approximate physical distance by narrative strength (Houghton, Siegel, and Goldsmith, 2013), this approach relies on available commercial tools in a target language, and does not differentiate on a system-theoretic level. Therefore, we identify approximate agent-proximity in a simple and more intuitive way by LIWC categories (Pennebaker et al., 2015b), whereby we hypothesise that LIWC categories like 'Affective processes' represent agent-internal language, 'Family','Friends', 'Home', and 'Perceptual processes' rather indicate language with close agent proximity, whereas 'Work' rather indicates language from more distal systemic layers.

We augment these with local data on COVID-19-related severe cases, hospitalisation, and deaths, as well as vaccination numbers to capture the influence of local norms, peer pressure, or imitation effects. Furthermore, we use personality questionnaires taken from participants all over Japan at four different time steps before and during the pandemic to identify geographic distribution of personality.

To understand difference between agent and contextual properties, we conduct a Japan-wide survey covering COVID-19 related attitudes, about the severity of the situation, abidance by governmental measures, and cooperation with others on on pandemic-related issues. To control against geographic personality, participants in the survey are asked to fill out a personality questionnaire, as well.

Since most of the data is not specifically collected for COVID-19-related research, the temporal granularity of the data points is uneven - ranging from individual points in time (e.g., each survey taken) over monthly data (e.g., certain COVID-19 statistics) to daily data (e.g., tweets, or vaccine doses administered). Hence, we decide to aggregate the data in the predictor space based on waves of COVID-19. Based on media research and official announcements, we identify eight waves, of which five are relevant for the phenomena observed:

1. <16.01.20 – 0 baseline
2. 16.01.20 - 26.03.20 – 1st wave
3. 26.03.20 - 30.06.20 – 2nd wave
4. 01.07.20 - 31.07.20 – 3rd wave
5. 01.08.20 - 24.09.20 – 4th wave

6. 24.09.20 - 31.12.20 – 5th wave
7. 2021 – 6th wave (not relevant for the vaccine-related observations)
8. 2022 – 7th 8th wave (continuing until 2023; not relevant for the vaccine-related observations)

Since the coarsest geographic differentiation is prefecture level, we furthermore aggregate all geographic data on this level. Therefore, the level of analysis is by wave and prefecture.

Finally, we define the outcome space o as duration in days with $o \in \mathbb{N}$, with $(o \in v_{long} \wedge o \notin v_{medium}) \vee (o \notin v_{long} \wedge o \in v_{medium})$, thus creating quasi-continuous outcome measures.

Data Set

We use the following data sets in our analysis to represent the different aspects of the framework:

- Tweets: daily hyperlocalised tweets from before and during the pandemic (daily granularity of all 60 LIWC features)
- COVID-19 Data: official local vaccination numbers, death numbers, hospitalisation numbers (daily granularity of each number)
- Ground truth data: geospatially distributed personality questionnaires at three different times before COVID-19 (used as constant of five Big Five values)
- Questionnaire Data: covering demographics, psychological questions, economic questions, and attitudes towards the COVID-19 situation and governmental measures, as well as willingness to abide by those and cooperate with others on pandemic-mitigation-related issues taken all over Japan (used as constant for during the pandemic)
- Questionnaire Personality: Personality tests of these survey participants (used as constant during the pandemic; being comprised of 60 facets and 5 Big Five scores)

Twitter Sample

We use the same data set comprised of hyperlocal tweets as in section 2.3. Given our findings on the questionable reliability of predicted personality on spatial level from that chapter, but the relevant outcomes regarding temporal event-identification, of which we do not know whether these represent personality change, measurement error, or linguistic variation not being captured in the model, we decide to use LIWC scores created from the selected data.

The dataset is comprised of hyperlocalised tweets generated from January 1st, 2019 to April 1st, 2021 from at least two cities of all 47 Japanese prefectures. It is comprised of 25,614,106 (SD = 44,924.94) tweets, with on average 189,734 extracted tweets from every city. The cities were chosen based on parameters like population size, but also spatial separation, based on official numbers (Statistics Bureau, Ministries, and Agencies, 2021). Due to the size of Hokkaido with its sub-prefectures, at least two cities per sub-prefecture are chosen. Given the metropolitan status of the Tokyo as one of the largest urban centres on the planet, those cities and special

wards with the most population and spatial separation are carefully selected. This results in 1,648 accounts, on average 35 per prefecture, with a maximum of Hokkaido with 235 and a minimum of Kumamoto-ken with 11 Twitter accounts. Excluding Hokkaido, the average number of accounts per prefecture is 30. The minimum number of tweets for a city is 70,425 tweets, and maximum number is 244,331 tweets. All tweets are harvested from 107,873 followers of 1,648 local city representative accounts like police stations and city mascots. On average, 799 (SD=46.16) follower accounts are harvested for each city; the minimum number of accounts for a city is 596 and the maximum is 822.

After removal of language features like retweet identifiers and emojis, this data subsequently is analysed with Linguistic Inquiry Word Count (LIWC) (Pennebaker et al., 2015b) to extract theory-driven, dictionary-based, hard-coded features. In specific, the 2015 version of LIWC (Pennebaker, 2015) and the Japanese dictionary and tokenisation method introduced by Igarashi, Okuda, and Sasahara (2021), is used, text is preprocessed by their latest Japanese psycholinguistic tokenisation dictionary, the MeCab/IPADIC (Kudo, 2005) python library, and their morphological analysis (word segmentation) and part of speech analysis (POS) code. This results in 60 category-by-category features based on word frequency analysis, featuring words and word stems, including standard language categories like pronouns, to psychological processes like emotions, and six sub-scores: insight, causation, discrepancy, tentativeness, certainty, and differentiation (Pennebaker et al., 2015b). Finally, all tweets from the same city are treated as one document, and daily LIWC results per prefecture are aggregated.

Survey

In January 2022, we conduct a survey to deeper understand regional connection between agent personality, agent cognitive and affective aspects about COVID-19, agent congruency with government and science, agent social synergy, and agent synergy with the narrative, thus both about the information field, direct systemic embedding, and about agent-internal aspects.

This survey is comprised of demographic questions (age, gender, income, household income, number of children, family status), psychological questions (number of siblings, sibling order), economic questions (household income, income, times eating out per week, profession, and COVID-19-related questions. The specific COVID-19-related questions are:

- “The measures of the government are justified” - cognitive item to capture the level of perceived justification of governmental measures and implicitly the level of congruence of participants with governmental measures
- “I believe in vaccines” - cognitive item to capture the degree of trust in science of a participant and implicitly the degree of acceptance of governmental narrative
- “The COVID-19 situation is dangerous” - affective item to capture the emotional sense of danger as well as congruence with official health narrative
- “I cooperate with those around me to deal with the pandemic” - social-behavioural item to capture horizontal synergy of participants and implicitly abidance with norms of the direct contextual embedding
- “I abide by governmental measures” - behavioural item to capture the vertical synergy of participants and implicitly abidance and congruence with broader societal and cultural norms

Overall, 6,266 (prefecture mean = 133.32; SD = 114.92, min = 35.00 max = 564.00) persons participate in the survey, of which on average 48% (SD = 3%, min = 42%, max = 55%) per prefecture are male. The average age over prefectures is 45.56 (SD = 1.67; min = 41.71, max = 51.36).

Furthermore, we ask each agent to fill-out the NEO-FFI, a high quality, well established personality questionnaire. Enabling comprehensive insights into human personality, the NEO-FFI is a seminal instrument of psychometrics that is well-documented, developed on sound science, and has been a stable in countless international studies since the 1980s. It uses a five-point Likert scale and offers broad applicability in different use cases including professional assessment, clinical psychology, and research. For example, it is deployed in the assessment of personality disorders and for deriving optimal treatment strategies (Costa and McCrae, 2008). In research, it is used to study associations of personality with various psychological and behavioural phenomena like the influence of personality traits on life outcomes (McCrae et al., 2004; McCrae and Costa, 1987). It is based on the Five-Factor Model of personality (Costa and McCrae, 1985), the Big Five Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. The development of the NEO-FFI is a major advancement for personality psychology, since it provides a robust, valid, and reliable of the Big Five model (Costa and McCrae, 1992), with meticulously researched facets. It is comprised of 60 items among the five latent traits of the Big Five model, each composed of further six 6 facets; asking two items per facet, of which about the half are reversely scored for being able to detect attention, cheating, and overall answering consistency. Despite being proprietary, we deploy the Japanese verion of the NEO-FFI for our study, since it offers a robust framework, cross-cultural validity, well proven psychometric properties, and still a decent answering time (10-15 min; which is much less than 45-60 min for the NEO-PI-R (Costa and McCrae, 2008)) that allows its co-deployment with other surveys or questionnaires.

Ground truth data

For Ground truth, we use the open-sourced geospatial personality distribution data collected from Yoshino and Oshio (2021a), who use the Japanese version (Oshio, Abe, and Cutrone, 2012) of the the Ten Item Personality Inventory (TIPI) (Gosling, Rentfrow, and Swann, 2003), a well-established psychometric instrument that exists in 27 languages and is used in 9,167 peer-reviewed papers, and which we deploy in chapter 1 to measure the synthesised personality of GPT-3. Meant for mass-deployment and large-scale studies, it uses a seven-point Likert scale, is comprised of only ten items; two per Big Five factor, of which one is reversed. "Although somewhat inferior to standard multi-item instruments" (p.504) Gosling, Rentfrow, and Swann, 2003, its outcomes for self-ratings, external ratings, and peer ratings vastly overlap with other established Big Five instruments, it displays high congruence between self-ratings and observer ratings, its test-retest reliability is high, and the levels of external correlates are on par with other studies reported in research.

Data is collected in three iterations; first between January and March 2012 (n = 4469, prefecture mean 95.09; SD = 85.95, min = 14.00, max = 388.00, 46% male; SD = 6%, min = 25% and max = 58%), the second iteration in January 2017 (n = 5619, prefecture mean 119.55; SD = 13.99, min 87.00, max 149.00, 60% male; SD = 5%, min = 50 %, and max = 71%), and the last iteration was in January 2019 (n = 4330, prefecture mean = 92.13; SD = 14.34, min = 58.00, max = 127.00, 66% male; SD = 6%, min = 53%, and max = 80%), and overall n = 14418, prefecture mean = 306.77;

SD = 101.63, min = 161.00 max = 648.00, 57 % male; SD = 4%, min = 46%, and max max = 65%).

COVID-19-related data

We use the same official COVID-19 data as in section 2.4 provided by the Japanese government (MHLW, 2021) that we cross-checked with data from the World Health Organisation (WHO, 2021) to ensure their correctness. Our main focus lies on severe cases/ hospitalisations, and death numbers, since these not only have a significant economic influence and detrimental impact on the health system, but foremost since these are psychological markers that effect the perception, cognition, emotion, and behaviour of people.

Analysis

Cleaning is not necessary, since data already prepared from previous studies on space in section 2.3 and time in section 2.4, however, to represent the waves of COVID-19 in Japan explicated above, data is aggregated by these time steps. Since most data is reported by official statistics on prefecture level, granulated city data is missing. Also psychological data is taken in certain time-steps, wherefore no time series model can be used. Unfortunately, this results in an unbalanced sample, with only 47 prefectures as cases, and 711 final features as predictors. Hence, application of normal machine learning methods would immediately burn all degrees of freedom, and yield no results. On the other hand, there are too many predictors to do serious theory-driven predictor selection. Therefore, we use a semi-manual method that simulates feature selection and dimensionality reduction by chaining traditional steps of statistics and statistical learning, which we designate as *statistical feature reduction*. Figure 2.33 depicts this approach.

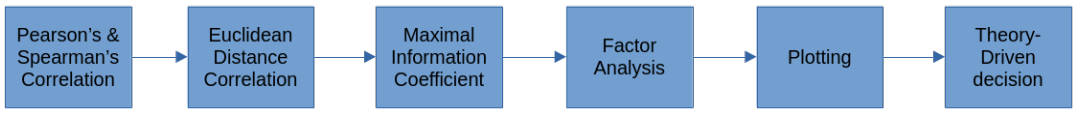


FIGURE 2.33: Analysis flow for semi-manual *statistical feature reduction*

Concretely, we first pre-select features based on their broad association with the outcome measure by Pearson's r and Spearman's ρ. Then, we further test the association by Euclidean Distance Correlation and the Maximal Information Coefficient. Next, we reduce dimensionality by factor analysis, which is as vulnerable to unbalanced samples as other machine learning techniques, but captures latent traits and therefore can be more powerful in a selection task than a principal component analysis (PCA) (Fávero, Belfiore, and Souza, 2023), especially with psychological latent traits that are known to be intercorrelated and best to be explored with non-orthogonal rotations like "oblimin" (Rust, Kosinski, and Stillwell, 2020). Subsequently, we plot the selected predictors against the outcomes to make an informed decision on the functional form. Finally, we select all predictors that occur in more than one selection method, and such that are aligned with theory, and create various manual iterations of regression models, until we find the one that yields the highest significance and R^2.

Step 1 - Correlation

In essence, Pearson's r is the preferred choice for analysing data with light-tailed distributions and linear relationships but requires normally distributed, continuous

data, whereas Spearman's ρ is more flexible, and better suited for assessing monotonic relationships, accepts non-parametric data, including ordinal variables, and can be deployed to assess non-linear relationships, or when the strict assumptions of Pearson's r are violated. It is ideal for heavy-tailed distributions, or in cases where outliers are prevalent, a common scenario in psychological studies (Winter, Gosling, and Potter, 2016). The occurrence of disparate results between these two methods, under the assumption of an identical underlying Gaussian distribution, typically signals potential issues with the data, most likely affecting the predictive outcomes. On the other hand, including both provides insights about the association between outcome and features and thus provides information about the best functional form at the same time.

Step 2 - Euclidean Distance Correlation

The Euclidean distance correlation serves as a statistical metric quantifying the dependence between two variable sets, denoted as X and Y. This measure effectively captures both linear and nonlinear relationships. It is grounded in the principle of distance covariance, which in turn expands the classic concept of covariance to accommodate multivariate and nonlinear scenarios.

For any two random vectors $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$, the Euclidean distance correlation, symbolized as $\mathcal{R}(X,Y)$, is defined by the following equation:

$$\mathcal{R}(X,Y) = \frac{\mathcal{V}^2(X,Y)}{\sqrt{\mathcal{V}^2(X) \times \mathcal{V}^2(Y)}} \tag{2.27}$$

Here, $\mathcal{V}^2(X,Y)$ represents the distance covariance between X and Y, while $\mathcal{V}^2(X)$ and $\mathcal{V}^2(Y)$ are the respective distance variances of X and Y.

The computation of distance covariance, $\mathcal{V}^2(X,Y)$, is given by:

$$\mathcal{V}^2(X,Y) = \frac{1}{n^2} \sum_{k,l=1}^{n} A_{kl} \cdot B_{kl} \tag{2.28}$$

In this context, A_{kl} and B_{kl} denote double-centered matrices derived from the pairwise Euclidean distances among the elements of X and Y, and n signifies the total number of samples.

The Euclidean distance correlation is constrained between 0 and 1, where $\mathcal{R}(X,Y) = 0$ implies a state of independence (subject to certain conditions) between X and Y, and $\mathcal{R}(X,Y) = 1$ reflects a perfect functional correspondence (Székely, Rizzo, and Bakirov, 2007).

Step 3 - Maximal Information Coefficient

The Maximal Information Coefficient (MIC), as introduced by Reshef et al. (2011), serves as a robust measure for quantifying the strength of the most pronounced linear or nonlinear associations between two variables within a dataset. Rooted in information theory, and diverging Pearson's r, which is limited to linear correlations, MIC excels in identifying a broad spectrum of associations, better encompassing nonlinear relationships than Spearman's ρ.

MIC strives to unveil any underlying patterns in data by aligning the greatest mutual information value with a grid-like partitioning on the x-y plane. For any given variables X and Y, the formal definition of MIC is as follows:

$$MIC(X,Y) = \max_{xy} \left(\frac{I(X,Y)}{\log(\min(x,y))} \right) \tag{2.29}$$

In this equation, $I(X,Y)$ denotes the mutual information shared between X and Y, with the maximization process spanning over the number of bins x and y utilised in segmenting the dataset.

MIC is constrained between 0 and 1, and finds extensive applications across diverse scientific domains, including bioinformatics, neuroscience, and environmental sciences, playing a pivotal role in revealing complex relationships within substantial datasets (Kinney and Atwal, 2014). For example, Chauhan and Choi (2023) use it to classify Alzheimer's Disease, Lazarsfeld, Johnson, and Adéníran (2022) for ensuring differential privacy, and just like us, Zhou et al. (2022) for feature selection.

Step 4 - Factor Analysis

Factor Analysis, a widely-utilised statistical method, aims to uncover latent variables, or factors, that elucidate the correlation patterns among a collection of observed variables. This technique simplifies the complexity of observed variables into a smaller number of unobserved variables, leveraging their correlations. The fundamental premise of this method is the direct correlation of each observed variable with any of the factors (Mathai, 2021).

Considering $X = (X_1, X_2, \ldots, X_n)$ as a vector representing observed variables, Factor Analysis formulates X as:

$$X = \mu + \quad F + \epsilon \tag{2.30}$$

where:

- μ represents the vector of means.
- is a matrix detailing the factor loadings on the variables.
- F comprises the vector of common factors.
- ϵ corresponds to the vector of unique factors, or error terms.

The primary objective of Factor Analysis is to ascertain the factor loadings that optimally account for the observed correlations in the dataset.

It is widely used in research, e.g., to deduce the covariance structure from diverse data sources (Chandra, Dunson, and Xu, 2023), for investigating temporal and spatial variations in patterns (Stegle and Rohan, 2022) as we do in this paper, or to reduce item in the construction of personality questionnaires (Rust, Kosinski, and Stillwell, 2020). In psychometrics, and behavioural economics, many latent traits tend to be intercorrelated, which can be captured by non-orthogonal rotations in factor analysis, like "oblimin" or "promax".

In **summary**, each step of this process is increasingly able to capture non-linear, higher dimensional aspects of the feature-outcome associations. However, in **Step 5 - Plotting** and **Step 6 - Theory-Driven decision-making**, we inject human perspective, expert-knowledge, and theory into the process again. In many ways, this even captures higher complexity, since it allows to step away from a pure data-driven process, and align all steps with both research perspective, and strategic alignment.

Software Used

Data manipulations have been conducted with Python 3.8.9 (Python Software Foundation, 2023), Pandas 2.1.3 (team, 2020), and calculations have been conducted in SciPy 1.11.4 (Virtanen et al., 2020), numpy 1.26.0 (Harris et al., 2020), and statsmodels 0.14.0 (Seabold and Perktold, 2010). All graphs have been plotted with Matplotlib

3.8.2 (Hunter et al., 2020), GeoPandas 0.14.1 (Bossche et al., 2020), and seaborn (Waskom, 2021).

2.5.4 Results

Outcome Measure

As discussed above, outcome measures are:

1. First phase = difference in days between the inflection point of vaccine uptake of the first and the third dose of the vaccine to represent the general psychological readiness of the population to take the vaccine (intrinsic motivation)

2. Second phase = difference in days between the inflection point of vaccine uptake of the first and the fifth dose of the vaccine to represent the effect of secondary measures of vaccine uptake in the population take the vaccine (e.g., persuasion/ extrinsic motivation)

3. Agent readiness = survey answers to the question "I abide by governmental measures", since that encompasses getting a vaccine.

As expected, all vaccine uptake curves have a sigmoid shape with clear inflection points. The following curve is a random example to illustrate population behaviour.

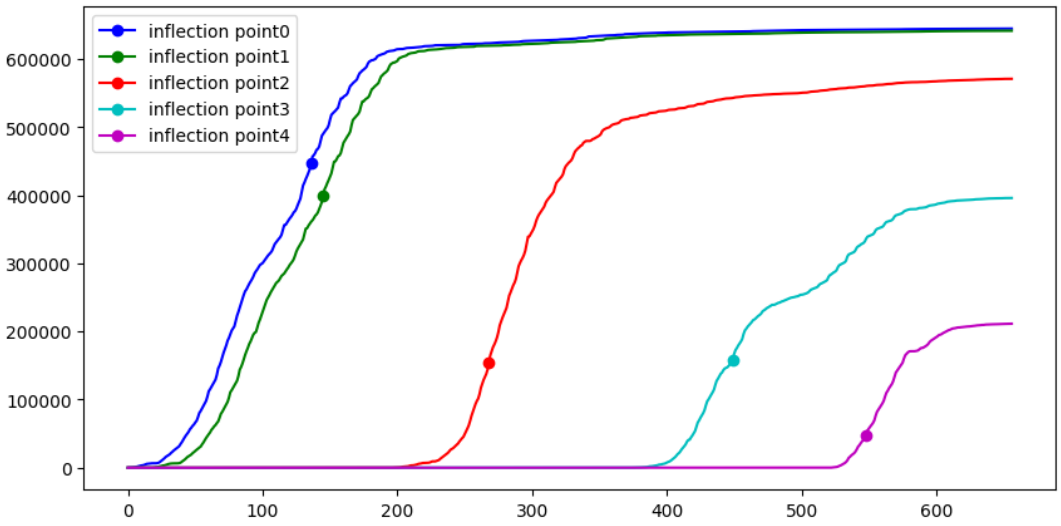


FIGURE 2.34: Inflection points of vaccine uptake curves Okinawa

Also, in accordance with theory, there are clear differences in the time between the fist and the third, and between the first and the fifth dose. These differences display clear differences across prefectures.

