Computer-aided Psychometrics with Natural Language Processing

Am Fachbereich Informatik
der Universität Hamburg
eingereichte

Dissertation

zur Erlangung des akademischen Grades
Doctor rerum naturalis
(Dr. rer. nat.)

vorgelegt von

Dirk Johannßen, M.Sc.

geboren in Elmshorn am 23. Juli 1990

Hamburg 2022

Dedicated to my beloved husband Kai, who stood by my side and believed in me.

Acknowledgements

This dissertation project has been a long journey with some steep climbs but also beautiful views at its summits. Without a myriad of people supporting this very project and me, none of this would have been achievable.

First and foremost I want to thank my supervisor Chris Biemann. Despite not having a classical academic background and having graduated from a university of applied sciences, Chris nonetheless believed in me and the project without hesitation. It was his hard work and dedication to support us and to gently push us beyond what was expected that has enabled this challenging balance between research, teaching, and employment. Chris has shown me some of the depths of our field and opened a fascinating world to be discovered. In times of joy and times of despair, Chris was the best of friends to share those emotions with for which I am greatly thankful.

Furthermore, I also thank my co-supervisor David Scheffer for his tireless support. Whenever there were psychological effects to be investigated, data to be collected, or interdisciplinary boundaries to be extended, David gave his wisdom, his time, and his effort to support me as if it were a matter of course.

Even though every Ph.D. candidate works towards this very day and everything is being planned way ahead, in the end, the thesis submission and defense organization had accelerated vastly. Hence I want to thank my Ph.D. committee for helping me organize and schedule these final steps. Namely, I want to thank Nicola Baumann, Janick Edinger, Mathias Fischer, and Anna Leffler for their support and selfless care.

All he wants from life is to be happy. His joyful attitude, the capability of seeing the bright side, and his dedication to his family, to his friends, and to me – his husband – were beyond care. Kai is one of the most selfless people I have ever met. This dissertation project has been his project and (almost) as much his mountain to overcome, as it was mine. More than once, I had to put myself and thus him through difficult times. He stood by my side. He believed in me – unconditionally. Your love is beyond comprehension. This thesis is dedicated to you, Kai.

Some might know, that the first spark of my curiosity was lit when I was given the opportunity to visit the USA after a rather unsuccessful yearning to find my place in society. My parents ensured me, that I had changed during this year abroad and have ever since strived for wisdom. Without the emotional and financial support of my parents, I could not have come this far. I am and will be forever thankful.

My colleagues at the language technology group have been the most amazing fellow researchers one can hope for. Time and again I complimented Chris on the LT members as being hard-working, empathetic, and amicable. The time spent was inspiring, joyful, and memorable. The LT members will still remember my times of despair in 2020, when it was our LT team, that cared, that helped, and that listened.

I thank my friends, who lent me their ears, gave me company, and tirelessly kept asking, how everything is going. These small occasions had yet a large effect on me. Many of you have read my final draft, given comments, listened through practice presentations, and given everything you know to help me through this marathon of a project.

At the beginning of this dissertation project, the balance between private life, academia, and work had been insurmountable. So much so, that I had to take drastic measures to move on to a new working place. The tireless support of my colleagues at the effective WEBWORK firm was not even questioned for a moment. When exhaustion, deadlines, or simply life had taken a toll, it were my colleagues who understood and helped. A special thank you to Matthias Finck, Josephine Kraus, and Annkristin Petri for their empathy and support.

Last but not least, I want to thank my students. The decision to pursue a doctoral degree came from those early occasions when I was allowed and able to teach. This, I knew, is what I want to do in life. Whilst learning about the world, my personality, my passions, and my world views have changed. It is this enthusiasm that I dream to pass on to my students. Academic education can be exhausting, confusing, and onerous. Whenever I was able to lower these obstacles through excitement, humor, or through profound explanations, i felt fullfilled and knew this is what I want to do in life. I also learned never to underestimate some of the young brilliant minds and to never judge a book by its cover. Dear students, you are the future. I am grateful to be a small part of it.

There are countless further people I could and should thank. To all of whom are not mentioned just yet, rest assured that your support has not been forgotten and will be held in my dearest memory.

List of Publications

The following works were published during and as part of the dissertation project:

Johannßen, D., Biemann, C., and Scheffer, D. (2022): Classification of German Jungian Extraversion and Introversion Texts with Assessment of Changes during the COVID-19 Pandemic. In Proceedings of the LREC22 workshop on Resources and Processing of linguistic, para-linguistic and extra-linguistic Data from people with various forms of cognitive / psychiatric / developmental impairments (RaPID-4). European Language Resources Association (ELRA), pages 31-40, Marseille, France.

Johannßen, D. and Biemann, C. (2020): Social Media Unrest during the COVID-19 Pandemic: Neural Implicit Motive Pattern Recognition as psychometric Signs of Sever Crises. In Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES). The COLING 2020 Organizing Committee, pages 74–86, Zurich, Switzerland.

Johannßen, D., Biemann, C., and Scheffer, D. (2020b): Ethical considerations of the GermEval20 Task 1. IQ assessment with natural language processing: Forbidden research or gain of knowledge? In Proceedings of the GermEval 2020 Task 1 Workshop in conjunction with the 5th SwissText & 16th KONVENS Joint Conference 2020, pages 30–44, Zurich, Switzerland (online).

Johannßen, D., Biemann, C., Remus, S., Baumann, T., and Scheffer, D. (2020a): GermEval 2020 Task 1 on the Classification and Regression of Cognitive and Motivational style from Text. In Proceedings of the 5th SwissText & 16th KONVENS Joint Conference 2020, pages 1–10, Zurich, Switzerland (online).

Johannßen, D. and Biemann, C. (2019): Neural classification with attention assessment of the implicit-association test OMT and prediction of subsequent academic success. Proceedings of KONVENS 2019, pages 68-78, Erlangen, Germany.

Johannßen, D., Biemann, C., and Scheffer, D. (2019): Reviving a psychometric measure: Classification and prediction of the Operant Motive Test. Proceedings of CLPsych 2019, pages 121–125, Minneapolis, MN, USA.

Johannßen, D. and Biemann, C. (2018): Between The Lines: Machine Learning for Prediction of Psychological Traits – a Survey. In: Proceedings of CD-MAKE 2018, pages 192-211, Hamburg, Germany.

Abstract

In the course of this dissertation, two state-of-the-art (SOTA) models were crafted for three psychological metrics, namely implicit motives, self-regulatory emotional levels, and the Jungian psychology types of extraversion and introversion. Novel and costly hand-annotated psychological data was published for free use and a shared task in the domain of aptitude diagnostics and implicit motives were conducted and its data distributed.

Not only does the automation of those metrics promise low-cost validation research in the domain of psychology. The chosen approaches of employing a feature-engineered logistic model tree (LMT) and a bi-directional long short-term memory network (Bi-LSTM) with attention mechanism paired with investigations of algorithmic decision-making push the psychological methodology towards human-like classification performance and greater explainability, compared with intuition-based annotation or word-based rule systems.

These models were applied to behavioral data sources, demonstrating predictability on the field of aptitude diagnostics, towards social unrest pattern recognition, and for the identification of individuals at risk of pandemic isolation fostering computer-aided psychology diagnostic empiricism.

As a result of this work, a prototypical implementation of these SOTA models was achieved, promising applicable tool support for empirical psychology. A funded and professional project for crafting a more sophisticated open-sourced NLPsych platform was initiated.

Lastly, steps in the direction of psychological pragmatics were taken. For the first time, the existence and alteration of a recently discovered fourth implicit motive *freedom* was verified by the utilization of the proposed NLP models, extending the underlying theory of this projective metric. Furthermore, intercorrelations between the researched metrics were measured and analyzed, extending the current knowledge in the field of psychology diagnostic empiricism.

x

Zusammenfassung

Im Rahmen dieser Dissertation wurden zwei Modelle auf dem Stand der Technik (State-of-the-Art, SOTA) für drei psychologische Metriken, namentlich implizite Motive, selbstregulierende emotionale Ebenen und die Jung'schen psychologischen Typen Extraversion und Introversion, entwickelt. Neuartige und kostspielige handannotierte psychologische Daten wurden zur freien Verwendung veröffentlicht und eine Shared Task im Bereich der Eignungsdiagnostik und der impliziten Motive wurde organisiert und die zugehörigen Daten veröffentlicht.

Die Automatisierung dieser Metriken verspricht nicht nur eine kostengünstige Validierungsforschung auf dem Gebiet der psychologischen Diagnostik. Die gewählten Modelle eines logistischen Entscheidungsbaums (Logistic Model Tree, LMT) und eines bi-direktionalen Long Short-Term Memory (LSTM) Netzwerks mit Aufmerksamkeitsmechanismus, gepaart mit Untersuchungen zur algorithmischen Entscheidungsfindung, vollziehen Klassifizierungen auf dem Gebiet der psychologischen Diagnostik auf annähernd menschlichen Niveau und verbessern die Erklärbarkeit im Vergleich zu intuitiven Annotationen oder wortbasierten Regelsystemen. Diese Modelle wurden auf behavioristische Daten angewendet, wobei die Vorhersagbarkeit auf dem Gebiet der Eignungsdiagnostik, der Erkennung von sozialen Unruhen und der Identifizierung von Personen mit dem Risiko einer pandemischen Isolation nachgewiesen wurde, wodurch die Forschung auf dem Gebiet der computergestützte psychologisch-diagnostischen Empirie vorangetrieben wird. Als weiteres Ergebnis dieser Arbeit wird eine prototypische Implementierung dieser SOTA-Modelle erwirkt, die eine Software-Tool-Unterstützung für die empirische Psychologie verspricht. Ein professionelles Projekt zur Entwicklung einer solchen Open-Source-NLPsych-Plattform wurde finanziert und initiiert.

Abschließend wurden erste Schritte in Richtung einer psychologischen Pragmatik unternommen. Zum ersten Mal wurde die Existenz eines jüngst entdeckten vierten impliziten Motivs *Freiheit* durch NLP-Modelle verifiziert, wodurch die zugrunde liegende Theorie dieser projektiven Metrik erweitert wurde. Darüber hinaus wurden Interkorrelationen zwischen den untersuchten Metriken gemessen und analysiert, wodurch der aktuelle Forschungs- und Wissensstand auf dem Gebiet der psychologischen diagnostischen Empirie erweitert wurde.

XII

Contents

I Introduction, Background, and Ethical Considerations

II   Personality Assessment in the Application Field of NLPsych

III Empirical Research of NLPsych for Aptitude Diagnostics, Social Unrest, and Pandemical Isolation

IV Psychological Pragmatics of NLPsych

V Conclusion and Outlook

List of Tables

XXI

XXII

List of Figures

List of Abbreviations

ACM	Association for Computing Machinery
AC	Assessment Center
ACL	Association for Computational Linguistics
AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformer
CNN	Convolutionary Neural Network
CTT	Classic Test Theory
DAG	Directed Acyclic Graph
DL	Deep Learning
GDP	Gross Domestic Product
GPT	Generative Pre-Trained Transformer
IG	Information Gain
IPT	Implicit Personality Test
IQ	Intelligence Quotient
IQt	Techniacl Intelligence Quotient
IRT	Item Response Theory
LIWC	Linguistic Inquiry and Word Count
LMT	Logistic Model Tree
LSTM	Long Short-Term Memory
MBTI	Myers-Briggs Type Indicator
MIX	Motive Index
ML	Machine Learning
MLM	Masked Language Modelling
NLP	Natural Language Processing
NLPsych	Natural Language Processing for Psychology
NLTK	Natural Language Toolkit

NSP	Next Sentence Prediction
OCEAN	Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism
OMT	Operant Motive Test
POS	Part-of-Speech
PSE	Picture Story Excercise
RNN	Recurrent Neural Network
SOTA	State-of-the-Art
SRL	Semantic Role Labeling
STTS	Stuttgart-Tübingen-Tagset
TAT	Thematic Apperception Test
ViQ	Vissual Questionnaire
XAI	Explainable Artificial Intelligence

Part I

Introduction, Background, and Ethical Considerations

CHAPTER 1

Introduction

Since the deep learning (DL) framework Tensorflow was made available under the Apache-2.0 Open Source Licence in 2015 by Abadi et al. (2016), not only has deep learning experienced unprecedented popularity and brought up a myriad of novel systems, which some would call omnipotent in their respective task, but artificial intelligence (AI) as a whole (Görz et al., 2020, p. 9). For the time being, humanity has entered the third wave of AI. Systems are capable of not only competing with humans in board games, which were thought to be too complex for machines to sufficiently play (e.g. the Chinese game of Go), but have even defeated world-leading players (Silver et al., 2016). Those players called said system Alpha Go *'God like'*. Strategies had been employed by reinforcement learning, which had never been utilized in observed Go matches (Görz et al., 2020, p. 9). Equally impressive, those systems riddle even the creators, since their explainability is a yet challenging and unsolved task (Danilevsky et al., 2020). Autonomous driving is on the brink of maturity and possesses – among others – superior systems for real-time imagery processing (Pisarov & Mester, 2020). Deep fakes approach human indistinguishable qualities (Westerlund, 2019) and (thus far in-domain) AI phone calls pass even extended Turing Tests by producing human-like voices, which even optimistic researches thought would be able to achieve in two decades (O'Leary, 2019; Russell & Norvig, 2016).

Natural language processing (NLP) takes a part in many of these novel developments and is concerned with processing large amounts of natural language data (often textual, i.e. *corpora*) with the goal of understanding and producing human language. From 2015 onward, NLP also experienced substantial breakthroughs and has fostered systems and applications not only for the NLP community, but many interdisciplinary scientific and professional fields. Modern NLP systems like the Generative Pre-Trained Transformer 3 (GPT3, see Subsection 2.3.4) by OpenAI[1] (Brown et al., 2020) with its 175 billion parameters are able to solve tasks that require

---

[1] https://openai.com

on-the-fly reasoning, which also was thought to be solvable in decades from 2015, not mere years (Russell & Norvig, 2016, p. 1021 ff.).

Despite those breakthroughs in many AI disciplines and NLP in particular, new approaches, new models, and some answered research questions have evoked even more problems to be solved. The limited capabilities of NLP prior to the third wave of AI, systems could solve linguistically interesting and complex problems, but were of limited use for other scientific fields. Nowadays, deep neural natural language models have shown such utilization flexibility and processing power, that they have become even expected research methodologies for scientific application fields such as empirical psychology, which slowly but surely adapt and embrace those methodologies. Yet, recent surveys come to the conclusion, that – despite becoming expected standard procedures – NLP for empirical psychology can still be considered a novel niche.

1.1 Current Challenges and Dissertation Contributions in the Fields of Psychology and NLP

NLP has experienced major breakthroughs in terms of approaching human-level performances on natural language tasks that were thought to be hardly solvable over the past few years. Furthermore, a paradigm shift took place in the area of computational linguistics and NLP, now resulting in large-scale transformer models with billions of parameters for solving most NLP problems.

The scientific field of psychology has experienced major paradigm shifts as well, reaching from early projective tests to behaviorism or empiricism and even neurobiological clinical psychology. Language has played an important role from the beginnings of psychology e.g. psychoanalysis.

The interdisciplinary cross-domain approach of performing NLP for psychological evaluations (NLPsych) promises a better understanding of the connection between psychology and language and evaluation objectivity hardly achievable for many psychological metrics. Furthermore, NLP research profits from approaches toward data sparseness, domain specificity, and decision explainability. NLP already investigates complex problems such as word-sense disambiguation, sarcasm detection, or colloquialisms. Psychology adds another layer of complexity to this already ambiguous research object of natural languages.

However, not all natural language problems are solvable with said models or at least remain highly challenging. If the uttered natural language differs in its predictive signals for target labels from the broad majority of tasks (i.e. the meaning of certain words differs greatly from the common usages, being highly domain specific), or if necessary labeled data is both complex and sparse, even large scale transformer models fail to sufficiently model those tasks.

Such a highly specific domain, that both require labeled data due to diverging language meaning and suffers from data sparseness, is the psychological domain. The scientific field of psychology has experienced major paradigm shifts as well, reaching from early projective tests to behaviorism or empiricism and even neurobiological clinical psychology with brain signals and electroencephalographies nowadays. Language has played an important role from the beginnings of psychology with e.g. psychoanalysis. In the middle of the 19th century, language declined in its perceived importance but attracts new attention again due to its empirically researched links to brain functions.

1.1.1 Challenges automating Implicit Methods

Modeling implicit methods (see Chapter 5) is challenging. Previous approaches utilized lexicon-based rule models, which could not satisfy psychological quality criteria and achieved a Pearson correlation coefficient (see Subsection 2.1.4) of $r = .2$ (Schultheiss, 2013). Machine learning approaches were able to reach moderate convergences between automatically classified and manually labeled implicit motives with correlations reaching from $r = .22$ to $r = .5$ per motive (Pang & Ring, 2020). However, these scores are still too low and they have been achieved on an outdated implicit test called the thematic apperception test (TAT, see Section 5.1). A differentiation into motives paired with self-regulatory levels was not considered. Automation research on more modern and condensed diagnostic projective tests like the Operant Motive Test (OMT, see Section 5.2) has not been conducted.

These rather unsatisfying approaches for automating implicit motive classifications or codings indicate the challenge of this pursuit.

1.1.2 Hand-annotated Psychology Data Sparseness

Psychology adds another layer of complexity to this already ambiguous research object of natural languages (Johannßen & Biemann, 2018).

However, hand-annotated psychological data is sparse and mostly expensive to gather, label, or obtain. Many valuable annotated data sets especially from clinical psychologists can not be shared with the broader research community due to strict data protection laws and the sensitivity of these information (Rainey et al., 2020).

1.1.3 Dissertation Contributions

In the course of this dissertation, two state-of-the-art (SOTA) models were crafted for three psychological metrics, namely implicit motives, self-regulatory levels, and the Jungian psychology types of extraversion and introversion. Novel and costly hand-annotated psychological data was published for free use and a shared task in the domain of aptitude diagnostics and implicit motives was conducted and its data distributed.

Not only does the automation of those metrics promise low-cost validation research in the domain of psychology. The chosen approaches of employing a feature-engineered logistic model tree (LMT) and a bi-directional long short-term memory network (Bi-LSTM) with attention mechanism paired with investigations of algorithmic decision-making pushes the psychological methodology towards human-like classification performance and greater explainability, compared with intuition-based annotation or word-based rule systems.

These models were applied to behavioral data sources, demonstrating predictability on the field of aptitude diagnostics, towards social unrest pattern recognition, and for the identification of individuals at risk of pandemic isolation fostering computer-aided psychology diagnostic empiricism.

Lastly, steps in the direction of psychological pragmatics were taken. For the first time, the existence and alteration of a recently discovered fourth implicit motive *freedom* was verified by the utilization of the proposed NLP models, extending the underlying theory of this projective metric. Furthermore, intercorrelations between the researched metrics were measured and analyzed, extending the current knowledge in the field of psychology diagnostic empiricism.

1.2   The Study of Language

Some of the fascination for NLP researchers in this scientific AI field stems from the challenge that the modeling and formalization of this ambiguously and organically formed means of communication poses a substantial intellectual challenge. For instance, in contrast to formal (i.e. mathematical) languages, the number of ways how to continue a started sentence approaches infinity. Nonetheless, there even exist well-formalized statistic natural language processing approaches for quantifying and estimating probabilities for possible next-word candidates. Language is complex and hard to analyze, but it is not unquantifiable (see e.g. Section 2.2.4).

1.2.1   Psychology as a Further Layer of Natural Language Complexity

From an NLP point of view, most psychometrics can be considered as classification tasks. Oftentimes, metrics for measuring cognitive phenomena aim to reduce the complexity of a subject's observable mind to manageable discrete categories or classes. Initially, this stigmatic simplification might appear wasteful to computer linguists, since such an approach disregards many signals. For psychologists, however, this reduction of information is necessary to reduce the system's complexity to such an extent as to be able to discover patterns and recognize cognitive processes beyond what is observable (e.g. verbalized utterances or language). Even with this reduction, psychometrical labels and such supervised learning of NLP for psychology add a substantial complexity layer on top of already challenging classification tasks.

The difficulty lies in the signal, that a machine learning model needs to pick up in order to solve the task. For many NLP tasks, the mere existence of sufficient signals can manually be observed by the designer of a model, e.g. sentiment tasks, hate speech detection, or spam classification. Even though it is challenging to craft state-of-the-art (SOTA) models on those and other tasks, and even though there will be many instances, where annotators will have difficulty identifying or agreeing upon a label, for most instances a human annotator would have a strong sense of which label to choose. As for psychology and textual information that might or might not contain relevant signals for the task at hand, even identifying signals and choosing the correct label can be seen as a challenge itself. Štajner & Yenikent (2021) analyzed stumbling blocks in modeling the Meyers-Briggs Type Indicator (MBTI, see Section 4.2), and identified diverging content and style towards concurring labels as one of those stumbling blocks amongst others. Accordingly, people can consciously alter the content of answers, but not the style (Manning & Schütze, 1999) thus making it difficult to measure cognitive processes only by analyzing content – which is the main approach of most psychology diagnostic metrics (Johannßen & Biemann, 2018).

This, in turn, is what makes this application domain so appalling for challenge-seeking NLP scientists, as well as progressive empirical psychologists.

1.2.2   NLPsych as Impactful Application Domain

In the majority of cases, empirical psychology utilizes psychometrics for its diagnostics. In turn, the combination of NLP and psychology (NLPsych) mostly performs supervised machine learning. There is a vast amount of human-produced textual data available. However, only a fraction of this textual data comes with sufficient labels. An even smaller fraction of this fraction comes with labels valuable for empirical psychologists. Mostly, the scientific community is forced to produce psychometrically labeled data manually and with high expertise – and thus, for high costs. Empirical psychological data with sufficient labels is sparse and hard to come by. For this research hinder some but necessary data protection measurements further tighten the data availability.

For empirical psychology, the collection, processing, and annotation of data often serves one assessment only and needs to be repeated each time there is the demand for new insights. The interdisciplinarity of this work between NLP and psychology promises to support and resource procurement for empirical researchers (not to be mistaken with the complete automation of psychometrics, which should be viewed critically, as further elaborated in Section 3.2).

Since the application domain of psychology adds a complexity layer to already challenging NLP problems, computer linguists and NLP researchers extend their knowledge and available set of methodologies by embracing this application domain. For psychologists, the novel and

potent methodology offered by NLP models allow for an elevated potential for new research findings.

Lastly, NLPsych is concerned with impactful real-world tasks and challenges, such as dream language assessment, aptitude diagnostics, depression detection, organizational psychology, or mass psychology up to panic, disarray, or social unrest.

All of those challenges will be met and fathomed by answering the following three research questions.

1.3 Research Questions

In this section, the central research questions will be stated and elaborated upon.