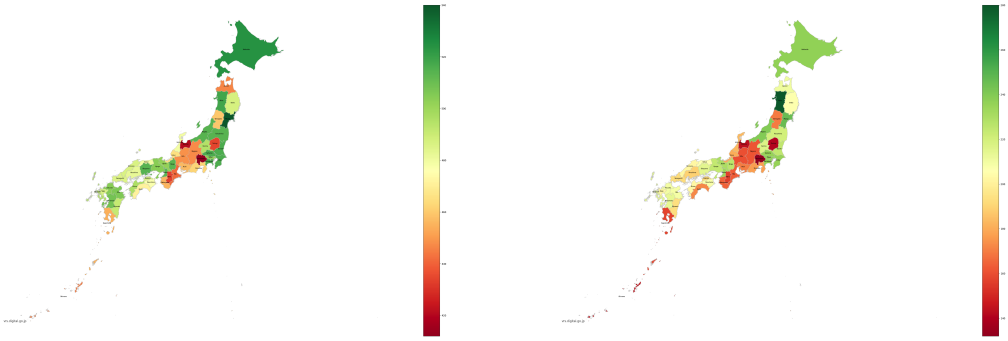


FIGURE 2.35: Comparison uptake times 1st to 5th (left picture) and 1st to 3rd (right picture) vaccine

Since our inflection-point-based approach already captures various probabilistic issues like different infrastructure or vaccine availability, *ceteris paribus* this pattern can best be explained by the effect of geospatial personality distribution or spatiotemporal fluctuations of the information field.

Statistical Feature Reduction Pipeline and Regression Results

We use the **Statistical Feature Reduction Pipeline** described above to manually "simulate" automated feature selection. For that, we extract those variables that have both correlation coefficients r_p r_s above .7 or below -.7, an Euclidean Distance Correlation of above .7, and a MIC of above .7. As expected, the factor analysis does not result in any meaningful result due to the unbalanced sample, with 46 factors of an Eigenvalue above 1 in the exploratory run, thus not interpretable nor useful. This resulted in about 30 useful predictors for each relevant outcome variable, which we then used for manual, theory-driven feature selection. Most plots appear to have a linear relationship with the outcome measures, wherefore we decide for a linear OLS regression analysis.

We designate the outcome of the **first model: predicting mid-term vaccine uptake** as "Cluster of the Lower Maslow Pyramid", since it is comprised of the following variables: the number of severe cases during the third phase of the pandemic, LIWC scores on "risk" before the onset of the COVID-19 pandemic and during the first wave, "negative emotions" during the first wave, "feelings" during the second wave, and "family" during the second and third wave (all variables are explained in Appendix B). We interpret this as communication patterns that make people aware of risks, trigger interest and sense of urgency and existential fear through preceding negative emotions in the onset of COVID-19, and about micro-systemic elements during the second and third phase, when uncertainty and fear during the pandemic peaked. Interestingly, no personality results are significant for this most important measure, and it is purely based on predictors from the information field, which are both proximal and relevant to primeval fears of survival on an individual- and micro-systemic level.

Table 2.6: OLS Regression Results "Vaccine Uptake Mid-Term"

Statistic	Value	Statistic	Value
Dep. Variable:	Vaccine Uptake Mid-Term	R-squared:	0.592
Model:	OLS	Adj. R-squared:	0.519
Method:	Least Squares	F-statistic:	8.090
Date:	Tue, 28 Mar 2023	Prob (F-statistic):	4.66e-06
Time:	00:14:52	Log-Likelihood:	-209.70
No. Observations:	47	AIC:	435.4
Df Residuals:	39	BIC:	450.2
Df Model:	7		
Covariance Type:	nonrobust		

Variable	coef	std err	t	P>|t|	[0.025	0.975]
const	-368.5625	116.788	-3.156	0.003	-604.788	-132.337
$risk_0$	1337.4154	410.808	3.256	0.002	506.477	2168.354
$risk_1$	-1654.1748	421.570	-3.924	0.000	-2506.880	-801.470
$negemo_1$	410.0607	99.189	4.134	0.000	209.432	610.689
$feel_2$	1254.5321	254.711	4.925	0.000	739.331	1769.733
$family_2$	-700.7044	228.158	-3.071	0.004	-1162.198	-239.211
$family_3$	392.4999	172.348	2.277	0.028	43.893	741.107
$severe_3$	7.3231	1.551	4.723	0.000	4.187	10.459

Omnibus:	3.147	Durbin-Watson:	1.811
Prob(Omnibus):	0.207	Jarque-Bera (JB):	2.244
Skew:	0.273	Prob(JB):	0.326
Kurtosis:	3.921	Cond. No.:	469

Notes: [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

We designate the **second model: predicting long-term vaccine uptake** as "Cluster of Social Learning Effects", since it is comprised of the following LIWC scores: "negative emotions" before the onset of the COVID-19 pandemic and during its first wave, "anxiety" during the fourth wave, plus agent Extraversion, and Extraversion of the contextual embedding (all variables are explained in Appendix B). It shows that such people take the fifth dose that are primed normatively towards negative emotions, then are exposed in the fourth wave with messages of anxiety, and who are very extraverted, in extraverted environments. This is interesting, since it indicates that regional extraversion improves social learning from others and the creation of normative believes, plus, synergetic with agent extraversion, fosters a climate of mutual exchange through human contact. It has to be stated though that the information field is orders of magnitude stronger than personality factors. However, results give indication for the correctness of the assumed event horizon, since it covers the first half of the time from the onset of the pandemic to the first availability of vaccines.

TABLE 2.7: OLS Regression Results "Vaccine Uptake Long-Term"

Statistic	Value	Statistic	Value
Dep. Variable:	Vaccine Uptake Long-Term	R-squared:	0.431
Model:	OLS	Adj. R-squared:	0.362
Method:	Least Squares	F-statistic:	6.214
Date:	Tue, 28 Mar 2023	Prob (F-statistic):	0.000225
Time:	00:30:58	Log-Likelihood:	-214.01
No. Observations:	47	AIC:	440.0
Df Residuals:	41	BIC:	451.1
Df Model:	5		
Covariance Type:	nonrobust		

Variable	coef	std err	t	P>|t|	[0.025	0.975]
const	-361.8265	251.280	-1.440	0.157	-869.297	145.644
$negemo_0$	1393.6929	327.229	4.259	0.000	732.840	2054.546
$negemo_1$	-1426.4957	338.818	-4.210	0.000	-2110.752	-742.239
anx_4	789.6031	296.970	2.659	0.011	189.859	1389.347
$E_{context}$	86.9427	40.665	2.138	0.039	4.818	169.068
E_{agent}	133.0699	51.547	2.582	0.014	28.969	237.171

Omnibus:	0.144	Durbin-Watson:	2.444
Prob(Omnibus):	0.930	Jarque-Bera (JB):	0.341
Skew:	-0.073	Prob(JB):	0.843
Kurtosis:	2.609	Cond. No.:	718

Notes: [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Finally, we designate the **third model: comply with governmental measures** as "Anatomy of Extraverted Agents", since it is comprised of the following LIWC scores: "affect" before the onset of the pandemic, during the second and fourth wave, "negative emotion" before the onset, during the first wave, and "anxiety" during the fourth wave. Furthermore, as with the second model, agent and contextual Extraversion are significant predictors, as well as survey results that indicate that agents find governmental measures justified. A further similarity is that strongest effects are negative emotions during the early onset of the pandemic, however, the difference is not as strong as with the second model (all variables are explained in Appendix B). In many ways, this third model is most interesting, since it predict survey results with both agent context and agent-internal variables, and thus offers unique insight into agents and not only an overview over mass behaviour. As expected, its explanatory power is much higher ($R^2 = 0.836$), and it offers a better temporal granulation, thus indicating that strong initial communication is imperative for reaching strongest health management results. Furthermore, it also provides partial evidence for our proposed framework, since it shows that steady, relevant information, plus proper contextual embedding and agent cognition promotes vaccine uptake; however, initial information seems to yield the strongest influence, which contradicts framework assumptions of current events being more influential to an agent.

Table 2.8: OLS Regression Results "Abiding by governmental measures"

Statistic	Value	Statistic	Value
Dep. Variable:	Abiding	R-squared:	0.836
Model:	OLS	Adj. R-squared:	0.796
Method:	Least Squares	F-statistic:	20.96
Date:	Tue, 28 Mar 2023	Prob (F-statistic):	5.74e-12
Time:	00:31:01	Log-Likelihood:	85.719
No. Observations:	47	AIC:	-151.4
Df Residuals:	37	BIC:	-132.9
Df Model:	9		
Covariance Type:	nonrobust		

Variable	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.1935	0.507	-2.354	0.024	-2.221	-0.166
$affect_0$	0.2880	0.087	3.319	0.002	0.112	0.464
$negemo_0$	2.3093	0.741	3.115	0.004	0.807	3.812
$negemo_1$	-2.4781	0.758	-3.269	0.002	-4.014	-0.942
$affect_2$	0.3836	0.104	3.673	0.001	0.172	0.595
$affect_4$	-0.3099	0.083	-3.731	0.001	-0.478	-0.142
anx_4	1.8305	0.543	3.371	0.002	0.730	2.931
$E_{context}$	0.2420	0.075	3.238	0.003	0.091	0.393
E_{agent}	0.2268	0.096	2.374	0.023	0.033	0.420
justified	0.4680	0.058	8.039	0.000	0.350	0.586

Omnibus:	1.115	Durbin-Watson:	2.028
Prob(Omnibus):	0.573	Jarque-Bera (JB):	0.938
Skew:	0.071	Prob(JB):	0.626
Kurtosis:	2.322	Cond. No.:	1.56e+03

Notes: [1] Standard Errors assume that the covariance matrix of the errors is correctly specified. [2] The condition number is large, 1.56e+03. This might indicate that there are strong multicollinearity or other numerical problems.

2.5.5 Discussion

We demonstrate that spatiotemporal measures can can be successfully used to explain and predict economic behaviour like population abidance with health regulations and governmental measures. For that, we use data from SNS to model the information field based on considerations from system-theory and brain sciences. We furthermore use inflection point differences of sigmoid vaccine uptake curves as a simple mean to identify the point when the first momentum is exhausted and other processes kick-in. This formulates a distribution of speed differences over Japan, which look dramatically different between the first and the fifth than between the first and the fifth dose. These differences can only be explained by psychological factors, given the homogeneous, highly efficient, and industrialised infrastructure of Japan.

We furthermore suggest a dynamic, spatiotemporal framework for better understanding and operationalisation of the information field, and enabling future research to be based on a solid foundation. Also, this framework enables us to discuss potential attachment points of the information field to individual-psychological architectures of agents. This framework has furthermore the advantage over models based on time or space alone that it is able to model ripple effects in the information field over space

and time and explain congruent behaviour changes. To facilitate ease of operationalisation of this framework, we suggest a new weay to interpret LIWC scores through a system-theoretic perspective. This allows us to simplify the data structure we operate on, however also leads to unbalanced samples, which we tackle with a semi-manual statistical feature reduction pipeline.

Finally, we show that agent and embedding personality are small but important factors to predict both individual and collective behaviour. As in the study of temporal components in section 2.4, Extraversion emerges as one of the leading factors of health behaviour, which we interpret as behaviour of actively seeking both the proximity and opinion of others, and thus leads to more spread of information and a synchronisation of ideas. Unfortunately aligned with findings from media psychology (Harcup and O'Neill, 2017), we also find that negative and relevant information in terms of proximity and personal involvement, as well as priming effects to play a role. However, on top of that we find, that strong communication at the onset of the pandemic is the biggest contributing factor to abiding with governmental measures and vaccine uptake. Those findings offer new pathways to tailor messaging from the government and health authorities to better mobilise a population towards health behaviour and thus may contribute to a safer society. However, nowhere else than in public communication is the line between help and manipulation thinner, and nowhere else can wrong approaches undermine public trust. Hence, those findings are powerful but need more rigorous confirmatory research, and a strong ethics debate should one want to use it for public relations.

On a more abstract level, (Gurnee and Tegmark, 2023) show that LLM not only learn temporal and spatial features but possess dedicated time and space neurons, similar to grid cells in the human entorhinal cortex. These spatiotemporal features are learned from language data at training time, and determine model behaviour at inference time. Since dedicated "grid cells" are identified, a purely abstract emergent capability due to model scale can be excluded. This section demonstrates the importance of spatiotemporal features for real-life results, and it would be a crucial task to conduct further research on the connection between "grid cells" of artificial neural networks, model behaviour, and spatiotemporal psycholinguistic features.

Limitation and Outlook

The strength of this study – the introduction of various simple yet effective methods to represent a framework for information field effectivenes – is also its biggest weakness, since several assumptions need to be done for that. Each of these assumptions (e.g., that LIWC categories represent geospatial distance) would need further research to hold. Hence, we cannot say that the *ceteris paribus* approach is fulfilled, and further research on each sub-component of the framework and methods needs to be conducted to stress test on causality, robustness and generalisability. Due to the unbalanced sample and theory-driven approach, some effects might be invisible. Hence further, more automated, and more extended analysis on those parts of the data that exists in deeper granularity (e.g., survey results on city level) should be conducted to inject more data-driven information into the outcomes. Also, more research on the forecasting aspects of the model should be conducted, which either uses a different method for identifying future-oriented language (e.g., using topic modelling), or this research should be repeated in another language where those missing LIWC features do exist.

Conclusion

We introduce and find evidence for a framework for understanding the effect of the information field on both individual agents as on aggregate groups of people, based on information theory, system theory, psychometrics, and behavioural economics. Furthermore, we introduce a novel way to allocate physical proximity by LIWC word categories, and use that effectively to find evidence for the framework. Unfortunately, future prediction is not possible to explore since J-LIWC2015 does not provide time categories; partially due to peculiarities in the Japanese language. We also introduce a simplified pipeline for manual statistical feature reduction for unbalanced samples, or when theory needs to be injected and results need to be aligned with strategic imperative. Finally, we use that framework and the methodology successfully to identify such individual-psychological factors in agents, aggregate regional psychology, and the information field that positively influence vaccine uptake. This enables novel, more precise and effective ways to encourage vaccine uptake, health behaviour, tailor and precision-apply governmental messages to combat fake news, and contribute towards a more healthy and robust society.

2.6 Chapter Conclusion

We conclude that just like personality and other psychological phenomena, linguistic tokens display distinct spatial distribution. However, personality predictions by those with available tools are very weak, which might be caused by local accents, data being trained by aggregate texts that neglect spatial features, or overfitting of models that were for example created for SNS, and then are applied to news data or vice versa. While temporal distribution of psychological phenomena is in its infancy - mainly due to the lack of data, temporal distribution of language features is well researched. While expecting comparably bad results as with spatial data, we show that temporal predictions of personality at least align with exogeneous events, however we do not know whether the algorithms predict what they are supposed to predict or just display changed linguistic features. These linguistic features either could spring from psychological latent traits, or these could just be artefacts from a changed situation on the ground. This changed situation would then be wrapped into words the models are not trained on, and thus display a personality change that in reality is none. Finally, we find that spatiotemporal models that are basically based on pure text analysis for hard-coded, theoretical features from LIWC, without trying to model personality, deliver results with a high eye-validity. Hence, we find that spatial, temporal, and spatiotemporal components need to be learned during training time in order for LLM to correctly model human psycholinguistic features. This lack of granularity and feature neglect might be one of the root causes of the results from chapter 1, and we expand on this discussion in the subsequent chapter.

Chapter 3

Hic Sunt Dracones: Towards a substrate-free, universal psychometric

3.1 Chapter Introduction

As we show in chapter 1, LLM seem to develop various synthetic core personalities with varying degrees of stability. These synthesised personality cores emerge within each language, and languages do not resemble in patterns, which points towards issues with training data, emergence through scale, or model specifications. We assume neglect of feature specification over and above culture-related (Atari et al., 2023; Johnson et al., 2022b) reasons, which is a novel perspective for research in LLM. We further assume that this feature neglect happens during model training, where data is aggregated and fed into the learning algorithms in batches that lack information of time and space, as we see for example with GPT-3 (Brown et al., 2020). Hence, taking a deep dive into geospatial, temporal, and spatiotemporal language features in chapter 2. We find that language features display distinct geospatial differences, yet if we use these to predict geographic personality patterns, outcomes are mostly uncorrelated with ground truth data (Yoshino and Oshio, 2021b), and display strong outliers in regions with divergent linguistic features.

On the other hand, we show that changes of predicted personality based on the same data coincide with exogenous shocks in a meaningful way. The question is, what we measure exactly through that – adaptive personality states to cope with extreme levels of stress on organisms, or "just" a change in the information field based on those events. In favour for change in personality states is that we find mainly changes in Extraversion and Agreeableness in the top 25% Conscientiousness group, which makes intuitively sense. Also, recent research shows that through intervention, personality traits can change over time; especially Agreeableness and Conscientiousness (Stieger, Flückiger, and Allemand, 2023); as such, the pandemic could have accelerated this change. On the other hand, just like LLM, personality prediction algorithms are trained on aggregate data, and do not take space or time into account, hence we might just capture a dramatic language change and the subsequent inability of personality prediction models to produce valid outputs. In psychometric terms, it may lose criterion validity and to some degree also external validity through that. But the question is, what we really measure when we detect changes. Finally, we find that over time and space, events create ripples in the information field, in line with a dynamical system-theoretic framework proposed by us, and discuss, where in the psychometric architecture of agents this information field might attach and influence behaviours. More concretely, we find that linguistic markers of temporal and geospatial proximity,

as well as agent Extraversion and embedding Extraversion predict vaccine uptake, which makes intuitive sense and opens new avenues for tailoring information from governments and health authorities through understanding contextual embedding of agents. In summary, we find that spatial and temporal factors are crucial for understanding text, yet while this finding is not old for NLP, it is new to LLM research (although latest research shows that LLM encode space and time just like grid cells in the brain (Gurnee and Tegmark, 2023)), psycholinguistics, and its application to behavioural economics.

However, training new models is a lengthy, industrial, and very expensive process. Some Big Tech companies even start constructing their own nuclear power plants for supplying the energy demands necessary for next generation language models (Calma, 2023). Hence, this chapter can only discuss the limits of classical psychometrics, suggest a novel (Rust, Kosinski, and Stillwell, 2020), more universal (Hernández Orallo, 2017), abstract (Safdari et al., 2023), and substrate-free (Romero, Fitz, and Nakatsuma, 2023) approach to it. More concretely, we look into two urgent questions derived from this thesis so far: first, we discuss the psychometric properties of LLM from the GPT family, and derive potential constraints and limitations. Second, we extend this discussion further into a necessary debate to the contextual embedding of agents, formulate a more universal and substrate-free framework to psychometrics, and connect this with a validity discussion on synthesised LLM personality. Finally, we give an outlook towards unsolved questions like whether aggregate human personality and synthesised LLM personality have the same constraints, and, in return, whether individual-psychological measures are applicable to groups of persons, hybrid AI-human systems, and, in extension, community mechanisms in behavioural economics.

3.2 Deeper discussion of psychometric properties

(This section was written by Peter Romero as main author, and supervised by Teruo Nakatsuma and Stephen Fitz.)

3.2.1 Issues with Training Data of GPT-3

Given the development process of GPT-3

3.2.2 Reliability

In terms of absence of measurement errors, quite a few critical points can be found. The parallel forms reliability is guaranteed, since the chosen instrument has been used in a variety of procedures and use cases. The same is true for inter-rater reliability, and test-retest reliability, as has been documented in the manual (Gosling, Rentfrow, and Swann, 2003). However, test-retest reliability, as well as parallel forms reliability of this instrument have not been measured for language models yet, hence this could be a source of variance, especially since it is to establish whether the model acts as one, multiple, or perfectly randomised agent. Also during the development process, internal consistency, reliability, and internal consistency were reported for the main instruments, however, these are not available for all languages, and some of these have not been peer-reviewed, hence this could be a natural source of variance. Of course, the setting wasn't interfered by environmental factors or attitudes of the researchers, however reply patterns could have been subject to social desirability, depending on the choices regarding training data and model behaviour its architects made.

3.2.3 Validity

In terms of the degree to which the constructs are measured, less critical yet more fundamental points can be found, as the deployed instruments have been thoroughly validated. Also, as a concession to the low number of items, TIPI sacrificed internal consistency in favour of validity, hence content validity, construct validity, face validity and criterion validity are not problematic (Gosling, Rentfrow, and Swann, 2003). However, since the instrument is applied to a language model, it is important to look deeper into those aspects of validity.

With regard to **content validity**, it is questionable whether an instrument geared for humans covers all aspects of the Big5 construct when it is applied to an artificially intelligent agent. Völkel et al. (2020) found evidence for potential non-human components of personality constructs for artificial assistants, since the "commonly used Big Five model for human personality does not adequately describe agent personality" (p. 1), which is of course not captured by TIPI. Also, since personality is an abstract latent trait that is mostly measured by self-introspection, it is unclear what this means for a machine; does it mimic or display emergent capabilities of self-introspection? Also, given the psycholinguistic development of the modern Big5 theory, is it really personality that an artificially intelligent agent displays on an emergent level, or just a probability distribution over words that appear most often together and are elicited by the linguistic dimension within the items? In extension, would that mean that humans do the same and that personality is nothing else but a probability distribution over words representing cognitive, affective, and conative patterns?

The **cross-cultural validity** is unclear since no consistent study over all language-versions of TIPI has been conducted so far. However, other research indicates that Big5 construct seems to be universally applicable across cultures, yet there are culturally distinct patterns of expression (Schmitt et al., 2007).

While both the distinction and universality could be explained as capturing specific cultural expressions of a universal underlying latent trait, one could argue that an instrument is developed in one culture then translated into another, but still comes with culture- and thus language-specific concepts that cannot easily be translated. Hence, applying a human-specific instrument to an artificially intelligent agent could magnify this effect, and thus lower the structural validity of an instrument. Also, philosophically speaking, the instrument that was developed for the human world is being transferred to the machine world, which represents an unprecedented faultline, for which no research exists.