1.3.1 RQ1: Can NLP systems model psychological metrics?

Due to recent and rapid advancements in the field of machine learning (ML, see Section 2.2), the capabilities of NLP models for capturing complex signals and world knowledge have been elevated. On the one hand, models such as the third Generative Pre-Trained Transformer (GPT3) (Brown et al., 2020) show human-like capabilities for solving reasoning tasks. On the other hand, the application domain of utilizing NLP for modeling psychometrics mostly requires a task understanding even beyond those capabilities of humans, which were not professionally trained and usually require supervised machine learning, i.e. sophisticated labels. Therefore, the first research question is:

> *RQ1: Can NLP Systems model psychological metrics?*

1.3.2 RQ2: Do modeled psychometrics predict behavioral observations?

For every empirical discipline, there are quality criteria to quantify the success of conducted research. For NLP, an established measure is the F_1 metric (see Subsection 2.2.9), which calculates the harmonic mean between precision and recall. If a model achieves a respectable F_1 measure on a held-out test set, reporting it to the community paired with further metrics and illustrations such as learning curves or a confusion matrix, is sufficient for convincing the community of the model's high qualities. As for the scientific field of psychology, displaying a tool's ability to reproduce the results consistently – the so-called reliability – is not enough to justify its utilization for measures. To convince psychological researchers of their validity, tools or models have to be tested on controlled randomized behavioral experiments.

Whilst psychology in the US formed out of philosophy and thus embraced theoretical concepts such as consciousness and the self, European philosophy was rather concerned with clinical laboratory conditions (Collin et al., 2012). The introductory work on conditioning animals by Pavlov (Pavlov, 1906) popularized psychological empiricism in Europe and the United

States (US) from the early 20th century (Clark, 2004). Even though the scientific landscape has become more diverse, psychology is still greatly concerned with empiricism.

Since this dissertation aims to explore NLP for automating and researching psychometrics, any proposed model of this work not only needs to fulfill the quality criteria on the field of NLP, but oughts to demonstrate its validity paired with further requirements from the psychological domain such as explainability and understandability, leading to the second research question:

> *RQ2: Do modeled psychometrics predict behavioral observations?*

1.3.3 RQ3: Do automated psychomerics correlate in their assessment on similar texts?

As stated in RQ1, measurable psychological signals are mostly too vast and diverse to discover underlying cognitive processes. Therefore, empirical psychology reduces complexity by introducing simplified metrics, which aim to categorize and stigmatize such mental processes. In practice, Freudian psychoanalysis, where a practitioner loosely listens to patients and tries to discover early lifetimes experiences, is nowadays considered untargeted. Instead, questionnaires and manual-driven behavior observations are employed in accordance with psychometrics.

Nonetheless, psychological researchers with a focus on language have identified linguistic markers, which empirically correlate with real-world observations. One example is the work by Pennebaker et al. (2014), which identified function words as linguistic markers for cognitive processes.

Time and again well-established psychometrics such as the five-factor inventory (also called *Big Five*, see Section 6.3) were substituted by- and correlated with linguistic markers or other psychometrics (Rammstedt et al., 2018). This leads to the assumption, that there are explanatory variables or signal generators beyond those metrics. The following third research question aims for broadening the available signals from psychometrics to language in general:

> *RQ3: Do automated psychomerics correlate in their assessment on similar texts?*

1.4 Dissertation Structure

This thesis is structured in five main parts. **Part I** provides an introduction with this work's main impact, research questions, and related published work (Chapter 1). Background information are given in Chapter 2. It consists of fundamentals of personality diagnostics, an overview of machine learning, NLP, and the application domain of performing NLPsych. Since the combination of automation, text processing, and psychology holds the potential of being misused, Chapter 3 is concerned with the ethical aspects of this work.

Part II focusses on the personality assessment in the application field of NLPsych and lays the fundamentals for the experimental studies of this work. Chapter 4 introduces the reader to the concept of Jungian psychological types, which found an influential paradigm of early European psychology. Implicit motives, a type of self-reflective projective testing procedure, is the subject of Chapter 5.

The proposed model architectures and psychometrics are utilized for broad validation studies in Part **III**. Firstly, the utilization of NLP for performing aptitude diagnostics (paired with a recommended reference to the ethics in Chapter 3) is described in Chapter 7. Social unrest as detected and analyzed by the combination of NLP methods and psychometrics is described in Chapter 8, and Chapter 9 contains research on Jungian psychology types and feelings of isolation during the COVID-19 pandemic.

This thesis not only aims to automize and empirical research and psychometrics, it furthermore aims to provide linguistic markers and discovered patterns to enhance the utilized and research psychometrics. This part of pragmatics and discourse is described in Part **IV**. Its Chapter 10 covers underlying similarities between all of the proposed psychometrics, linking linguistic signals directly to validation studies and thus approaching pragmatics.

Finally, this work comes to a conclusion in **Part V**, which summarizes the psychometrical assertion of personality traits, the empirical evidence collected over the course of the dissertation project, recapitulates findings on overreaching metric similarities and answers the proposed research questions. Lastly, future directions of the ongoing scientific activity of NLPsych are discussed and an outlook is provided.

Following this introduction we first lay the necessary background information in the following Chapter 2. It covers both, information concerning psychology, and information concerining NLP.

Background

This chapter contains necessary terminology, definitions, and background information. It provides a collection of thereafter utilized principles and is structured into the sections *psychological personality diagnostics* (Section 2.1), *machine learning* (Section 2.2), NLP (Section 2.3), NLPsych, and related work (Section 2.4). The broader research objects, namely psychological diagnostical testing procedures – and thus fundamentals and backgrounds as well – are explained in more detail in their respective chapters (Psychological Archetypes in Chapter 4, Implicit Motives in Chapter 5, and Personality Questionnaires in Chapter 6).

2.1 Psychological Personality Diagnostics

The understanding and definition of the term *psychological diagnostics* has changed over the past years and has not yet found a set consensus. Schmidt-Atzert et al. (2018) collect some aspect from related work and define the term as a subdiscipline of psychology, which aims to answer questions that describe, classify, or predict human behavior and experience of individuals or groups. Furthermore, psychological diagnostics collects information (i.e. data) and interprets them with psychological knowledge and under the utilization of psychological methods, which satisfy scientific standards.

Accordingly, one key aspect is the consensus of the psychological community as to which methods satisfy scientific standards. This, at times, excludes approaches with methodologies that have not (yet) been established. Some major methodological breakthroughs in NLP in the past years were achieved by novel architectures, word representations, and paradigm shifts from empiricism and psychoanalysis to clinical and neurobiological psychology nowadays (see Chapter 1). However, these novel architectures have not yet been established as research methodology in the field of empirical psychology due to what psychologists call the *Classic Test Theory* (CTT). This CTT formulates five basic axioms, which psychological tests ought to fullfill in order to be acknowledged as scientifically sound. Due to this characteristic of

axiomatic validation of diagnostical tests, everything that does not hold up to that axiom can not – in the consensus of psychology – produce trustworthy results.

2.1.1   Classic Test Theory (CTT)

Psychology originated at the end of the 19th century. During the enlightenment of the 18th century, natural sciences exceedingly explained many world phenomena, as well as biological processes. René Descartes Descartes (1956) first radically thought of and theorized about mind-body dualism. This dualism states that mental phenomenons are to be separated from the physical form of a being or the body (Grankvist et al., 2016). Whilst the bodily processes could be described and researched, the mind remained an unknown entity. The pursuit of knowledge after those initial mind-body dualism thoughts was the pursuit of what is now known as psychology.

As with many scientific fields, different schools of thought and paradigms emerged over the course of psychological research. While central European psychology mainly focused on strict laboratory and measurable research, which aimed to explain neurological processes within the mind, psychologists in the US focused on Behaviorism, which can be thought of as situation-reaction observations and predictions as research objectives without investigating the underlying neural processes (Collin et al., 2012).

As described by Schmidt-Atzert et al. (2018), Charles Spearman, an English psychologist, focused his work on statistical evaluations of empirical research in a vast quantity of publications, including one where Spearman described the *rank correlation coefficient* (Spearman, 1904). In 1950, Spearman's work was summarized and contextualized by Gulliksen (1950). Especially the work by Lord et al. (1968) fostered empirical psychological work in central Europe and around the world and became the basis for the so-called classical test theory (CTT).

The CTT is a theory of reliability and justifies the measurement precision of psychological testing evaluations. This CTT reliability has become a cornerstone of trustworthy psychological research besides the also impactful item response theory (IRT, Steinberg & Thissen Steinberg & Thissen (2013)), which is more applicable to implicit methods. The main premise of the CTT is that every psychological test is prone to errors and will contain errors. There are true values of each psychological trait for the individual. However, those true values can never be measured with such precision and without errors that this true value emerges. Therefore, the basic questions in CTT become the questions of significance, conditional probabilities, and distributions.

The CTT consists of five main axioms or principles. Every other statistical evaluation of test results emerges from those axioms. From those axioms, subsequent evaluation statistics and metrics can be derived and are described – amongst others – in Section 2.1.

1st Axiom: $X_i = T_i + E_i$

This first axiom states simply that the observed value X_i for person i is the sum of the true value T_i added with an error E_i. Even though this equation appears to be self-explainatory, it nonetheless renders one of the main issues addressed by the CTT: the observed value is never equal to the true value and always contains an error, which obscures the observed value. The error E_i can be both positive and negative, thus the observed value might be higher or lower as compared to the true value. However, the CTT states that E_i can never be exactly 0.

The true value T_i is immutable and stays consistent with every testing procedure being performed. Collin et al. (2012) presents an example: assuming that two groups of psychologists each create a very precise intelligence quotient (IQ) test. One participant takes both tests but for one of those tests, the participant achieves an IQ of 120, for the other this participant scores 130. Besides both tests' high precision, the values differ. When performing both tests with e.g. 100 participants, both tests could e.g. correlate with a Pearson population correlation coefficient (see Subsection 2.1.4) of $\rho = .6$. The explanation for this delta lies within the first axiom: each test measures a slightly different *intelligence*. The true IQ score can not be measured, since there is no amount of intelligence, but simply an observable operationalized manifestation of the personality trait that one would call *intelligence* (the term is debated upon and might better be called *situational skill*).

The first axiom also implies that $T_i = E(X_i)$, with T_i being the true value and $E(X_i)$ being the expected value of X_i. This equation states that the more often a test is being conducted, the closer the average true value, which exists for every test performed on a person, and the expected value become.

2nd Axiom: $E(E_i) = 0$

The second axiom states that the error E_i for a test being performed with person i is expected to be 0. This second axiom apparently contradicts the first axiom, which states that an error has to be expected for every test and that this error is not 0.

However, the second axiom does not state that an error is 0 for a single execution of a test is performed, but rather that if a test is being performed for an infinite amount of times, it will – on average – eventually approach 0.

The main idea behind this second axiom is that every influential variable introduced to a test is unsystematic. Introduced unsystematic variables (i.e. variables that can not be measured or predicted and that are variable per execution) might include mistakes during the test construction (e.g. including questions or items that are ambiguous), during the execution (e.g. differing lighting or noises), or during the evaluation (e.g. poor standardization of annotators).

Since the error E_i approaches 0 with an increasing number of executions, it can also be stated that: $E(E_i) = 0 \rightarrow E(X_i) = E(T_i) + E(E_i) = E(T_i) + 0 = E(T_i)$, which reads that if a testing procedure is conducted with an increasingly large sample size, the error becomes continually

smaller. The more often a testing procedure is performed on one individual, the closer the averaged results come to the true value T.

3rd Axiom: $P(E|T) = P(E)$

This axiom states that the error E_i is independent from the true value T_i. In other words, the probability for the error E given the true value T is equal to the probability of the error E. This axiom can be understood in terms of the influence of an error. If e.g. an error occurs during the testing procedure, distracting a participant by loud noises, then each person should be distracted individually and independently to one another by those noises – this distraction would be independent of the true value and should impact both persons, even if their true potential would be very different (e.g. during an IQ test, an average person with a true IQ of 100 would lose 2 IQ points due to the noise as well as a person with a true IQ of 120, which would be assumed to lose 2 IQ points during this disturbed IQ testing procedure, as well – independent of one another and only depending on the noise and distraction).

4th Axiom: $Corr(E_A, T_B) = 0$

The 4th axiom is closely related to the 3rd axiom and states that the correlation between the error of the first test execution i) and the error of the second test exection ii) is equal to 0. In other words, the true value of the whole testing procedure with all its executions is independent of the errors per execution, if this error is not systematic (i.e. a flaw in the testing design) but exogenous. This axiom thus states that if e.g. two participants perform similarly in a multitude of testing procedures and react similarly distracted by noise during one testing procedure, this does not correlate to any results of other testing procedures and thus with increasing repetitions, this error approaches 0 (see axiom 2). If an error was observed during a testing procedure, this does not justify omitting any other testing procedure no matter the type of the exogenous error.

5th Axiom: $Corr(E_A, E_B) = 0$

Lastly, the 5th axiom states that the errors E per testing procedure are independent from one another. That is, even if e.g. the evaluation of the results of a testing procedure might be erroneous, the magnitude and type of this error would not correlate with any other type of error being introduced the subsequent times a test is being conducted.

2.1.2 Aptitude Diagnostics

Psychological diagnostics is an application of empirical psychology, which has become one of the most broadly utilized knowledge transfer domains. Amongst the available testing procedures, aptitude diagnostics are surveyed by Roth & Herzberg (2008) to be the second most popular type of testing procedure, only being exceeded by personality testing (see Subsection 2.1.3).

The term *aptitude* is just an indicator. As the classical test theory (CTT) in Subsection 2.1.1 states, that a true value can only be indicated and always differs from a tested value by an unknown but assumable testing error. Aptitude indicates capabilities, skills, and knowledge. Besides those three main traits, the existence of attention and focus are discussed but at times subsumed under intelligence and are considered constructs or traits (Schmidt-Atzert et al., 2018). A psychological construct is an identifier or name for a broader set of behaviors, which might be situational and is mostly learned over time, whilst a trait or personality trait rather describes the underlying attitudes, which stay stable over time and are innated at birth (Fried, 2017).

Aptitude: capabilities, skills, and knowledge

Aptitude can be understood as performance per time unit or period. The three indicated traits capabilities, skills, and knowledge can be further categorized. Capabilities describe the potential of acquiring skills or gaining knowledge and thus are trainable. Skills and knowledge are moderated by the capabilities of an individual. The lines between those three indicated traits are blurred and highly correlated.

Aptitude diagnostical testing procedure differ from personality tests (see Subsection 2.1.3) in that whilst the personality tests aim to measure the default mode and behavior of an individual, aptitude diagnostics aim to measure the peak performance per given period. The participants have to be informed that their peak performance will be measured and how it will be measured. As Schmidt-Atzert et al. (2018, p. 183) summarize, Hausknecht et al. (2007) surveyed a practice effect for aptitude tests. All in all, practice of a certain type of test only influenced subsequent tests with $r = .26$ with r being the Peasron correlation coefficient (see Subsection 2.1.4), which is not a large effect. However, when the same test is being undertaken three consecutive times, this effect grows to $r = .56$. So-called *coaching* increased it to $r = .7$. In order to challenge this advantage of experienced participants, the investigator could provide all participants with as much information as possible on the testing procedure and the test items themselves, to nihilate those advantages towards a fairer aptitude testing procedure (Schmidt-Atzert et al., 2018, p. 183).

Attention and focus

Besides those three indicators for aptitude, capabilities, skills, and knowledge, there are the two constructs attention and focus, which are debated upon. Both constructs are of importance during aptitude diagnostical procedures. However, they rather moderate or influence the other indicators or traits. It is unquestionable that attention and focus are necessary to develop the full potential during aptitude testing. However, it remains an open question whether or not they should be viewed as part of either capabilities, skills, or knowledge. Since aptitude diagnostics aim to measure potential, which would be of great importance during certain professional careers, for school grades, or in extreme situations, attention and focus have to be considered as important and worthwhile being measured (Schmidt-Atzert et al., 2018, p. 188).

Testing procedures that measure attention mostly present participants with a stimulus and track the time it took participants to react to said stimulus. Furthermore, it can also be measured whether the reaction was correct. Whether or not there is an incorrect reaction to a given stimulus depends on the type of test. Some tests provide participants with different stimuli to which one out of many possible reactions is considered correct. Table 2.1 displays an overview of different sub-types of attention, namely alertness, focussed attention, shared attention, permanent attention, and vigilance.

Attention type	Test principle	Example
Alertness	Simple stimuli to be reacted upon quickly	X on the screen
Focussed attention	React to simple stimuli of a certain class amongst few stimuli	Present some patterns of which two are critical stimulus patterns
Shared attention	React to at least two stimuli of very distinct classes	Different Xs in a 4x4 matrix and acoustic high/low noises
Permanent attention	Focussed shared attention over a long period	Five to seven ever-changing triangles with peak to the top or to the bottom for up to 35 minutes
Vigilance	React to rare stimuli over a long period of time	High illuminated points appear on a circular track for up to 70 minutes

Table 2.1: Different types of attention testing procedures mainly differ in terms of the number of stimuli to react to and the duration of the testing procedure (Schmidt-Atzert et al., 2018, p. 188).

Focus tests aim for measuring the ability to ignore distractions during an error-prone and oftentimes tedious task. The distraction does not necessarily emerge from exterior stimuli, but rather from pronounced similarities of wrong items to those items to be selected. Focus tests most often contain search tasks, where participants are asked to quickly identify items qualified in accordance of shape, quality or other patterns, whilst ignoring (or at times crossing out) distracting items that share some features but do not qualify as being correct. Some focus tests employ simple calculus. The Cronbach's α as a measurement of reliability is of great importance for focus tests, as well as a good test-retest reliability (see Subsection 2.1.4) (Schmidt-Atzert et al., 2018, p. 197).

Intelligence testing

Intelligence quotient (IQ) testing procedures are considered to be the most sophisticated and best-researched types of general aptitude diagnostical tests. IQ tests correlated with school grades and subsequent job performances with $r = .5$, which is considered high amongst aptitude tests. Even though IQ tests contain the term *intelligence* in their name, the term is not properly defined. Intelligence in terms of IQ testing can most likely be translated to skills or reasoning. Some IQ tests aim to test this *core intelligence* by testing reasoning capabilities. Others aim to test skills in a more broad sense and include testing components such as calculus, linguistic skills, or spatial thinking (Schmidt-Atzert et al., 2018, p. 203).

IQ testing, even though amongst the most valid, reliable, and stable testing procedures known to psychological diagnostics, has been heavily criticized in terms of fairness and correctness (see Sections 3.3 and 7.5). Firstly, fairness can be achieved by carefully selecting the

base population to be as representative as possible for individuals participating in a test. Furthermore, cultural fairness can be approached by focusing on so-called fluid intelligence rather than crystallized. Fluid intelligence is situational and requires reasoning skills, whilst crystallized intelligence can mainly be acquired by being trained in school or by the use of language (often dependent on the upbringing household) (Schmidt-Atzert et al., 2018, p. 204).

School aptitude

Besides even more generalized approaches of measuring the whole spectrum of aptitude, including capabilities, skills, knowledge, attention, focus intelligence, lexical knowledge, memory, and different stimuli such as acoustic or visual by e.g. Carroll (2010), which did not emerge to be broadly utilized, another important area for aptitude diagnostics is the aptitude for attending or graduating (mostly high-) schools. Those types of aptitude diagnostical procedures can be divided into school acceptance tests and school aptitude tests. Acceptance tests aim for identifying the necessary capabilities and skills (but not the knowledge) for acquiring taught subjects during the school years. Those tests mostly include calculus, logic, and reading. If a participant is too young for e.g. already reading, symbol recognition can be measured. School aptitude tests try to remove subjective aspects introduced by teachers during school grading and employ more standardized testing procedures. Most of those tests are not broadly applicable, since young students develop rapidly. A skill set suitable for a school year can be insufficient just a year after. The most valid school aptitude tests are developed by including the school's curriculum for a certain year (Schmidt-Atzert et al., 2018, p. 237).

2.1.3   Personality Testing

The goals of psychological personality diagnostics can be contrary: on the one hand, psychologists aim for a precise characterization of traits that predict and explain behavior. On the other hand, however, those characterizations should not be too detailed, since behavioral observations, development implications, or predictions of e.g. group behavior would suffer from fragmentation. Therefore, personality testing and diagnostics have developed tools and methods that have reduced possible target classes to a minimum without losing much of their descriptive power. Those methods are referred to as psycholexical. The well-known five factor inventory or *Big Five* (for details see Chapter 6) is a psycholexical questionnaire procedure, which consists of the OCEAN target classes: O - Openness, C - Conscientiousness, E - Extraversion, A - Agreeableness, and N - Neuroticism (McCrae & Costa Jr., 1999; Goldberg, 1981).

Observations and questionnaires

Mostly, personality tests emerged from empirical observation studies. Assessment centers (ACs) are mostly employed for human resource aptitude diagnostics and recruitments and measure aptitude by asking participants to engage in exercises, observing their behavior, and

evaluating this behavior, oftentimes paired with (self-) questionnaires. Since ACs focus on observations made by professionally trained psychologists and are highly standardized, they are suited for developing and validating personality tests. For measuring personality, different paradigms have emerged: questionnaires (either self-administered or administered by professionals), diagnostic interviews, or implicit tests (for implicit tests, see Chapter 5).

The most broadly applied method for measuring personality are self-conducted questionnaires. Those are usually highly standardized, document every relevant information, are cost-effected, and widely accepted by both, the evaluators and the participants. Since most answers are pre-formulated and offer little to no interpretation, those questionnaires can be analyzed by well-known and highly comparable descriptive statistics. Furthermore, questionnaires are said to possess high evaluation objectivity, which is a quality criterion in psychology for similar results independend of the individual evaluating the results.

One of the major setbacks of (especially self-) questionnaires is the necessity for participants to be able to reflect on themselves and on their behavior without any bias or misconception. However, many studies have shown that so-called explicit methods suffer from biases. The term explicit refers to participants being forced to consciously decide upon their answers. A severe bias is the socio-desirability or socio-expectation bias. This bias states that individuals can not neutrally reflect upon direct questions on personality traits, which they associate with normative values, since they unconsciously reflect how peers would react (e.g. negatively) if they knew about their opinions and thoughts (Paulhus, 1984; Schüler et al., 2015; Brunstein, 2008). As Schmidt-Atzert et al. (2018, p. 237) point out, a study conducted by Post et al. (2008) revealed a delta of 9.4% between answers given by pregnant women on their smoking behavior as compared to their observed smoking behavior.

One way of counteracting this socio-desirability bias is a forced-choice format, where participants do not react with *yes* or *no* to items, but rather have to choose between two descriptions, which are aimed to be similarly desirable. A more promising approach is the employment of implicit methods, which – per design – avoid the explicit nature of questions and thus avoid this socio-desirability bias (Schultheiss & Brunstein, 2010).

Implicit methods

A different paradigm of measuring personality to explicit personality tests are implicit personality tests (see Chapter 5 for details). Explicit tests employ explicit methods, which are self-attributed and are conscious, direct responses to social incentives. Implicit methods, however, measure task-intrinsic incentives (Schultheiss & Brunstein, 2010, p. 16) and are measured via projective or associative testing procedures. During an early type of such implicit tests, the Thematic Apperception Test (TAT) by Murray (1943), participants are presented with ambiguous imagery and are asked to associate its content and situation, resulting in a *story*. Subsequent tests include the Operant Motive Test (OMT, by Kuhl & Scheffer (1999)), which condenses the test. The main goal of those tests is the implicit effect that if participants see

or even just imagine a social situation and are asked to interpret the situation, emotions, and intentions of involved persons, they project their unconscious desires upon those persons. Those projections and resulting stories can be interpreted in terms of personality constructs or traits (Sarges & Scheffer, 2008; Baumann & Scheffer, 2010).