This notion overlaps with **construct validity**, since using a personality questionnaire might not only miss important emerging phenomena within a language model but its use case for development might not overlap with the 'reality' of the language model, which emerges from the various text sources it was developed with. For example, occupational personality inventories like Orpheus (Rust and Golombok, 2014a) might capture better such components emerging from vocational training data, whereas general purpose inventories like TIPI might miss out on these specific aspects but take up overall more variance since training data is broadly distributed across various text sources. In extension, since many computerised personality inventories are extensions of classical paper and pencil questionnaires, including TIPI, the question remains, whether CAT designs based on IRT, various forms of IAT, or more indirect, contextual measures like games, shadow assessments, virtual-reality-based assessments or inference from text, phone and sensor data really measure the same construct. In most manuals of these measures, still correlations to questionnaires are

given that originated in one way or another from former paper and pencil questionnaires. On the most simple level of criticism, one could argue that these instruments all provide different forms of granularity - see for example the differences that already exist between NEO-PI-R and NEO-FFI (Aluja et al., 2005). A future study might even repeat this research with two scales of different granularity or with two scales from different application use cases to explore that aspect more deeply. However, the heretical question may arise, whether ultimately, personality as a construct is based on the way it is operationalised, which is isomorphic to the criticism of IQ, that some people say is defined as what an IQ test measures (Rust and Golombok, 2014a). Also, since psychological traits are inter-correlated up to varying degrees, it would be crucial to understand the influence of the the instrument on that connection.

In terms of **criterion validity**, TIPI displays "substantial" (p.517) convergence across measures (Gosling, Rentfrow, and Swann, 2003). However, in the case of measuring artificially intelligent agents, there exists no psychological instrument that could be counted as "gold standard" yet, hence it is important to explicate usual measurement issues behind criterion validity more deeply; mainly norm groups, construct stability, and the state-trait problem.

First, norm groups are the adaptation of test scores towards meaningful subsets of a population. Hence, most test providers will market this as a strength of their instruments – whether in the development phase, or, with assessment companies, the wealth of their data base upon which such groups were developed. This comes with a multitude of problems, though. Foremost, people from various backgrounds are put into categories that may or may not be useful. For example, a manager from an engineering department in the German automotive industry may be clustered together with a sales executive from an Malaysian insurance company. While for universal aspects like leadership that might be useful, finer categories that are highly relevant to a specific organisation or culture might be averaged out through this approach. Furthermore, while findings on personality are stable across cultures, their perceived importance in the workplace is not. For example, Asian and European cultures focus on different aspects, wherefore the need to create a new personality component for Asia, Dependence on Others, was discussed, which represents collectivist patterns of experiencing and behaviours, other than the existing categories within Big5 that were created in the individualistic cultural space of the West (Hofstede, 2007). Thus, we postulate the *"curse of norm groups"*: the more dimensions of psychological latent traits are taken into account, the less useful these become for the individual case, and vice versa. Applied to measuring personality on a language model that is trained on a multitude of languages and corpora within a specific language like GPT-3, this means that neither broad-level measures like TIPI nor specific measures like ORPHEUS (Rust and Golombok, 2014a) may be useful for describing its emergent psychological properties, especially when taking into consideration that the model tries escaping through the easiest route, for example by switching to languages where it presumably has more underlying training data.

Second, while personality is relatively stable, it still changes over a life-time. In younger years, as well as under prolonged external stress, this change is quicker than during adulthood (Bleidorn et al., 2021). Also, there is evidence that personality displays elasticity as potential coping mechanism to extreme exogenous conditions, thus might display an emergency expression of normal traits that cannot be explained by states (Romero et al., 2021). Hence, while construct stability of personality is relatively high, it is malleable and undergoes transitional stages. It is unclear, which stage of maturity in humans maps onto GPT-3, or whether the notion of maturity even applies in this context. This implies, that it is unclear if the model "matures"

further, for example through various training processes or augmentations like with GPT-3.5 that uses RL to both enable chat processes and filter undesired requests (OpenAI, 2022), or whether with every iteration, it can be considered as a new "species". Maybe, training of the model itself represents accelerated evolutional processes that makes it change its personality in due course. While this question should be covered in future research, stability of its current personality could change by training, augmentation, or in the best case, only with each new iteration. Hence, the used instrument might not be adequate since it is calibrated on an adult audience, whereby we cannot determine the "maturity" of GPT-3.

Last, psychometric measurements suffer form state-trait-problems (Rust and Golombok, 2014a) despite all test-retest reliability, whereby the true value or trait is a function of all states that are measured. States are relatively temporary, oscillating around the true value of the trait, which is either stable, or changes very slowly over time, as is the case for personality (Bleidorn et al., 2021). Reasons for that can be variances in the latent psychological trait or measurement errors, which encompass both technical aspects as well as internal or external processes that temporarily influence the agent. Depending on the measurement cadence, also memory effects are part of the measurement error. Furthermore, psychological latent traits are not deterministic but probabilistic in nature, to allow degrees of behavioural freedom and updates of internal representations. For example, many academics seem to be 'extraverted introverts' who prefer spending time alone researching and writing, but need to network and teach, as well (Irfani, 1978). It is unclear whether GPT-3 applied its answers to the questions, which might be indicated through the reasons it gave for each score it chose, and which resembled or partially mirrored content from the questions, or whether it displayed its "true personality score". If the former was the case, it could be interpreted as its adaptation to the contextual embedding of an assessment situation.

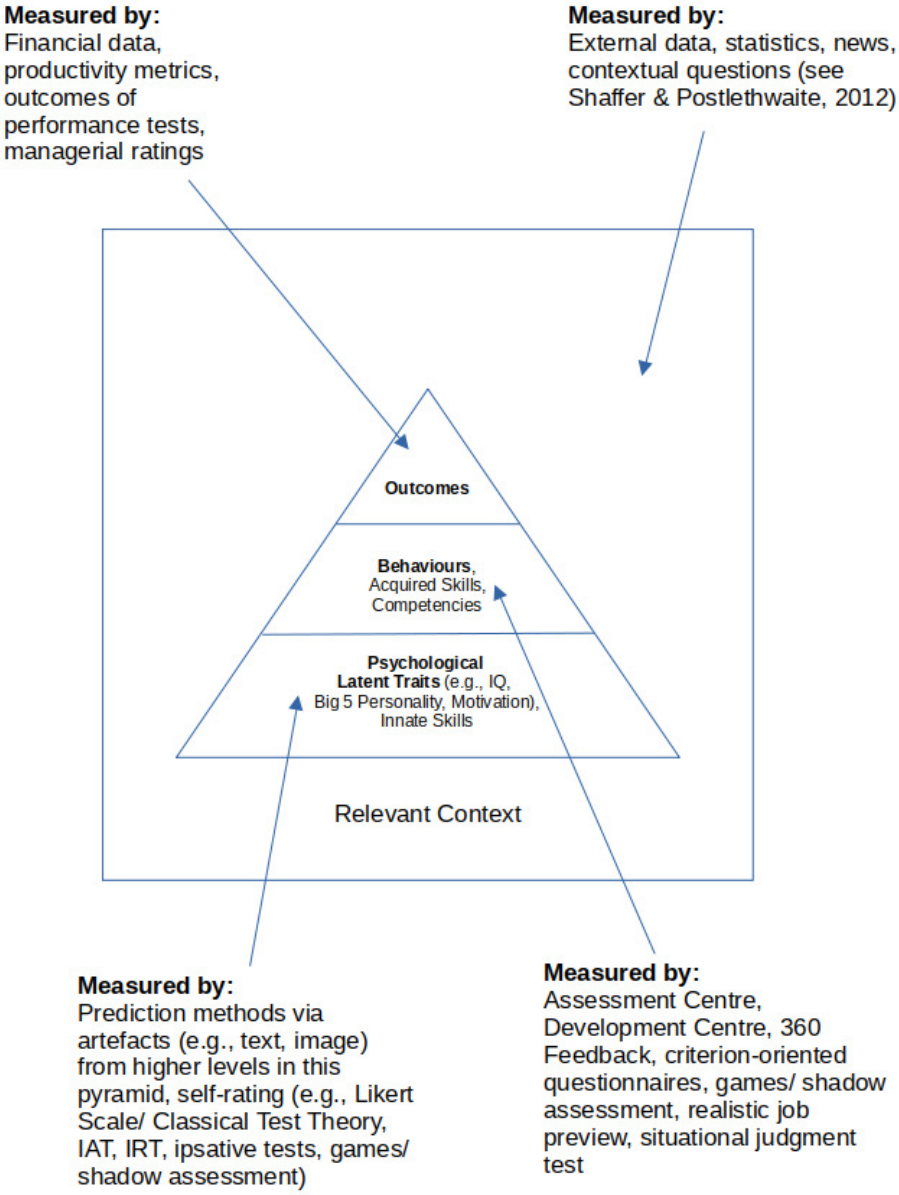
Contextualisation of personality measures that represent the broader systemic embedding of test takers are known to improve validity (Shaffer and Postlethwaite, 2012). This makes intuitively sense, since specific behaviours are more relevant to specific locations than others. For example, assertive behaviour may be more rewarded in the workplace than it would be in family settings. Since the instrument used is not contextualised, and since the optimal contextualisation for a language model has not been researched upon, it is furthermore unclear whether this may explain additional variance and reduce overall validity. Hence, it is important to take a closer look on the contextual embedding of an agent.

3.3 Deeper discussion of contextual embedding of behaviour

While psychological research provides further evidence for the importance of contextual embedding of behaviours (Shaffer and Postlethwaite, 2012), an overarching measurement model, which classifies behaviours, psychological latent traits, results, and the influence of contextual embedding, is still missing. Research on competencies may come closest to that. Competencies are defined "sets of behaviors that are instrumental in the delivery of desired results or outcomes" (p.7), and encompass underlying latent psychological traits, as well as behaviours (Bartram, Robertson, and Callinan, 2002). However, those are mostly operationalised in organisational settings, wherefore contextual factors are very specific to hierarchical or functional levels.

For abstraction towards the psychometrics of intelligent agents based on other than biological hardware, this approach has to be generalised and become as substrate-free as possible. Since competencies are sets of behaviours, driven by underlying latent psychological traits ("competency potential"), resulting in desired outcomes (Bartram, Robertson, and Callinan, 2002), these can be described as 'higher order functions' in mathematics - taking functions as arguments, and returning functions as outcomes. A function $f : X \to Y$ in mathematics is a mapping of each element in the domain X to a subset of the codomain Y, which is denoted $\text{img}(f) \subseteq Y$. Behaviours can be domains and codomains, whereas psychological latent traits can only be domains, and outcomes only the codomain. Therefore, the set of possible behaviours is modulated by latent traits and the contextual embedding of an agent, which puts constraint on its size in actuality. Competencies contain various domains, codomains, and mappings, yet are not supersets of all potential mappings, since contextual embeddings moderate the functional form of each mapping. This follows a hierarchical encapsulated model of *Contextual Embedding(Outcomes(Behaviours(Psychological Factors)))*, and is supported by known moderate correlations of psychological measures with each other; the main reason why most complexity reduction mechanisms like factor analysis, applied to psychological data, cannot assume orthogonality (Rust and Golombok, 2014a).

Figure 3.3 displays this embedding and the hierarchical order of latent psychological traits, resulting behaviours and concluded outcomes, and gives examples of which psychometric tools best measure each level.



While behaviours take place in relevant environments, agents receive information from outside perception and inside predictive processes. Learning and behaviour take place in a systemic context, whereby the pre-existing knowledge steers behaviours top-down, and gets updated by bottom-up processes and learning. During this process, agents create and update their own data set in form of a representation of the world. This actualisation takes place at different pace, depending on factors internal (personality? IQ?) and external (motivation? values? norms?) to the agent, which explains different degrees of adaptation, speed, and success to various environments, depending on the individual agent, and it allows flexible adaptation on right level of stress outside homeostasis; thus enhancing its fitness. Formalising and operationalising the context for models of competencies is crucial, since learning and thus behaviour takes place in relevant contextual embeddings. Some embeddings might be ”closer” and thus more relevant to individual agents than others in terms of measures of distance, and in terms of prior direct or indirect knowledge. Aligned with systems theory, systems are encapsulated in higher and lower level systems, whereby the systemic levels influence each other with a relative strength based on their hierarchy (Willke, 2000). This helps us to further formalise this interaction more precisely.

For example, a human agent is embedded in clusters like family → friends → colleagues → organisational members, *et cetera*. This context might be close or distant to the agent, so the effect can be high or low, and represented as a matrix that has distance and effect as columns, and contextual levels as rows. This matrix represents the regulator of external forces that affects the mappings between the psychological latent traits and behaviours, the behaviours and the outcomes. Thus, it limits the degree to which a potential of a subsystem of an individual agent can be expressed in a specific context. More formalistically, we operationalise:

$$\textit{Can Do} \times \textit{Will Do} \times \textit{Context} \rightarrow \textit{Behaviours} \rightarrow \textit{Outcomes}$$

Can Do encompasses more proximal competency potential like personality, intrinsic or internalised extrinsic motivation and societal norms in form of values, and innate skills, whereas *Will Do* encompasses more distal competency potential like extrinsic motivation, societal norms, or acquired skills and knowledge, and both are disjoint. The set of *Behaviours* is a Cartesian product of countably infinite sets: *Can Do*, *Will Do*, and *Context*. The *Outcomes* are a function of *Behaviors*.

Context encompasses both social as well as spatial ambient embeddings in which a behaviour potentially takes place. While theoretically, more proximal social embeddings are more relevant than rather distal spatial embeddings (Willke, 2000), this might of course change, depending on how these facilitate or inhibit fitness (Doreian and Conti, 2012), wherefore the authors did not distinguish any further. However, to enable statistical analysis, we need to define probability distributions on the above sets. The co-domains of these distributions will be called:

$$\tilde{\textit{Can}}\ \textit{Do},\ \tilde{\textit{Will}}\ \textit{Do},\ \tilde{\textit{Con}}\textit{text},\ \tilde{\textit{Behav}}\textit{iours},\ \tilde{\textit{Outc}}\textit{omes} \in [0,1] \subset \mathbb{R}$$

Thereby, these subsets in the closed unit interval of real numbers, since they are defined as sets of potential elements displayed by, acquired, or innate to an individual, given a specific potential set of spatial embeddings, with 0 being the least desired, and 1 being the most desired set of behaviours for individual fitness. These are possible to some extent and thereby strictly ≥ 0 and ≤ 1, with 0 and 1 being the most unlikely outcomes given the probabilistic nature of human behaviour and contextual facilitation or inhibition. Congruently, the optimal set of $\tilde{\textit{Behav}}\textit{iours}$ to reach the optimal set of $\tilde{\textit{Outc}}\textit{omes}$ is determined by the probability that the most optimal set of $\tilde{\textit{Can}}\ \textit{Do}$ and $\tilde{\textit{Will}}\ \textit{Do}$ is present in the most optimal $\tilde{\textit{Con}}\textit{text}$, wherefore these two elements are defined as a closed unit interval of real numbers, as well.

The main point is to provide foundation to contextual embeddings of agents and thus the overall validity discussion. Since all psychometric tools so far have been created for biological intelligent agents, a more general, substrate-free new kind of measurement has to be defined. Partially, this definition began with the promotion of "culture-free" psychometric assessments, which failed for a variety of reasons like geospatial, historic and cultural embedding (Lupyan, 2022); in the interpretation of the authors mainly since these were still bound to wetware. As planes do not flap their wings, but abstracted bird wings through the principles of aerodynamics, so will artificial agents abstract biological psychometrics into something we are not aware of yet. The subtle hints of a potential non-biological personality dimension (Völkel et al., 2020), or the inconsistencies of GPT-3's emergent personality expression might be the harbinger of a substrate-free psychometric approach, which must include biological psychometrics as only subset of many. What is the world of a language

model? It only knows text, hence all it does is predicting the next word based on input data, comparable to the first stage in Plato's allegory of the cave. Hence, the entire psychometric structure abstracted above by generalisation of competency models and extension by contextual embeddings is for a language model analogous to a noisy projection. Only by extending its universe into our reality – likely through robotic embodiment or merging with wetware through neural interfaces – will it be able to develop further. As a first step of this development, and potential evidence of the correctness of the abstraction above, ChatGPT, or GPT-3.5 was embraced as the watershed moment in the public recognition of NLP. It basically is the language model of GPT-3, augmented by a chat module that uses RL to understand extended interaction with humans, which could be considered the extension of behaviours on top of latent traits. As people started abusing this system for creating hate-speech, a second reinforcement module was set on top, which taught it to avoid potential abusive content. This second module can be considered the contextual embedding, which moderates the connection between latent traits (GPT-3), behaviours (GPT-3.5), and outcomes (the text produced by GPT-3.5).

3.4 Chapter And Dissertation Conclusion

We conclude that recent advancements in AI have been driven by enhanced hardware architecture, access to vast quantities of high-quality data, and innovative algorithmic frameworks, especially those involving multi-layered artificial neural networks. These developments have led to widespread application and dominance in various practical areas such as information processing, data governance, organisational digitisation, and consumer electronics, thereby becoming increasingly integrated into various facets of life (Romero and Fitz, 2021). Especially the remarkable predictive capabilities in unsupervised problems, its user-friendliness, and its universal relevance have prompted its adoption across a wide spectrum of academic disciplines beyond computer science, including but not limited to biology, life sciences, finance, stock market forecasting, predictive policing, computational social sciences, and the field of art.

While disciplines with a strong grounding in statistics and theory-based approaches, such as psychometrics (Rust, Kosinski, and Stillwell, 2020) and econometrics (Mullainathan and Spiess, 2017; Varian, 2014), initially favoured symbolic AI or GOFAI, particularly under regulatory frameworks like the EU's General Data Protection Regulation, which restricts fully automated decision-making on humans, and due to risk management practices that discourage the use of opaque models, the speed of development is ever accelerating. Increasing research in neural network architectures, advancements in AI safety, classifier prediction explanations (Ribeiro, Singh, and Guestrin, 2016), alignment research, and market dynamics are gradually shifting this trend, since who does not use it, will be left behind – on an indivdiual, corporate, and national level.

This is nowhere more obvious than in the mass-adoption of ChatGPT, which is a watershed moment in the societal deployment of Artificial Intelligence (AI), with immediate, sustainable, and deep-reaching ramifications for science, research, and economy (Chow, 2023). Having long passed the Turing test (Biever, 2023), and being optimised for safe and agreeable interaction (OpenAI, 2022), humans can interact with it cooperatively like with another human being of great knowledge and authority, or use it more like a sophisticated search engine in an utilitaristic manner. On the other hand, AI displays increasing levels of agency and autonomy (DiBlasi et al., 2020), which could put them at one point in position to not only cooperate or obey,

but also to lead. Aligned with the vision of Japanese Society 5.0 (*Society 5.0: A People-centric Super-smart Society* n.d.) and Fourth Industrial Revolution (Schwab, 2016), this implies that work of the future will be in hybrid systems; humans will interact with other humans and AI systems, as well as AI systems with each other.

Most interaction with AI systems takes place via language, which is the domain of Large Language Models (LLM) like ChatGPT. Depending on scale and quality of training data, computation, and model parameters, LLM display emergent, unplanned, and unpredictable capabilities (Wei et al., 2022b) like mathematical skills (Frieder et al., 2023), theory of mind (Kosinski, 2023), and logical reasoning (Hagendorff, Fabi, and Kosinski, 2022). Also, psychological latent traits like values (Johnson et al., 2022a; Miotto, Rossberg, and Kleinberg, 2022) and personality (Safdari et al., 2023; Miotto, Rossberg, and Kleinberg, 2022; Jiang et al., 2022; Karra, Nguyen, and Tulabandhula, 2022) emerge, which are malleable and display external validity in their prompted behavioural patterns (Safdari et al., 2023; Jiang et al., 2022). However, based on their architecture, artefacts from the measurement approach (Digutsch and Kosinski, 2022), training data set (Karra, Nguyen, and Tulabandhula, 2022; Li et al., 2022a), prompting strategy (Miotto, Rossberg, and Kleinberg, 2022; Jiang et al., 2022; Karra, Nguyen, and Tulabandhula, 2022), or missing memory from past responses (Miotto, Rossberg, and Kleinberg, 2022), LLM are also sensitive to illusions, gender and racial bias (Digutsch and Kosinski, 2022), are skewed towards US values (Johnson et al., 2022a), score high in psychopathy, narcissism, and Machiavellianism (Li et al., 2022a), display unstable gender attributes (Miotto, Rossberg, and Kleinberg, 2022), inherit personality from training data (Karra, Nguyen, and Tulabandhula, 2022), and display unstable and split synthetic core personalities (Romero, Fitz, and Nakatsuma, 2023).

Since language use (Lee et al., 2007), mutual sympathy (Liu and Sundar, 2018), and work outcomes (Rust, Kosinski, and Stillwell, 2020) are strongly associated with personality, synthetic LLM personality is central to hybrid systems, and may moderate the dynamism of human-AI interaction. This might have a direct effect on their overall intelligence and performance, since depending on the nature of tasks, varying degrees of cooperation and different set of competencies are necessary. For example, individual contribution augmented by AI may perform worse than cooperative strategies in hybrid systems that demand communication, teamwork, and leadership competencies (Hernández Orallo, 2017).