2.1.4 Quality Criteria in Psychology

Many quality criteria in the field of psychological diagnostics directly result from the five main axioms of the classical test theory (see Subsection 2.1.1). The quality of diagnostic tests does not solemnly consist of the results, but also of the testing procedure itself (e.g. whether it is easy to conduct). Furthermore, the quality of a testing procedure as a whole depends on the quality of each unique item of this test. Lastly, a psychological diagnostic test can also be evaluated on its quality during the creation of the procedure, i.e. how scientifically sound the test was constructed.

The most basic and over the course of centuries unchanged quality criteria are objectivity, reliability, and validity. Other quality criteria have changed over time. An overview of the most important quality criteria by Schmidt-Atzert et al. (2018, p. 145) is provided in Table 2.2.

Quality criterion	Objective
Objectivity	How much does the result depend on the person that conducts, evaluates, and interprets the test?
Reliability	How precise is the measurement and how much do results differ during multiple iterations?
Validity	How well does the test measure what it is supposed to measure?

Table 2.2: Objectivity, reliability, and validity are the most important quality criteria for psychological diagnostical procedures (Schmidt-Atzert et al., 2018, p. 131).

Objectivity

A testing procedure is *objective* when its results are independent of the person or entity that conducts, evaluates, or interprets the test and its results. Those three aspects, conduction, evaluation, and interpretation can be viewed as types of errors in accordance with the CTT (see Subsection 2.1.1).

The objectivity of application describes, how a diagnostic test oughts to be conducted in terms of receiving the same results independent from the conducting entity. This can be achieved by documenting the testing procedure as thoroughly as possible. Documentation usually involves testing manuals, questionnaires, software systems, or even specialized pens, if they differ from the usual (Kuhl & Scheffer, 1999). Every other necessary material has to be described in detail. The most important aspect of objectivity of application is the standardization and training of the investigators or experimenter.

The evaluation objectivity is the highest when similar reactions of a participant lead to similar results. This can be achieved when testing material only allows for controllable reactions and provide unique and well-defined evaluations for each possible and expectable answer or reaction to a testing item. Those definitions of how a reaction needs to be evaluated are documented in testing manuals. The evaluation objectivity can be determined by measuring the coefficient of the variance between the testing protocols S_A^2 related to the empirical variance of all test score values S_x^2:

$$r = \frac{S_A^2}{S_x^2}$$

This coefficient r is not to be confused with the Pearson correlation coefficient r defined in Subsection 2.1.4. The overall variance of all test values S_x^2 is a combination of an array of varriances, formally denoted as:

$$S_x^2 = S_A^2 + S_B^2 + S_C^2 + S_X^e$$

with

S_A^2 = the variance of all testing protocolls

S_B^2 = variance between all the experimenters

S_C^2 = variance of the interactions between evaluators and protocolls, and

S_e^2 = variance of the situational testing errors (Schmidt-Atzert et al., 2018, p. 136).

Lastly, the interpretation objectivity can be understood as the standardized transition from raw test result values to the interpretation of those results. The interpretation objectivity is the highest, when all experimenters reach the same conclusion of a participant, regardless of who conducts the experiment. High interpretation objectivity can be achieved similarly to the evaluation objectivity: the test manual should thoroughly describe this transition from raw results to interpretation as unambiguously as possible (Schmidt-Atzert et al., 2018, p. 136).

Reliability

Reliability is the consistency of a testing procedure. A testing procedure is highly reliable when its result is not easily disturbed by unsystematic errors. A reliable test will produce consistent results for multiple executions of the same test for the same tested object (or mostly participant). This reliability can be measured by a reliability coefficient. The closer this reliability coefficient is to one, the higher the reliability of the test. There are different methods for approximating the overall reliability.

The test-retest reliability describes the correlation between two subsequent executions of a test for the same participant. Two tests do not have to be conducted immediately one after another. But the later a retest is conducted, the stronger the influences of stability and trait

alterations of a participant. As Schmidt-Atzert et al. (2018, p. 137) summarize, Charter (2003) surveyed the reliability coefficients of previous tests and found that personality tests scored a standard deviation of $\sigma = .79$ and thus show a very high test-retest reliability.

The split-half reliability splits a testing scale into multiple chunks and measures the correlation between those chunks. If a test is consistent, the correlation should be high. This split-half reliability can be measured by the Kudar-Richardson-Formula, which is a predecessor of the Cronbach's α (Cronbach, 1951). For calculating the Cronbach's α, a test per participant is split in as many chunks as there are items. The sum of the variances of all items s_i^2 is put into relation with the variance of the test results s_t^2:

$$\alpha = \frac{m}{m-1}\left(\frac{\sum_{i=1}^{m} s_i^2}{s_t^2}\right)$$

with

α = the Cronbach's α

m = the number of items

s_i^2 = the variance of the ith item, and

s_t^2 = the variance of all items of the test (Cronbach, 1951).

Cronbach's α describes the inner consistency of a test or homogeneity. Whether or not this type of reliability should always be high, depends on the testing procedure. It is not always desirable to have a high Cronbach's α. A high homogeneity can sometimes indicate a large testing error. Additionally, heterogeneous traits would result in a low Cronbach's α, even though they are desired (heterogeneous tests are those that test multiple dimensions, e.g. during aptitude diagnostics). It is important to note, that the Cronbach's α is not applicable on implicit motives due to the dynamic response theory (Runge et al., 2016).

Validity

The validity determines whether testing results correspond with what a test is supposed to measure. This especially means that thought-to-be-measured traits correspond to observable behavior outside of the testing procedure.

Heterogeneous testing procedures with a low Cronbach's α oftentimes show higher validity scores. Furthermore, some tests could score low in reliability or objectivity but still achieve a high validity. In general, a high validity is the most important quality criterion of a diagnostic procedure and can in and of itself legitimize a test's utilization. One example is the *Big Five* (see Chapter 6), which initially showed low scores in objectivity and reliability, but has proven its validity in countless sophisticated studies.

Since the validity is determined by observations, it can only be measured empirically. There are three different types of validity: i) content validity, ii) criterion validity, and iii) construct validity.

Content validity describes the representability of the test items. It can be assumed that many different items (i.e. questions in a questionnaire or problems to be solved during a math test) could measure a construct such as extraversion. The content validity is high if the set of chosen items can be assumed to represent all possible items. Achieving a high content validity is not trivial and involves identifying every expression, behavior, or symptom for a trait together with a set of items (tests, questions, or tasks). This item collection is reduced to one item per manifestation or expression and should contain the smallest necessary but representative set of items for measuring a trait (Schmidt-Atzert et al., 2018, p. 145).

The criterion validity is simply the idea that a criterium to be measured could be measured outside of this test by a valid additional testing procedure. E.g. if a test is said to measure the risk of alcoholism, the criterion is alcoholism and could furthermore be measured by the amount of alcohol being consumed. Usually, the criterium validity is specified by a correlation coefficient between the test results and additional out-of-test measurements.

The third type of validity, the construct validity, is by far the most important. Psychological constructs can not be directly observed – they are abstract concepts for a domain of behavior. Intelligence is an example of a construct: as described in Subsection 2.1.2 on aptitude, intelligence is a collection of skills, which mostly require reasoning or the acquisition of knowledge. The construct validity would be, whether an intelligence testing procedure measures intelligence or a different construct, mistaken as being part of intelligence (e.g. extraversion). As Schmidt-Atzert et al. (2018, p. 148) point out, Cronbach & Meehl (1955) described constructs as basic principles anchored in nomological networks, which can be observed, theoretically described, and interact differentiable with other constructs.

Measuring construct validity is challenging. In order to ensure that a construct is not mistaken with another, usually, two assumed to be similar expressions or behaviors are both measured and thereafter correlated. This second, not to be measured expression, is called convergent and the resulting validity is the convergent validity. Furthermore, a third expression is taken into account that is assumed to be close to the original construct but emerges from a different construct. This construct is called divergent validity. A construct is constructed valid if the convergent validity is higher than the divergent validity. Those values usually are plotted into a matrix.

Correlation coefficient

Frequently used are so-called correlation coefficients. These measures are the expression of the linear relationship between variables. The decisive factor is the scaling of the available data. Thus, values must be at least ordinally scaled. Furthermore, for ordinally scaled variables the rank correlation coefficient according to Spearman is utilized. The correlation coefficient according to Pearson for samples is denoted as r and for populations the Pearson correlation coefficient is denoted as ρ. With bivariate statistics, studies of differently scaled variables can always be compared with the investigations for the lower-ranking scaling. Therefore, if two

variables are ordinally scaled and proportionally scaled, all studies for ordinal scaled variables can be performed for these two variables (Schira, 2012, p. 92).

The Spearman rank correlation coefficient is being computed as follows:

$$\rho_s = 1 - \frac{6 * \sum_{i=1}^{n}(r_i - s_i)^2}{n^3 - n}$$

with

r_i = rank position within the variable X of the i-th [entity]

s_i = rank position within the variable Y of the i-th [entity], and

n = number of [entities] (Schira, 2012, p. 527).

The correlation coefficient according to Pearson is used far more frequently than the rank correlation coefficient according to Spearman. It can be used for all at least cardinally scaled variables. For populations X and Y it is calculated as follows:

$$\rho_{X,Y} := \frac{cov_{XY}}{\sigma_X * \sigma_Y}$$

with

cov_{XY} = the covariance of both populations

σ_X = the standard deviation of population X, and

σ_Y = the standard deviation of population Y (Schira, 2012, p. 95).

Pearson point biseral correlation coefficient

In case of two variables being dichtomous, meaning either belonging to one class or not belonging to this class similar to a coin flip, the point biseral correlation coefficient r_{pb} can be calculated, resulting in more precise correlation estimates than e.g. the well-known Pearson product-moment correlation coefficient (known as r).

In case of a dichotomous variable being divided into two groups X and Y only containing one of the manifestations of the variable, r_{pb} can be calculated as follows:

$$r_{pb} = \frac{M_1 - M_0}{s_n} \sqrt{\frac{n_1 n_0}{n^2}}$$

with

s_n = the standard deviation for every member of the population,

M_1 = the mean value for all data points in group 1,

M_0 = the mean value for all data points in group 2,

n_1 = the number of data points of group 1,

n_0 = the number of data points of group 2, and

n = the number of data points in the population (Schmidt-Atzert et al., 2018, p. 120).

After this background section on psychological personality diagnostics, the next section covers machine learning (ML), which provides information on the fundamental workings of how machines learn. First, the position of ML in the field of AI is described, as well as the basic functionality of learning algorithms. A few algorithms, information representations, and neural architectures are presented, including decision trees, feature engineering, neural networks, and attention. Lastly, this section ends with a brief description of technical biases and evaluation measures for ML approaches.

2.2 Machine Learning

The fields of natural language processing, text mining, and computational linguistics have experienced a rapid shift in paradigms over the past decades and even years. Whilst originally, natural language processing was mainly concerned with statistical methods, with a growth in available data, calculation speed, and methodological advancement, they more recently have shifted towards utilizing machine learning (ML). The shift towards neural networks fostered the development of novel NLP methods (e.g. different types of word embeddings), architectures (e.g. BERT, (Devlin et al., 2019)), and research objectives (e.g. explainable language models).

2.2.1 Positioning in the Field of AI

ML is part of a broader scientific field of artificial intelligence (AI). AI subsumes a broad variety of fields, methods, and applications. Figure 2.1 provides an overview of the AI map. Fields include robotics, artificial life (e.g. evolutionary algorithms), knowledge bases (e.g. expert systems, inferences), and pattern recognition. Applications can not be specified any further. Methods include logic, approximation (e.g. Taylor polynomials), planning systems (e.g. PDDL), and – most importantly – machine learning, which in turn can be divided into *classical* or non-neural machine learning and deep learning. In popular parlance, AI and DL are used interchangeably (Russell & Norvig, 2016, 728).

2.2.2 Functioning of Learning Algorithms

Machine learning is the discipline of giving a machine the ability to learn a target function (mapping its given inputs to desired outputs) without being specifically programmed in terms
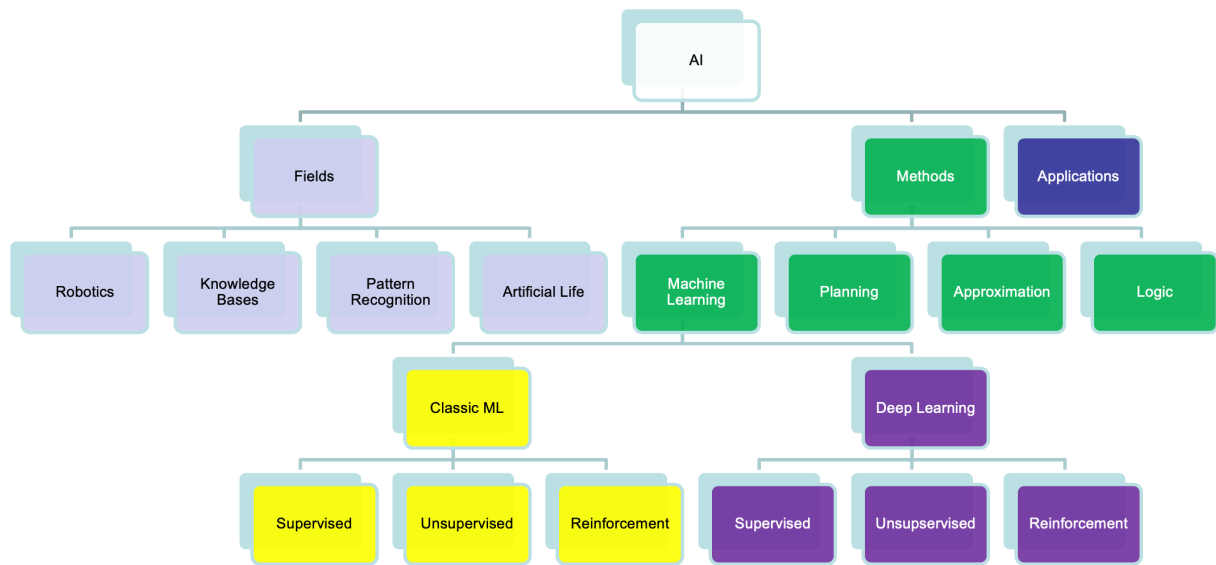
Figure 2.1: The field of artificial intelligence (AI) is broad. Even though machine learning is nowadays used interchangeably with the term AI, it also consists of fields such as robotics, planning approaches, or approximation (adapted from Russell & Norvig (2016)).

of how to solve a problem. A machine is learning, if it improves its performance on future tasks based on provided observations about an observable world (Russell & Norvig, 2016, p. 693), which is also called *training*. Even though different forms of learning exists, so-called supervised learning comes with the advantage of easier trainable models with fewer data instances. A machine learning model learns supervisedly if it is provided with information of discrete target classes per data instance. That information on the target classes per instance is called *label*. Supervised learning can be divided into regressions, where one variable is explained by one or multiple explaining variables, and classification, where the goal is to assign one of n many discrete target classes. Thus, during supervised learning, a model gets to *know* what an instance is in contrast to so-called unsupervised learning or clustering, where such information is not available during training. Supervised learning can be formalized as:

"Given a training set of N example input-output pairs

$$(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$$

where each y_j was generated by the unknown function $y = f(x)$, discover a function h that approximates the true function f" (Russell & Norvig, 2016, p. 695).

In order for a machine to learn, an algorithm producing a so-called model (comparable to mathematical functions) needs to alter the model's parameters describing a possible solution towards minimizing a loss, which is calculated by a loss function.

Assume we have two target classes of pairs (x, y), plotted on a two-dimensional canvas, and assume we can linearly separate those two classes (e.g. stars and circles as shown in Figure 2.2). The regression line separating those two classes is the target function, which the machine learning model is supposed to learn by itself (even though this is a very simple math problem, easily solvable via calculus).
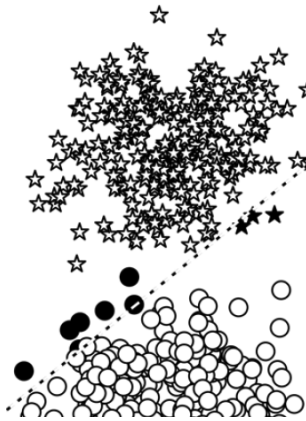
Figure 2.2: A simple linear regression, separating two target classes *star* and *circle*. Linear regression can be learned by machines via alterations of b (the y-interception) and m (the slope) and the loss function *mean squared error*. However, the regression line can also be found via calculus (Rao et al., 2019, p. 84).

The regression line, like every linear function, can be expressed as $y = mx + b$, whereas y is the interception with the ordinate (or y-axis) and m is the slope of the line. Those two parameters, m and b, ought to be altered to minimize the so-called loss, calculated by a loss function. A loss function calculates an error or punishment for the model. When the loss with respect to parameter manifestations is interpreted as a curve, the model tries to traverse this curve to a local (or better global) minimum of this function. The direction that the model needs to move (uphill or downhill) is determined by the steepness of the tangent of point x on the loss-function curve f. This is determined by calculating derivatives of the influential parameters (in this example, m and b). The loss function of linear regression is the root-mean-square deviation or mean square error (MSE), displayed in Figure 2.3.

From the current linear function to all the instances (data points with (x, y)) squares can be drawn. The line approaches the target function best, when the sum of the areas of the squares is minimal:

$$MSE = \sum_{i=1}^{n}(y_i - \bar{y})^2$$
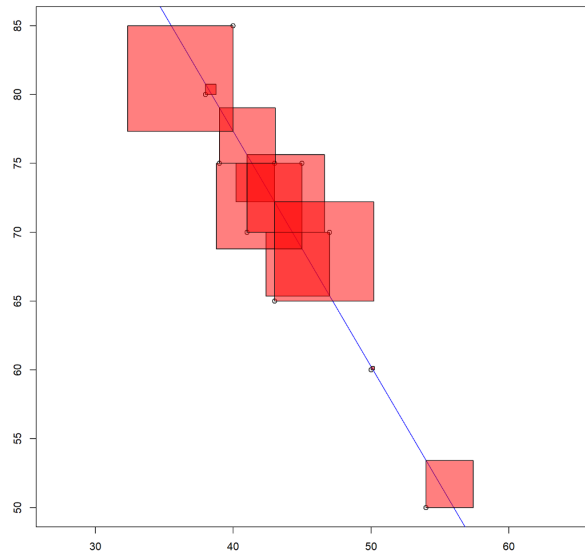
Figure 2.3: Illustration of the total sum of squares, utilized as the medium square error (MSE) loss function (Hildebrandt & Köhler, 2022)

Machine learning tweaks the available parameters (e.g. *m* and *b*) in one direction as long as the loss is minimized. As soon as the loss increases, and thus the model performs worse than before, the tweaking is reduced, until the resulting function oscillates. This tweaking is being done in small steps, determined by a so-called hyperparameter, the step size. Too large step sizes can lead to a model not finding an ideal solution, whilst a too small step size leads to very slow progress up to unfeasible calculation times. The step size is just one example of a hyperparameter, which can come in large numbers, depending on the chosen learning algorithm. Besides linear or other regressions (e.g. quadratic, logistic, geometric, etc.) there are e.g. decision trees, Bayesian networks, or neural networks. One main task when utilizing machine learning besides data pre-processing or feature selection is the empirical exploration of ideal hyperparameters, called parameter tuning.

2.2.3 Decision Trees

A decision tree is a directed acyclic graph (DAG). A graph is a data structure, consisting of *edges* and *nodes*. It is defined as:

"'A *graph* $G = (V, E)$ consists of an [...] amount $V = \{v_1, v_2, ..., v_m\}$ of nodes [...] and an [...] amount $E = \{e_1, e_2, ..., e_n\}$ of edges [...]"' (Owsnicki-Klewe, 2002, S. 38).

A DAG is shown in Figure 2.4. To meet the requirement of an undirected graph, the following property must be fulfilled:

$$\{\exists \quad a \quad path \quad from \quad v_i \quad to \quad v_j\} => \{v_i \prec v_j\}$$

This property states that there may not be a directed edge from v_i to v_j, if the node v_i is the predecessor or parent of v_j. A node without a predecessor in a DAG is called a root. Nodes without successors in a DAG are called leaves. All nodes that lie in the sequence from v_i to v_j are called *predecessor* of v_j are called. If there is exactly one directed edge from v_i to v_j, v_i is called *parent* of v_j, whilst v_j is called *child* of v_i (Scutari & Denis, 2014, vgl. S. 175).
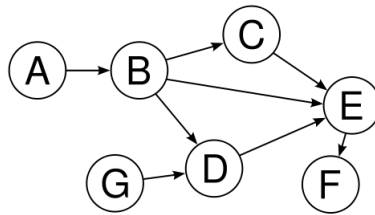
Figure 2.4: Directed acyclic graph (DAG) (Russell & Norvig, 2012, p. 511)

Decision trees are predictive models, which utilize divide and conquer techniques for formulating graph dependend decision strcutre. Each node represents an attribute and each edge represents a decision, directing the flow towards a leave, which determines the decision. A logistic model tree (LMT, Landwehr et al. (2005)) is displayed in Figure 2.5. An LMT is a decision tree, which performs logistic regressions at its leaves (for regressions, see Subsection 2.2.2).
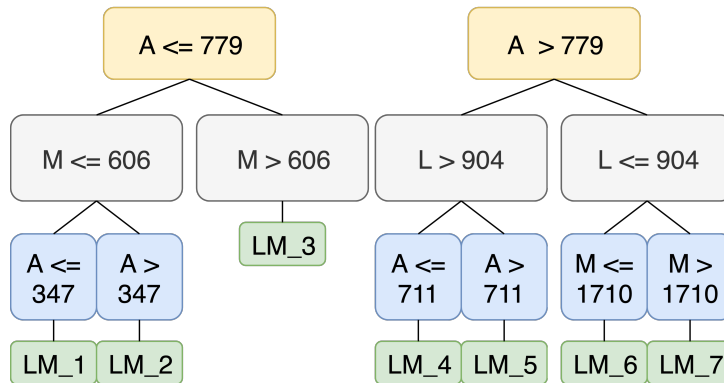
Figure 2.5: Illustration of a Logistic Model Tree (LMT), which performs a information gain split at its root and further decision splist, but finalizes the classification decision via logistic regressions at its leaves (Landwehr et al., 2005).

For the training set $\mathbb{S}_{train} = \{(x^{(1)}, y^{(1)}), \dots, (x^{(n)}, y^{(n)})\}$ with statistically independend input and output tuples, the input and output subsets are denoted as $X = \{x^{(1)}, \dots, x^{(n)}\}$ and $Y =$

$\{Y^{(1)}, \dots, Y^{(n)}\}$. The goal of a decision tree is to find a parameter with the maximum likelihood approach that maximizes

$$p(Y|X, \theta) = \prod p(y|x; \theta)$$

with

θ = denoting the decided attribute splits of the decision tree (Görz et al., 2020, p. 463).

Many decision trees apply a *greedy* approach for determining the best attribute splits. The set $\mathbb{X} = \{R_1, \dots, R_m\}$ denotes the splits. For the first execution of the greedy approach, this set only contains one element or split, thus $\mathbb{X} = \{R_1\}$. In order to determine all the other splits, the greedy approach i) disassembles R_j into subsets in accordance with

$$R_{j,0} = \{x \in R_j | x_i \leq x_{i,s}\}$$

and

$$R_{j,1} = \frac{R_j}{R_{j,0}}$$

ii) estimates the values for the class distribution $\hat{p_j}$ for both subsets, and iii) calculates $-log(p(Y|C, \theta))$ for both subsets in which R_j is replaced by $R_{j,0}$ and $R_{j,1}$ respectively. These three steps are repeated until all target variables in R_j are identical or a stop criterion is reaeched (e.g. for limited-depth approaches). Since the resulting *greedy* decision trees can become very large, pruning is applied for reducing its size and enhancing the prediction power (Görz et al., 2020, p. 464).