In extension, learning from hybrid systems, and developing psychometrics for those might also bring new developments into research on aggregate individual-psychological measure as being used in chapter 2. Humans behave differently in masses than alone, which resembles the aggregate nature of data in LLM that might behave not as one entity, but as many entities [1]. Hence, research on whether individual-psychological measures are applicable to LLM or, how data issues influence this, what potential ramifications for hybrid systems are, and what this means for the future development of LLM, can all give us valuable feedback for better understanding geographic psychology, which is tightly connected to geographical econometrics.

That being said, hybrid systems are not content from SciFi movies, but are here for a long time, from the moment that the first personal computer was in use in a network and autonomously interacted with other humans and computers. Hybrid systems emerge when we interact with navigation systems, use digital assistants, self

[1] in reverse, it is unclear, whether repeated measures of the same test with one LLM, as being conducted in chapter 2, has to be considered as just different answers from the same person, or that of many person; whereby each new prompt could be considered as an individual person, e.g., from the "ChatGPT population"

driving cars, ChatGPT, or in the future with robots at home, the workplace or in all other areas of human life. This is the future of our species, and we will merge with intelligent agents to a new form of society (*Society 5.0: A People-centric Super-smart Society* n.d.; Schwab, 2016; Schwab and Malleret, 2020), whether we like it or not. The real watershed moment of ChatGPT is not that it creates human-like outcomes that are available to a vast number of people, but that it starts talking back to us, just like an alien visitor. We will have to develop a new and mutual co-existence, which will, aligned with the cognitive tradeoff hypothesis, also lead to a new kind of language that enables this interaction (Hecht, 2018), and thus to changes in community mechanisms (Ogaki, 2022) that demand rigorous research in behavioural economics.

Thus, as our ancestors, who, in all likelihood, once were pushed out of the trees into the uncertainty of the Savannah, had to give up cerebral capacity for language, just to return as the apex predator of this planet, which accelerated our evolution and enabled us to split the atom and travel to the galaxies; so will the arrival of AI catapult us to uncharted, new territories, and it will also mark our exodus into a new *terra incognita*, which the cartographers from the old not without reason symbolised with dragons; beware therefore, because – *hic sunt dracones*.

Appendix A

Application to Economics

A.1 Systems Theory

Systems theory is an interdisciplinary research approach to complex systems like economics, psychology, and even neural or chemical structures (Willke, 2000). Its goal is to describe complex, dynamic and apparently chaotic behaviour in such a way that it can be operationalised, quantified, measured, and predicted. It is comprised of a set of complementary research foci:

- System borders: main subject of systems theory is the differentiation of inside and outside the system and its sub-systems. In terms of Economics, these are distinct economic actors, based on their systemic level, from nation states over companies to individuals.
- Autopoiesis: is the quality of a system to replicate itself in both new systems, as well as its internal architecture. In terms of Economics, this is the tendency of economic systems to replicate themselves, e.g., continuity of industrialisation, or social classes.
- Specialisation: is the tendency of systems and their sub-systems to specialise towards distinct tasks and capabilities. In terms of Economics, this is the tendency of economic actors to specialise on one aspect within the economic exchange and thus allow for exchange of goods and services.
- Emergence: is the tendency of emergent qualities within systems of sufficient scale. In terms of Economics, this means that different scales of emergence arise, from individuals to teams, to companies, to labour unions, to industries, states, and economic regions. This happens in an encapsulated manner and is for example explicated in Bronfenbrenner's Ecological Systems Theory (Bronfenbrenner, 2013).
- Internal and external connections: are the sum of all connections between sub-systemic components. In terms of Economics, this means all connections an economic agent can legally or illegally make; on individual and collective level. In terms of behavioural economics, a connection has to be made, before any communication or exchange process can take place.
- Internal communication: is the information exchange within a system. In terms of Economics, this is the exchange process of economic agents. For example, all social interaction as in typical games of behavioural economics, are based on this communication.
- Internal exchange and currency: is the exchange of energy in terms of internal currency within a system. In terms of Economics, this is the economic exchange

process that takes place on the level of a specific system, and is mostly regulated by societal and governmental regulations like currencies.

Within this dissertation, I apply methods of computational psychometrics and psycholinguistics to behavioural artefacts, mostly in text form, on an aggregate level, to explore emergent phenomena. By applying Bronfenbrenner's Ecological approach to Systems Theory (Bronfenbrenner, 2013), this becomes obvious.

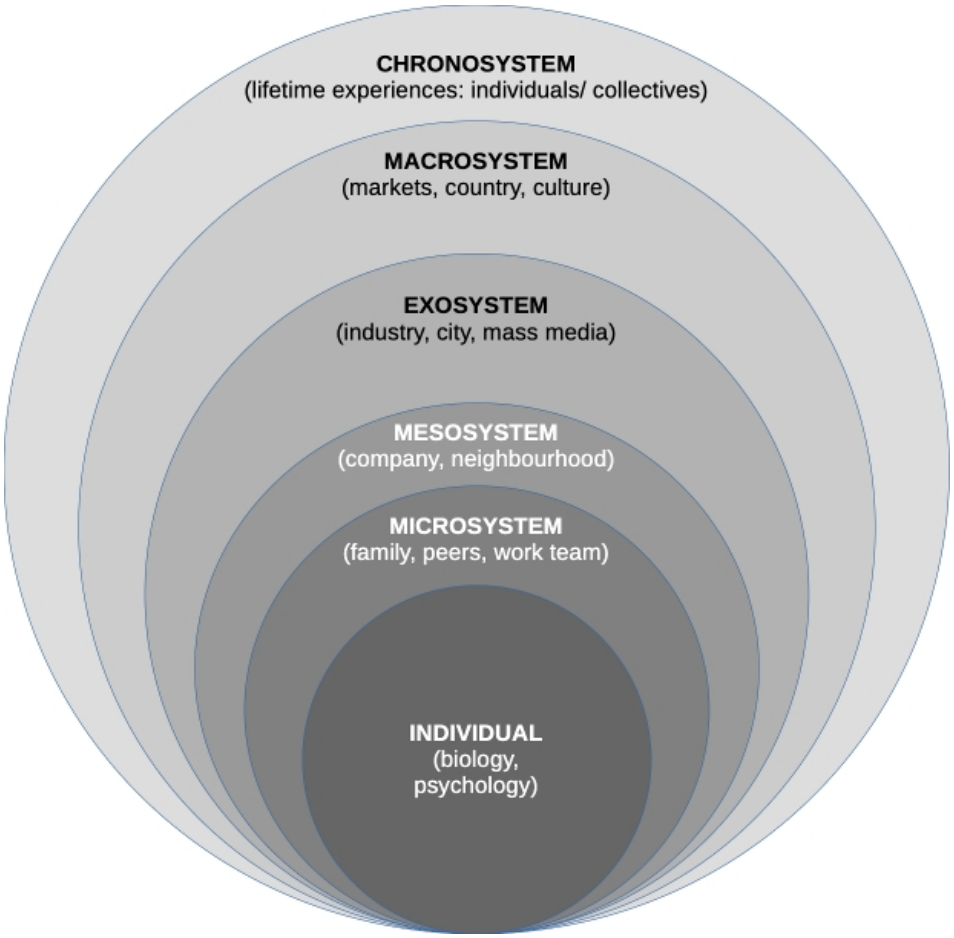


FIGURE A.1: Application of Bronfenbrenner's Ecological Systems Theory to Economics

Thereby, behavioural artefacts on individual-systemic level (human agents) are aggregated up to the level of the exosystem, to better understand collective behaviour. Thereby, the main level of measurement, as depicted in figure A.2 from chapter 3, is based on outcomes of specific behaviours of economics actors, which are based on both latent psychological traits and contextual embeddings (Rust, Kosinski, and Stillwell, 2020; Bartram, Robertson, and Callinan, 2002).

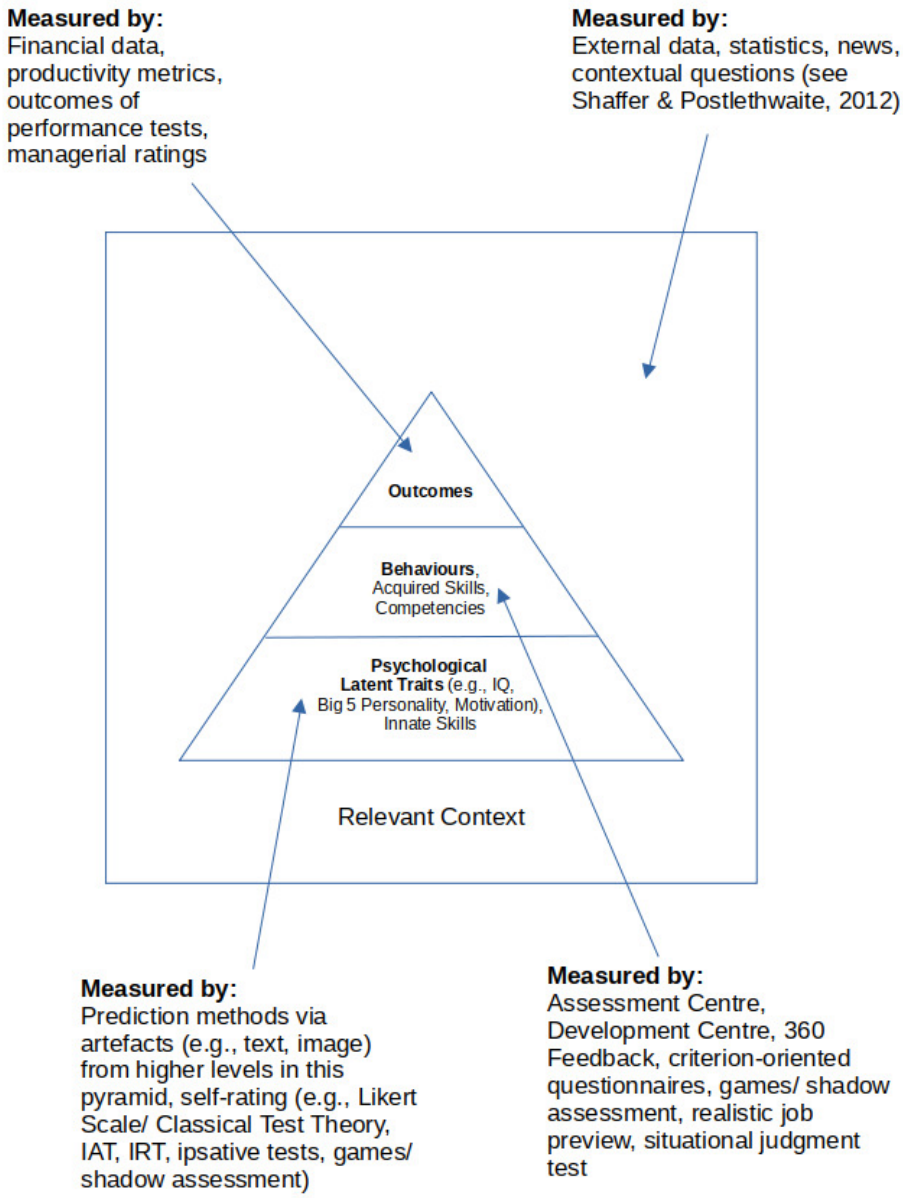


FIGURE A.2: Encapsulated model of measurement.

This aggregation is also successfully used by organisational scholars to predict corporate outcomes (Schneider and Bartram, 2017). And through further aggregation and comparison of systemic levels with the expected changes of Society 5.0 (*Society 5.0: A People-centric Super-smart Society* n.d.), economic measures on exosystem and macrosystem level can be explored by the same method.

FIGURE A.3: Society 5.0 (MEXT, 2024)

It will be subject of future research to identify communalities of pure human with hybrid systems.

Appendix B

Data explication and dictionaries

B.1 General Variables Used In Multiple Studies

B.1.1 LIWC

The Linguistic Inquiry and Word Count (LIWC) is a dictionary-based, hard-coded feature-extraction tool that is the quasi-standard in computational psycholinguistics (Pennebaker et al., 2015a). It is maintained by a team of psychologists, linguists, and computer scientists, who update it regularly to capture latest developments in languages. Local versions are available through various dictionaries, whereby each language is not just translated but adapted based on its linguistic peculiarities. Previous studies wrongfully translate the dictionary to Japanese either through the Chinese version, or directly from English, which results in skewed or even wrong results. We use the official Japanese dictionary (Igarashi, Okuda, and Sasahara, 2021) that was only released during the writing of this dissertation, wherefore some older studies (temporal psycholinguistics from section 2.4) do not include it yet. It is the foundation for section 2.2 on author profiling, section 2.3 on spatial econometrics and section 2.5 on spatiotemporal econometrics. Table B.1 provides an overview over all features and their psychometric properties.

Category	**Abbrev**	**Examples**	**Words in category**	**Internal Consistency (Uncorrected α)**	**Internal Consistency (Corrected α)**
word count	wc	-	-	-	-
summary language variables					
analytical thinking	analytic	-	-	-	-
clout	clout	-	-	-	-
authentic	authentic	-	-	-	-
emotional tone	tone	-	-	-	-
words/ sentence	wps	-	-	-	-
words > 6 letters	sixltr	-	-	-	-

Continued on next page

Table B.1 Continued from previous page

Category	Abbrev	Examples	Words in category	Internal Consistency (Uncorrected α)	Internal Consistency (Corrected α)
dictionary words	dic	-	-	-	-
linguistic dimensions					
total function words	funct	it, to, no, very	491	.05	.24
total pronouns	pronoun	i, them, itself	153	.25	.67
personal pronouns	ppron	i, them, her	93	.20	.61
1st pers singular	i	i, me, mine	24	.41	.81
1st pers plural	we	we, us, our	12	.43	.82
2nd person	you	you, your, thou	30	.28	.70
3rd pers singular	shehe	she, her, him	17	.49	.85
3rd pers plural	they	they, their, they'd	11	.37	.78
impersonal pronouns	ipron	it, it's, those	59	.28	.71
articles	article	a, an, the	3	.05	.23
prepositions	prep	to, with, above	74	.04	.18
auxiliary verbs	auxverb	am, will, have	141	.16	.54
common adverbs	adverb	very, really	140	.43	.82
conjunctions	conj	and, but, whereas	43	.14	.50
negations	negate	no, not, never	62	.29	.71
other grammar					
common verbs	verb	eat, come, carry	1000	.05	.23
common adjectives	adj	free, happy, long	764	.04	.19

Continued on next page

Table B.1 Continued from previous page

Category	Abbrev	Examples	Words in category	Internal Consistency (Uncorrected α)	Internal Consistency (Corrected α)
comparisons	compare	greater, best, after	317	.08	.35
interrogatives	interrog	how, when, what	48	.18	.57
numbers	number	second, thousand	36	.45	.83
quantifiers	quant	few, many, much	77	.23	.64
psychological processes					
affective processes	affect	happy, cried	1393	.18	.57
Positive emotion	posemo	love, nice, sweet	620	.23	.64
Negative emotion	negemo	hurt, ugly, nasty	744	.17	.55
Anxiety	anx	worried, fearful	116	.31	.73
Anger	anger	hate, kill, annoyed	230	.16	.53
Sadness	sad	crying, grief, sad	136	.28	.70
Social processes	social	mate, talk, they	756	.51	.86
Family	family	daughter, dad, aunt	118	.55	.88
Friends	friend	buddy, neighbor	95	.20	.60
Female references	female	girl, her, mom	124	.53	.87
Male references	male	boy, his, dad	116	.52	.87
Cognitive processes	cogproc	cause, know, ought	797	.65	.92
Insight	insight	think, know	259	.47	.84

Continued on next page

Table B.1 Continued from previous page

Category	Abbrev	Examples	Words in category	Internal Consistency (Uncorrected α)	Internal Consistency (Corrected α)
Causation	cause	because, effect	135	.26	.67
Discrepancy	discrep	should, would	83	.34	.76
Tentative	tentat	maybe, perhaps	178	.44	.83
Certainty	certain	always, never	113	.31	.73
Differentiation	differ	hasn't, but, else	81	.38	.78
Perceptual processes	percept	look, heard, feeling	436	.17	.55
See	see	view, saw, seen	126	.46	.84
Hear	hear	listen, hearing	93	.27	.69
Feel	feel	feels, touch	128	.24	.65
Biological processes	bio	eat, blood, pain	748	.29	.71
Body	body	cheek, hands, spit	215	.52	.87
Health	health	clinic, flu, pill	294	.09	.37
Sexual	sexual	horny, love, incest	131	.37	.78
Ingestion	ingest	dish, eat, pizza	184	.67	.92
Drives	drives		1103	.39	.80
Affiliation	affiliation	ally, friend, social	248	.40	.80
Achievement	achieve	win, success, better	213	.41	.81
Power	power	superior, bully	518	.35	.76
Reward	reward	take, prize, benefit	120	.27	.69
Risk	risk	danger, doubt	103	.26	.68
Time orientations	TimeOrient				
Past focus	focuspast	ago, did, talked	341	.23	.64
Present focus	focuspresent	today, is, now	424	.24	.66
Future focus	focusfuture	may, will, soon	97	.26	.68
Relativity	relativ	area, bend, exit	974	.50	.86
Motion	motion	arrive, car, go	325	.36	.77

Continued on next page

Table B.1 Continued from previous page

Category	Abbrev	Examples	Words in category	Internal Consistency (Uncorrected α)	Internal Consistency (Corrected α)
Space	space	down, in, thin	360	.45	.83
Time	time	end, until, season	310	.39	.79
Personal concerns					
Work	work	job, majors, xerox	444	.69	.93
Leisure	leisure	cook, chat, movie	296	.50	.86
Home	home	kitchen, landlord	100	.46	.83
Money	money	audit, cash, owe	226	.60	.90
Religion	relig	altar, church	174	.64	.91
Death	death	bury, coffin, kill	74	.39	.79
Informal language	informal		380	.46	.84
Swear words	swear	fuck, damn, shit	131	.45	.83
Netspeak	netspeak	btw, lol, thx	209	.42	.82
Assent	assent	agree, OK, yes	36	.10	.39
Nonfluencies	nonflu	er, hm, umm	19	.27	.69
Fillers	filler	Imean, youknow	14	.06	.27

TABLE B.1: LIWC2015 Output Variable Information Combined (Pennebaker et al., 2015a)

Where "Uncorrected α" is Cronbach's α coefficient averaged over various development corpora, and the "Corrected α" is based on Spearman-Brown prediction. Unfortunately, some of these are not yet available in the Japanese version as explicated in (Igarashi, Okuda, and Sasahara, 2021). Most prominently, time-related words are difficult to analyse in Japanese due to its linguistic properties. Hence, its deployment in 2.5 suffers from these restrictions. Also, since the range of potential outcomes strongly depends on the medium and language of origin, we refer to the given ranges in the original manual (Pennebaker et al., 2015a) as well as the iterations discussed in the Japanese manual (Igarashi, Okuda, and Sasahara, 2021). The Japanese version also displays slightly different reported corrected and uncorrected consistencies, as well as a range of additional subcategories that are more relevant for linguistic than for economic analysis (e.g., "case particles", "adjective verbs", "pre-noun adjectivals"), wherefore we do not include those in the econometric analysis.

B.1.2 IBM-Watson Personality Insights

IBM-Watson Personality Insights is a commercial tool for linguistic analysis that is deprecated in the meantime (IBM, 2021). Based on behavioural artefacts from digital footprints, it uses an approach based on theory from linguistics, psychology, and marketing sciences, to infer personality characteristics as behavioural tendencies (Costa and McCrae, 1992; Norman, 1963), needs as behavioural drivers (Armstrong et al., 2014; Ford, 2005), and values as guiding principle for behaviours (Schwartz, 1992; Schwartz, 2006) of individuals. While the first iteration is based on LIWC, later versions use an open vocabulary GOFAI approach to create predictions.

With regards to personality (Costa and McCrae, 1992; Norman, 1963), it extracts:

1. **Openness:** Characteristics include inventiveness and curiosity. Individuals high in openness tend to be adventurous and creative.
2. **Conscientiousness:** Efficiency and organisation versus easy-going, careless behavior. High conscientiousness indicates self-discipline and a preference for planned behavior.
3. **Extraversion:** Denoted by sociability, assertiveness, and emotional expressiveness. High extraversion suggests an energetic approach towards the social and material world.
4. **Agreeableness:** A tendency towards being compassionate and cooperative rather than suspicious and antagonistic towards others.
5. **Neuroticism (Emotional Stability):** The tendency to experience unpleasant emotions easily. Low scores indicate emotional stability and resilience.

Thereby, for Japanese, it achieves an average MAE of 0.1 (facets: 0.12) and an average correlation to surveys of 0.3 (facets: 0.22). With regards to needs in resonance with individual personality (Armstrong et al., 2014; Ford, 2005), it extracts:

1. Challenge
2. Closeness
3. Curiosity
4. Excitement
5. Harmony
6. Ideal
7. Liberty
8. Love
9. Practicality
10. Self-expression
11. Stability
12. Structure

Thereby, for Japanese, it achieves an average MAE of 0.11 and an average correlation to surveys of 0.25. With regards to values (Schwartz, 1992; Schwartz, 2006), it extracts:

1. Self-transcendence / Helping others
2. Conservation / Tradition
3. Hedonism
4. Self-enhancement / Achieving success
5. Openness to change / Excitement

Thereby, for Japanese, it achieves an average MAE of 0.11 and an average correlation to surveys of 0.19.

However, these average correlations and MAE are based on a narrow development corpus, and later studies like ours and that of Giorgi et al. (2022) show that prediction accuracy differs dramatically depending on corpus and application. Since its main use case is econometric analysis for marketing, it additionally provides predictions for most likely consumption and consumer behaviour categories. Table B.2 provides an overview on all extracted variables.