2.2.4   Feature Engineering

As stated in Section 2.2, a machine is learning, if it improves its performance on future tasks based on provided observations about an observable world (Russell & Norvig, 2016, p. 693). One key aspect concerns *observations*. A machine is able to process numbers, more precisely values representable in bits, naturally by design. Thus, a machine can process e.g. sensor data directly drawn from a mechanical machine without much hindrance (i.e. the necessity for engineers to pre-process or digitalize the data). A data point can be thought of as e.g. tuples in a coordinate system for regression tasks, as displayed in Figure 2.2.

As for NLP when viewed as an ML application domain, major differences to other application domains are the ambiguity of language and the necessity to transform linguistic signals into processable information. There are different approaches to achieving this, one being *word embeddings*, which are described in Subsection 2.3.3. Another approach is *feature engineering*. Features can be thought of as rules. This rule-based approach is one of three paradigms for
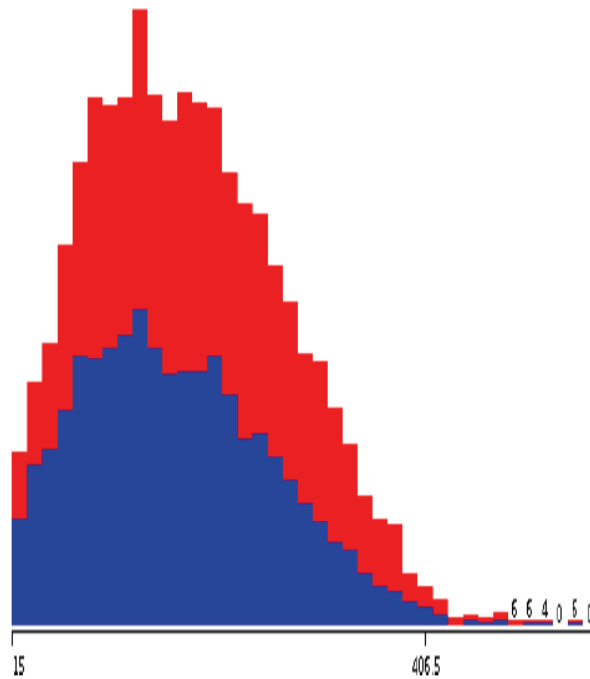
Figure 2.6: Assuming that two features correspond with the target classes (*red histogram* and *blue histogram*), then those histogram areas that do not intersect (i.e. just blue or just red) offer a pattern for a machine learning algorithm to differentiate between those two target classes based on this very feature (Witten et al., 2011, p. 408)

representing textual resources, with the others being statistical approaches (see Subsection 2.3.2) and neural approaches (see Subsection 2.3.3), whereas all three approaches are still relevant and have been utilized for empirical studies, as described in Part III (Biemann et al., 2022, p. 74).

The goal of machine learning is to approach a function, which determines for a set of for the problem relevant information the most probable solution for a given problem. This set of relevant information can be understood as characteristics of instances to be learned from. An instance can be on the document level (e.g. emails, articles, websites) or on the word level (e.g. paragraphs, sentences, words). The goal of feature engineering is to provide machine learning algorithms with characteristics of a given textual input. As for the rule-based approach, the feature engineer needs to define a rule, which translates a text into those characteristics – mostly, by writing a program (Biemann et al., 2022, p. 74).

The output of a rule-based feature function has to be numeric, has to provide information on the input text, should be applicable for most of the expected input texts (i.e. should not result to 0 for most inputs), and should be normalized to account for different lengths of input

texts. For a document, a feature could be the number of words it contains. For words, a feature could be whether a word is capitalized (Biemann et al., 2022, p. 74).

Features usually do not differentiate target classes by themselves but are rather combined with other features to form a feature vector (Russell & Norvig, 2016, p. 866). This vector should only contain features, which are highly relevant to the task. The choice of how this relevance is determined depends on the task at hand and which value types the inputs are (i.e. numerical or categorical). For supervised ML, the importance is measured mostly as the correlation between features in terms of importance towards the target label (Kuhn & Johnson, 2013, p. 488).

During supervised learning, a feature offers relevant information, if it differentiates target classes the most. Figure 2.6 illustrates, what is considered good differentiation: assuming that two features correspond with the target classes *red* and *blue*, then those histogram areas that do not intersect (i.e. just blue or just red) offer a pattern for a machine learning algorithm to differentiate between those two target classes based on this very feature. The displayed histogram feature distributions in Figure 2.6 also intersect. If this feature calculates an instance to be in the area of the interception, then this very feature can not inform the ML model, which target class this instance would belong to (but possibly, to which extent it belongs to either red or blue). However, since for feature engineering a variety of features is calculated and provided to the ML model or algorithm in a feature vector, there might be other features that still differentiate this instance to be either blue or red. This feature vector can be formalized as

$$F(x) = (f_1(X), f_2(X), f_3(X), \dots, f_n(X))$$

with $F(X)$ being the feature vector and $f_n(X)$ being the nth feature for the given set of inputs X (Biemann et al., 2022, p. 80).

Rule-based features mostly fail to capture the complexity and ambiguity of natural languages. They were popular in the 1980s, but are on the decline nowadays. There are some situations, where rule-based features are applicable to the task at hand: i) in those cases, where a differentiable characteristic of a text can be expressed by a simple rule, ii) in those cases, where unsupervised textual data needs to be processed quickly and as a rapid prototype or iii) when easily explainable and transparent rules are more important than linguistic precision (Biemann et al., 2022, p. 79). More sophisticated and precise features utilize supervised linguistic statistics and are described in Subsection 2.3.2, including language modeling and perplexity.

2.2.5 Neural Networks

Neural networks perform machine learning and thus inherit all the described characteristics of machine learning: Those data structures change their inner structure with respect to the

loss calculated by a loss function by observing provided (in this case labeled, and thus supervised) data instances, and reduce the calculated loss step-wise by updating parameters, until a satisfying minimum of the loss-function is reached and the training results oscillate or the training gets stopped early (Russell & Norvig, 2016, p. 727 ff.).

Artificial neural networks consist of a number of units or cells, aligned in a layer, which in turn are arranged. Even though a network with one input layer, one so-called *hidden* (hence not directly observable) middle layer, and one output layer technically already forms an artificial neural network, those networks are only called *deep*, if there are at least two hidden layers. Each unit of one layer (in the case of a so-called fully connected network) is connected with each unit of the subsequent layer. Those connections are called weights (w) and essentially are the main parameters to be tweaked during training processes. Units have previously been called cells since early neural networks shared some characteristics with neural connections of a brain. The decision made by a neural network in essence is determined by the way information gets passed through the network structure. The weights connecting cells and determining the information flow are updated and adjusted during the training phase via an algorithm called *backpropagation*, which honors the overall loss or error and determines, how much each involved weight has to change in order to improve the network's performance. A unit is displayed in Figure 2.7.
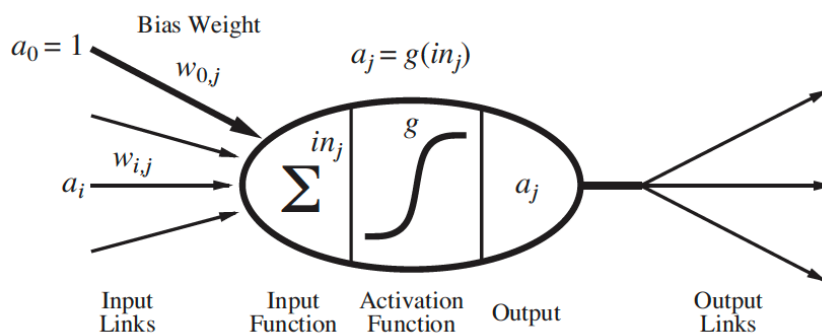
Figure 2.7: Illustration of a neural unit. 'The output function is $a_i = g(\sum_{i=0}^{n} w_{i,j} a_i)$, where a_i is the output activation of unit i and $w_{i,j}$ is the weight on the link from unit i to this unit.' (Russell & Norvig, 2016, p. 728)

Each unit itself can be seen as a combination of matrices of the sum of all weight-vectors multiplied with their input-vectors and added by a bias value. The activation of the $i - th + 1$ layer can formally be described as $a^{(i+1)} = \sigma(Wa^{(i)} + b)$, with σ being the activation function, W the weight vector, $a^{(i)}$ the activation of the $i-th$ layer, and b the vector of biases (all represented as matrices (Görz et al., 2020, p. 509)). If that sum surpasses a threshold defined by an activation function, which is either elevated or lowered by the bias value, the information is passed on to the subsequently connected cells. The structure is displayed in Figure 2.8.
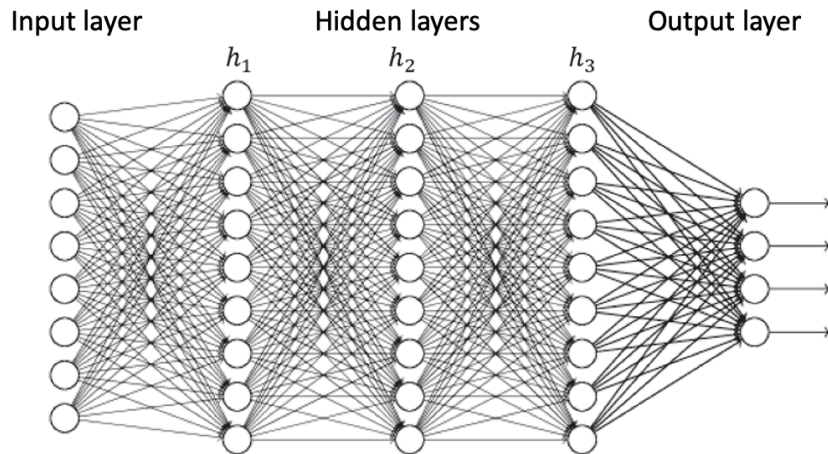
Figure 2.8: Illustration of a neural network. The circles symbolize units or cells, the lines symbolize their connecting weights. This example consists of one input layer, one output layer, and three hidden layers (Görz et al., 2020, p. 509).

2.2.6 Recurrent Neural Network Architectures

A long short-term memory neural network (LSTM, (Hochreiter & Schmidhuber, 1997)) is a type of Recurrent Neural Network (RNN) which, in turn, is a deep neural network architecture, that allows for the neural cells to access other cells of the same recurrent layer with a time delay and thus develop a so-called memory. An LSTM furthermore employs memory cells that allow storing information of an arbitrary time horizon. Forget and update gates allow for those cells to purposely omit information and control which information gets altered. LSTMs have successfully solved the issues of vanishing or exploding gradients present in general RNNs (Hochreiter, 1998). LSTMs have been successfully utilized for classifying short texts.

Lai et al. (2015) designed a recurrent convolutional neural network (RCNN) for text classification with promising results. A RCNN is a RNN with a max-pooling layer as its output. The main advantage of a RCNN in comparison with RNNs is the enhanced selection of targets or regions to have an impact on algorithmic decision-making.

The model displayed in Figure 2.9 consists of a bi-directional LSTM combined with an attention mechanism (see Subsection 2.2.7).

Bi-directional refers to the direction in which an input is being processed. Usually, textual input is processed token by token (i.e. words) from left to right. Thus, a one-directional network can only take previous tokens into account, when deciding upon the impact or meaning of a token. A bi-directional network combines both directions of input and concatenates the impacts of a token in dependence on the previous and following context of this token. Lastly, the attention mechanism (Bahdanau et al., 2014) models the algorithmic importance of a net-
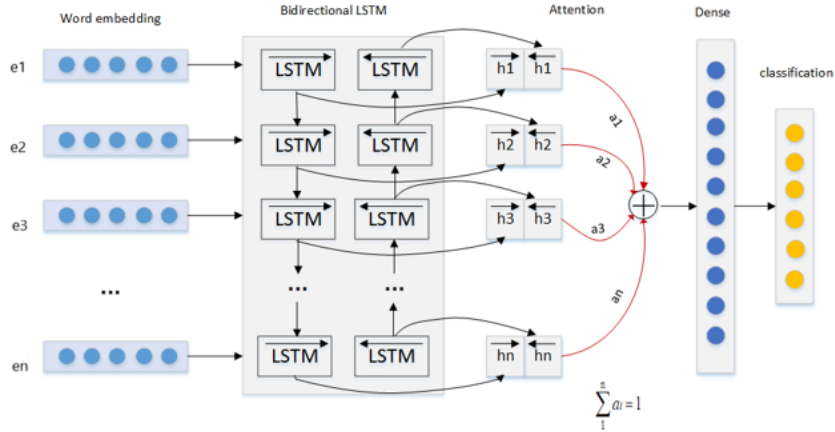
Figure 2.9: The employed model is a bi-LSTM with attention mechanism (image by Zhou & Wu (2018)). This type of architecture allows for the model to observe the input from both sides, left and right. The attention supports algorithmic decisions made and at times allows for an analysis of more algorithmic important parts of input or instance.

work by multiplying hidden states with an alignment score to create a context vector, which then gets concatenated with a previous output.

2.2.7 Attention Mechanism

Young et al. (2018) found attention mechanisms as part of decoder-encoder-architectures to be among those recent advancements in their survey on recent trends in DL based NLP. Accordingly, attention mechanisms allow for decoders to assess their memory by referring back to their input sequence, which can enhance the network's performance. The idea of employing attention to a seq2seq encoder-decoder system originated from Bahdanau et al. (2014).

With a sequence of annotations h_i being $(h_1, \ldots, h_{(T_x)})$, a context vector c_i represents the weighted sum of the annotations via:

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j \tag{2.1}$$

The weight α_{ij} is computed as:

$$\alpha_{ij} = \frac{exp(e_{ij})}{\sum_{k=1}^{T_x} exp(e_{ik})} \tag{2.2}$$

whilst $e_{ij} = a(s_{i-1}, h_j)$, with $a(\ldots)$ being a score function describing how well two words are aligned.

In other words, the system translates an input sequence (this could be e.g. a certain language or a whole text to be summarized) into a context vector. This context vector together with hidden states functions as input for the attention mechanism, which computes attention weights and passes this context vector together with the attention weights onto the output layer. This process is illustrated in Figure 2.10.
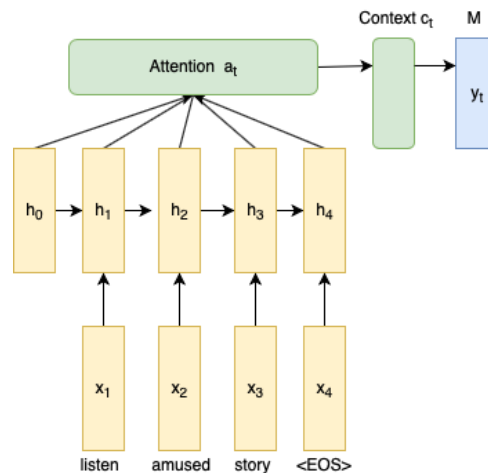
Figure 2.10: Illustration of the LSTM with a self-attention mechanism. The LSTM receives hidden states and attention weights as inputs in order to output a corresponding context vector, which thereafter gets fed to a softmax output layer. Figure based on Bahdanau et al. (2014) and `https://bzdww.com/article/250330/`

Attention mechanisms were successfully employed for various tasks, demonstrating the universal benefit. Gupta et al. (2018) utilized a CNN on group images for learning the global representation of the image and employed an attention mechanism for merging faces in order to learn local representations of only the faces, thus leading to a network capable of detecting emotions from entire groups of people. For this, the authors employed a so-called sequence-to-sequence system (Seq2Seq) with an attention mechanism (originally proposed by Vaswani et al. (2017)). Images received automated descriptions by using a CNN encoder, an attention layer, and an LSTM decoder by Xu et al. (2015). Furthermore, the authors were able to project the attention weights onto the images, visualizing the gaze of the network. Speech has been analyzed for detecting emotions utilizing an attention mechanism by Ramet et al. (2018).

In terms of textual data, attention mechanisms have enhanced both, classification performance and comprehension tasks. Hermann et al. (2015) advanced automated reading comprehension and comprehension question answering for texts with minimal prior knowledge. So-called self-attention was the enabler of semantic role labeling (SRL) for Tan et al. (2018). Self-attention is a special case of an attention mechanism that only requires a single sequence to compute its representation. Vinyals et al. (2015) showed that a seq2seq model with an at-

tention mechanism could enhance syntactic constituency parsing to SOTA performance even on unoptimized CPUs, implying the strong optimization of an attention mechanism can contribute to a task.

2.2.8 Transformers

Vaswani et al. (2017) introduced the transformer architecture, which extended the idea of encoder-decoder networks. These transformers are of importance for the creation and utilization of contextualized embeddings (see Subsection 2.3.4). The authors saw the shortcommings of RNNs, LSTMs or other gated recurrent neural networks in their sequentiality. Recurrent models calculate hidden states h_t by processing the previous hidden state h_{t-1} and the input for position t. Especially longer sequences and context require more memory and larger batching than mostly feasible. As a result, the authors proposed tarnsformer models, which rely entirely on an attention mechanism (see Subsection 2.2.7).

The schematic model architecture proposed by Vaswani et al. (2017) is displayed in Figure 2.11. This architecture consists of six stacked layers each for the encoder and decoder parts. Each layers consists of two sub-layers, the first is a multi-head self-attention mechanism, and the second sub-layer is a position-wise fully connected feed-forward network. Instead of sequentially processing information, this architecture can encode window-sized input sequences to be calculated in parallel.

The authors named their approach *scaled dot-product attention*. For the self-attention, the authors map a query and key-value pairs to an output, where the query, keys, and values are vectors arranged in matrices Q, K, and V. The matrix of outputs is computed as

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

with

Q = the queries arranged in a matrix

K = the keys arranged in a matrix

V = the values arranged in a matrix

d_k = queries and keys dimension (Vaswani et al., 2017).

Furthermore, the authors performed a linear projection of query, keys, and values h times with different learned linear projections to d_k and d_v. The attention function is performed in parallel on these projected versions, resulting in d_v-dimensional output values displayed in Figure 2.12.

Lastly, the proposed transformer architecture consists of a feed-forward network stack as second sublayer per layer stack. This fully connected feed-forward network is applied to each position separately, making it a position-wise network. The output is passed through a softmax
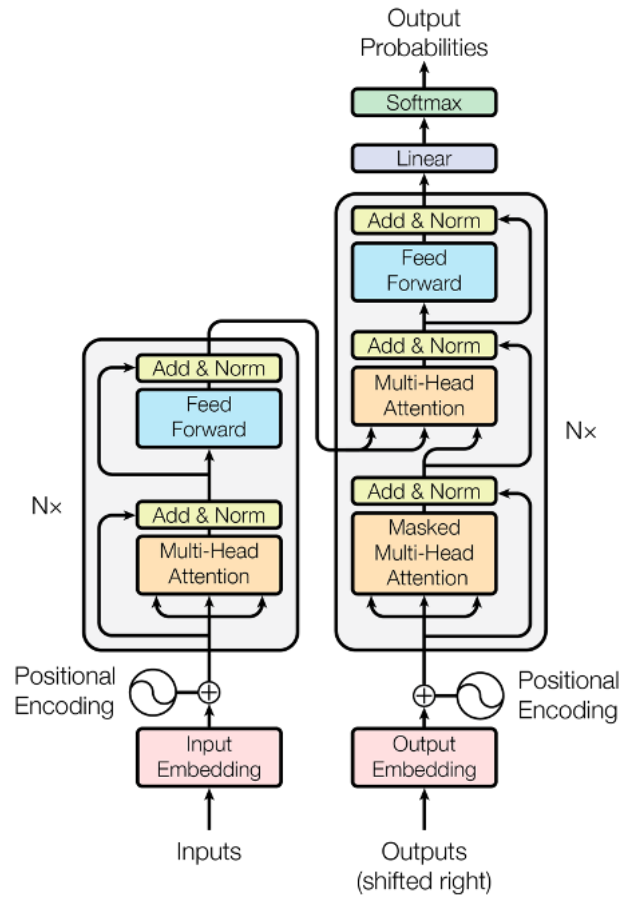
Figure 2.11: The transformer model architecture proposed by Vaswani et al. (2017) utilizes self-attention instead of recurrency and is composed of six encoder and decoder layers each.

function, resulting in semi-interpretable probability-like output information (Vaswani et al., 2017).

2.2.9 Evalation Measures

Same as for the formal definition of a decision tree in Subsection 2.2.3, the training set $\mathbb{S}_{train} = \{(x^{(1)}, y^{(1)}), \dots, (x^{(n)}, y^{(n)})\}$ with statistically independent input and output tuples are disassembled into the input and output subsets $X = \{x^{(1)}, \dots, x^{(n)}\}$ and $Y = \{Y^{(1)}, \dots, Y^{(n)}\}$.

In addition to $\mathbb{S}_{train}$, machine learning approaches require the creation of a testing set of instances $\mathbb{S}_{test}$. This $\mathbb{S}_{test}$ also contains statistically independent input and output tuples. Usually, $\mathbb{S}_{train}, \mathbb{S}_{test} \in \mathbb{X}$, meaning that both the training and test sets are subsets of all available instances. Thus both sets emerged from the same data source. Furthermore, both sets should be as statistically identical as possible. Lastly, these two sets $\mathbb{S}_{train}$ and $\mathbb{S}_{test}$ should be statistically independent, as well. The test set $\mathbb{S}_{test}$ is not to be utilized during the machine learning
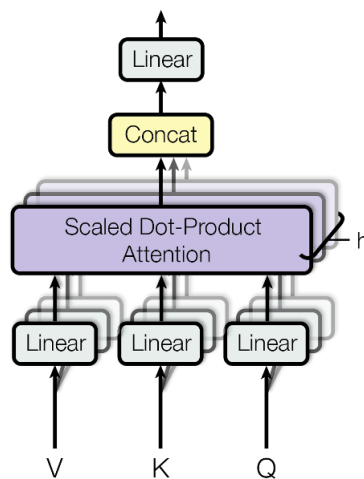
Figure 2.12: The values (V), keys (K), and queries (Q) are aligned in matrices and are passed first through a linear transformation and then through an attention mechanism multiple times, where h denotes one attention head. The independent attention outputs are then concatenated and once again linearly transformed, giving the multi-head attention module the capability of attending to different representations (Vaswani et al., 2017)

model's training phase as to not having the models parameters on any of the instances contained in $\mathbb{S}_{test}$ (Görz et al., 2020, p. 436).

Such an approach is referred to as *held-out test set*. Hiding the test set from the experimental training phase as well as even from the researchers conducting the experiments are meant to approximate the subsequent productive utilization of the resulting model.

The test set $\mathbb{S}_{test}$ can be utilized for testing a resulting model in terms of assumable performance on new, yet unseen data from a comparable data source as $\mathbb{S}_{test}$. Even though a loss function (see Subsection 2.2.2) does hint towards the performance of a model during the training phase, it does not allow for the interpretation of assumable productive performance of the resulting model.

With the test set $\mathbb{S}_{test}$ multiple performance measures can be calculated:

$$Accuracy = \frac{\text{Number of correctly classified } x \in \mathbb{S}_{test}}{\text{Number of all } x \in \mathbb{S}_{test}} \qquad (2.3)$$

$$Precision_i = \frac{\text{Number instances } x \in \mathbb{S}_{test} \text{ correctly classified as class i}}{\text{Number of all } x \in \mathbb{S}_{test} \text{ predicted as class i}} \qquad (2.4)$$

$$Recall_i = \frac{\text{Number instances } x \in \mathbb{S}_{test} \text{ correctly classified as class i}}{\text{Number of all } x \in \mathbb{S}_{test} \text{ belonging to class i}} \qquad (2.5)$$

$$F_i = 2 * \frac{Precision_i * Recall_i}{Precision_i + Recall_i} \qquad (2.6)$$

The accuracy is a global measure of how well all classes were predicted on average. However, the accuracy is prone to misleading performance numbers on highly imbalanced data sets. As an example, an almost perfect accuracy can easily be achieved if the major class accounts for 99% of all instances.

Since the precision and recall metrics account for the prediction of one specific target class, they do account well for imbalanced data sets. Precision measures which proportion of predicted class i instances do in fact belong to class i. Recall measures which proportion of instances belonging to class i were predicted correctly. The F_1 score (also F-measure or f) is the harmonic mean of precision and recall.

Usually, high F_1 scores are preferable for most classification tasks. However, depending on the task, either precision or recall is more desirable to be scored higher than the F_1 metric. If e.g. a classifier aims to identify rare diseases, it is rather preferable to identify some instances as positive, which are in fact negative, than to miss any cases. In this example, a high recall score is preferable to a high precision metric (Görz et al., 2020, p. 443).

Table 2.3 displays a further evaluation approach, called confusion matrices. These matrices contain a more detailed depiction of the type of errors a model made on a held-out test set $\mathbb{S}_{test}$. The described evaluation metrics of accuracy, precision, recall, and F_1 can be calculated in terms of proportions of either true or false positives or negatives (abbreviated as e.g. TP for true positive). Confusion matrices can provide valuable information on a model's misconceptions (e.g. if certain classes are often confused with specific other classes, hinting towards shared characteristics). Furthermore, such a table provides easily identifiable visual cues toward the prediction performances of a model. Usually, the corresponding table cells are colored in deeper shades for higher proportions of instances per cell. If e.g. the diagonally arranged cells from *true positives* to *true negatives* appear darker, this visualization hints towards better overall results (Witten et al., 2011, p. 164).

		Predicted	
		yes	**no**
Actual	**yes**	true positive	false negative
	no	false positive	true negative

Table 2.3: A confusion matrix contains more details on the types of misclassifications and can provide valuable information on a model's misconceptions (e.g. if certain classes are often confused with specific other classes, hinting towards shared characteristics).

2.2.10 Technical Biases

A bias is a phenomenon where a system introduces systematic prejudices due to false assumptions (Görz et al., 2020, p. 918). However, these prejudices can differ in their descriptive or normative assessment (see Subsection 3.1.3). It is a misconception that algorithmic decisions would be fair (Görz et al., 2020, p. 918) in the sense that they reach balanced decisions on the basis of even unbalanced data sources (Blodgett et al., 2020).

The occurrence of biases does not first and foremost cause harm and are technical phenomena rather than inherently *bad*. However, since the occurrence of biases and the assessment of a bias being undesirable are intertwined in most of the literature on this topic (Blodgett et al., 2020; Görz et al., 2020; Fine et al., 2014; Hovy & Prabhumoye, 2021; Savoldi et al., 2021), a more in-depth description of the occurrence of biases, and their possible normative hurdles and countermeasures is presented in Chapter 3, more precisely in Section 3.2.

After this background section on machine learning, the next section covers natural language processing (NLP), which is one possible application for machine learning. First, a psychological dictionary text analysis is presented. Thereafter, language models and word embeddings are described. Lastly, the following section briefly presents contextualized embeddings. All of these aspects of NLP are utilized for the empirical research presented in Part III.

2.3 Natural Language Processing

The vast majority of human knowledge is documented in written language (Jurafsky & Martin, 2008, p. 42). In 2009, more than a trillion pages of information were available on the internet (Ehrlich, 2009). In order to acquire knowledge, machines need to process natural languages. Similar to formal languages such as programming languages, natural languages consist of words, structures, and grammar. One difference from formal languages (e.g. programming language) to natural (human) languages is the ambiguity of natural languages. The goal of natural language processing (NLP) is to process natural languages to communicate with humans or to extract knowledge from textual or acoustic natural language resources. As

a subfield of AI, NLP utilizes large bodies of text, called corpora, to analyze, understand, and produce natural language. A text corpus is a – usually large – body or collection of texts of a certain domain or with certain similarities. NLP is interdisciplinary between linguistics and computer science (and – at times – humanities) (Russell & Norvig, 2016, p. 861).

2.3.1   Linguistic Inquiry and Word Count (LIWC)

The tool *Linguistic Inquiry and Word-Count* (LIWC) was developed by Pennebaker et al. (2007) for the English language and has been transferred to other languages such as e.g. German by Wolf et al. (2008). The tool was psychometrically validated and can be considered a standard in the field. LIWC stands for a tool that operates with recorded dictionaries of word lists and a vector of a number of categories metrics (depending on the version and language).

When analyzing a text, LIWC increments category counts (i.e. positive emotions, cognitive processes, or anxiety) based on matching validated dictionary terms per category. E.g. the category *family* contains words such as sister, father, mother, mom, etc. The counts per category then get normalized over the length of the input. The results are percentages of words belonging to each category. The German LIWC allows for 96 categories to be assigned to each token, ranging from rather syntactic features such as personal pronouns to rather psychometric values such as familiarity, negativity, or fear. Listing 2.1 displays some of the 96 categories with corresponding word prefixes. If a word beginns with one of these prefixes, LIWC increments the corresponding category value to be displayed at the end of the analysis.

The core dictionary and tool with its capability of calculating a feature vector for capturing properties of language samples is well established and can be categorized as a method of choice in psychological language inquiry. Even though the tool appears rather simple from an NLP point of view, it has a long tradition to be utilized for content research in the field of behavioral psychology. The importance of LIWC stems from its validation rather than its linguistic methodology. Studies utilizing LIWC have shown that function words are valid predictors for long-term developments such as academic success (Pennebaker et al., 2014). Furthermore, it has been shown that LIWC correlates with the Big Five inventory (McCrae & Costa Jr., 1999). Importantly, the writing style of people can be considered a trait, as it has shown high stability over time, which means that it is not dependent on one's current mood, the time of day, or other external conditions (Pennebaker & King, 2000).

```
17        Anxiety
18        Anger
19        Sad
20        Cognitivemechanism
21        Cause
22        Insight
...
ab              10        37        41
abbrach*        38
abbreche        39
abbrich*        39
abend*          37
abendessen*     60        63
```

Listing 2.1: Examples of the German LIWC dictonary depicting some categories with corresponding word prefixes. The numbers following a word prefix determine the corresponding LIWC category.

The way LIWC is used is very common. However, researchers usually focus on some selected aspects of the feature vector in order to grasp psychological effects. Coppersmith et al. (2015) used LIWC for differentiating the use of language of healthy people versus people with mental conditions and diseases. Hawkins & Boyd (2017) and Niederhoffer et al. (2017) researched the language landscape of dream narratives. Scores, such as the LIWC sadness score were the basis of the work of Homan et al. (2014) on depression symptoms. Morales et al. (2017) also surveyed the broad use of LIWC in depression detection systems. Pennebaker et al. (2014), who partly developed LIWC, used the tool to research word usage in connection with college admission essays. Reece et al. (2017) captured the general mood of participants by using LIWC and Shen & Rudzicz (2017) surveyed the language of a personal crisis with LIWC.

2.3.2 Language Models

A language model (LM) is a function, which calculates a probability for a given sequence of input tokens. Thus, a language model is a probability distribution over sentences. The function quantifies to which extent an input sentence (or sequence of tokens) is represented by this distribution.

Those observations are called *events*. One key question in terms of probabilities is, how often an event has been observed and how probable a certain (seen or unseen) event is. As for natural languages, a sequence of tokens needs to be translated into an observable sequence of events. As with many natural language statistics, the context of a given document (be it words, tokens, or sentences) is of importance. Corpora statistics, which utilize large bodies of texts, utilize this context principle. Translating a sequence of tokens into a sequence of observable events can be achieved by forming so-called n-grams. An n-gram is a model of the probability

distributions of n-letter (or token) sequences (Russell & Norvig, 2016, p. 861). Given a word sequence $w_1^n = w_1 \dots w_n$ and with the use of the chain rule of probability, formalized as

$$P(w_1^n) = P(w_1)P(w_2|w_1)P(w_3|w_1^2) \dots P(w_n|w_1^{n-1}) = \prod P(w_k|w_1^{k-1})$$

we can formalize e.g. the bigram approximation as

$$P(w_1^n) = \prod P(w_k|w_1^{k-1})$$

and a more generalized n-gram approximation as

$$P(w_1^n) = \prod P(w_k|w_{k-1}^{k-N+1}).$$

As with other forms of stochastics, probabilities can be calculated by simply counting seen events and normalizing those by the amount of all events. This is called the maximum likelihood estimation (MLE). The maximum likelihood for bigrams in accordance to the formalized bigram approximation can thus be simplified to:

$$P(w_n|w_{n-1}) = \frac{count(w_{n-1}w_n)}{\sum_w count(w_{n-1})}$$

or, since the sum of all bigram counts, which start with a given word w_{n-1} must be equal to the unigram count for that word, this equation can be simplified further to (Jurafsky & Martin, 2008, p. 99):

$$P(w_n|w_{n-1}) = \frac{count(w_{n-1}w_n)}{count(w_{n-1})}$$

where n is the index of a word in a given sequence of tokens and *count* is a function, which simply counts the occurrences of words, which appear at the index n. This equation can be read as the probability of a word at index n given that the word at index $n-1$ appears is approximated by counting how often those two words appeared together in a given corpus and dividing this frequency by the number of appearances of the word at position $n-1$. This is a very simple bigram MLE and a basic language model.

Lastly, the information theory and entropy by Shannon (1948) offers a valuable measure for evaluating language models, called the *perplexity*. Firstly, the entropy in information theory calculates as

$$H(X) = -\sum_{x \in X} p(x) log\ p(x))$$

and is a measure of homogeneity of a message. This measure states how data can be minimally reduced or compressed for the transmission over a communication channel without

losing its original message. Thus, this formula states how many information packages (e.g. bits) are absolutely necessary for transmitting a message.

A measure of how well a given input sequence suits a given probability, which in turn can be a language model, is the so-called *perplexity*. One statistic and expressive feature is the calculation of the perplexities for multiple target classes based on the n-gram language model probabilities for a given input. The perplexity has an inverse relationship to entropy. Thus, the higher the perplexity, the less likely a given input suits the probability distribution at hand. As a feature, language models trained on the totality of all given instances per target class would be measured in their perplexity for a given input. The language model of a class with the lowest perplexity could be assumed to be the most suitable for this input sequence. The feature would be the calculated perplexities per target class language model.

The perplexity of a model q is:

$$2^{H(X)}$$

and thus:

$$2^{-\frac{1}{N} \sum_{i=1}^{N} \log_2 q(x_i)}$$

The perplexity (PP) can also be calculated as (Jurafsky & Martin, 2008, p. 106):

$$PP(W) = P(w_1 w_2 \dots w_N)^{-\frac{1}{N}}$$

which equals:

$$PP(W) = \sqrt[N]{\frac{1}{P(w_1 w_2 \dots w_N)}}$$

The chain rule expands the probability to:

$$PP(W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_1 w_2 \dots w_N)}}$$

Thus for bigrams:

$$PP(W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i | w_{i-1})}}$$

During the course of the dissertation project, language models have been successfully employed. In Johannßen et al. (2019), an LMT was combined with calculated perplexities for language models per target class. Those perplexities were the most influential features of the experiment and became the root of the LMT. The experiment is described in Chapter 7.

2.3.3 Word Embeddings

Embeddings are representations of lexical units such as words, sentences, or texts. Those representations embed those lexical units in high-dimensional vector spaces. Those real-valued vectors can be interpreted as a point in this vector space. Embeddings are useful for determining distributed similarities between lexical units and are of importance for translating those symbolic representations into continuous numerical values (Biemann et al., 2022, p. 218).

One technique for embedding lexical units is the bag-of-words model (BOW). A text gets represented as a vector of frequencies of words. Since the mere frequencies get modeled but not the positions of words, the information of context is not modeled (Biemann et al., 2022, p. 218).

Context and the modeled information of context are of importance in NLP. The principle of words with similar meanings sharing similar contexts is called the *distributional hypothesis* and has been the main principle of many word embedding approaches and statistical semantics (Harris, 1954; Sahlgren, 2008). The way word embeddings are trained is closely connected to language modeling (see Subsection 2.3.2). One advantage of incorporating word embeddings is the possibility of fine-tuning pre-trained embeddings and utilizing those vector representations for multiple NLP tasks such as sequence predictions, language modeling or classification.

One of the most broadly utilized statistical word embedding approaches is called Word2Vec (i.e. word to vector) by Mikolov et al. (2013a,b). Word2Vec can be trained efficiently on even large corpora and assign n-dimensional vectors to words (Biemann et al., 2022, p. 218). Word2Vec is an unsupervised algorithm, which produces so-called dense vector representations. As for so-called sparse vectors, the representations consist mainly of zeros with only *hot* ones where the index or feature applies for the represented lexical unit. Those sparse *one-hot* representations, where BOW is a part, have the disadvantage that a model needs to train and tune many irrelevant zeros, whilst Word2Vec produces dense or *rich* vectors, which do not suffer from this issue.

Word2Vec utilizes the continuous BOW (CBOW) model and a continuous skip-gram model (Mikolov et al., 2013a). This Word2Vec architecture is displayed in Figure 2.13.

For CBOW, a word oughts to be predicted with only the limited (windowed) context of this word, whereas the order of the context words is irrelevant – hence the *bag-of-words*. As input, multiple context words are provided to a shared projection layer, which sums and projects the input words onto a single output. The skip-gram part of the architecture can be taught of as the inverse of CBOW: for the skip-gram model, the context of a given word (i.e. words left and right within a window) ought to be predicted but with a relevant order or index (Mikolov et al., 2013a). The training objective of the skip-gram model for the given word sequence $w_1, w_2, w_3, \dots, w_T$ is:
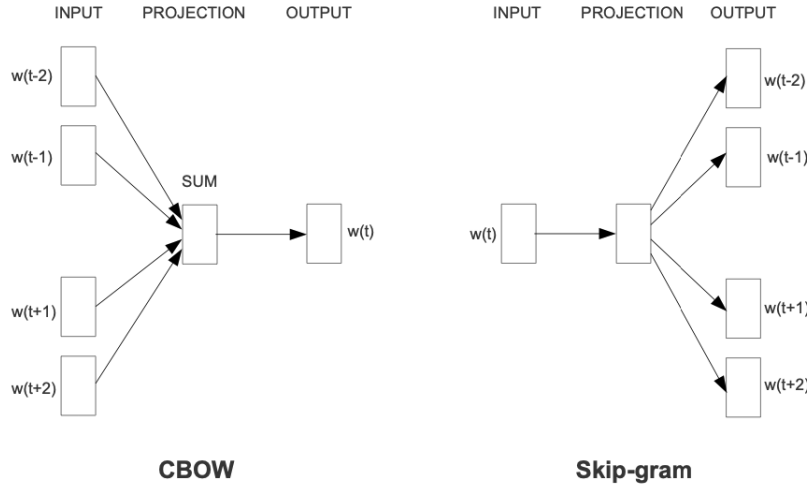
Figure 2.13: The Word2Vec approach combines a continuous bag-of-words model (CBOW) with a continuous skip-gram model. For CBOW a given and orderless context oughts to be utilized for predicting a target word, whilst for skip-gram, a target word oughts to be utilized for predicting an ordered context – both of a defined window of mostly two words in each direction (Mikolov et al., 2013a).

$$\frac{1}{T} \sum_{t=1}^{T} \sum_{-c \leq j \leq c, j \neq 0} log \ p(w_{t+j}|w_t)$$

with c being the size of the training context and T the number of words. In order to receive $p(w_{t+j}|w_t)$, the skip-gram model utilizes the softmax function:

$$p(w_{t+j}|w_t) = \frac{exp(v'^T_{w_o} v_{w_I})}{\sum_{w=1}^{W} exp(v'^T_{w_o} v_{w_I})}$$

where v_w and v'_w are the input and output vector representations of w, and W is the number of words in the vocabulary (Mikolov et al., 2013b).

FastText is an embedding approach by Bojanowski et al. (2017), which extends the ideas of Mikolov et al. (2013a). Instead of a softmax function as the skip-gram model probability, which only predicts one single target context word w_c, FastText alters this approach towards a set of independent binary classification tasks. The goal thus becomes to independently predict the presence or absence of context words. The positive target context words to be predicted get extended by sampled negative words from a dictionary to be avoided. As a binary logistic loss, this results in the negative log-likelihood:

$$log(1 + e^{-s(w_t, w_c)}) + \sum_{n \in N_{t,c}} log(1 + e^{s(w_t, n)})$$

With the denoted logistic loss function $l : x \mapsto log(1 + e^{-x})$, the original objective of the continuous skip-gram model can be rewritten to:

$$\sum_{t=1}^{T}[\sum_{c\in C_t} l(s(w_t, w_c)) + \sum_{n\in N_{t,c}} l(-s(w_t, n))]$$

In other words, the altered objective creates a continuous skip-gram model with negative sampling. FastText introduces a vector representation for each word. Each word w is represented as a bag of character n-gram (Bojanowski et al., 2017). This leads to the properties of FastText being less sensitive to errors. Words with spelling mistakes or so-called *typos* (i.e. an incorrect letter key on the keyboard was mistakenly pressed) receive a very similar vector representation as the correct and correctly spelled word. Furthermore, FastText represent word compositions correctly as a real-valued vector representation of the combination of the words it is composed of. This property is especially helpful for e.g. German language, which utilizes word compoundings frequently (Biemann et al., 2022, p. 222).

2.3.4 Contextualized Embeddings

As described in Subsection 2.3.3, word embeddings are dense vector representations of words, and thus low dimensional representations in a higher dimensional vector space. Whilst rather static embeddings such as Word2Vec or FastText fostered NLP research and allowed for a semantic representation of language, *contextualized embeddings* additionally embed context. Ethayarajh (2019) estimate that for common words with multiple meanings in dependence of the context such as *'mouse'* as depicted by Figure 2.14, static embeddings can only account for 5% of the varriance, whilst contextualized embeddings can account for all of the varriance. These contextualized embeddings have fostered research and have become the SOTA methodology for most NLP tasks and was achieved mainly by the introduction of two architectures: i) Bidirectional Encoder Representations from Transformer (BERT, Devlin et al. (2019)), and ii) the Generative Pre-trained Transformer (GPT, Radford et al. (2018)).
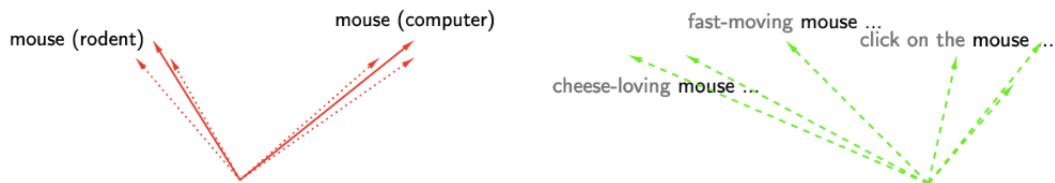
Figure 2.14: Ethayarajh (2019) estimate that contextualized embeddings from BERT are able to explain all of the variance of multi-purpose words such as mouse, whilst static embddings such as those produced by Word2Vec can only account for 5% of the variance on average.

The early transformer-based (for transformers, see Subsectino 2.2.8) contextualized language model GPT released by the company OpenAI[2] utilized only the decoder part of an encoder-decoder architecture with multi-headed self-attention and resulted in a uni-directional language model as to not lessen the model's generalization by having a word to be predicted *seen* itself beforehand (Radford et al., 2018). The subsequent models GPT-2 and GPT-3 are language models, which can be considered SOTA (Brown et al., 2020).

For the dissertation project, BERT language models have been explored and utilized for conducted research (Johannßen & Biemann, 2020). BERT is a transformer neural network architecture (see Subsection 2.2.8) as well. For BERT the authors utilized the encoder part of a transformer architecture and made it bi-directional, giving it the capability to process tokens from left to right and from right to left. To tackle the issue of words to be predicted seeing themselves, the authors introduced the novel methodology of masking parts of the prediction task (Devlin et al., 2019). 15% of words were masked and thus hidden with only their positional information for the model to predict them during the pre-training phase (masked language modelling, MLM). Furthermore, the authors mixed incorrect and correct words per mask during the training time, as to enhance generalization of the language model.
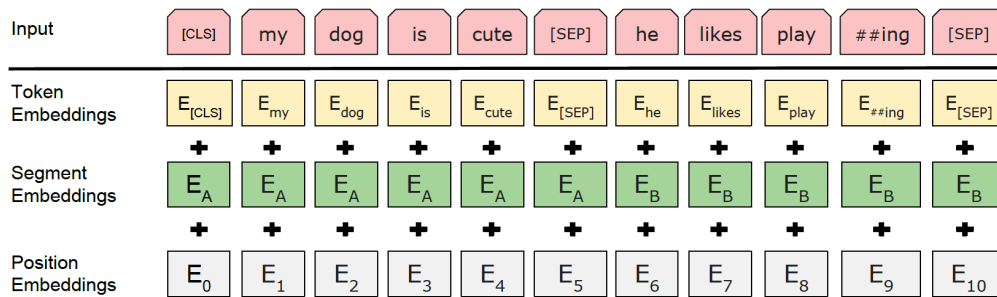
Figure 2.15: Devlin et al. (2019) introduced the concept of masking words to be predicted by inferring from position information during the pre-training phase.

Figure 2.15 illustrates the masking technique, where the network is provided with an input containing masks, as well as token embeddings (identification information from the vocabulary), sentence embeddings (identification of sentences), and position embeddings (the position of a word in a sequence) for inferring the masked words to be predicted. [*CLS*] marks the first token of every sequence. [*SEP*] marks a delimiter utilized during the pre-training phase for next sentence predicitions. [*MASK*] marks masked tokens to be predicted during the pre-training phase.

Another technique employed during the pre-training phase is called next sentence predicitions (NSP). During the NSP pre-training phase, two sentence segments separated by the [*SEP*]

---

[2]https://openai.com

are provided to the model. The second sentence is either successor of the first sentence or randomly chosen. The goal of the model is to identify, whether the sentence is the successor (Devlin et al., 2019).

The BERT Base model architecture consists of 12 transformer blocks containing encoders. After fine-tuning BERT models on a specific task, empirically the concatenated sum of the last 4 layers from the 12 hidden encoders are best to be utilized as contextualized embeddings for downstream tasks (Devlin et al., 2019).