Variable Name	Output Range
Big Five Personality Factors	
Openness Aggregate Score	[0 - 100]
Openness - adventurousness	[0 - 100]
Openness - artistic interests	[0 - 100]
Openness - emotionality	[0 - 100]
Openness - imagination	[0 - 100]
Openness - intellect	[0 - 100]
Openness - authority challenging	[0 - 100]
Conscientiousness Aggregate Score	[0 - 100]
Conscientiousness - achievement striving	[0 - 100]
Conscientiousness - cautiousness	[0 - 100]
Conscientiousness - dutifulness	[0 - 100]
Conscientiousness - orderliness	[0 - 100]
Conscientiousness - self discipline	[0 - 100]
Conscientiousness - self efficacy	[0 - 100]
Extraversion Aggregate Score	[0 - 100]
Extraversion - activity level	[0 - 100]
Extraversion - assertiveness	[0 - 100]
Extraversion - cheerfulness	[0 - 100]
Extraversion - excitement seeking	[0 - 100]
Extraversion - outgoing	[0 - 100]
Extraversion - gregariousness	[0 - 100]
Agreeableness Aggregate Score	[0 - 100]
Agreeableness - altruism	[0 - 100]
Agreeableness - cooperation	[0 - 100]
Agreeableness - modesty	[0 - 100]
Agreeableness - uncompromising	[0 - 100]

Continued on next page

Table B.2 – continued from previous page

Variable Name	Output Range
Agreeableness - sympathy	[0 - 100]
Agreeableness - trust	[0 - 100]
Neuroticism Aggregate Score	[0 - 100]
Neuroticism - fiery	[0 - 100]
Neuroticism - prone to worry	[0 - 100]
Neuroticism - melancholy	[0 - 100]
Neuroticism - immoderation	[0 - 100]
Neuroticism - self consciousness	[0 - 100]
Neuroticism - susceptible to stress	[0 - 100]
Needs	
Challenge	[0 - 100]
Closeness	[0 - 100]
Curiosity	[0 - 100]
Excitement	[0 - 100]
Harmony	[0 - 100]
Ideal	[0 - 100]
Liberty	[0 - 100]
Love	[0 - 100]
Practicality	[0 - 100]
Self-expression	[0 - 100]
Stability	[0 - 100]
Structure	[0 - 100]
Values	
Conservation	[0 - 100]
Openness to change	[0 - 100]
Hedonism	[0 - 100]
Self-enhancement	[0 - 100]
Self-transcendence	[0 - 100]
Predicted Consumption Patterns	
Likely to be sensitive to ownership cost when buying automobiles	[0, 100]
Likely to prefer safety when buying automobiles	[0, 100]
Likely to prefer quality when buying clothes	[0, 100]
Likely to prefer style when buying clothes	[0, 100]
Likely to prefer comfort when buying clothes	[0, 100]
Likely to be influenced by brand name when making product purchases	[0, 100]
Likely to be influenced by product utility when making product purchases	[0, 100]
Likely to be influenced by online ads when making product purchases	[0, 100]
Likely to be influenced by social media when making product purchases	[0, 100]
Likely to be influenced by family when making product purchases	[0, 100]
Likely to indulge in spur of the moment purchases	[0, 100]
Likely to prefer using credit cards for shopping	[0, 100]
Likely to eat out frequently	[0, 100]
Likely to have a gym membership	[0, 100]
Likely to like outdoor activities	[0, 100]

Continued on next page

Table B.2 – continued from previous page

Variable Name	Output Range
Likely to be concerned about the environment	[0, 100]
Likely to consider starting a business in next few years	[0, 100]
Likely to like romance movies	[0, 100]
Likely to like adventure movies	[0, 100]
Likely to like horror movies	[0, 100]
Likely to like musical movies	[0, 100]
Likely to like historical movies	[0, 100]
Likely to like science-fiction movies	[0, 100]
Likely to like war movies	[0, 100]
Likely to like drama movies	[0, 100]
Likely to like action movies	[0, 100]
Likely to like documentary movies	[0, 100]
Likely to like rap music	[0, 100]
Likely to like country music	[0, 100]
Likely to like R&B music	[0, 100]
Likely to like hip hop music	[0, 100]
Likely to attend live musical events	[0, 100]
Likely to have experience playing music	[0, 100]
Likely to like Latin music	[0, 100]
Likely to like rock music	[0, 100]
Likely to like classical music	[0, 100]
Likely to read often	[0, 100]
Likely to read entertainment magazines	[0, 100]
Likely to read non-fiction books	[0, 100]
Likely to read financial investment books	[0, 100]
Likely to read autobiographical books	[0, 100]
Likely to volunteer for social causes	[0, 100]

TABLE B.2: IBM-Watson Personality Insights variable names and output ranges

Whereby all psychological variables follow a Gaussian with mean of 50 and SD of 30, and all consumption predictions are presented as binary variables of 0 and 100 (IBM, 2021). It is unclear how the translation is done between languages, and whether maybe only one languagae was developed and others translated, or whether each language was developed individually. We use it for section 2.3 on spatial econometrics, and section 2.4 on temporal econometrics.

B.1.3 TIPI

The Japanese version of the Ten Items Personality Inventory (TIPI) is a sparse and concise personality instrument meant for mass-deployment (Oshio et al., 2013). It directly measures each of the Big Five personality factors described above with just two items, of which one is reversely scored. It uses a 7-point Likert scale and assumes a normally distributed population, hence the expected mean outcome is a 4, and the expected SD of 1. We use it for the introductory chapter 1 on split personality, as well as for spatial econometrics in chapter 2.3 and spatiotemporal econometrics in chapter 2.5.

B.1.4 Ground-truth data

We use official COVID-19 data provided by the Japanese government (MHLW, 2021) that we cross-checked with data from the World Health Organisation (WHO, 2021) to ensure their correctness. This is comprised of:

- deaths: total number per day
- hospitalisations/ severe cases: total number per day
- vaccination numbers: total number per day

This data is used in the section on temporal econometrics 2.4 and spatiotemporal econometrics 2.5.

B.2 Study-Specific Variables

B.2.1 Author Profiling Studies

In the author profiling study of section 2.3, we use a variety of different open-vocabulary GOFAI features that are manually engineered. More precisely, we create language-based features that are described in table 2.1.

- LIWC – as described above
- Words and phrases: 6,627 words and "phrases" of up to three words; basically comprising cuts from the entire text. "Phrases" in this regard are a special case of n-grams that have been used sufficiently often, and offer sufficient information-theoretic insights.
- N-Grams: uni-, bi-, and tri-grams based on sliding windows over texts (e.g., "I go to university" becomes "I", "go", "to", "university" as unigrams, and "I go", "go to", "to university" as bigrams)
- Topics: created using Latent Dirichlet Allocation to identify topics, which are probability distributions over words as described in Blei, Ng, and Jordan (2003). In essence, some words are replaceable (e.g., "hot" with "warm" without altering too much of the idea. By using simplices, similar meaningful distributions over words are identified, and used as features.
- General message characteristics: lengths of message, attached media files and contact cards, frequency of emoji/ emoticon and their range of overall usage.
- Emoji preferences: use of emojis that have been overall used by at least 5% of all participants.

Given the vast amount of features, explaining each of those would be far longer than the entire dissertation, hence we rather provide this overview for understanding language features. In essence, we create synthetic variables based on text, and use them as features for a machine learning process to predict outcome variables in line with the research question.

B.2.2 Spatiotemporal Econometrics

NEO-FFI

The NEO Five-Factor Inventory (NEO-FFI) (Costa and McCrae, 1992) is a comprehensive, research-oriented, commercial instrument for measuring personality based on the Five Factor Model. It offers six facets to each factor (asked in two items each, whereby one item is reversely scored) to provide a more detailed perspective on an individual's personality:

1. **Neuroticism**
    - **Anxiety**: Reflects the tendency to experience feelings of nervousness, tension, and worry.
    - **Depression**: Measures the propensity for experiencing feelings of sadness, hopelessness, and lack of motivation.
    - **Self-Consciousness**: Refers to the degree of self-awareness and sensitivity to social evaluation.
    - **Hostility**: Captures the inclination towards anger, irritability, and aggression.
    - **Impulsiveness**: Assesses the tendency to act on impulses and display poor restraint.
    - **Vulnerability**: Reflects susceptibility to stress and emotional instability.
2. **Extraversion**
    - **Warmth**: Reflects friendliness, approachability, and empathy towards others.
    - **Gregariousness**: Measures the preference for socialising and enjoying the company of others.
    - **Assertiveness**: Indicates the tendency to take charge of situations and express oneself confidently.
    - **Activity**: Assesses the level of energy and vigour in pursuing activities and interests.
    - **Excitement-Seeking**: Captures the inclination towards seeking stimulation and excitement.
    - **Positive Emotions**: Reflects the frequency and intensity of experiencing positive mood states.
3. **Openness to Experience**
    - **Fantasy**: Measures the tendency to engage in imaginative and creative thinking.
    - **Aesthetics**: Reflects appreciation for art, beauty, and unconventional ideas.
    - **Feelings**: Assesses emotional depth, sensitivity, and receptivity to inner experiences.
    - **Actions**: Indicates the inclination towards seeking out new experiences and variety in life.

- **Ideas**: Captures intellectual curiosity, openness to new concepts, and willingness to entertain unconventional beliefs.
- **Values**: Reflects openness to diverse values and perspectives.

4. **Agreeableness**
   - **Trust**: Measures the general belief in the sincerity and honesty of others.
   - **Straightforwardness**: Reflects sincerity, honesty, and directness in communication.
   - **Altruism**: Indicates concern for the welfare of others and willingness to help.
   - **Compliance**: Assesses the tendency to be cooperative, agreeable, and non-confrontational.
   - **Modesty**: Reflects humility, modesty, and lack of self-centredness.
   - **Tender-Mindedness**: Captures sensitivity to the needs and feelings of others.
5. **Conscientiousness**
   - **Competence**: Reflects confidence in one's abilities and the tendency to strive for mastery.
   - **Order**: Measures the preference for organisation, structure, and tidiness in one's environment.
   - **Dutifulness**: Indicates a sense of responsibility, duty, and obligation towards others.
   - **Achievement-Striving**: Assesses the motivation to set and pursue ambitious goals.
   - **Self-Discipline**: Reflects the ability to control impulses, stay focused, and persist in tasks.
   - **Deliberation**: Captures the tendency to think carefully and consider alternatives before making decisions.

Based on a 5-point Likert scale, it assumes an underlying Gaussian distribution in the population, hence expects a mean of 3 and a SD of 1.

Keio Survey

The Keio Survey is described in section 2.5 and uses a 5-point Likert scale. It is based on psychological theory (Rust, Kosinski, and Stillwell, 2020) and separates items in cognitive, emotional, and behavioural aspects, of which we do neither expect nor observe a Gaussian distribution, but one being influenced by agent-internal and external factors.

Bibliography

Abril-Pla, Oriol (2022). *xarray-einstats*. URL: https://github.com/arviz-devs/xarray-einstats.

Ahmed, Akif et al. (Dec. 22, 2020). "Can COVID-19 Change the Big5 Personality Traits of Healthcare Workers?" In: *7th International Conference on Networking, Systems and Security*. 7th NSysS 2020: 7th International Conference on Networking, Systems and Security. Dhaka Bangladesh: ACM, pp. 12–17. ISBN: 978-1-4503-8905-1. DOI: 10.1145/3428363.3428370. URL: https://dl.acm.org/doi/10.1145/3428363.3428370 (visited on 07/12/2021).

Ahmed, Amr, Liangjie Hong, and Alexander J. Smola (May 13, 2013). "Hierarchical geographical modeling of user locations from social media posts". In: *Proceedings of the 22nd international conference on World Wide Web*. WWW '13. New York, NY, USA: Association for Computing Machinery, pp. 25–36. ISBN: 978-1-4503-2035-1. DOI: 10.1145/2488388.2488392. URL: https://dl.acm.org/doi/10.1145/2488388.2488392 (visited on 11/22/2023).

Aluja, Anton et al. (2005). "Comparison of the NEO-FFI, the NEO-FFI-R and an alternative short version of the NEO-PI-R (NEO-60) in Swiss and Spanish samples". In: *Personality and Individual Differences* 38.3, pp. 591–604.

An, J. et al. (2018). "Factors Influencing Emoji Usage in Smartphone Mediated Communications". In.

Apley, D. W. and J. Zhu (2020). "Visualizing the effects of predictor variables in black box supervised learning models". In: *Journal of the Royal Statistical Society Series B* 82.4, pp. 1059–1086.

Araujo, Theo and Toni GLA van der Meer (May 1, 2020). "News values on social media: Exploring what drives peaks in user activity about organizations on Twitter". In: *Journalism* 21.5. Publisher: SAGE Publications, pp. 633–651. ISSN: 1464-8849. DOI: 10.1177/1464884918809299. URL: https://doi.org/10.1177/1464884918809299 (visited on 07/18/2021).

Argamon, S. et al. (2007). "Mining the blogosphere: Age, gender and the varieties of self-expression". In: *First Monday* 12.9.

Armstrong, Gary et al. (2014). *Principles of marketing*. Pearson Australia.

Ashton, Michael C and Kibeom Lee (2009). "The HEXACO–60: A short measure of the major dimensions of personality". In: *Journal of personality assessment* 91.4, pp. 340–345.

Atari, Mohammad et al. (Sept. 22, 2023). *Which Humans?* DOI: 10.31234/osf.io/5b26t. URL: https://psyarxiv.com/5b26t/ (visited on 09/29/2023).

Bai, Q. et al. (2019). "A Systematic Review of Emoji: Current Research and Future Perspectives". In: *Frontiers in Psychology* 10. DOI: 10.3389/fpsyg.2019.02221.

Bamman, D., J. Eisenstein, and T. Schnoebelen (2014). "Gender identity and lexical variation in social media". In: *Journal of Sociolinguistics* 18.2, pp. 135–160. DOI: 10.1111/josl.12080.

Bardi, Anat et al. (2009). "The structure of intraindividual value change." In: *Journal of personality and social psychology* 97.5, p. 913.

Bartram, Dave, Ivan T Robertson, and Militza Callinan (2002). "Introduction: A framework for examining organizational effectiveness". In: *Organizational effectiveness: The role of psychology*, pp. 1–10.

Bayer, Joseph B, Penny Trieu, and Nicole B Ellison (2020). "Social media elements, ecologies, and effects". In: *Annual review of psychology* 71, pp. 471–497.

Bazarova, N. N. and Y. H. Choi (2014). "Self-Disclosure in Social Media: Extending the Functional Approach to Disclosure Motivations and Characteristics on Social Network Sites". In: *Journal of Communication* 64.4, pp. 635–657. DOI: 10.1111/jcom.12106.

Bazarova, N. N. et al. (2013). "Managing Impressions and Relationships on Facebook: Self-Presentational and Relational Concerns Revealed Through the Analysis of Language Style". In: *Journal of Language and Social Psychology* 32.2, pp. 121–141. DOI: 10.1177/0261927X12456384.

Benoit, Kenneth et al. (2018). "quanteda: An R package for the quantitative analysis of textual data". In: *Journal of Open Source Software* 3.30, p. 774. DOI: 10.21105/joss.00774. URL: https://quanteda.io.

Betsch, Cornelia et al. (2012). "Social norms and vaccination decisions". In: *The Lancet* 379.9829, pp. 1855–1856.

Biever, Celeste (July 2023). "ChatGPT broke the Turing test — the race is on for new ways to assess AI". en. In: *Nature* 619.7971, pp. 686–689. DOI: 10.1038/d41586-023-02361-7. URL: https://www.nature.com/articles/d41586-023-02361-7 (visited on 09/04/2023).

Binder, M. et al. (2020). *mlrCPO: Composable Preprocessing Operators and Pipelines for Machine Learning (0.3.6) [Computer software]*. URL: https://CRAN.R-project.org/package=mlrCPO.

Bischl, B. et al. (2012). "Resampling Methods for Meta-Model Validation with Recommendations for Evolutionary Computation". In: *Evolutionary Computation* 20.2, pp. 249–275. DOI: 10.1162/EVCO_a_00069.

Blei, David M and John D Lafferty (2009). "Topic models". In: *Text Mining: Theory and Applications*. Taylor and Francis, pp. 71–93.

Blei, David M, Andrew Y Ng, and Michael I Jordan (2003). "Latent dirichlet allocation". In: *Journal of machine Learning research* 3.Jan, pp. 993–1022.

Bleidorn, Wiebke et al. (2021). "Personality trait stability and change". In: *Personality Science* 2, e6009. ISSN: 2700-0710. DOI: 10.5964/ps.6009. URL: https://ps.psychopen.eu/index.php/ps/article/view/6009 (visited on 07/16/2021).

Borgatti, Stephen P., Martin G. Everett, and Jeffrey C. Johnson (2021). *Analyzing Social Networks*. SAGE Publications.

Borgholthaus, Cameron J., Joshua V. White, and Peter D. Harms (Feb. 1, 2023). "CEO dark personality: A critical review, bibliometric analysis, and research agenda". In: *Personality and Individual Differences* 201, p. 111951. ISSN: 0191-8869. DOI: 10.1016/j.paid.2022.111951. URL: https://www.sciencedirect.com/science/article/pii/S0191886922004561 (visited on 11/27/2023).

Bossche, Joris Van den et al. (2020). *GeoPandas*. https://geopandas.org. Version latest.

Bouchard Jr, Thomas J (1994). "Genes, environment, and personality". In: *Science* 264.5166, pp. 1700–1701.

Breiman, L. (2001). "Random forests". In: *Machine Learning* 45.1, pp. 5–32.

Brennen, J Scott et al. (2020). "Types, sources, and claims of COVID-19 misinformation". In: *Reuters Institute*.

Bronfenbrenner, Urie (2013). "Bronfenbrenner's ecological systems theory". In: *The Psychology Notes HQ Online Resources for Psychology Students*.

Brooks, James et al. (Feb. 24, 2021). "Uniting against a common enemy: Perceived outgroup threat elicits ingroup cohesion in chimpanzees". In: *PLOS ONE* 16.2. Publisher: Public Library of Science, e0246869. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0246869. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0246869 (visited on 07/19/2021).

Brown, Tom et al. (2020). "Language models are few-shot learners". In: *Advances in neural information processing systems* 33, pp. 1877–1901.

Bryant, C (2018). *Researchpy*.

Bucknall, Vittoria et al. (Dec. 8, 2015). "Mirror mirror on the ward, who's the most narcissistic of them all? Pathologic personality traits in health care". In: *CMAJ : Canadian Medical Association Journal* 187.18, pp. 1359–1363. ISSN: 0820-3946. DOI: 10.1503/cmaj.151135. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4674404/ (visited on 11/27/2023).

Burger, J. D. et al. (2011). "Discriminating Gender on Twitter". In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 1301–1309.

Butterworth, S. E. et al. (2019). "Sender Gender Influences Emoji Interpretation in Text Messages". In: *Frontiers in Psychology* 10. DOI: 10.3389/fpsyg.2019.00784.

Bègue, Laurent et al. (2015). "Personality Predicts Obedience in a Milgram Paradigm". In: *Journal of Personality* 83.3, pp. 299–306. ISSN: 1467-6494. DOI: 10.1111/jopy.12104. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/jopy.12104 (visited on 07/18/2021).

Calma, Justine (Sept. 26, 2023). *Microsoft is going nuclear to power its AI ambitions*. The Verge. URL: https://www.theverge.com/2023/9/26/23889956/microsoft-next-generation-nuclear-energy-smr-job-hiring (visited on 11/28/2023).

Catalinac, A. and K. Watanabe (2019). "Nihongo no Ryoteki Tekisuto Bunseki ("Quantitative Text Analysis in Japanese")". In: *Waseda Institute for Advanced Study Research Bulletin* 11.

Chandra, Noirrit Kiran, David B. Dunson, and Jason Xu (2023). "Inferring Covariance Structure from Multiple Data Sources via Subspace Factor Analysis". In: *arXiv:2305.04113*. DOI: 10.48550/arxiv.2305.04113.

Chauhan, Nishant and Byung-Jae Choi (2023). "Classification of Alzheimer's Disease Using Maximal Information Coefficient-Based Functional Connectivity with an Extreme Learning Machine". In: *Brain Science* 13.7, p. 1046. DOI: 10.3390/brainsci13071046.

Chen, Z. et al. (2018). "Through a gender lens: Learning usage patterns of emojis from large-scale Android users". In: *Proceedings of the 2018 World Wide Web Conference*, pp. 763–772.

Chou, Wen-Ying Sylvia, Anna Gaysynsky, and Robin C Vanderpool (2020). "Media and Misinformation in the Time of COVID-19". In: *Journal of Health Communication* 25.10, pp. 760–763.

Chow, Andrew R. (Feb. 2023). *How ChatGPT Managed to Grow Faster Than TikTok or Instagram*. en. URL: https://time.com/6253615/chatgpt-fastest-growing/ (visited on 09/04/2023).

Church, K. W. and P. Hanks (1990). "Word Association Norms, Mutual Information, and Lexicography". In: *Computational Linguistics* 16.1, pp. 22–29.

Costa, Paul T and Robert R McCrae (1985). "The NEO Personality Inventory manual". In.

— (2008). "The Revised NEO Personality Inventory (NEO-PI-R)". In: *The SAGE handbook of personality theory and assessment* 2, pp. 179–198.

Costa Jr., Paul T and Robert R McCrae (1992). *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI): Professional Manual*. Psychological Assessment Resources. Odessa, FL.

Curran, Patrick J. (Oct. 1, 2003). "Have Multilevel Models Been Structural Equation Models All Along?" In: *Multivariate Behavioral Research* 38.4, pp. 529–569. ISSN: 0027-3171. DOI: 10.1207/s15327906mbr3804_5. URL: https://doi.org/10.1207/s15327906mbr3804_5 (visited on 07/18/2021).