After having learned about the fundamentals of psychological personality, diagnostics, machine learning, and NLP the subsequent section contains related work regarding the interdisciplinary combination of utilizing NLP for psychological textual data. This is important for the broader scientific context of this dissertation project, which fills the gap of automating otherwise manual psychometric assessments.

2.4 NLPsych & Related Work

Mental health is the most common problem domain for approaches that use NLP to characterize psychological traits (Johannßen & Biemann, 2018).

Depression detection systems. Morales et al. (2017) summarized different depression detection systems in their survey and show an emerging field of research that has matured. Those depression detection systems often are linked to language and therefore have experienced gaining popularity among NLP in clinical psychology. Morales et al. (2017) described and analyzed utilized data sources as well. *The Distress Analysis Interview Corpus (DAIC)*[3] offers audio and video recordings of clinical interviews along with written transcripts on depression and thus is less suitable for textual approaches that solemnly focus on textual data but can be promising when visual and speech processing is included. The *DementiaBank* database offers different multi-media entries on the topic of clinical dementia research from 1983 to 1988. *The ReachOut Triage Shared Task* dataset from the SemEval 2004 Task 7 consists of more than 64,000 written forum posts and was fully labeled for containing signs of depression. Lastly, *Crisis Text Line*[4] is a support service, which can be freely used by mentally troubled individuals in order to correspond textually with professionally trained counselors. The collected and anonymized data can be utilized for research.

Suicide attempts. In their more recent work, Coppersmith et al. (2016) investigated mental health indirectly by analyzing social media behavior prior to suicide attempts on Twitter. *Twitter*[5] is a social network, news- and micro-blogging service and allows registered users to

---

[3] http://dcapswoz.ict.usc.edu/
[4] https://www.crisistextline.org/
[5] https://twitter.com/

post so-called tweets, which were allowed to be 140 characters in length before November 2017 and 280 characters after said date. As before in Coppersmith et al. (2015), the Twitter users under observation had publicly self-reported their condition or attempt.

Crisis. Besides depression, anxiety or suicide attempts, there are more general crises as well, which Kshirsagar et al. (2017) detect and attempt to explain. For their work, they utilized a specialized social network named *Koko*[6] and used a combination of neural and non-neural techniques in order to build classification models. Koko is an anonymous emotional peer-to-peer support network, used by Kshirsagar et al. (2017). The dataset originated from a clinical study at the Massachusetts Institut of Technology (MIT) and can be implemented as a chatbot service. It offers 106,000 labeled posts, with and some without crisis. A test set of 1,242 posts included 200 crisis labeled entries, i.e. $\sim 16\%$. *Reddit*[7] is a community for social news rather than plain text posts and offers many so-called subreddits, which are sub-forums dedicated to certain, well-defined topics. Those subreddits allow for researchers to purposefully collect data. Shen & Rudzicz (2017) detected anxiety on Reddit by using depression lexicons for their research and training Support Vector Machine (SVM, (Cortes & Vapnik, 1995)) classifiers, as well as Latent Dirichlet Allocation (LDA, (Blei et al., 2003)) for topic modeling. Those lexicons offer broad terms that can be combined with e.g. LIWC (see Subsection 2.3.1) features in order to identify different conditions in order to be able to distinguish those mental health issues. Shen & Rudzicz (2017) used an API offered by Reddit in order to access subreddits such as r/anxiety or r/panicparty.

Dementia. In their recent work, Masrani et al. (2017) used six different blogs to detect dementia by using different classification approaches. Especially the lexical diversity of language was the most promising feature, among others.

Multiple mental health conditions. Coppersmith et al. (2015) researched the detection of a broad range of mental health conditions on Twitter. Coppersmith et al. (2015) targeted the good discriminability of language characteristics of the following conditions: attention deficit hyperactivity disorder (ADHD), anxiety, bipolar disorder, borderline syndrome, depression, eating disorders, obsessive-compulsive disorder (OCD), post-traumatic stress disorder (PTSD), schizophrenia, and seasonal affective disorder (SAD) – all of which were self-reported by Twitter users.

Dream language. Niederhoffer et al. (2017) researched the general language of dreams from a data-driven perspective. Their main targets are linguistic styles, differences between waking narratives and dream narratives, as well as the emotional content of dreams. In order

---

[6]https://itskoko.com/
[7]https://www.reddit.com/

to achieve this, they used a community named DreamsCloud. *DreamsCloud*[8] is a social network community dedicated to sharing dreams in a narrative way, which also offers the use of data for research purposes. There are social functions such as 'liking' a dream narrative or commenting on it, as Niederhoffer et al. (2017) describe in their work. There are more than 119,000 dream narratives from 74,000 users, which makes this network one of the largest of its kind. An advantage of utilizing DreamsCloud for the assessment of dream language is its high specialization in contrast to more generalized social networks such as Facebook [9].

LIWC and personality traits. Hawkins & Boyd (2017) laid their focus on LIWC characteristics especially and a correlation with the personality of a dreamer. Data was collected by clinical studies in which Hawkins & Boyd (2017) gathered dream reports from voluntary participants. Their work is more thorough in terms of length, depth, and rate of conducted experiments on LIWC features. Dreams could be distinguished from waking narratives, but – as of said study – correlations with personality traits could not be found.

Mental changes and mental health problems are seemingly connected. However, natural changes such as growth or life-changing experiences can alter the use of language as well.

Data generation and life-changing events. Oak et al. (2016) pointed out that the availability of data in clinical psychology often is difficult for researchers. The application scenario chosen for a study on data generation for clinical psychology are life-changing events. Oak et al. (2016) aimed to use NLP for tweet generation. The BLEU score measures n-gram precision, which can be important for next character- or next word predictions, as well as for classification tasks. Another use case of this measure is the quality of machine translations. Oak et al. (2016) use the BiLingual Evaluation Understudy (BLEU) score to evaluate the quality of their n-grams for language production of their data generation approach of life-changing events. Even though the generated data would not be appropriate to be used for e.g. classification tasks, Oak et al. (2016) nonetheless proposed useful application scenarios such as virtual group therapies. 43 percent of human annotators thought the generated data to be written by real Twitter users.

Changing language over the course of mental illnesses. A study by Reece et al. (2017) revealed that language can be a key for detecting and monitoring the whole process from onsetting mental illnesses to a peak and a decline as therapy shows positive effects on patients. participants involved in the study had to prove their medical diagnosis and supply their Twitter history. Different techniques were used to survey language changes. The crowdsourcing marketplace Amazon Mechanical Turk (MTurk[10]), which allows researchers to define manual tasks and quality criteria, was used for labeling their data. Reece et al. (2017) were able to show

---

[8]https://www.dreamscloud.com/

[9]http://www.facebook.com

[10]https://www.mturk.com

a correlation between language changes and the course of a mental disease. Furthermore, their model achieved high accuracy in classifying mental diseases throughout the course of illness.

Language decline through dementia and Alzheimer's. It is known that cognitive capabilities decline during the course of the illness dementia. Masrani et al. (2017) were able to show that language declines as well. Lancashire & Hirst (2009) researched the possibility of approaching Alzheimer's of the writer Agatha Christie by analyzing novels written at different life stages from age 34 to 82. The first 50,000 words of included novels were inquired with a tool named TACT, which operates comparable to LIWC and showed a decline in language complexity and diversity. During their research, Masrani et al. (2017) detected dementia by including blogs from medically diagnosed bloggers with and without dementia. Self-reported mental conditions, as it is often used for the research of social networks, are at risk of being incorrect (e.g. pranks, exaggeration, or inexperience).

Development. Goodman et al. (2008) showed that the acquisition and comprehension of words and lexical categories during the process of growth corresponds with frequencies of parental usage, depending on the age of a child. Whilst the acquisition of lexical categories and comprehension of words correlates with the frequency of word usage of parents later on in life, simple nouns are acquired earlier. Thus, whether words were more comprehensible was dependent on known categories and a matter of similarity by the children.

Emotions and motivations are less common problem domains. Some approaches aim at detecting general emotions, further researchers focus on strong emotions such as hate speech, others try to provide valuable resources or access to data.

Distant emotion detection. In order to better understand the emotionality of written content, Pool & Nissim (2016) used emotional reactions of Facebook users as labels for classification. *Facebook* offers insightful social measurements such as richer reactions on posts (called *emoticons*) or numbers as friends, even though most available data is rather general.

Hate speech. Serrà et al. (2017) approached the question of emotional social network posts by surveying the characteristics of hate speech. In order to tackle hate speech usually containing a lot of neologism, spelling mistakes and out-of-vocabulary words (OOV), Serrà et al. (2017) constructed a two-tier classification that firstly predicts the next characters and secondly measures distances between expectation and reality. Other works on hate speech include that of Benikova et al. (2018), Warner & Hirschberg (2012), and Schmidt & Wiegand (2017).

Motivational dataset. Since data sources for some sub-domains such as motivation are sparse, Pérez-Rosas et al. (2016) contributed a motivational interviewing (MI) dataset by including 22,719 utterances from 227 distinct sessions, conducted by 10 counselors. Pérez-Rosas et al. (2016) used MTurk for labeling their short texts by crowdsourcers. They achieved a high Intraclass Correlation Coefficient (ICC) of up to .95. MI is a technique in which the topic

'change' is the main object of study. Thus, this dataset could also contribute to early mental disease detection. MI is mainly used for treating drug abuse, behavioral issues, anxiety, or depression.

Emotions. Pool & Nissim (2016) summarized in their section on emotional datasets some highly specialized databases on emotions, which the authors analyzed thoroughly. *The International Survey on Emotion Antecedents and Reactions* (ISEAR)[11] dataset offers 7,665 labeled sentences from 3,000 respondents on the emotions of joy, fear, anger, sadness, disgust, shame, and guilt. Different cultural backgrounds are included. *The Fairy Tales*[12] dataset includes the emotional categories angry, disgusted, fearful, happy, sad, surprised, and has 1,000 sentences from fairy tales as the data basis. Since fairy tales usually are written with the intention to trigger certain emotions of readers or listeners, this dataset promises potential for researchers. *The Affective Text*[13] dataset covers news sites such as Google news, NYT, BBC, CNN and was composed for the SemEval 2007 Task 14. It offers a database with 250 annotated headlines on emotions including anger, disgust, fear, joy, sadness, and surprise.

Few researchers in NLPsych have approached a connection between language and academic success. Some challenges are lack of data and heavy biases as some might assume that an eloquent vocabulary, few spelling mistakes or sophisticated use of grammar indicate a cognitively skilled writer. Pennebaker et al. (2014) approached the subject in a data-driven fashion and therefore less biased. Data was collected by accessing more than 50,000 admission essays from more than 25,000 applicants. The college admission essays could be labeled later academic success indicators such as grades. The study showed that rather small words such as function words correlate with subsequent success, even across different majors and fields of study. Function words (also called closed class words) are e.g. pronouns, conjunctions, or auxiliary words, which tendentially are not open for expansion, whilst open class words such as e.g. nouns can be added during productive language evolution.

After this section on related work, the following last section of this chapter describes a best-practice approach, which can be identified in most works discussed in this previous related work section. The empirical research presented in Part III follows this best-practice approach.

2.5 NLPsych Best-Practice Approach

During a survey on NLPsych conducted by Johannßen & Biemann (2018) a broad variety of fragmented research works were similar in that they adopted a best-practice approach for crafting NLPsych systems for mostly classification tasks.

---

[11] http://emotion-research.net/toolbox/toolboxdatabase.2006-10-13.2581092615

[12] https://github.com/bogdanneacsa/tts-master/tree/master/fairytales

[13] http://web.eecs.umich.edu/~mihalcea/downloads/AffectiveText.Semeval.2007.tar.gz
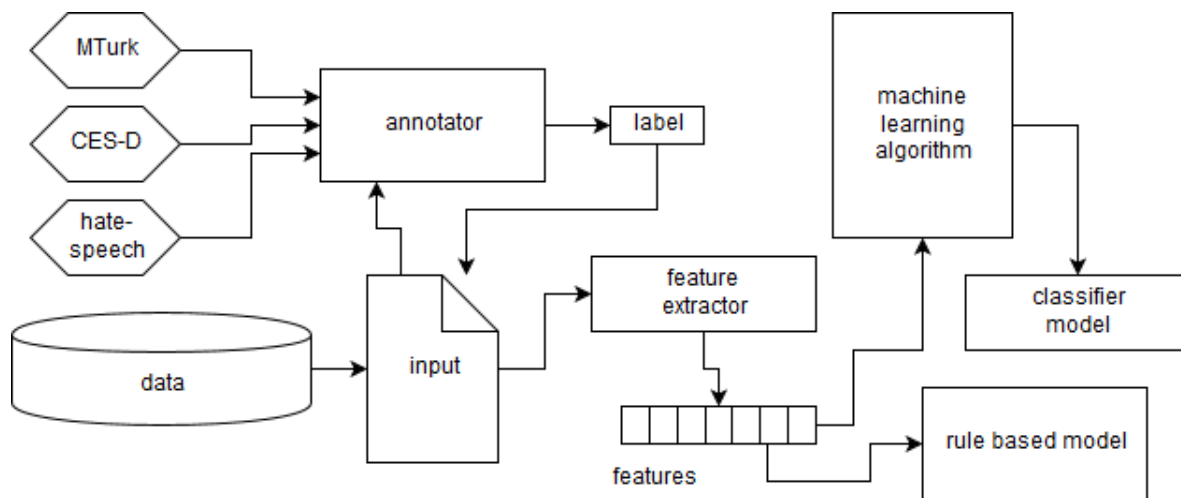
Figure 2.16: A general setup for classification tasks in NLPsych

Figure 2.16 illustrates this best-practice classification system approach. Firstly, after having collected data, pieces of information are read and function as input. Different measures or techniques can be applied to the data by an annotator to assign labels to the input. Whether or not annotation takes place, depends on the task and origin of the data.

Secondly, after separating training, test, and sometimes development sets, features get extracted from those data items, e.g. LIWC category counts, or part-of-speech (POS) tags. A feature extractor computes a nominal or numerical feature vector.

Thirdly, depending on the approach, this feature vector is directly fed into rule-based models such as e.g. defined LIWC scores that correlate with dream aspects, as Niederhoffer et al. (2017) did. A different approach uses the feature vector on a machine learning algorithm in order to compute a classifier model that thereafter can be used to classify new instances of information, as Reece et al. (2017) demonstrated in their work.

Finally, for both of the approaches, the accuracy of the classification task is determined and researchers analyze and discuss the consequences of their findings, as well as use the models for classification tasks.

This chapter laid the background on psychological personality diagnostics, machine learning, and NLP. Furthermore, this chapter presented related work on the interdisciplinary field of NLPsych and concluded with a best-practice approach.

The next chapter is concerned with ethics and ethical considerations. It contains the fundamentals of ethics as a discipline, describes NLPsych ethics in more detail, and discusses a critical assessment of a conducted shared task. Since the automation of textual processing via NLP for modeling psychological metrics, an array of possible ethical issues emerges that ought to be discussed.

Ethical Considerations of NLPsych

This chapter contains the ethical fundamentals in Section 3.1, a more detailed assessment on NLPsych ethics in Section 3.2, and finally presents and discusses the ethical consideration of a conducted shared task in Section 3.3.

Ethical considerations of NLP research have become a growing concern. Codes of ethics did exist in most institutions and in most scientific fields (e.g. for the Association for Computing Machinery (ACM), the first code was released in 1992), but their strict compliance has only recently been demanded. The ACM released a major revision of its code of ethics in 2018 together with dedicated sessions and mandatory ethical evaluations for submitted research papers (Wolf et al., 2019). The code was adopted by the Association for Computational Linguistics (ACL) (Schütze, 2020).

The ACM code of ethics (Gottenbarn et al., 2018) is organized in four major parts, 1) general ethical principles, 2) professional responsibilities, 3) professional leadership principles and 4) compliance with the code. The principals and guidelines of the code help researchers audit their research in terms of ethical soundness. However, some aspects such as dual-use or even forbidden research, which were points raised in a debate on the GermEval 2020 conducted shared task 1 on the classification and regression of cognitive and motivational style from text, rooted deeper in normative and meta-ethics than most rather checkbox-oriented guidelines can provide. This chapter provides necessary ethical and applicatory fundamentals to discuss ethics of NLPsych research in more detail.

3.1 Ethical Fundamentals

First, a few of the basic terminologies such as ethics, morals, or justice will be clarified. Thereafter, selected and for this work, relevant schools of thought are described. Finally, the basic idea and challenges of an ethical dilemma are explored.

3.1.1   Ethics, Morals, & Justice

Ethics, ethos, and morals are often used and discussed interchangeably. The basic principles of all three terms stem from the same, European-centered idea of what is considered to be *good*. The ACM code of ethics adopts this idea of doing good and interprets it as contributing to society and to human well-being (Gottenbarn et al., 2018). However, being good does not possess a manifestation in and of itself, but rather finds its meaning in the context of society and from a point of view (Altman, 2011, p. 23). Europe and the United States developed an individualistic point of view, whilst African cultures and societies (despite Africa being a continent consisting of 54 countries and thus multiple cultures) lean towards collectivism (Brey et al., 2015).

The European and 'western' consensus of ethics stems from Kant and his Kantian deonto-logical idea of ethics as being instrumentally read as *what is the right thing to do?* (Waluchow, 2003, p. 67). A finer distinction between the terms can be drawn when not ethics, but an ethical theory in contrast to moral philosophy is considered (Waluchow, 2003, p. 15). Accordingly, the ethical theory is a branch of philosophy, which aims to understand and guide the practice of morality. Moral philosophy, on the other hand, is the practice of doing good and thinking about moral dilemmas or making moral judgments. Whilst ethics is concerned with the philosophy of a good life and right actions, ethos is the basic individual attitude towards one's own ethics. To view an act or thought as ethical is associated with the thought process for an individual situation. Morality is the sum of an individual's norms and values in a society or group of people (Pflege & Menche, 2014).

Justice, or within the application domain of ethics and morals *social justice*, is concerned with a community's obligations to correct conditions, that are detrimental to individuals or groups (Waluchow, 2003, p. 196). In turn, justice can be understood as sanctioning by the sovereign (i.e. a state or the people) through the utilization of norms (i.e. by morality) under threat of punishments or consequences.

Now that the terms ethics, morals, and justice were clarified, it is important to note that the everyday enacted morals of individuals might not be shared as ethics of a community. That is: practices or situations might be viewed as unethical, even though individuals would not act upon correcting them. Furthermore, it is important to note, that even commonly viewed unethical and immoral circumstances might not be corrected by the collective or sovereign, e.g. when practices or wrongdoing are not against the law and thus not illegal ('acting above the law') (Folger et al., 2005, p. 2019).

3.1.2   Descriptive, Normative, and Meta Ethics

Descriptive ethics is concerned with the manifestation of morals and ethos in a society with the ethos being the fundamental belief of an individual, differentiated into attitudes, positions, and the own meaning of life.

Normative ethics search for righteous lived customary nominal statements. It makes statements on how humans should act, what the standards of actions are, and what duties emerge. Furthermore, normative ethics discuss life goals, desires, and what to strive for. Lastly, virtue is being investigated.

Metaethics does not make content-wise statements on virtue or the moral good, rather than investigating the ethics of society itself (Pflege & Menche, 2014, p. 18).

3.1.3 Ethical Schools of Thought

The study of cognitive processes can reveal deeply personal and sensitive information. Previous research has concluded that the use of natural langauge can reveal cognitive processes beyond of what is narratively being said – this is called psychological pragmatics (Boyd & Schwartz, 2021). Since learning machines can pick up nuanced signals, which humans might miss, these machine learning systems can cause a multitude of dangers and cause ethical dilemmas (see Subsection 3.1.4) between valuable utility and potential misuse. It is this duality and dilemma, which renders an assessment of ethics for this dissertation project tremenduously important. This importance is emphasized by a substantial debate which had emerged on the conducted GermEval20 shared task 1 described in Section 7.5 (Johannßen et al., 2020b).

Time and again, philosophers established novel ideas, contributed those first by mere speeches and through conversations and later on mainly over written texts. Many of those influential philosophers became scholars, founded their own schools or institutions, and passed on their philosophical views and ideas to students. Over the course of centuries, some so-called philosophical schools of thought emerged in a distinct, differentiable niche of philosophy (Weischedel, 2005).

Three main ethical schools for ethical analysis exist, which are i) consequentialism (and thus utilitarian, which is part of this school), ii) deontology, and iii) virtue ethics (Kaptein & Wempe, 2002; Gensler, 2011; Graham, 2010). Some scholars also include iv) contractualism in their consideration of relevant ethical schools (Werner, 2020). An overview of those schools of ethic is provided in Figure 3.1.

Consequentialism

Consequentialism is an ethical school of thought, which is mainly concerned with behavioral and observable acts. It is therefore part of normative ethics, which can also be measured and described empirically. For consequentialists, the mere desirability of an outcome determines whether an action or behavior is viewed as being *good*. An individual's mere duty is to do whatever promises the best consequences.

Consequentialism can be further differentiated into two more detailed forms: i) classical utilitarism and ii) rule utilitarism (Gensler, 2011, p. 110 f.).

Classical Consequentialism states that one ought to act in a way that results in the best balance of pleasure over pain for everyone that might be affected by an action. Whilst Classical

Consequentialism is mainly thought of and transported by strict rules, Utilitarianism (a more precise form of Consequentialism) adopts the mindset of ensuring everyone's happiness with the aim of maximizing it.

One influential representative of Utilitarianism was John Stuart Mill, which argued with principles of the New Testament and the so-called Golden Rule. This rule states the well-known principle that one should only treat others as one consent to being treated in the same situation. Mill viewed many teachings of the New Testament as honoring the Golden Rule, such as love your neighbor as yourself (Mill, 1871, p. 22).

Deontology

In opposition to Consequentialism, Deontology (also called Nonconsequencialism) views some actions as wrong in and of themselves and not just because an action's consequence is bad. With this, Deontology is rather a-priori in its evaluation of *good*, whilst Consequentialism is a-posteriori and evaluates the consequences of an action as *desirable* or *undesirable* (Gensler, 2011, p. 110).

Modern Deontology is shaped by the teachings of Immanuel Kant (Waluchow, 2003, p. 173). In Kant's teachings, Deontology consists of absolute, exceptionless, or strict principles, which can not be circumvented, whilst decisions ought to be free of any feeling, thoughts of consequences, or moral judgments. According to Kant, reason must be in the center of those deontological rules. This reason is the human capacity to act upon valid reasons of actions, making us human (the so-called Categorical Imperative) (Waluchow, 2003, p. 174).

Virtue Ethics

Virtue Ethics is a broad adapted form of normative ethics (see Subsection 3.1.2). In contrast to the duty or rules central to Deontology or the consequences of an action (Consequentialism and Utilitarianism), Virtue Ethics emphasizes virtues or moral character. Instead of relying on fundamental rules such as the Golden Rule, Virtue Ethics would rather advise helping others for the sake of being charitable (Hursthouse & Pettigrove, 2018).

To find the best possible (normative) action, Virtue Ethics rely on the idea of what *good* truly manifests to, the laws of nature or intersubjective communication (Gensler, 2011, p. 139). The foundations of all of those aspects – the idea of good or human nature – have been laid by Plato and Aristoteles, which both are the main representatives of Virtue Ethics.