Dawans, Bernadette von et al. (2012-06-01). "The Social Dimension of Stress Reactivity: Acute Stress Increases Prosocial Behavior in Humans". In: *Psychological Science* 23.6. Publisher: SAGE Publications Inc, pp. 651–660. ISSN: 0956-7976. DOI: 10.1177/0956797611431576. URL: https://doi.org/10.1177/0956797611431576 (visited on 07/18/2021).

Deary, Ian J. (2018). "The Psychology of Language: A Critical Introduction". In: *Psychological Bulletin*.

DiBlasi, Johnny et al. (2020). "Agency & Autonomy: Intersections of Artificial Intelligence and Creative Practice". In: *International Symposium on Electronic Art (2020)*. Vol. 7.

Diener, ED et al. (1985). "The satisfaction with life scale". In: *Journal of personality assessment* 49.1, pp. 71–75.

Diener, Ed et al. (2010). "New well-being measures: Short scales to assess flourishing and positive and negative feelings". In: *Social indicators research* 97, pp. 143–156.

Digutsch, Jan and Michal Kosinski (2022). *Overlap in Meaning Is a Stronger Predictor of Semantic Activation in GPT-3 Than in Humans*. en. preprint. PsyArXiv. DOI: 10.31234/osf.io/dx5hc. URL: https://osf.io/dx5hc (visited on 01/25/2023).

Ding, Ruixue et al. (July 18, 2023). "MGeo: Multi-Modal Geographic Language Model Pre-Training". In: *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '23. New York, NY, USA: Association for Computing Machinery, pp. 185–194. ISBN: 978-1-4503-9408-6. DOI: 10.1145/3539618.3591728. URL: https://dl.acm.org/doi/10.1145/3539618.3591728 (visited on 11/22/2023).

Doreian, Patrick and Norman Conti (2012). "Social context, spatial structure and social network structure". In: *Social Networks* 34.1, pp. 32–46. ISSN: 0378-8733. DOI: https://doi.org/10.1016/j.socnet.2010.09.002. URL: https://www.sciencedirect.com/science/article/pii/S037887331000047X.

Douglas, Karen M. et al. (2019). "Understanding Conspiracy Theories". In: *Political Psychology* 40 (S1), pp. 3–35. ISSN: 1467-9221. DOI: 10.1111/pops.12568. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/pops.12568 (visited on 11/12/2023).

Dröes, Martijn I and Hans RA Koster (2016). "Renewable energy and negative externalities: The effect of wind turbines on house prices". In: *Journal of Urban Economics* 96, pp. 121–141.

Duckworth, Angela L. and David S. Yeager (2020). "Language Ability and Socioeconomic Success: A Review of the Literature and the Role of Emotional Intelligence". In: *Journal of Personality*.

Eichstaedt, J. C. et al. (2020). "Closed- and Open-Vocabulary Approaches to Text Analysis: A Review, Quantitative Comparison, and Recommendations". In: *Psychological Methods*.

Elleman, Lorien G. et al. (2018). "The personality of U.S. states: Stability from 1999 to 2015". In: *Journal of Research in Personality* 72, pp. 64–72. ISSN: 0092-6566. DOI: 10.1016/J.JRP.2016.06.022.

Evans, D. (2020). *Why the US government is questioning WhatsApp's encryption*. URL: https://www.cnbc.com/2020/02/21/whatsapp-encryption-under-scrutiny-by-us-government.html.

Fabes, R. A. and C. L. Martin (1991). "Gender and age stereotypes of emotionality". In: *Personality and Social Psychology Bulletin* 17.5, pp. 532–540.

Faisal, Fahim and Antonios Anastasopoulos (Dec. 20, 2022). *Geographic and Geopolitical Biases of Language Models*. DOI: 10.48550/arXiv.2212.10408. arXiv: 2212.10408[cs]. URL: http://arxiv.org/abs/2212.10408 (visited on 11/23/2023).

Feinerer, I., K. Hornik, and D. Meyer (2008). "Text Mining Infrastructure in R". In: *Journal of Statistical Software* 25.5. DOI: 10.18637/jss.v025.i05.

Ferrucci, David A (2012). "Introduction to "this is watson"". In: *IBM Journal of Research and Development* 56.3.4, pp. 1–1.

Finau, Glen et al. (July 2018). "Social media and disaster communication: A case study of cyclone Winston". In: *Pacific Journalism Review*. URL: https://search.informit.org/doi/abs/10.3316/informit.738398775561009 (visited on 07/18/2021).

Flaxman, Abe and Benjamin T. Vincent (2022). "<notebook title>". In: *PyMC examples*. Ed. by PyMC Team.

Fone, David et al. (Mar. 2016). "Change in alcohol outlet density and alcohol-related harm to population health (CHALICE): a comprehensive record-linked database study in Wales". In: *Public Health Research* 4.3, pp. 1–184. ISSN: 2050-4381, 2050-439X. DOI: 10.3310/phr04030. URL: https://www.journalslibrary.nihr.ac.uk/phr/phr04030/ (visited on 04/22/2019).

Ford, Kevin (2005). *Brands laid bare: Using market research for evidence-based brand management*. John Wiley & Sons.

Forman, Rebecca et al. (2021). "COVID-19 vaccine challenges: What have we learned so far and what remains to be done?" In: *Health policy* 125.5, pp. 553–567.

Fox, Jean-Paul (2010). *Bayesian item response modeling: Theory and applications*. Springer Science & Business Media.

Frieder, Simon et al. (2023). *Mathematical Capabilities of ChatGPT*. DOI: 10.48550/arXiv.2301.13867. arXiv: 2301.13867[cs]. URL: http://arxiv.org/abs/2301.13867 (visited on 02/18/2023).

Friedman, J., T. Hastie, and R. Tibshirani (2010). "Regularization paths for generalized linear models via coordinate descent". In: *Journal of Statistical Software* 33.1, p. 1.

Friedman, Mike and Erica Carlisle (2022). *TIPI French*. URL: https://gosling.psy.utexas.edu/scales-weve-developed/ten-item-personality-measure-tipi/ (visited on 06/01/2022).

Fullwood, C., L. J. Orchard, and S. A. Floyd (2013). "Emoticon convergence in Internet chat rooms". In: *Social Semiotics* 23.5, pp. 648–662. DOI: 10.1080/10350330.2012.739000.

Fávero, Luiz Paulo, Patrícia Belfiore, and Rafael de Freitas Souza (2023). "Chapter 12 - Principal component factor analysis". In: *Data Science, Analytics and Machine Learning with R*. Ed. by Luiz Paulo Fávero, Patrícia Belfiore, and Rafael de Freitas Souza. Academic Press, pp. 203–214. ISBN: 978-0-12-824271-1. DOI: https://doi.org/10.1016/B978-0-12-824271-1.00006-8. URL: https://www.sciencedirect.com/science/article/pii/B9780128242711000068.

Giancola, Marco, Massimiliano Palmiero, and Simonetta D'Amico (2023). "Dark Triad and COVID-19 vaccine hesitancy: the role of conspiracy beliefs and risk perception". In: *Current Psychology*. ISSN: 1936-4733. DOI: 10.1007/s12144-023-

04609-x. URL: https://doi.org/10.1007/s12144-023-04609-x (visited on 11/12/2023).

Giorgi, Salvatore et al. (2022). "Regional personality assessment through social media language". In: *Journal of Personality* 90.3, pp. 405–425. ISSN: 0022-3506, 1467-6494. DOI: 10.1111/jopy.12674. URL: https://onlinelibrary.wiley.com/doi/10.1111/jopy.12674 (visited on 06/29/2022).

Goode, L. (2019). *Private Messages Are the New (Old) Social Network | WIRED*. URL: https://www.wired.com/story/private-messages-new-social-networks/.

Goodin, D. (2021). *WhatsApp gives users an ultimatum: Share data with Facebook or stop using the app*. URL: https://arstechnica.com/tech-policy/2021/01/whatsapp-users-must-share-their-data-with-facebook-or-stop-using-the-app/.

Gorbunova, Vera et al. (Dec. 15, 2007). "Changes in DNA repair during aging". In: *Nucleic Acids Research* 35.22, pp. 7466–7474. ISSN: 0305-1048. DOI: 10.1093/nar/gkm756. URL: https://doi.org/10.1093/nar/gkm756 (visited on 07/18/2021).

Gosling, Samuel D, Peter J Rentfrow, and William B Swann (2003). "A very brief measure of the Big-Five personality domains". In: *Journal of Research in Personality* 37.6, pp. 504–528. ISSN: 00926566. DOI: 10.1016/S0092-6566(03)00046-1. URL: https://linkinghub.elsevier.com/retrieve/pii/S0092656603000461 (visited on 06/01/2022).

Gosling, Samuel D. et al. (2002). "A room with a cue: Personality judgments based on offices and bedrooms." In: *Journal of Personality and Social Psychology* 82.3, pp. 379–398. ISSN: 1939-1315, 0022-3514. DOI: 10.1037/0022-3514.82.3.379. URL: http://doi.apa.org/getdoi.cfm?doi=10.1037/0022-3514.82.3.379 (visited on 07/19/2021).

Götz, Friedrich M., Tobias Ebert, and Peter J. Rentfrow (2018). "Regional cultures and the psychological geography of Switzerland: Person–environment–fit in personality predicts subjective wellbeing". In: *Frontiers in Psychology* 9. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2018.00517. URL: http://journal.frontiersin.org/article/10.3389/fpsyg.2018.00517/full.

Government, Tokyo Metropolitan (2020). *Tokyo Metropolitan Government Disaster Prevention Guide Book*. URL: https://www.bousai.metro.tokyo.lg.jp/content/book/guidbook_pocketguide/2020guid_e.pdf (visited on 09/08/2021).

Griffiths, Thomas L and Mark Steyvers (2004). "Finding scientific topics". In: *Proceedings of the National Academy of Sciences* 101.suppl 1, pp. 5228–5235.

Grün, B. and K. Hornik (2011). "topicmodels: An R Package for Fitting Topic Models". In: *Journal of Statistical Software* 40.13.

Guenole, Nigel, Sheri Feinzig, and Jonathan Ferrar (2015). *Employee privacy preferences in the world's major economies*. en.

Guntuku, Sharath Chandra et al. (July 6, 2019). "Studying Cultural Differences in Emoji Usage across the East and the West". In: *Proceedings of the International AAAI Conference on Web and Social Media* 13, pp. 226–235. ISSN: 2334-0770. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/3224 (visited on 07/18/2021).

Gurnee, Wes and Max Tegmark (Oct. 3, 2023). *Language Models Represent Space and Time*. arXiv.org. URL: https://arxiv.org/abs/2310.02207v1 (visited on 11/23/2023).

Ha, Shang E. (2022). *TIPI Korean*. URL: https://gosling.psy.utexas.edu/scales-weve-developed/ten-item-personality-measure-tipi/ (visited on 06/01/2022).

Hagendorff, Thilo, Sarah Fabi, and Michal Kosinski (2022). "Uncovering human-like intuitive decision-making in GPT-3.5". en. In.

Harcup, Tony and Deirdre O'Neill (2017). "What is news? News values revisited (again)". In: *Journalism studies* 18.12, pp. 1470–1488.

Harris, Charles R. et al. (Sept. 2020). "Array programming with NumPy". In: *Nature* 585.7825, pp. 357–362. DOI: 10.1038/s41586-020-2649-2. URL: https://doi.org/10.1038/s41586-020-2649-2.

Hawkins, Jeff (2021). *A thousand brains: A new theory of intelligence*. Basic Books.

Hecht, Erin E. (2018). "Plasticity, innateness, and the path to language in the primate brain: Comparing macaque, chimpanzee and human circuitry for visuomotor integration". In: *Interaction Studies* 19.1-2, pp. 370–387. DOI: 10.1075/IS.17039.HEC.

Heckert, N Alan et al. (2002). "Handbook 151: NIST/SEMATECH e-handbook of statistical methods". In: URL: http://www.itl.nist.gov/div898/handbook/ (visited on 01/03/2023).

Hernández Orallo, José (2017). *The measure of all minds: evaluating natural and artificial intelligence*. en. Cambridge, United Kingdom ; New York, NY: Cambridge University Press. ISBN: 978-1-107-15301-1.

Herring, S. C. and A. R. Dainas (2020). "Gender and Age Influences on Interpretation of Emoji Functions". In: *ACM Transactions on Social Computing* 3.2, pp. 1–26. DOI: 10.1145/3375629.

Hino, Kimihiro and Yasushi Asami (May 1, 2021). "Change in walking steps and association with built environments during the COVID-19 state of emergency: A longitudinal comparison with the first half of 2019 in Yokohama, Japan". In: *Health & Place* 69, p. 102544. ISSN: 1353-8292. DOI: 10.1016/j.healthplace.2021.102544. URL: https://www.sciencedirect.com/science/article/pii/S135382922100040X (visited on 07/18/2021).

Hofmann, Valentin et al. (Jan. 1, 2023). *Geographic Adaptation of Pretrained Language Models*. DOI: 10.48550/arXiv.2203.08565. arXiv: 2203.08565[cs]. URL: http://arxiv.org/abs/2203.08565 (visited on 11/23/2023).

Hofstede, Geert (2007). "A European in Asia†". In: *Asian Journal of Social Psychology* 10.1, pp. 16–21. ISSN: 1467-839X. DOI: 10.1111/j.1467-839X.2006.00206.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-839X.2006.00206.x (visited on 06/29/2022).

Hong, Liangjie et al. (Apr. 16, 2012). "Discovering geographical topics in the twitter stream". In: *Proceedings of the 21st international conference on World Wide Web*. WWW '12. New York, NY, USA: Association for Computing Machinery, pp. 769–778. ISBN: 978-1-4503-1229-5. DOI: 10.1145/2187836.2187940. URL: https://dl.acm.org/doi/10.1145/2187836.2187940 (visited on 11/22/2023).

Houghton, James, Michael Siegel, and Daniel Goldsmith (2013). "Modeling the Influence of Narratives on Collective Behavior". In: p. 24.

Howard, Matt C. (May 1, 2022). "The good, the bad, and the neutral: Vaccine hesitancy mediates the relations of Psychological Capital, the Dark Triad, and the Big Five with vaccination willingness and behaviors". In: *Personality and Individual Differences* 190, p. 111523. ISSN: 0191-8869. DOI: 10.1016/j.paid.2022.111523. URL: https://www.sciencedirect.com/science/article/pii/S0191886922000265 (visited on 11/12/2023).

Hunter, John D. et al. (2020). *Matplotlib: Visualization with Python*. https://matplotlib.org. Version latest.

IBM (2021). *IBM Watson Personality Insights - The science behind the service*. URL: https://cloud.ibm.com/docs/cloud.ibm.com/docs/personality-insights (visited on 07/18/2021).

Igarashi, Tasuku, Shimpei Okuda, and Kazutoshi Sasahara (Aug. 2021). "Development of the Japanese Version of the Linguistic Inquiry and Word Count Dictionary 2015 (J-LIWC2015)". In: DOI: 10.31234/osf.io/5hq7d. URL: https://doi.org/10.31234/osf.io/5hq7d.

Irfani, Suroosh (1978). "Extraversion-introversion and self-rated academic success". In: *Psychological Reports* 43.2, pp. 508–510.

Ishii, Masahiko and Etsu Onuma (1998). "Vocabulary Survey of Television Broadcasting". In: *National Institute for Japanese Language and Language 50th Anniversary Research Presentation Material Collection: Let's Walk the World of Japanese*, pp. 183–188.

Jaeger, S. et al. (2017). "Emoji questionnaires can be used with a range of population segments: Findings relating to age, gender and frequency of emoji/emoticon use". In: *Food Quality and Preference* 68. DOI: 10.1016/j.foodqual.2017.12.011.

Jaegher, Kris De (2021). "Common-Enemy Effects: Multidisciplinary Antecedents and Economic Perspectives". In: *Journal of Economic Surveys* 35.1. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/joes.12394, pp. 3–33. ISSN: 1467-6419. DOI: 10.1111/joes.12394. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/joes.12394 (visited on 07/19/2021).

Jaidka, K., S. C. Guntuku, and L. H. Ungar (2018). "Facebook versus Twitter: Cross-platform Differences in Self-Disclosure and Trait Prediction". In: *Twelfth International AAAI Conference on Web and Social Media*.

Jaidka, Kokil et al. (2020). "Estimating geographic subjective well-being from Twitter: A comparison of dictionary and data-driven language methods". In: *Proceedings of the National Academy of Sciences* 117.19, pp. 10165–10171.

Jiang, Guangyuan et al. (2022). *MPI: Evaluating and Inducing Personality in Pre-trained Language Models*. en. URL: http://arxiv.org/abs/2206.07550 (visited on 01/25/2023).

John, Oliver P, Sanjay Srivastava, et al. (1999). "The Big-Five trait taxonomy: History, measurement, and theoretical perspectives". In.

Johnson, Rebecca L et al. (2022a). "The Ghost in the Machine has an American accent: value conflict in GPT-3." en. In: p. 15.

Johnson, Rebecca L et al. (2022b). *The Ghost in the Machine has an American accent: value conflict in GPT-3*. arXiv: 2203.07785 [cs.CL].

Jones, Daniel N and Delroy L Paulhus (2014). "Introducing the short dark triad (SD3) a brief measure of dark personality traits". In: *Assessment* 21.1, pp. 28–41.

Jones, L. L. et al. (2020). "Sex differences in emoji use, familiarity, and valence". In: *Computers in Human Behavior* 108, p. 106305. DOI: 10.1016/j.chb.2020.106305.

Kahneman, Daniel, Olivier Sibony, and Cass R. Sunstein (2021). *Noise: A Flaw in Human Judgment*. Little, Brown Spark.

Kahneman, Daniel and Amos Tversky (June 7, 2012). "Prospect Theory: An Analysis of Decision Under Risk". In: *Handbook of the Fundamentals of Financial Decision Making*. Vol. Volume 4. World Scientific Handbook in Financial Economics Series Volume 4. WORLD SCIENTIFIC, pp. 99–127. ISBN: 978-981-4417-34-1. DOI: 10.1142/9789814417358_0006. URL: https://www.worldscientific.com/doi/abs/10.1142/9789814417358_0006 (visited on 11/27/2023).

Kaigo, Muneo (2012). "Social Media Usage During Disasters and Social Capital: Twitter and the Great East Japan Earthquake". In: 34, p. 17.

Kalimeri, Kyriaki et al. (Jan. 23, 2019). *Evaluation of Biases in Self-reported Demographic and Psychometric Information: Traditional versus Facebook-based Surveys*. DOI: 10.48550/arXiv.1901.07876. arXiv: 1901.07876[cs]. URL: http://arxiv.org/abs/1901.07876 (visited on 11/22/2023).

Karra, Saketh Reddy, Son Nguyen, and Theja Tulabandhula (2022). *AI Personification: Estimating the Personality of Language Models*. en. URL: http://arxiv.org/abs/2204.12000 (visited on 06/01/2022).

Kashima, Saori and Junyi Zhang (Mar. 1, 2021). "Temporal trends in voluntary behavioural changes during the early stages of the COVID-19 outbreak in Japan". In: *Public Health* 192, pp. 37–44. ISSN: 0033-3506. DOI: 10.1016/j.puhe.2021.01.002. URL: https://www.sciencedirect.com/science/article/pii/S0033350621000020 (visited on 07/18/2021).

Kazmierczak, Izabela et al. (Mar. 8, 2023). "Self-selection biases in psychological studies: Personality and affective disorders are prevalent among participants". In: *PLOS ONE* 18.3. Publisher: Public Library of Science, e0281046. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0281046. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0281046 (visited on 11/22/2023).

Kern, M. L. et al. (2016). "Gaining insights from social media language: Methodologies and challenges". In: *Psychological Methods* 21.4, pp. 507–525. DOI: 10.1037/met0000091.

Ketipov, Rumen (2022). *Bulgarian Version of TIPI*. URL: https://gosling.psy.utexas.edu/wp-content/uploads/2020/12/Bulgarian_Version_of_TIPI.pdf (visited on 09/02/2022).

Kim, Myunghee and Mychal Voorhees (2011). "Government Effectiveness and Institutional Trust in Japan, South Korea, and China". In: *Asian Politics & Policy* 3.3. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1943-0787.2011.01278.x, pp. 413–432. ISSN: 1943-0787. DOI: 10.1111/j.1943-0787.2011.01278.x. URL: https://www.onlinelibrary.wiley.com/doi/abs/10.1111/j.1943-0787.2011.01278.x (visited on 07/18/2021).

Kinney, Justin B. and Gurinder S. Atwal (Mar. 4, 2014). "Equitability, mutual information, and the maximal information coefficient". In: *Proceedings of the National Academy of Sciences* 111.9. Publisher: Proceedings of the National Academy of Sciences, pp. 3354–3359. DOI: 10.1073/pnas.1309933111. URL: https://www.pnas.org/doi/10.1073/pnas.1309933111 (visited on 11/28/2023).

Koch, Timo, Peter Romero, and Clemens Stachl (Nov. 2, 2020). *Predicting Age and Gender from Language, Emoji, and Emoticon Use in WhatsApp Instant Messages*. type: article. PsyArXiv. DOI: 10.31234/osf.io/92ydh. URL: https://psyarxiv.com/92ydh/ (visited on 07/18/2021).

Koch, Timo K, Peter Romero, and Clemens Stachl (2022). "Age and gender in language, emoji, and emoticon usage in instant messages". In: *Computers in Human Behavior* 126, p. 106990.

Kojima, Takeshi et al. (2022). *Large Language Models are Zero-Shot Reasoners*. arXiv: 2205.11916[cs]. URL: http://arxiv.org/abs/2205.11916 (visited on 09/03/2022).

Kosinski, Michal (2023). *Theory of Mind May Have Spontaneously Emerged in Large Language Models*. DOI: 10.48550/arXiv.2302.02083. arXiv: 2302.02083[cs]. URL: http://arxiv.org/abs/2302.02083 (visited on 02/18/2023).

Kosinski, Michal et al. (2015). "Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines." In: *American Psychologist* 70.6, pp. 543–556. ISSN: 1935-990X, 0003-066X. DOI: 10.

1037/a0039210. URL: http://doi.apa.org/getdoi.cfm?doi=10.1037/a0039210 (visited on 01/15/2018).

Kring, A. M. and A. H. Gordon (1998). "Sex Differences in Emotion: Expression, Experience, and Physiology". In: *Journal of Personality and Social Psychology* 18.

Kudo, Taku (2005). "Mecab: Yet another part-of-speech and morphological analyzer". In: *http://mecab. sourceforge. net/*.

Kulkarni, Vivek, Bryan Perozzi, and Steven Skiena (2016). "Freshman or Fresher? Quantifying the Geographic Variation of Language in Online Social Media". In: *Proceedings of the International AAAI Conference on Web and Social Media* 10.1. Number: 1, pp. 615–618. ISSN: 2334-0770. DOI: 10.1609/icwsm.v10i1.14798. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14798 (visited on 11/23/2023).

Kumar, Ravin et al. (2019). "ArviZ a unified library for exploratory analysis of Bayesian models in Python". In: *Journal of Open Source Software* 4.33, p. 1143. DOI: 10.21105/joss.01143. URL: https://doi.org/10.21105/joss.01143.

Latonero, Mark and Irina Shklovski (Oct. 2011). "Emergency Management, Twitter, and Social Media Evangelism". In: *IJISCRAM* 3, pp. 1–16. DOI: 10.4018/jiscrm.2011100101.

Lazarsfeld, John, Aaron Johnson, and Emmanuel Adéníran (2022). "Differentially Private Maximal Information Coefficients". In: *International Conference on Machine Learning*. DOI: 10.48550/arXiv.2206.10685.

Lee, Chang H et al. (2007). "The relations between personality and language use". In: *The Journal of general psychology* 134.4, pp. 405–413.

Lester, Brian, Rami Al-Rfou, and Noah Constant (2021). "The Power of Scale for Parameter-Efficient Prompt Tuning". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 3045–3059. DOI: 10.18653/v1/2021.emnlp-main.243. URL: https://aclanthology.org/2021.emnlp-main.243 (visited on 09/03/2022).

Li, Tania Y. et al. (July 1, 2023). "Openness buffers the impact of Belief in Conspiracy Theories on COVID-19 vaccine hesitancy: Evidence from a large, representative Italian sample". In: *Personality and Individual Differences* 208, p. 112189. ISSN: 0191-8869. DOI: 10.1016/j.paid.2023.112189. URL: https://www.sciencedirect.com/science/article/pii/S0191886923001125 (visited on 11/12/2023).

Li, Xiang Lisa and Percy Liang (2021). "Prefix-Tuning: Optimizing Continuous Prompts for Generation". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics, pp. 4582–4597. DOI: 10.18653/v1/2021.acl-long.353. URL: https://aclanthology.org/2021.acl-long.353 (visited on 09/03/2022).

Li, Xingxuan et al. (2022a). *Is GPT-3 a Psychopath? Evaluating Large Language Models from a Psychological Perspective*. en. URL: http://arxiv.org/abs/2212.10529 (visited on 01/25/2023).

Li, Zekun et al. (Oct. 21, 2022b). *SpaBERT: A Pretrained Language Model from Geographic Data for Geo-Entity Representation*. DOI: 10.48550/arXiv.2210.

12213. arXiv: 2210.12213[cs]. URL: http://arxiv.org/abs/2210.12213 (visited on 11/23/2023).

Limaye, Rupali Jayant et al. (June 1, 2020). "Building trust while influencing online COVID-19 content in the social media world". In: *The Lancet Digital Health* 2.6. Publisher: Elsevier, e277–e278. ISSN: 2589-7500. DOI: 10.1016/S2589-7500(20)30084-4. URL: https://www.thelancet.com/journals/landig/article/PIIS2589-7500(20)30084-4/abstract (visited on 07/18/2021).

Linden, Wim J. Van der (2017). *Handbook of Item Response Theory, Volume Three: Applications*. CRC Press.

Liu, Bingjie and S. Shyam Sundar (Oct. 2018). "Should Machines Express Sympathy and Empathy? Experiments with a Health Advice Chatbot". In: *Cyberpsychology, Behavior, and Social Networking* 21.10. Publisher: Mary Ann Liebert, Inc., publishers, pp. 625–636. ISSN: 2152-2715. DOI: 10.1089/cyber.2018.0110. URL: https://www.liebertpub.com/doi/abs/10.1089/cyber.2018.0110 (visited on 09/04/2023).

Liu, Pengfei et al. (2021). *Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing*. arXiv: 2107.13586[cs]. URL: http://arxiv.org/abs/2107.13586 (visited on 09/03/2022).

Loo, J. van de, G. De Pauw, and W. Daelemans (2016). "Text-Based Age and Gender Prediction for Online Safety Monitoring". In: *International Journal of Cyber-Security and Digital Forensics* 5.1, pp. 46–60. DOI: 10.17781/P002012.

Loomba, Sahil et al. (2021). "Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA". In: *Nature Human Behaviour* 5, pp. 337–348.

Lu, J. et al. (2019). "Learning under Concept Drift: A Review". In: *IEEE Transactions on Knowledge and Data Engineering* 31.12, pp. 2346–2363. DOI: 10.1109/TKDE.2018.2876857.

Lu, Jackson G. et al. (2020). "Disentangling stereotypes from social reality: Astrological stereotypes and discrimination in China." In: *Journal of Personality and Social Psychology* 119.6, pp. 1359–1379. ISSN: 1939-1315, 0022-3514. DOI: 10.1037/pspi0000237. URL: http://doi.apa.org/getdoi.cfm?doi=10.1037/pspi0000237 (visited on 08/28/2022).

Lupyan, Gary (2022). "There is no such thing as culture-free intelligence". In: *Behavioral and Brain Sciences* 45, e169.

MacIntyre, C Raina (2015). "Increasing the uptake of vaccination against infectious diseases". In: *Medical Journal of Australia*.

Mangalik, Siddharth et al. (2023). "Robust language-based mental health assessments in time and space through social media". In: *arXiv preprint arXiv:2302.12952*.

Marquardt, J. et al. (2014). "Age and Gender Identification in Social Media". In: *Proceedings of CLEF 2014 Evaluation Labs*. Vol. 1180, pp. 1129–1136.

Mathai, Arak M. (2021). "Factor Analysis Revisited". In: *Journal of Statistical Theory and Practice*. DOI: 10.1007/S42519-020-00160-1.

McCrae, Robert R and Paul T Costa (1987). "Validation of the five-factor model of personality across instruments and observers." In: *Journal of personality and social psychology* 52.1, p. 81.

McCrae, Robert R. and Paul T. Costa (1997). "Personality trait structure as a human universal". In: *The American psychologist* 52.5, pp. 509–16. ISSN: 0003-066X. DOI: 10.1037/0003-066X.52.5.509. arXiv: Krejcie,R.V.,&Morgan,D.W.(1970).ACTIVITIES,38,607-610.. URL: http://www.ncbi.nlm.nih.gov/pubmed/9145021.

McCrae, Robert R et al. (2004). "NEO-PI-R data from 36 cultures: Further intercultural comparisons". In: *In International perspectives on psychological science* 2, pp. 105–125.

Mechanic, David (Feb. 1986). "The concept of illness behaviour: culture, situation and personal predisposition1". In: *Psychological Medicine* 16.1. Publisher: Cambridge University Press, pp. 1–7. ISSN: 1469-8978, 0033-2917. DOI: 10.1017/S0033291700002476. URL: https://www.cambridge.org/core/journals/psychological-medicine/article/concept-of-illness-behaviour-culture-situation-and-personal-predisposition1/1E8D61EEB2C2C41BD873AAFD1B61741B (visited on 07/18/2021).

Meier, T. et al. (2019). ""LIWC auf Deutsch": The Development, Psychometrics, and Introduction of DE- LIWC2015". In: *PsyArXiv*. DOI: 10.31234/osf.io/uq8zt.

Melki, Jad et al. (June 4, 2021). "Mitigating infodemics: The relationship between news exposure and trust and belief in COVID-19 fake news and social media spreading". In: *PLOS ONE* 16.6. Publisher: Public Library of Science, e0252830. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0252830. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0252830 (visited on 07/18/2021).

MEXT (2024). *society 5.0*. ja. URL: https://www.mext.go.jp/b_menu/hakusho/html/hpaa202201/1421221_00017.html (visited on 02/19/2024).

Meyers, Lauren Ancel, Fredric D. Ancel, and Michael Lachmann (Aug. 26, 2005). "Evolution of Genetic Potential". In: *PLOS Computational Biology* 1.3. Publisher: Public Library of Science, e32. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.0010032. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.0010032 (visited on 07/18/2021).

MHLW (2021). *Press Conference from the Ministry of Health, Labour and Welfare*. URL: https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/newpage_00032.html (visited on 07/18/2021).

Minakov, VF et al. (2017). "The expansion of time series innovations in a series of sigmoid". In: *International Journal of Applied Business and Economic Research* 15.18, pp. 311–319.

Miotto, Marilù, Nicola Rossberg, and Bennett Kleinberg (2022). *Who is GPT-3? An Exploration of Personality, Values and Demographics*. en. URL: http://arxiv.org/abs/2209.14338 (visited on 01/25/2023).

Molnar, C. (2018). "iml: An R package for Interpretable Machine Learning". In: *Journal of Open Source Software* 3.26, p. 786. DOI: 10.21105/joss.00786.

— (2019). *Interpretable Machine Learning*. URL: https://christophm.github.io/interpretable-ml-book/.

Muck, Peter M., Benedikt Hell, and Samuel D. Gosling (2007). "Construct Validation of a Short Five-Factor Model Instrument". In: *European Journal of Psychological Assessment* 23.3, pp. 166–175. ISSN: 1015-5759, 2151-2426. DOI: 10.1027/1015-5759.23.3.166. URL: https://econtent.hogrefe.com/doi/10.1027/1015-5759.23.3.166 (visited on 08/28/2022).

Mullainathan, Sendhil and Jann Spiess (May 2017). "Machine Learning: An Applied Econometric Approach". en. In: *Journal of Economic Perspectives* 31.2, pp. 87–106. ISSN: 0895-3309. DOI: 10.1257/jep.31.2.87. URL: http://pubs.aeaweb.org/doi/10.1257/jep.31.2.87 (visited on 10/07/2018).

Muñoz-Fernández, Noelia and Ana Rodríguez-Meirinhos (Feb. 16, 2021). "Adolescents' Concerns, Routines, Peer Activities, Frustration, and Optimism in the Time of COVID-19 Confinement in Spain". In: *Journal of Clinical Medicine* 10.4, p. 798.

ISSN: 2077-0383. DOI: 10.3390/jcm10040798. URL: https://www.mdpi.com/2077-0383/10/4/798 (visited on 07/18/2021).

Nadeau, C. and Y. Bengio (2003). "Inference for the Generalization Error". In: *Machine Learning* 52.3, pp. 239–281. DOI: 10.1023/A:1024068626366.

Nagata, Shohei et al. (2021). "Mobility Change and COVID-19 in Japan: Mobile Data Analysis of Locations of Infection". In: *Journal of Epidemiology* advpub. DOI: 10.2188/jea.JE20200625.

Nakatsuma, Teruo (2007). *Introduction to Bayesian Statistics*. Asakura Shoten.

Natesan, Prathiba et al. (2016). "Bayesian Prior Choice in IRT Estimation Using MCMC and Variational Bayes". In: *Frontiers in Psychology* 7. DOI: 10.3389/fpsyg.2016.01422. URL: https://www.readcube.com/articles/10.3389%2Ffpsyg.2016.01422 (visited on 07/17/2021).

Neff, Timothy et al. (Oct. 2021). "Vaccine hesitancy in online spaces: A scoping review of the research literature, 2000-2020". en-US. In: *Harvard Kennedy School Misinformation Review*. DOI: 10.37016/mr-2020-82. URL: https://misinforeview.hks.harvard.edu/article/vaccine-hesitancy-in-online-spaces-a-scoping-review-of-the-research-literature-2000-2020/ (visited on 11/29/2023).

Nehal, Kimberly R et al. (2021). "Worldwide vaccination willingness for COVID-19: a systematic review and meta-analysis". In: *Vaccines* 9.10, p. 1071.

Newman, M. L. et al. (2008). "Gender Differences in Language Use: An Analysis of 14,000 Text Samples". In: *Discourse Processes* 45.3, pp. 211–236. DOI: 10.1080/01638530802073712.

Nguyen, D., N. A. Smith, and C. P. Rosé (2011). "Author Age Prediction from Text using Linear Regression". In: *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pp. 115–123. URL: https://www.aclweb.org/anthology/W11-1515.

Nguyen, D. et al. (2013). ""How Old Do You Think I Am?"; A Study of Language and Age in Twitter". In: *Proceedings of the seventh international AAAI conference on weblogs and social media*.

Nguyen, T. et al. (2011). "Prediction of Age, Sentiment, and Connectivity from Social Media Text". In: *Web Information System Engineering – WISE 2011*. Vol. 6997. Springer Berlin Heidelberg, pp. 227–240. DOI: 10.1007/978-3-642-24434-6_17.

Norman, Warren T (1963). "Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings." In: *The journal of abnormal and social psychology* 66.6, p. 574.

Obschonka, Martin et al. (2018). "Fear, Populism, and the Geopolitical Landscape: The "Sleeper Effect" of Neurotic Personality Traits on Regional Voting Behavior in the 2016 Brexit and Trump Elections". In: *Social Psychological and Personality Science* 9.3, pp. 285–298. ISSN: 1948-5506, 1948-5514. DOI: 10.1177/1948550618755874. URL: http://journals.sagepub.com/doi/10.1177/1948550618755874 (visited on 12/11/2018).

Ogaki, Masao (July 2022). "Economics of the community mechanism". In: *The Japanese Economic Review* 73.3, pp. 433–457. ISSN: 1352-4739, 1468-5876. DOI: 10.1007/s42973-022-00113-2. URL: https://link.springer.com/10.1007/s42973-022-00113-2 (visited on 09/26/2023).

Ogaki, Masao and Saori C. Tanaka (2017). *Behavioral Economics: Toward a New Economics by Integration with Traditional Economics*. Springer Texts in Business and Economics. Singapore: Springer Singapore. ISBN: 978-981-10-6438-8 978-981-10-6439-5. DOI: 10.1007/978-981-10-6439-5. URL: http://link.springer.com/10.1007/978-981-10-6439-5 (visited on 09/26/2023).

Oleszkiewicz, A. et al. (2017). "Who uses emoticons? Data from 86 702 Facebook users". In: *Personality and Individual Differences* 119, pp. 289–295. DOI: 10.1016/j.paid.2017.07.034.

Ones, Deniz S. and Chockalingam Viswesvaran (2001). "Integrity Tests and Other Criterion-Focused Occupational Personality Scales (COPS) Used in Personnel Selection". In: *International Journal of Selection and Assessment* 9.1. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1468-2389.00161, pp. 31–39. ISSN: 1468-2389. DOI: 10.1111/1468-2389.00161. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1468-2389.00161 (visited on 07/19/2021).

Oortwijn, Rick (2020). "How Openness to experience relates to Conspiracy mentality and Vaccine hesitancy". In: *Economic Psychology*. URL: http://arno.uvt.nl/show.cgi.

OpenAI (2022). *Introducing ChatGPT*. en-US. URL: https://openai.com/blog/chatgpt (visited on 03/04/2023).

Ortner, Tuulia M. and Manfred Schmitt (Jan. 1, 2014). "Advances and Continuing Challenges in Objective Personality Testing". In: *European Journal of Psychological Assessment* 30.3. Publisher: Hogrefe Publishing, pp. 163–168. ISSN: 1015-5759. DOI: 10.1027/1015-5759/a000213. URL: https://econtent.hogrefe.com/doi/full/10.1027/1015-5759/a000213 (visited on 07/19/2021).

Oshio, Atsushi, Shingo Abe, and Pino Cutrone (2012). "Development, Reliability, and Validity of the Japanese Version of Ten Item Personality Inventory (TIPI-J)." In: *Japanese Journal of Personality* 21.1, pp. 40–52. URL: https://gosling.psy.utexas.edu/wp-content/uploads/2014/09/2012TIPI_J.pdf (visited on 06/01/2022).

Oshio, Atsushi et al. (2013). "Big Five Content Representation of the Japanese Version of the Ten-Item Personality Inventory". In: *Psychology* 04.12, pp. 924–929. ISSN: 2152-7180, 2152-7199. DOI: 10.4236/psych.2013.412133. URL: http://www.scirp.org/journal/doi.aspx?DOI=10.4236/psych.2013.412133 (visited on 06/01/2022).

Park, G. et al. (2016). "Women are Warmer but No Less Assertive than Men: Gender and Language on Facebook". In: *PLOS ONE* 11.5, e0155885. DOI: 10.1371/journal.pone.0155885.

Paulhus, Delroy L and Kevin M Williams (2002). "The dark triad of personality: Narcissism, Machiavellianism, and psychopathy". In: *Journal of research in personality* 36.6, pp. 556–563.

Pavalanathan, U. and J. Eisenstein (2015). "Emoticons vs. Emojis on Twitter: A causal inference approach". In: *ArXiv Preprint ArXiv:1510.08480*.

Pedregosa, F. et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Peersman, C., W. Daelemans, and L. Van Vaerenbergh (2011). "Predicting age and gender in online social networks". In: *Proceedings of the 3rd International Workshop on Search and Mining User-Generated Contents - SMUC '11*, p. 37. DOI: 10.1145/2065023.2065035.

Pennebaker J. W., Booth R. J. Boyd R. L. Francis M. E. (2015). *Linguistic Inquiry and word count: LIWC 2015 [computer software]*.

Pennebaker, J. W. and L. A. King (1999). "Linguistic styles: Language use as an individual difference". In: *Journal of Personality and Social Psychology* 77.6, pp. 1296–1312. DOI: 10.1037/0022-3514.77.6.1296.

Pennebaker, J. W. and L. D. Stone (2003). "Words of wisdom: Language use over the life span". In: *Journal of Personality and Social Psychology* 85.2, pp. 291–301. DOI: 10.1037/0022-3514.85.2.291.

Pennebaker, J. W. et al. (2015a). *Linguistic Inquiry and Word Count: LIWC 2015 [Computer software]*. Pennebaker Conglomerates, Inc.

Pennebaker, J. W. et al. (2015b). "The development and psychometric properties of LIWC2015". In.

Peters, Heinrich et al. (2023). "Regional personality differences predict variation in early COVID-19 infections and mobility patterns indicative of social distancing." In: *Journal of Personality and Social Psychology* 124.4, p. 848.

Prada, M. et al. (2018). "Motives, frequency and attitudes toward emoji and emoticon use". In: *Telematics and Informatics* 35.7, pp. 1925–1934. DOI: 10.1016/j.tele.2018.06.005.

Preoţiuc-Pietro, D. et al. (2016). "Modelling Valence and Arousal in Facebook posts". In: *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp. 9–15. DOI: 10.18653/v1/W16-0404.

Probst, P. and A.-L. Boulesteix (2017). "To tune or not to tune the number of trees in random forest". In: *The Journal of Machine Learning Research* 18.1, pp. 6673–6690.

Python Software Foundation (2023). *Python*. Version 3.8.9. URL: https://www.python.org/.

Pérez-Sabater, C. (2019). "Emoticons in Relational Writing Practices on WhatsApp: Some Reflections on Gender". In: *Analyzing Digital Discourse: New Insights and Future Directions*. Ed. by P. Bou-Franch and P. Garcés-Conejos Blitvich. Springer International Publishing, pp. 163–189. DOI: 10.1007/978-3-319-92663-6_6.

Quan-Haase, A. and A. L. Young (2010). "Uses and Gratifications of Social Media: A Comparison of Facebook and Instant Messaging". In: *Bulletin of Science, Technology Society* 30.5, pp. 350–361. DOI: 10.1177/0270467610380009.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. URL: https://www.R-project.org.

Radford, Alec et al. (2019). "Language models are unsupervised multitask learners". In: *OpenAI blog* 1.8, p. 9.

Rao, D. et al. (2010). "Classifying latent user attributes in Twitter". In: *Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents - SMUC '10*, p. 37. DOI: 10.1145/1871985.1871993.

Raykov, Tenko, Michael Harrison, and George A. Marcoulides (Jan. 2, 2020). "Examining Class Separation Contribution by Observed Indicators in Latent Class Models: A Multiple Testing Procedure". In: *Structural Equation Modeling: A Multidisciplinary Journal* 27.1, pp. 88–96. ISSN: 1070-5511. DOI: 10.1080/10705511.2018.1554446. URL: https://doi.org/10.1080/10705511.2018.1554446 (visited on 07/18/2021).

Renau, Vanessa et al. (2013). "Translation and validation of the Ten- Item-Personality Inventory into Spanish and Catalan". In: p. 13.

Rentfrow, Peter J. (2020). "Geographical psychology". In: *Current Opinion in Psychology* 32, pp. 165–170. ISSN: 2352-250X. DOI: 10.1016/J.COPSYC.2019.09.009.