Contractualism

Contractualism emerged as a political philosophy during the late seventeenth and early eighteenth century by influential philosophers such as Thomas Hobbes, John Locke, and Jean Jaques Rousseau. Their work mainly contained economical theories and thus the thought emerged, that ethics could be viewed as a contract between individuals, that has to be negotiated and settled at the ideal point between own desires and honoring the desires of the other

participating party. Law and society, therefore, are products of an actual or hypothetical social contract among rational, self-interested individuals (Waluchow, 2003, p. 121).

Thomas Hobbes coined the idea of liberal markets, the *Leviathan* as godly being, consisting of the sovereign of the people, and the invisible hand controlling the free markets to an equilibrium of interests between economic subjects. Therefore, the representatives of Contractualism believed in perfect fairness of a decision, when every involved party or subject acts in its own best interest with the goal of reaching an intersubjective agreement, that settles in the middle of those opposing poles (Ashford & Mulgan, 2018).

Metaethical consideration			
What constitutes an ethical statement? What characterizes it?			
Virtue ethics **Natural law theory** Platon, Aristoteles, Max Scheler, John Dewey Moral actions and norms are justified by values determinable by content. Values consists of • the idea of good • human nature • feelings • intersubjective communication	**Deontology** **Duty ethics** Kant, Habermas Moral action is formed by the subject. Norms can be discovered and communicated rationally. Norms consist of • Reason • interpersonal communication	**Utilitarianism** **Consequentialism** Jeremy Bentham, John Stuart Mill Moral action can be measured by their consequences. Norms consist of • weighing the maximum of happiness • lust • benefit • minimalization of suffering	**Contractualism** John Hobbes John Rawls Moral action is based on a contract. It is concluded between legal entities Norms consist of • will to survive • the ideal form of justice

Figure 3.1: Overview of different schools of ethics with their main representatives and basic ideas (adapted from Pflege & Menche (2014)).

3.1.4   Ethical Dilemmas

As defined in Subsection 3.1.1, ethos is the study of righteous living. The conception of what it means to live in the *right* way is difficult to define. Kant would view it with his categorical imperative, which – in a nutshell – asks, whether a righteous deed is done only for the sake of doing good and is viewed by every person as such. Therefore, for a problem to be viewed as ethically debatable or unsound, one must suspect this problem or its solution to violate or harm others. Altman (2011, p. 4) defines the Categorical Imperative as formular of universal law, which states that one should not adopt a subjective principle of action (which Kant calls maxim), that one cannot also will as a universal law.

The mentioning of Kant's Categorical Imperative is necessary in modern times, since some problems are marked as ethically questionable and are debated upon not for the sake of doing

good but at times for the sake of preventing sanctions or an enforced justice, which would not be a shared maxim of what is right or wrong amongst all people. If every problem with its solution is expected to be viewed as ethically questionable, then the Categorical Imperative does not exist anymore, as it does not differentiate between the maxim and the norm.

To honor the Categorical Imperative, one must first determine whether or not the problem at hand is an *ethical dilemma*. Only those ethical dilemmas should be viewed as ethically questionable. To mark a problem as ethically questionable, which are not ethical dilemmas, can hinder the problem's solution-finding process, and thus can dampen progress. Utilitarianism as a school of thought (see Subsection 3.1.3) and which Kant had embraced (Altman, 2011, p. 11), would evaluate this as an undesirable outcome, which should be avoided at all costs.

Braunack-Mayer (2001) define ethical dilemmas narrowly as situations, in which – on moral grounds – an individual both ought to do and not to do something. In those cases, where there is neither a choice of clearly right nor wrong action in a moral situation, this case thereafter becomes a moral dilemma. Accordingly, an ethical dilemma involves choices and conflicts between those choices or the outcomes thereof. However, at times problems are declared as ethical dilemmas solemnly on the idea of a conflict and not of an already occurred conflict itself. The issue lies within the predictability of outcomes. Open points of discussion are practices, which involve uncertain wrongful, or injustice outcomes, which are already viewed as ethical dilemmas and acted upon such as *predictive policing*, where sanctions are issued even before harm even could have been inflicted (Schlehahn et al., 2015).

After these fundamentals of ethics, especially a basic unerstanding of different perspectives such as descriptive vs. normative ethics and the different schools of ethics, the following section presents a more detailed assessment of ethics in the field of NLPsych.

3.2 NLPsych Ethics

As described in Subsection 3.1.4, even inventors with the normatively most preferable intentions in mind are faced with the ethical dilemma that, thus far, all impactful and paradigm-changing inventions have been utilized or evaluated for the possibility of being utilized for harmful and often militaristic purposes. This circumstance is called dual-use (see Subsection 3.3.4) (Williams-Jones et al., 2014).

In addition to that, for every invention, there might be unforeseeable and unintended negative consequences. It might be infeasible to analytically or empirically inspect every possible outcome or behavior of a system for every possible input. The more complex a system becomes (in terms of the number of parts and the number of changing relationships of those parts to each other), the harder it becomes to analyze those systems exhaustively. Modern NLP produces and utilizes complex systems and even language itself can be considered a complex system due to the sheer amount of possible utterances and the myriads of possible relations

those utterances can have, resulting in e.g. uncalculable conditional probabilities (Heyer et al., 2006, p. 122). At times, inventors know about those shortcomings or at least suspect them or those unfavorable outcome disparities are unbeknownst to the inventor. This is called a bias. A bias can be defined as a mismatch of an underlying distribution of the ground truth and the approximated distribution modeled by a system (Hovy & Prabhumoye, 2021).

In the case of so-called outcome disparity biases, there is a distinction between normative and descriptive disparities. Even though we might not normatively agree with the association that a certain nation might be predicted to be obese, this might very well be descriptively correct, if the body mass index does measure obesity on average. On the other hand, we might not normatively agree with the association of doctors being predicted to be male, even though this might descriptively be correct (Hovy & Prabhumoye, 2021).

Besides those priors and impacts on our informed decisions, such mismatches between ground through and prediction can lead to a number of unfavorable characteristics of such a system and its workings, such as a poor generalization and thus worse performances on yet unseen data (Shah et al., 2020).

Rather than researching narrow tasks and niches, some research have systematized the phenomenon of biases in NLP systems (Hovy & Prabhumoye, 2021; Shah et al., 2020; Blodgett et al., 2020). Firstly, this section summarizes systematized occurrences of biases in NLP systems. Thereafter, countermeasures are discussed. In order to comprehend the magnitude of distribution severities, the scientific understanding of *truth* is established. Lastly, interpretable, understandable, and explainable artificial intelligence (XAI) is described and ethically evaluated, which is of importance for the subsequent research shown in this thesis.

3.2.1 Occurences of Harmfull Biases in Normative Settings

In terms of machine learning biases, an important differentiation has to be made between descriptive and normative assessments (see Subsection 3.1.2). Not all biases are to be avoided. As described in Subsection 2.2.10, biases at times do describe real-world functional relationships and patterns. In these cases, a bias can be descriptively correct and thus be desired. However, some models tend to overestimate patterns adapted during the training process, leading to then non-descriptive biases. These overestimated non-descriptive biases are to be avoided since they lead to worse results during the testing phase (see Subsection 2.2.9) and poorer overall generalization of the resulting model.

Besides the assessment of whether a bias is descriptive, one can also evaluate the normativity of a bias. If a bias descriptively discriminates between job performances of men compared to women in a non-normative disadvantageous way for the women, designers of machine learning systems might want to alter the model's predictions away from the descriptiveness of previous data patterns (if e.g. the discrimination occurred due to poor annotator training) and towards the normatively desired outcomes.

In this second approach, the normative point of view, despite models reflecting descriptively correct patterns, can be considered as an ethical dilemma: the aim would be to curtail data correctness for a more normative utilization of machine learning models (for ethical dilemmas, see Subsection 3.1.4), with which this subsection is concerned.
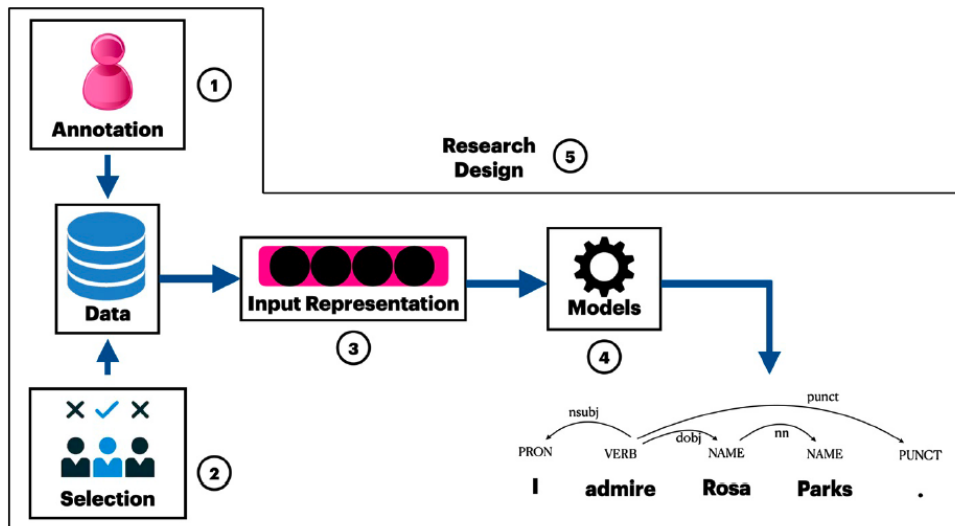
Figure 3.2: Hovy & Prabhumoye Hovy & Prabhumoye (2021) identified five bias sources in general NLP processing pipelines: 1) the procedure used for annotating the labels, 2) the labels chosen for training, 3) the choice of representation used for the data, 4) the choice of models or machine learning algorithms used, or 5) the entire research design process.

Hovy & Prabhumoye (2021) identified five bias sources in general NLP processing pipelines: 1) the procedure used for annotating the labels, 2) the labels chosen for training, 3) the choice of representation used for the data, 4) the choice of models or machine learning algorithms used, or 5) the entire research design process (see Figure 3.2).

Biases from annotations

The main reason for biases from annotations is a mismatch between the annotator population from the (inherent, unknown) base population of the data. Hovy & Prabhumoye (2021) name some possible reasons for *label bias*, which were not measured properly and appear to be rather speculative, namely that annotators are lazy, uninterested, or distracted. However, it is noted that a more harmful bias emerges from informed and highly motivated annotators, that disagree. Blodgett et al. (2020) recognize the issue of possible *label biases* by recommending the research question: "Data: How are datasets collected, preprocessed, and labeled or annotated?".

Bias from input representations

Jurafsky & Martin (2008) define language models as models that assign probabilities to sequences of words. As described in more detail in Section 2.3, both statistical linguistics as well

as word embeddings model the probabilities of events on the basis (i.e. the occurrences of words in a certain order) in accordance to corpora and the contexts of words in those corpora. Therefore, the subsequent statistical or technical representation of words depends on the way words have been used and the context they were in. Since a bias is defined as a mismatch between a sample and a population, differing contexts and word meanings between corpora and subsequent utilization can lead to representation biases.

This embedding mismatch can be investigated further in the field of societal semantic biases into descriptive and normative correctness of representations. As Shah et al. (2020) survey, societal attitudes are mostly captured by *normative truths*, whilst those can and often do differ from *descriptive truths*, creating amplifying effects by word embedding representations – e.g. women might be represented closer (angle-wise) to cooking than to mechanics compared to men, which might be descriptively correct (if that is the narrative most often found in corpora) but most likely normatively wrong, as this representation does not reflect our modern understanding of emanzipation (Garg et al., 2018; Kozlowski et al., 2019).

Not only do technical or statistical mismatches lead to biases. Especially during the phase of feature engineering (see Section 2.3), erroneous priors can result from a conscious selection of partial information to be considered when representing language and experimental inputs. This conscious selection suffers from (at times unconscious) stereotypes or incorrect assumptions. Shah et al. (2020) created the bias origination and occurrence overview displayed in Figure 3.3 and associate a total of three possible biases when engineering features: i) over-amplification, ii) label bias, and iii) selection bias.
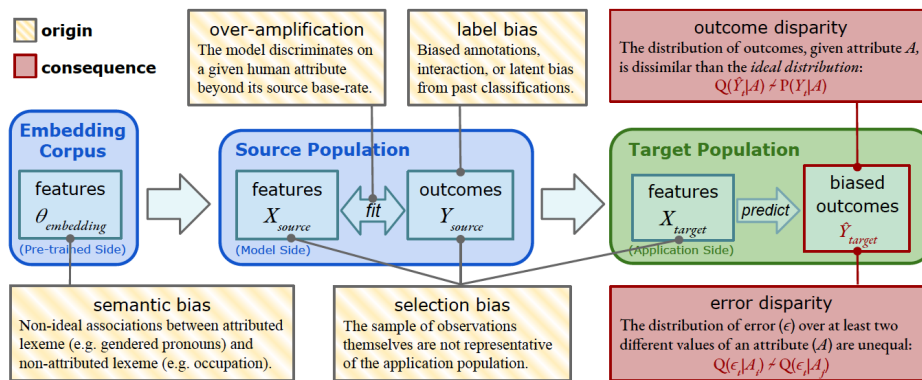
Figure 3.3: Blodgett et al. (2020) proposes a framework for identifying different bias reasons. The yellow crosshatched textboxes describe the possible origins of biases. The red textboxes describe the consequences. Noteworthy, the consequences mostly occur during the last step of an experiment, the predictions, even if they originated much early in the experimental pipeline.

Bias from data

Both prior bias sources, from annotations and input representations, can result in biases from

the utilized experimental data. Statistical and technical language models can only model and adapt signals, which were induced into datasets and are offered during the training phase. Some of the strongest language models such as GPT-3 (Brown et al., 2020) consist of 175 billion tuned parameters and was trained on 500 billion tokens, mostly obtained from *Common Crawl*.[14] As Hovy & Prabhumoye (2021) survey, many resources incorporated into large corpora emerge from utterances produced by white, middle-aged, educated, upper-class men (Garimella et al., 2019; Hovy & Søgaard, 2015).

Another hurdle identified by Hovy & Prabhumoye (2021) is the temporal divergence between many available corpora and the current broad use of natural language. Over time, the way people express themselves and use natural language can alter greatly from the utilized language of the past. The severity of this divergence depends on the dynamic of the gathered resource or medium. Short messaging services most likely adapt faster to a change in colloquial language or slang compared to e.g. natural language utilized in novels and literature. But even highly standardized writing sites e.g. journalism and news outlets can diverge in their use of language from once collected corpora (Hovy & Prabhumoye, 2021).

Furthermore, major events can alter the use of language in an impactful way over a very short amount of time. The research on the COVID-19 pandemic in Chapters 8 and 9 display a change in language (Johannßen & Biemann, 2020; Johannßen et al., 2022). The Leibniz-Institut for the German Language approximated 2,000 novel words, which have emerged during the pandemic and have been utilized by the broad majority of Germans frequently.[15]

Bias from models

Different from humans, machine learning models do not actively reach out for additional or different data, resources, or experimental alterations. Models can simply adapt patterns inherently provided by training data. Since machine learning models do not possess world knowledge, recognized patterns tend to be overvalued. Due to stigmata (i.e. descriptive data truths, which might not necessarily be reflected as normatively desirable) leading to better training performances, models tend to amplify patterns. This is called bias overamplification.

Blodgett et al. (2020) formalize and define this overamplification bias as follows: "in overamplification the predicted distribution $(Q(\hat{Y}_s|A_s))$ is dissimilar to the source training distribution $(Q(Y_s|A_s))$ with respect to a human attribute, A". It is, once again, the prior $P(A)$, which differentiates or mismatches the modeled signal from the signal inherent in the ground truth, which Blodgett et al. (2020) formalize as: $(Q(\hat{Y}_s|A_s)) \nsim (Q(Y_s|A_s)) \sim P(Y_t|A_t)$.

Bias from research design

Many of the proposed biases are well known. Even though it is challenging, those biases can be detected and countermeasures can be taken (see the next Subsection 3.2.2). However, the occurrences, the detection and the possible countermeasures of an NLP experiment and

---

[14]Common Crawl, `https://commoncrawl.org/`.

[15]`https://www.ids-mannheim.de/neologismen-in-der-coronapandemie/`

research are highly depended on the overall research design, the awareness of the authors and the experimental circumstances in which experiments are performed.

Hovy & Prabhumoye (2021) are mainly concerned with the research design bias emerging from the majority of linguistic resources coming from- and research being performed on the English language. This impairment hinders research being performed on other, less-resourced languages, which, in turn, limits the resource availability of those sparse resource languages. Blodgett et al. (2020), which illustrated different bias sources in Figure 3.3, did not address the research design as possible bias source.

Lastly, a resulting bias from an insufficient research design can also suffer from resource sparseness due to data protection laws. In order to analyze and countermeasure e.g. demographic biases, those demographic information ought to be available. In our ethical consideration of the GermEval shared Task 1 on Cognitive and Motivational Style from Text (Johannßen et al., 2020b) we addressed this lack of demographic knowledge due to the (righteously, we may add) strict data protection law in Germany, which mostly prohibits the inclusion of personal data collection without clear necessity.

3.2.2 Countermeasures for Biases

Biases from annotations

The most promising and easy countermeasure for annotation biases is firstly the detection of said bias. This can mainly be achieved by either comparing annotated label with gold labels (in case of those gold labels being present), or by having multiple annotators label the same instances for measuring their agreements Ragheb & Dickinson (2013).

In case of disagreements between multiple annotators, these differences could either be accepted as part of the ambiguity of the problem and simply be reported in a paper. Investigations of those differences could reveal multiple correct answers, which in turn could be employed in the models and enhance the overall results (Hovy & Prabhumoye, 2021).

Lastly, if the difference is too large and does not reflect upon multiple correct answers or the natural ambiguity of a task, countermeasures could include more thorough training of the annotators and a better calibration of the annotation process.

Noteworthy, our extensive work on implicit motives was achieved by costly and time-consuming training phases, reaching up to 20 hours of professional training per motive class (Schultheiss & Brunstein, 2010, p. 140).

Bias from input representations

First and foremost, the first countermeasure against biases from input, representations is the acknowledgment of their existence. Depending on the researched domain and specific research problem, the severity of mismatch between the intended language representation or ground truth and the embeddings can be measured. Debiasing word embeddings mostly focus

on specific, well-known societal biases such as normatively unfavorable gender differences (Prost et al. (2019) cited by Hovy & Prabhumoye (2021)).

Bias from data

Skewed and non-representative data can most effectively be debiased by revising representative and methodologically sound data collection. Exemplary, the American Psychological Association (APA)[16] provide well-defined guidelines for the information to be included in sample tables. Information includes gender, marital status, educational levels, or employment status. Even though that information can be critical and might fall under data protection laws, they help the scientific community and during the peer-review process to identify possible imbalances and representation issues of the underlying datasets.

Additional to the revision of the data collection methodology, existing data can be asserted in terms of its typical ground truth and available demographical soundness. E.g. Mohammady & Culotta (2014) extrapolated existing Twitter corpus data with demographics of countries in order to revise incorrect demographic data information (Hovy & Prabhumoye, 2021).

Bias from models

Since the inherent biases and flaws of machine learning models and NLP systems are well known, the well-established principles of sound scientific work should be honored. One of those principles emerged from the teachings of Karl R. Popper in his 1959 book "The Logic of Scientific Discovery" (Popper, 2002), in which Popper calls for the premise of falsificationism. This falsificationism states that – opposed to many researchers seeking confirmation and acknowledgment – sound science oughts to try to disprove its findings. For findings to be easily falsifiable, one must formulate clear and binding statements.

Transferred to NLP systems, this falsificationism can be achieved by always separating representative held-out test sets, calculating multiple performance measures such as the accuracy, recall, precision, F_1 score – to name but a few. Information on the inter-annotator agreement should be provided with as many measures as possible, such as percentage agreements or the Cohen's Kappa. In terms of psychological data, it should be shown that the classical test theory (see Section 2.3) was respected. In short, the capabilities and proper functioning of models should be doubted at any time and the scientific community should be provided with the uttermost information for critically evaluating the research's soundness.

Bias from research design

As for the sparse resources especially for non-English datasets and NLP systems, Hovy & Prabhumoye (2021) suggest a reflective question, whether researchers would investigate a research problem, if there was little to no data or barely a resource for solving the problem available. If this question is confirmed, it might be worthwhile to redirect the research interest

---

[16]`https://apastyle.apa.org/style-grammar-guidelines/tables-figures/sample-tables`

towards the problem that has barely been researched yet and thus foster the availability of novel resources.

Apart from that, it should always be stated clearly, which language is being researched, even if it is English. This so-called *Bender Rule* has become an established practice, which increases the awareness of language resources and availability.[17]

The subsequent section describes and assesses an international ethical discussion on a shared task crafted during this dissertation project. It applies the ethical fundamentals and NLPsych ethics from this chapter and evaluates the empirical work presented in Part III.

3.3 Ethical Consideration of the empirical GermEval20 Task 1

As described in Section 1.2, psychology adds a further layer of complexity to the already complex research object of natural languages. The related work on this matter as presented in Section 2.5 shows, that unsupervised approaches thus far are insufficient for capturing cognitive processes transported by natural language and even those approaches, that do not require hand-labeled data, label their data by the utilization of inventories (Basile et al., 2021) (also described in Section 2.5). Labeled data for this interdisciplinary domain is sparse. Reasons for this data sparseness are vast and include data protection laws, high cost for hand-labeled data by expert annotators or a lack of interdisciplinary research between psychology and computer linguists, which could result in the collection and annotation of larger corpora (Johannßen & Biemann, 2018).

One of the contributions of this dissertation project is the provision of freely available and hand-labeled psychological textual data. For distributing the largest available data set on the Operant Motive Test (OMT, see Section 5.2), we conducted a shared task at the GermEval 2020 workshop, which will be described in this section and was published (Johannßen et al., 2020a), from which the following excerpts are taken.

The task can be impactful and potentially cause harm. However, it also offers the opportunity to discuss broader ethical effects of the interdisciplinary field of natural language processing for psychological textual data. The GermEval shared task 1 on cognitive and motivational style is described in more detail in Section 7.5.

3.3.1 The NORDAKADEMIE Aptitude Test

The NORDAKADEMIE aptitude test is being conducted since 2011 at this private university of applied sciences in Germany. The test contains multiple parts from both, skill testing procedures and pscyhological assessments and takes up to 3 hours. The test takes place online.

---

[17]`https://thegradient.pub/the-benderrule-on-naming-the-languages-we-study-and-why-it-matters/`

Figure 3.4 illustrates some of the parameters with corresponding values from one of the participants during the aptitude testing procedure. The resulting data is rather noisy and non-standardized.

Parameter	Value	Parameter	Value	Parameter	Value	Parameter	Value
ID	12321	MIX_affiliation	84	ViQ-s	1.28	Math1-module	-1.70
Sex	M	Mix_power	27	ViQ-n	0.74	Math2-module	-1.52
School_german	12.75	IQt_language	44	ViQ-t	1.9	Progr-module	-0.26
School_english	9.75	IQt_memorization	52	ViQ-f	-1.37		
School_math	8.50	IQt_logic	76	ViQ-j	0.93		
NAK_engl_test	0.87	IQt_calculus	96	ViQ-e	0.83		
MIX_archievement	75.00	IQt_technique	52	NAK_math_test	90		

Figure 3.4: Exemplary parameters and values from one data instance emerging from the NORDAKADE-MEI aptitude test. The aptitude tests consists of multiple skill and psychology testing procedure. Resutling values are rather noisy and non-standardized.