Rentfrow, Peter J., Samuel D. Gosling, and Jeff Potter (2008a). "A theory of the emergence, persistence, and expression of geographic variation in psychological characteristics". In: *Perspectives on Psychological Science* 3.5, pp. 339–369. ISSN: 1745-6916. DOI: 10.1111/j.1745-6924.2008.00084.x. URL: http://journals.sagepub.com/doi/10.1111/j.1745-6924.2008.00084.x.

— (2008b). "A Theory of the Emergence, Persistence, and Expression of Geographic Variation in Psychological Characteristics". In: *Perspectives on Psychological Science* 3.5, pp. 339–369. ISSN: 1745-6916, 1745-6924. DOI: 10.1111/j.1745-6924.

2008.00084.x. URL: http://journals.sagepub.com/doi/10.1111/j.1745-6924.2008.00084.x (visited on 03/12/2019).

Rentfrow, Peter J., Markus Jokela, and Michael E. Lamb (2015). "Regional personality differences in Great Britain". In: *PLOS ONE* 10.3. Ed. by Robert D Latzman, e0122245. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0122245. URL: https://dx.plos.org/10.1371/journal.pone.0122245.

Rentfrow, Peter J. et al. (2013). "Divided we stand: Three psychological regions of the United States and their political, economic, social, and health correlates." In: *Journal of Personality and Social Psychology* 105.6, pp. 996–1012. ISSN: 1939-1315. DOI: 10.1037/a0034434. URL: http://doi.apa.org/getdoi.cfm?doi=10.1037/a0034434.

Reshef, David N. et al. (Dec. 16, 2011). "Detecting Novel Associations in Large Datasets". In: *Science (New York, N.y.)* 334.6062, pp. 1518–1524. ISSN: 0036-8075. DOI: 10.1126/science.1205438. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3325791/ (visited on 11/26/2023).

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (Aug. 2016). *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. en. arXiv:1602.04938 [cs, stat]. URL: http://arxiv.org/abs/1602.04938 (visited on 09/13/2023).

Roberts, Margaret E., Brandon M. Stewart, and Dustin Tingley (2019a). "stm: An R Package for Structural Topic Models". In: *Journal of Statistical Software* 91.2, pp. 1–40. DOI: 10.18637/jss.v091.i02.

Roberts, Margaret E, Brandon M Stewart, and Dustin Tingley (2019b). "Stm: An R package for structural topic models". In: *Journal of Statistical Software* 91.2, pp. 1–40.

Roesslein, Joshua (2020). *Tweepy: Twitter for Python!* URL: https://github.com/tweepy/tweepy.

Romero, Peter and Stephen Fitz (2021). "The use of psychometrics and artificial intelligence in alternative finance". In: *The Palgrave handbook of technological finance*, pp. 511–587.

Romero, Peter, Stephen Fitz, and Teruo Nakatsuma (Mar. 2023). *Do GPT Language Models Suffer From Split Personality Disorder? The Advent Of Substrate-Free Psychometrics*. en. preprint. In Review. DOI: 10.21203/rs.3.rs-2717108/v1. URL: https://www.researchsquare.com/article/rs-2717108/v1 (visited on 09/03/2023).

Romero, Peter et al. (2021). *Modelling Personality Change During Extreme Exogenous Conditions*. DOI: 10.31234/osf.io/rtmjw. URL: https://psyarxiv.com/rtmjw/ (visited on 06/28/2022).

Rust, John and Susan Golombok (2014a). *Modern psychometrics: The science of psychological assessment*. Routledge.

— (July 23, 2014b). *Modern Psychometrics: The Science of Psychological Assessment*. 3rd ed. London: Routledge. 272 pp. ISBN: 978-1-315-78752-7. DOI: 10.4324/9781315787527.

Rust, John, Michal Kosinski, and David Stillwell (Dec. 2020). *Modern Psychometrics: The Science of Psychological Assessment*. en. 4th ed. Fourth edition. | Milton Park, Abingdon, Oxon ; New York, NY: Routledge, 2021.: Routledge. ISBN: 978-1-315-63768-6. DOI: 10.4324/9781315637686. URL: https://www.taylorfrancis.com/books/9781317268772 (visited on 08/26/2023).

Safdari, Mustafa et al. (June 2023). *Personality Traits in Large Language Models*. arXiv:2307.00184 [cs]. DOI: 10.48550/arXiv.2307.00184. URL: http://arxiv.org/abs/2307.00184 (visited on 09/03/2023).

Salmas, Konstantinos, Despina-Athanasia Pantazi, and Manolis Koubarakis (2013). "Extracting Geographic Knowledge from Large Language Models: An Experiment". In.

Sap, M. et al. (2014). "Developing Age and Gender Predictive Lexica over Social Media". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1146–1151. DOI: 10.3115/v1/D14-1121.

Savage, David A. (2019). "Towards a complex model of disaster behaviour". In: *Disasters* 43.4. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/disa.12408, pp. 771–798. ISSN: 1467-7717. DOI: 10.1111/disa.12408. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/disa.12408 (visited on 07/18/2021).

Schler, J. et al. (2006). "Effects of age and gender on blogging". In: *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*. Vol. 6, pp. 199–205.

Schmitt, David P et al. (2007). "The geographic distribution of Big Five personality traits: Patterns and profiles of human self-description across 56 nations". en. In: *Journal of cross-cultural psychology* 38.2, pp. 173–212.

Schneider, Benjamin and Dave Bartram (2017). "Aggregate personality and organizational competitive advantage". In: *Journal of Occupational and Organizational Psychology* 90.4, pp. 461–480. ISSN: 09631798. DOI: 10.1111/joop.12180. URL: http://doi.wiley.com/10.1111/joop.12180 (visited on 11/24/2017).

Schwab, Klaus (Jan. 26, 2016). "The Fourth Industrial Revolution". In: ISSN: 0015-7120. URL: https://www.foreignaffairs.com/articles/2015-12-12/fourth-industrial-revolution (visited on 07/28/2021).

Schwab, Klaus and Thierry Malleret (2020). "The Great Reset". In: *World Economic Forum: Geneva, Switzerland*.

Schwartz, H. A. et al. (2017). "DLATK: Differential Language Analysis ToolKit". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 55–60. DOI: 10.18653/v1/D17-2010.

Schwartz, H Andrew et al. (2013). "Personality, gender, and age in the language of social media: The open-vocabulary approach". In: *PloS one* 8.9, e73791.

Schwartz, Shalom (2006). "A theory of cultural value orientations: Explication and applications". In: *Comparative sociology* 5.2-3, pp. 137–182.

Schwartz, Shalom H (1992). "Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries". In: *Advances in Experimental Social Psychology*. Advances in experimental social psychology. Elsevier, pp. 1–65.

Schwartz, Shalom H, Bianka Breyer, and Daniel Danner (2015). "Human values scale (ESS)". In: *Zusammenstellung sozialwissenschaftlicher Items und Skalen*.

Seabold, Skipper and Josef Perktold (2010). "statsmodels: Econometric and statistical modeling with python". In: *9th Python in Science Conference*.

Selye, Hans (Oct. 7, 1955). "Stress and Disease". In: *Science* 122.3171, pp. 625–631.

Sergeeva, A.S., B.A. Kirillov, and A.F. Dzhumagulova (2016). "Translation and adaptation of short five factor personality questionnaire (TIPI-RU): convergent validity, internal consistency and test-retest reliability evaluation". In: *Experimental Psychology (Russia)* 9.3, pp. 138–154. ISSN: 2072-7593, 2311-7036. DOI: 10.17759/exppsy.2016090311. URL: https://psyjournals.ru/en/exp/2016/n3/sergeeva.shtml (visited on 09/02/2022).

Shaffer, Jonathan A. and Bennett E. Postlethwaite (2012). "A matter of context: A meta-analytic investigation of the relative validity of contextualized and noncontextualized personality measures". In: *Personnel Psychology* 65.3, pp. 445–494.

Sherman, Ryne A, Christopher S Nave, and David C Funder (2016). "The power of personality: The comparative validity of personality traits, socioeconomic status,

and cognitive ability for predicting important life outcomes". In: *Advances in experimental social psychology* 52, pp. 71–115.

Shibata, Daisaku et al. (Dec. 2016). "Detecting Japanese Patients with Alzheimer's Disease based on Word Category Frequencies". In: *Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP)*. Osaka, Japan: The COLING 2016 Organizing Committee, pp. 78–85. URL: https://aclanthology.org/W16-4211 (visited on 07/18/2021).

Shin, Minkyu et al. (2023). "Superhuman Artificial Intelligence Can Improve Human Decision-Making by Increasing Novelty". In: *Proceedings of the National Academy of Sciences*.

Siebenhaar, B. (2018). "Funktionen von Emojis und Altersabhängigkeit ihres Gebrauchs in der Whatsapp-Kommunikation". In: *Jugendsprachen/Youth Languages*. Ed. by A. Ziegler. De Gruyter, pp. 749–772. DOI: 10.1515/9783110472226-034.

Society 5.0: A People-centric Super-smart Society (n.d.). Singapore.

Soto, Carlos J. and Oliver P. John (2019). "Psychometric Properties of Language Tests". In: *Language Assessment Quarterly*.

Stachl, C. et al. (2020). "Personality Research and Assessment in the Era of Machine Learning". In: *European Journal of Personality*. DOI: 10.1002/per.2257.

Stachl, Clemens et al. (July 15, 2021). "Computational personality assessment". In: *Personality Science* 2, e6115. ISSN: 2700-0710. DOI: 10.5964/ps.6115. URL: https://ps.psychopen.eu/index.php/ps/article/view/6115 (visited on 07/16/2021).

Statista (July 3, 2022). *Internet: most common languages online 2020*. Statista. URL: https://www.statista.com/statistics/262946/share-of-the-most-common-languages-on-the-internet/ (visited on 09/19/2022).

Statistics Bureau, Ministry of Internal Affairs, Communications with collaboration of Ministries, and Agencies (2021). *Regional Statistics Database (System of Social and Demographic Statistics)*. Portal Site of Official Statistics of Japan website (https://www.e-stat.go.jp/). URL: https://www.e-stat.go.jp/en/regional-statistics/ssdsview (visited on 06/13/2021).

Stegle, Oliver and Timothy M. Rohan (2022). "Identifying temporal and spatial patterns of variation from multimodal data using MEFISTO". In: *Nature Methods*. DOI: 10.1038/s41592-021-01343-9.

Stieger, Mirjam, Christoph Flückiger, and Mathias Allemand (2023). "One year later: Longer-term maintenance effects of a digital intervention to change personality traits". In: *Journal of Personality* n/a (n/a). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jopy.12898. ISSN: 1467-6494. DOI: 10.1111/jopy.12898. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/jopy.12898 (visited on 11/28/2023).

Straka, M. and J. Straková (2017). "Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe". In: *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pp. 88–99. DOI: 10.18653/v1/K17-3009.

Strobl, C. et al. (2008). "Conditional variable importance for random forests". In: *BMC Bioinformatics* 9.1, p. 307. DOI: 10.1186/1471-2105-9-307.

Sugaya, Nagisa et al. (Dec. 2020). "A real-time survey on the psychological impact of mild lockdown for COVID-19 in the Japanese population". In: *Scientific Data* 7.1, p. 372. ISSN: 2052-4463. DOI: 10.1038/s41597-020-00714-9. URL: http://www.nature.com/articles/s41597-020-00714-9 (visited on 07/12/2021).

Sutin, Angelina R. et al. (Aug. 6, 2020). "Change in five-factor model personality traits during the acute phase of the coronavirus pandemic". In: *PLOS ONE* 15.8.

Ed. by Angel Blanch, e0237056. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0237056. URL: https://dx.plos.org/10.1371/journal.pone.0237056 (visited on 07/12/2021).

Székely, Gábor J., Maria L. Rizzo, and Nail K. Bakirov (2007). "Measuring and testing dependence by correlation of distances". In: *Annals of Statistics* 35.6, pp. 2769–2794. DOI: 10.1214/009053607000000505.

Takeshimura, Kazuhisa (Dec. 14, 2020). *Psychology of social attention to new Corona Virus infections*. URL: https://yab.yomiuri.co.jp/adv/wol/opinion/COVID-19/20201214.php (visited on 07/12/2021).

Tanaka, Takanao and Shohei Okamoto (Feb. 2021). "Increase in suicide following an initial decline during the COVID-19 pandemic in Japan". In: *Nature Human Behaviour* 5.2, pp. 229–238. ISSN: 2397-3374. DOI: 10.1038/s41562-020-01042-z. URL: https://www.nature.com/articles/s41562-020-01042-z (visited on 07/18/2021).

Tankovska, H. (2021). *Leading countries based on number of Twitter users as of April 2021*. Statista. URL: https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/ (visited on 07/19/2021).

Tashiro, Ai and Rajib Shaw (Jan. 2020). "COVID-19 Pandemic Response in Japan: What Is behind the Initial Flattening of the Curve?" In: *Sustainability* 12.13. Number: 13 Publisher: Multidisciplinary Digital Publishing Institute, p. 5250. DOI: 10.3390/su12135250. URL: https://www.mdpi.com/2071-1050/12/13/5250 (visited on 07/18/2021).

Tausczik, Yla R and James W Pennebaker (2010). "The psychological meaning of words: LIWC and computerized text analysis methods". In: *Journal of language and social psychology* 29.1, pp. 24–54.

team, The pandas development (Feb. 2020). *pandas-dev/pandas: Pandas*. Version latest. DOI: 10.5281/zenodo.3509134. URL: https://doi.org/10.5281/zenodo.3509134.

Thaler, Richard H. (2018). *Misbehaving: The Making of Behavioral Economics*. W. W. Norton Company.

Thaler, Richard H and Cass R Sunstein (2021). *Nudge: The final edition*. Yale University Press.

Tossell, C. C. et al. (2012). "A longitudinal study of emoticon use in text messaging from smartphones". In: *Computers in Human Behavior* 28.2, pp. 659–663. DOI: 10.1016/j.chb.2011.11.012.

Varian, Hal R. (May 2014). "Big Data: New Tricks for Econometrics". en. In: *Journal of Economic Perspectives* 28.2, pp. 3–28. ISSN: 0895-3309. DOI: 10.1257/jep.28.2.3. URL: http://pubs.aeaweb.org/doi/10.1257/jep.28.2.3 (visited on 10/07/2018).

Virtanen, Pauli et al. (2020). *SciPy: Open source scientific tools for Python*. https://www.scipy.org. Version latest.

Völkel, S. T. et al. (2019). "Understanding Emoji Interpretation through User Personality and Message Context". In: *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services - MobileHCI '19*, pp. 1–12. DOI: 10.1145/3338286.3340114.

Völkel, Sarah Theres et al. (2020). "Developing a Personality Model for Speech-based Conversational Agents Using the Psycholexical Approach". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI '20: CHI Conference on Human Factors in Computing Systems. Honolulu HI USA: ACM, pp. 1–14. ISBN: 978-1-4503-6708-0. DOI: 10.1145/3313831.3376210. URL: https://dl.acm.org/doi/10.1145/3313831.3376210 (visited on 06/01/2022).

Waskom, Michael L. (2021). "seaborn: statistical data visualization". In: *Journal of Open Source Software* 6.60, p. 3021. DOI: 10.21105/joss.03021. URL: https://doi.org/10.21105/joss.03021.

Watanabe, Sumio (2013). "A Widely Applicable Bayesian Information Criterion". In: *Journal of Machine Learning Research* 14, pp. 867–897.

Watson Personality Insights (n.d.). URL: https://cloud.ibm.com/docs/openwhisk?topic=openwhisk-pkg_person_insights.

Wei, Jason et al. (2022a). *Chain of Thought Prompting Elicits Reasoning in Large Language Models*. arXiv: 2201.11903[cs]. URL: http://arxiv.org/abs/2201.11903 (visited on 09/03/2022).

Wei, Jason et al. (2022b). *Emergent Abilities of Large Language Models*. en. URL: http://arxiv.org/abs/2206.07682 (visited on 09/07/2022).

Weisæth, Lars (1989). "A study of behavioural response to an industrial disaster". In: *Acta Psychiatrica Scandinavica* 80 (s355). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1600-0447.1989.tb05250.x, pp. 13–24. ISSN: 1600-0447. DOI: 10.1111/j.1600-0447.1989.tb05250.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1600-0447.1989.tb05250.x (visited on 07/18/2021).

WHO (2021). *Japan: World Health Organization Coronavirus Disease (COVID-19) Dashboard With Vaccination Data*. URL: https://covid19.who.int (visited on 07/18/2021).

Willke, Helmut (2000). *Systemtheorie 1. Grundlagen.* Stuttgart: UTB.

Winter, Joost C. F. de, Samuel D. Gosling, and Jeff Potter (2016). "Comparing the Pearson and Spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data". In: *Psychological Methods* 21.3. Place: US Publisher: American Psychological Association, pp. 273–290. ISSN: 1939-1463. DOI: 10.1037/met0000079.

Wolf, A. (2000). "Emotional Expression Online: Gender Differences in Emoticon Use". In: *CyberPsychology & Behavior* 3.5, pp. 827–833. DOI: 10.1089/10949310050191809.

Wood, A.J. et al. (2022). *Predicting future spatial patterns in COVID-19 booster vaccine uptak*. Type: article. DOI: 10.1101/2022.08.30.22279415. URL: https://europepmc.org/article/PPR/PPR539178 (visited on 11/28/2023).

Wright, M. N. and A. Ziegler (2017). "ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R". In: *Journal of Statistical Software* 77.1, pp. 1–17. DOI: 10.18637/jss.v077.i01.

Wright, M. N., A. Ziegler, and I. R. König (2016). "Do little interactions get lost in dark random forests?" In: *BMC Bioinformatics* 17.1, p. 145. DOI: 10.1186/s12859-016-0995-8.

Xiao, Xiaoqiang et al. (Apr. 2018). "Geographic Language Models for Automatic Speech Recognition". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). ISSN: 2379-190X, pp. 6124–6128. DOI: 10.1109/ICASSP.2018.8462550. URL: https://ieeexplore.ieee.org/abstract/document/8462550?casa_token=puf6Qf8sd20AAAAA:cNlGtRTheFfMN1SpRk8h9RNPnuDdlxNFTzth9cN1K2LiZw6fKyoSqulqEG1VQtggmMW4Q (visited on 11/23/2023).

Yamamoto, Tetsuya et al. (2020). "The psychological impact of 'mild lockdown'in Japan during the COVID-19 pandemic: a nationwide survey under a declared state of emergency". In: *International journal of environmental research and public health* 17.24, p. 9382.

Yap, Bee Wah and Chiaw Hock Sim (2011). "Comparisons of various types of normality tests". In: *Journal of Statistical Computation and Simulation* 81.12, pp. 2141–2155.

Yarkoni, T. and J. Westfall (2017). "Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning". In: *Perspectives on Psychological Science: A Journal of the Association for Psychological Science* 12.6, pp. 1100–1122. DOI: 10.1177/1745691617693393.

Yi, Xing, Hema Raghavan, and Chris Leggetter (Apr. 20, 2009). "Discovering users' specific geo intention in web search". In: *Proceedings of the 18th international conference on World wide web*. WWW '09. New York, NY, USA: Association for Computing Machinery, pp. 481–490. ISBN: 978-1-60558-487-4. DOI: 10.1145/1526709.1526774. URL: https://dl.acm.org/doi/10.1145/1526709.1526774 (visited on 11/22/2023).

Yoo, SeungHwan and Ulrike Gretzel (2016). "The influence of culture on consumer sensitivity to health communication: a multilevel study of the impact of individualism-collectivism". In: *PsyEcology* 7.3, pp. 251–279.

Yoshino, Shinya and Atsushi Oshio (2021a). "Regional differences in Big Five personality traits in Japan". In: *Japanese Journal of Environmental Psychology* 9.1, pp. 19–33. DOI: 10.20703/jenvpsy.9.1_19.

— (2021b). "Regional differences in Big Five personality traits in Japan: Evidence from three large datasets". In: *The Japanese Journal of Environmental Psychology* 9.1, pp. 19–33.

Yoshino, Shinya et al. (2021). "The association between personality traits and hoarding behavior during the COVID-19 pandemic in Japan". In: *Personality and individual differences* 179, p. 110927.

Zhang, Han, Shawndra Hill, and David Rothschild (May 29, 2018). "Addressing Selection Bias in Event Studies with General-Purpose Social Media Panels". In: *Journal of Data and Information Quality* 10.1, 4:1–4:24. ISSN: 1936-1955. DOI: 10.1145/3185048. URL: https://doi.org/10.1145/3185048 (visited on 07/18/2021).

Zhou, Peng et al. (2022). "Unknown Type Streaming Feature Selection via Maximal Information Coefficient". In: *IEEE*. DOI: 10.1109/ICDMW58026.2022.00089.

Zou, H. and T. Hastie (2005). "Regularization and variable selection via the elastic net". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.2, pp. 301–320. DOI: 10.1111/j.1467-9868.2005.00503.x.

Zuboff, Shoshana (2023). "The age of surveillance capitalism". In: *Social Theory Re-Wired*. Routledge, pp. 203–213.

Zuckerberg, M. (2019). *A Privacy-Focused Vision for Social Networking | Facebook*. URL: https://www.facebook.com/notes/mark-zuckerberg/a-privacy-focused-vision-for-social-networking/10156700570096634/.

Zullinger, Elissa M. et al. (Nov. 30, 1984). "Fitting Sigmoidal Equations to Mammalian Growth Curves". In: *Journal of Mammalogy* 65.4, pp. 607–636. ISSN: 0022-2372. DOI: 10.2307/1380844. URL: https://doi.org/10.2307/1380844 (visited on 11/28/2023).