English, Math, and Memory Test

The English test of the NORDAKADEMIE is divided into two parts: English grammar questions and English text comprehension questions. In the first part, applicants read a text and then answer comprehension questions.

The mathematics test contains a total of ten questions that test basic knowledge from the advanced technical college entrance qualification. These include problems with the rule of three, percentage calculations and fractions. The level was deliberately restricted to simple university entrance qualification content. The main reason for this is to filter out potential dropouts rather than to identify the best candidates.

Even the knowledge test provided for the respective subject area of the is limited to simple basics and does not aim to reward particularly well versed prior knowledge.

Motive Index (MIX)

Implicit motives and the MIX are covered in more detail in Chapter 5. In the aptitude test of the NORDAKADEMIE, the MIX motives are obtained by confronting participants with blurred images of multiple persons in ambiguous situations where they interact socially. Participants are asked to answer open questions on e.g. who the presumed main character is and what he or she feels.

Visual Questionnaire (ViQ)

The Visual Questionnaire (ViQ) is covered in more detail in Subsection 5.3.1. The ViQ according to Scheffer et al. (2016) and Scheffer & Loerwald (2008) is a psychological indicator for

intrinsic desires that is primarily measured visually. Applicants for the NORDAKADEMIE participating in the aptitude test are presented with a choice between usually two shapes. Those shapes differ mostly in geometrics, texture, or symmetry. The participants decide which shape appears more appealing to them. According to Scheffer & Loerwald (2008, p. 54) different cognitive stimuli reflected by the personality are measurable by the use of the ViQ.

Technical IQ (IQt)

The IQt is determined by a classical intelligence test with a selection of mathematical-technical IQ examinations (for details on IQ tests, see Subsection 2.1.2). The areas examined are language, memory, logic, arithmetic and technology. The IQt values are recorded as percentages of the test persons, who scored lower than the examinee (Plate, 2016).

The language comprehension is tested among other things by word permutations. For example, the subject is asked which of the four following words corresponds with a bird (SELMA for Amsel, a blackbird). Other questions give a proverb, to which another proverb of the same meaning from a selection of four proverb must be chosen.

For the understanding of the technique, the subject is presented with, for example, a selection of steel beams of different shapes, from which they are to select the most stable one. Gear wheel constellations and the question as to which system could system could rotate correctly are part of the IQt technique value. Examples of this are shown in Figure 3.5.

Simple arithmetic problems are used to test the IQt arithmetic value, which are to be solved correctly in a short time without the aid of a calculator. The main challenge here is the time limit of just a few seconds per problem.

Logical understanding is tested, among other things, through the continuation of numerical series. The challenge here is to recognize schemata particularly quickly and to and to apply them to the series of numbers. The logical logical continuation of graphics and symbols is also tested in this way.

The memorization ability of the test persons is tested by a detailed address of a person being shown for a few seconds and then hidden, whereafter detailed questions are asked, e.g. about the house number or zipcode. or postal code are asked. Word fields are also briefly displayed in this way and then ask which letter did not appear.

Ethical Discussion

Even though parts of this test are questionable and are currently under discussion, no single part of this test leads to an application being rejected. Only when a significant amount of those test parts are well below a threshold, applicants may not enter the second stage of the application process, which is applying at a private company due to the integrated study program the college offers. Roughly 10 percent of all applicants get rejected based on their aptitude test results. Furthermore, every applicant has the option to decline the data to be utilized for research purposes and still can apply to study at the NORDAKADEMIE. All anonymized data instances emerged from college applicants that consented for the data to be utilized in

this type of research setting and have the opportunity to see any stored data or to have their personal data deleted at any given moment (e.g. sex, age, the field of study).

Any research performed on this aptitude test or the annually conducted assessment center (AC) at the NORDAKADEMIE is under the premise of researching methods of supporting personnel decision-makers, but never to create fully automated, stand-alone filters (Binckebanck, 2019). First of all, since models might always be flawed and could inherit biases, it would be highly unethical. Secondly, German law prohibits the use of any – technical or non-technical – decision or filter system, which can not be fully and transparently explained. Aptitude diagnostics in Germany are legally highly regulated.

The most debated part of the aptitude test is the IQ. Intelligence in psychology is understood as results measured by an intelligence test (and thus not the intelligence of individuals itself). Furthermore, intelligence is always a product of both, genes and the environment. Even though there are hints that the IQ does not measure intellectual ability but rather cognitive and motivational style (DeYoung, 2011), it is defined and broadly understood as such.

3.3.2 IQ Testing

Mainly companies in Europe employ IQ tests for selecting capable applicants. In the United Kingdom, roughly 69 percent of all companies utilize IQ. In Germany, the estimate is 13 percent (Nachtwei & Schermuly, 2009).

Since IQ tests only measure the performance in certain tasks that rather ask for skill in certain areas (logics, language, problem-solving) than cognitive performance, such intelligence tests should rather be called comprehension tests. Due to unequal environmental circumstances and measurements in non-representative groups, minorities can be discriminated by a bias (Rushton & Jensen, 2005). One result of research on the connection between implicit motives and intelligence testing could help to improve early development and guided support.

It is this bias, which leads to unequal opportunities, especially in countries where there is a rich diversity among the population. Intelligence testing has had a dark history. Eugenics during the great wars e.g. in the US by sterilizing citizens (Lombardo, 2010) or in Germany during the Third Reich are some of the most gruesome parts of history.

But even in modern days, the IQ is misused. Recently, IQ scores have been used in the US to determine which death row inmate shall be executed and which might be spared. Since IQ scores show a too large variance, the Supreme Court has ruled against this definite threshold of 70 (Roberts, 2014). However, Sanger (2015) has researched an even more present practice of 'racial adjustment', adjusting the IQ of minorities upwards to take countermeasures on the racial bias in IQ testing, resulting in death row inmates, which originally were below the 70 points threshold, to be executed.

There is an ethical necessity to carefully view, understand and research the way intelligence testing is conducted and how those scores are – if at all – correlated with what we

understand as 'intelligence', as they might be mere cognitive and motivational styles. Further valuable research can be conducted to investigate connections between other personality tests such as implicit motives with intelligence or comprehension tests. Racial biases are measurable, variances are large and many critics state that IQ scores reflect upon skill or cognitive and motivational style rather than real intelligence as it is broadly understood.

It is important to note, however, that this shared task is not about automating IQ predictions from text or to research the IQ, but to conduct basic research on the possibility to predict psychological traits by text with a focus on implicit motives.

A more detailed ethical evaluation of this task and a recommended read have been formulated by Johannßen et al. (2020b).

The Intelligence Quotient (IQ) is considered to be one of the most valid, stable, and reliable psychometrics in the whole scientific field of diagnostical psychology and has been established for more than 100 years (Benson, 2003). Validity studies in a professional context have shown that those cognitive abilities asserted by an IQ testing procedure function as predictors for professional success (Schmidt & Hunter, 1998; Ones et al., 2017; Kramer et al., 2009).

However, the utilization of IQ testing procedures is also controversial. The scientific consensus criticizes the IQ test for introducing racial or socioeconomic biases (Turkheimer et al., 2003).

3.3.3 Misinterpreted Main Title

The original title of a shared task proposed to the GermEval workshop at the joint conference SWISSTEXT & KONVENS 2020 with the title *Classification and Regression of Cognitive and Emotional Style from Text* employed textual data from the Operant Motive Test (OMT, see Section 5.2), paired with aptitude diagnostical psychometrics such as IQ scores, as described by Johannßen et al. (2020a).
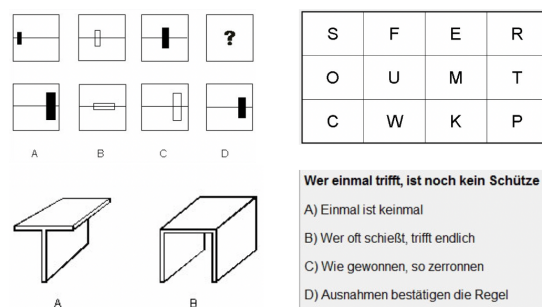
Figure 3.5: Different parts from an IQ test utilized at the Nordakademie. Upper left: logical comprehension, upper right: memory skills, lower left: technical comprehension, lower right: linguistic comprehension.

This section reflects upon and extends some of the aspects discussed in the published ethical consideration of said shared task (Johannßen et al., 2020b).

It can be debated, whether researchers should focus on theoretical tasks or if a very practical focus is legit. The NORDAKADEMIE is a university for applied sciences and the whole context of the GermEval20 Task 1 aims for researching implicit motives in the very application-oriented field of aptitude diagnostics.[18]

Even if there are very good and strong arguments against aptitude diagnostics, assessment centers, the consideration of socioeconomically biased high school grades or personal job interviews, it is a very common practice in Germany and Europe to examine all of those approaches for decisions on whom to employ.

Academic research has the responsibility to benefit society. Even though the organizers of the GermEval20 Task 1 do not focus on IQ testing but on the implications of implicit motives, since IQ testing is part of the conducted practice in Germany and Europe, there is an academic responsibility to research its implications. Furthermore, science nowadays is called upon making efforts towards findings that are closely related to everyday society, as Bornmann (2013) points out.

3.3.4 Dual Use

The first general principle to acknowledge is that knowledge is not harmless. There are many examples of theoretical research being utilized for destructive follow-up research or dangerous utensils directly. Exemplarily, Alfred Nobel did not invent dynamite to be used for war, but rather for mining. Historians assume that Nobel included a peace dedicated Nobel Price in his last will due to his invention being misused for war purposes.[19]

This is an example of a so-called dual use of inventions. When inventions intentioned for civil uses is misused without the consensus of the inventor for military purposes, this is called dual use. Williams-Jones et al. (2014) describe dual use more generally as being used for good and bad either intentionally or unintentionally by the inventors.

Furthermore, the authors describe the dilemma of this dual use, as there is rarely any impactful research that could not be considered dual use. Most meaningful findings could be utilized for the good and the bad. Moreover, at times it is not even possible to imagine the negative or bad dual use of one's inventions, as further research has not been conducted yet and novel products have yet not been seen (Williams-Jones et al., 2014).

One infamous example of dual use that was not necessary imaginable is nuclear energy and its characteristics, which has to lead to a lot of scientific progress (e.g. research on cancer treatments), civilian use (e.g. nuclear energy), but also great destructions and threats (e.g. nuclear weapons) (Tucker, 2012, p. 74 ff.).

---

[18]https://idw-online.de/de/news492748

[19]https://www.nobelprize.org/alfred-nobel/alfred-nobels-thoughts-about-war-and-peace/

3.3.5 Neutrality

During the long history of research in the field of IQ testing, many mistakes were made and investigated. Aptitude diagnosticians have spent decades challenging and correcting the strong socioeconomic biases, that were present in most of the earliest IQ tests. Nowadays, there are many different variants and approaches to IQ testing.

In Germany, there is little diversity among private college applicants. Even though researchers at the NORDAKADEMIE try to actively challenge those socioeconomic biases by employing implicit motives, that are known to be less biased than other metrics in the field of aptitude diagnostics, the employed IQ test also accounts for the little diversity of the participant population.

The NORDAKADEMIE utilized the IST 200 R intelligence structural test by Liepmann et al. (2007), which was normalized on high school graduates.[20] Since only about a third of students attend high school in Germany, the base population of this IQ test accounts for the little diversity of most applicants at the NORDAKADEMIE, which already experienced a socioeconomic filter. Even though this filter is discrimination already, the employed IQ test objectively accounts for the type of the basic population that takes the test and thus challenges this bias.

3.3.6 Evaluation Objectivity

IQ scores are prone to pseudoscientific settings and are not easily distinguishable from serious and sophisticated settings, thus masking the overall utility of IQ testing.

Hansson, a Swedish philosopher, first differentiates science from pseudoscience in that scientists enjoy common *raison d'être* to provide the reader with the most epistemically warranted statements (Hansson, 2013, p. 62 ff.) by employing known and broadly respected methods for finding those statements.

Furthermore, Hansson describes the correspondence between different scientific fields and disciplines that are interconnected. No given statement violates statements made by other disciplines and fields.

As for pseudoscience, authors are mostly divided as to which characteristics define pseudoscience. However, two major characteristics appear to be agreed upon by most authors: i) Non-science posing as science and ii) doctrinal components (Hansson, 2017).

For pseudoscience to be posed as science paramount effort is undertaken to mask statements as being made with those scientific principles, even if they are not. As science offers advantages of describing true phenomena and reality, pseudoscientists strive for acceptance

---

[20]In Germany, the secondary school tier consists of three types of schools: The Hauptschule (practice-oriented vocational education), the Realschule (theory-oriented vocational education) and the Gymnasium (high school, preparations for pursuing a college education). Only about 30% of graduates go to college in Germany (Fernandez-Kelly, 2015)

by readers with statements, that normally would not hold the thorough process of scientific work.

For pseudoscience to be of deviant doctrine, the pseudoscientists put sustained effort to promote standpoints different from those that have scientific legitimacy. Thus, pseudoscientists disregard major principles of scientific work, like correspondence, consensus, and consistency, as well as transparent methodology, replicability or intersubjectivity (Sahakian & Sahakian, 1993).

As for the GermEval20 Task 1, one could assume that either non-scientific work is being presented as a scientific one or a doctrine, disregarding established methods from corresponding scientific fields, which are NLP and psychology. The main arguments for calling the shared task pseudoscience is most likely the view, that since IQ testing is viewed by many researchers as biased and unprecise, even asking for machine learning systems would be pseudoscientific. They view the methodology as not being reconcilable with established ones.

However, on the point of discussion ii), this criticism mistakenly assumes that the task is about building an automated system for ranking students or classifying IQ scores, whilst, in turn, it is only about researching the implicit motive theory. Furthermore,a s stated, IQ tests have had a century-long scientific history and are well-established.

3.3.7 Luhmann System Theory

The systems theory by Luhmann is a philosophical and sociological communication theory, that describes agents of an environment not as instances but in their relations to other agents. Communication, according to Luhmann, is the constructing principle of an environment and not just a mere tool. An agent is understood as an autonomous part of this environment, which offers its inner structure as a matter of communication to other agents (Görke & Scholl, 2006).

However, as the channel model of communication by Shannon (1948) describes, there is no communication between agents (sender and receiver) without being obscured and disturbed by noise.

One environment or system is science. Every scientific discipline can be described as an agent in this environment. Whenever there is incomplete knowledge of the inner state of an agent, any type of communication between those systems gets obscured by noise and thus assumptions of those inner states can range from approximations to mere guesses. In any case, the assumptions are flawed.

Applied to the GermEval20 Task 1 and the ethical dilemma of IQ testing at hand, it can be stated, that since the scientific field of applied NLP does not comprehend the inner state of the scientific field of psychology and aptitude diagnostics, assumptions of the implications, limits, and effects of IQ testing from any non-psychological researcher must be viewed with caution

Figure 3.6: The exemplary 2014 year of graduation from the NORDAKADEMIE illustrates the cultural homogenity, as the vast majority of graudates are white. In Germany, a strongly biased socioeconomical filter is already present at the high school level[21].

– especially, if no correspondence has been undergone, as truth is the interaction between correspondence, consensus and consistency (Sahakian & Sahakian, 1993).

3.3.8 Marketplace of Ideas

The concept of the *Marketplace of Ideas* was first described in 1859 by John Stuart Mills in his book *On Liberty* (Mill, 2011). Mill was a liberal philosopher. To him, freedom was the main endeavor of humanity. Accordingly, freedom is the absence of restrictions (Altman, 2011, p. 145). Mill demanded absolute freedom of opinion and sentiment on all subjects, practical or speculative, scientific, moral or theological (Mill, 2011, p. 15).

The metaphor of a marketplace stems from the liberal assumption, that ideas, statements, or thoughts should be presented to market participants. If almost perfect transparency, replicability, and reliability can be assured, the market participants will freely choose which (scientific) ideas they would accept. Mill assumes that only the truth will emerge from this marketplace of ideas. Furthermore, proponents of this idea demand a free exchange of ideas without the interference of governments or society, since novel approaches can not yet be determined or fully understood (Gordon, 1997).

The marketplace of ideas does not mean, that criticism may not occur. Even the opposite holds true. Ideas need to be presented as clearly as possible. Those, who are capable of conducting an informed and knowledgable discourse may intervene and participate. This way, a consensus can be formed.

Debates on in-domain forums or communication exchanges offer the opportunity to critically discuss possibly harmful research. Other mediums, such as Twitter, are less suited for academic debates, where uninformed arguments can quickly put strong social media pressure

---

[21]https://www.shz.de/lokales/elmshorner-nachrichten/lasst-die-huete-fliegen-id19354606.html

on scientific ideas and demand sanctions. Thus, the idea itself can not be discussed and debated upon, but rather a forceful stopping of the professional discourse can emerge.

Even though the impact of the ideas is hardly comparable, there has been substantial criticism of well-established ideas nowadays. When Charles Darwin proposed his Theory of Evolution in 1859 in his book *On the Origin of Species* (Darwin, 1859), he was heavily criticized not only by the broader public, but also by the scientific community. In accordance with the Marketplace of Ideas, Darwin presented all his evidence as clearly as possible but was not able to disprove himself (nor was he disproven).[22] After a scientific consensus had formed, the Theory of Evolution eventually replaced the antiquated idea of creationism or even Lamarckismus, which believed traits to be passed on to the offspring which are used more frequently.

In the aftermath of ethical consideration of the Shared Task (Johannßen et al., 2020b), an NLP + Society podium discussion took place at the SWISSTEXT and KONVENS 2020 shared conference[23], where indeed the scientific consensus was formed, that such shared tasks should be crafted carefully and be ethically audited, but that the GermEval20 Task 1 served a purpose greater than the expectable harm it might inflict.

3.3.9   Knowledge cannot be Restrained

As Gershon (1983) describes, multiple researchers announced to leave science, after having discovered the knowledge of isolating DNA fragments for the first time. They feared that this discovery would lead to political and social pressure. One of those scientists even formed a group, categorically pressuring any scientific work on this genetic field. Nonetheless, DNA sequencing has continued to be researched.

There are implications, that – at least basic – research discoveries can not be fully prevented or stopped, as the so-called *multiple discovery* or *simultaneous invention* principle calls for them to be made. This multiple discovery principle is the hypothesis, that most discoveries are made independently by multiple scientists at the same time, often internationally. The Nobel Price committee often recognizes this hypothesis by rewarding multiple scientists who, at that time, did not collaborate directly.

This hypothesis is thought to be observable, since discoveries, theories and scientific tools enable practicing scientists of a field to now make discoveries. As the circumstances are ideal in an internationally spread research community, simultaneous inventions are made possible. One example is radar technology, which was discovered by multiple countries independently and at the same time (Galati, 2015). Thus, many believe the suppression of scientific progress is not possible.

On the other side, Martin (1978) argues, that a development of science and technology emerging from that science independent from the thoughts and desires of single scientists

---

[22]https://en.wikipedia.org/wiki/Reactions_to_On_the_Origin_of_Species

[23]https://swisstext-and-konvens-2020.org/programme/

and pressure from society are historically incorrect. In his article, the author argues with selected examples, namely nuclear power, food additives, transport policy, genetic engineering, and automation – all of which are characterized as technologies, having emerged from basic research and having experienced pressure and concerns from the research community and society. What the author does not argue about, is the value of basic research itself. He states that the path of scientific and technological development is not usually predictable beforehand. Furthermore, Martin notes that concerns over scientific and technological development has almost always to do with *applications and implications* for the wider society.

At times, the research could have assumed what negative impact a discovery or invention could have on society, as Nobel, which invented the dynamite mainly for supporting mining, could have imagined the use for military purposes. Nonetheless, the individuals utilizing dynamite to build weaponry are rather to blame than Nobel himself, even if he greatly regretted, that his discovery was used for such.[24]

3.3.10 Utilitarianism

Whilst the US has spent 4,545.7 million dollars (Pece, 2020) in research and development (R&D) of computer sciences and mathematics, the US Department of Defense controlled an R&D budget of 52,973.3 million dollars, which is more than 40% of the total US R&D budget. Some of the most influential advancements in computer science has been researched *behind closed doors* for military purposes such as the RSA cryptosystem, which was already invented by the GCHQ four years before the later patented peer-reviewed method[25] or the predecessor of the internet, the ARPANET, which was developed by the U.S. Airforce in 1969 (O'Neill, 1995).

Some private companies possess comparably large R&D budgets as well: Alphabet, the parent company of Google corp. spent 26,018 billion dollars on R&D.[26] Even though the most recent scientific advancements were made open-sourced and have been peer-reviewed, such as the bidirectional encoder representations from transformers (BERT, (Devlin et al., 2019)) and Tensorflow 1.0.0 (Fujita et al., 2017, p. 564), earlier developments, such as the Google PageRank algorithm, were kept hardly reproducible, despite patents describing the basic procedure (Lindberg, 2008).

One causality and risk of violations of the marketplace of ideas is that researchers, which experience pressure, might leave the public academia to pursue research in the private sector, which does not necessarily publish research to be reviewed, discussed, and criticized by the public. This could lead to knowledge monopolies, as well as fraudulent or misconducted research.

---

[24]https://www.nobelprize.org/alfred-nobel/alfred-nobels-thoughts-about-war-and-peace/

[25]https://www.wired.com/1999/04/crypto/

[26]https://abc.xyz/investor/

This is further reflected by the recent development, that influential technology companies have caused a so-called AI brain drain, meaning, that many countries experience the emigration of AI researchers. A national brain drain is observable from the public research sector and academia to private firms due to higher salaries, greater funding, and at times more academic freedom (Kunze, 2019).

After having described and explored the ethical aspects of this dissertation project, the following part lays the fundamentals of three empirically researched and automated psychometrics, namely Jungian psychology types, implicit motives, and personality questionnaires. The subsequent chapter lays the fundamentals of the Jungian types, which are one of the empirically automated metrics described in Part III.

Part II

Personality Assessment in the Application Field of NLPsych

CHAPTER 4
Psychology Types

In this chapter, the origin and the fundamentals of the Jungian psychology types or archetypes are laid out. Thereafter the theory itself and utilization of the theory – the Meyers-Briggs Type Indicator (MBTI) – are described. Lastly, the functioning and the introspective questionnaires are presented.

The landscape of available, researched, and practiced psychometrics emerged from more than a century of discoveries, observations, and developments (Segal & Coolidge, 2001; Archer et al., 2006). During this century of (mostly empirical) psychological research, trends, and paradigms had shifted and thus the currently most popular procedures, tests, and metrics had changed as well. Those paradigm changes can be major. For example, psychoanalysis, developed by Sigmund Freud in the early 20th century, was once viewed as a breakthrough and enabled many more psychological principles to be discovered. Nowadays, however, the teachings of Freud are viewed as inaccurate, obsolete, and outdated by many scholars (the reality lies within and some teachings of Freud are still being researched and validated, whilst others have indeed been substituted by other, more reliable and more valid procedures).

When analyzing currently utilized procedures, even the most novel ones employ theories and principles, which had been discovered a century ago. As an example, Deci & Ryan (2000) developed the *self determination theory*, which employs intrinsic motivation closely connected to the implicit motive theory (see Chapter 5) and has to be determined by traversing five hierarchical levels, of which the second level moderates those subsequent from it, which is a characteristic closely connected to the Jungian Psychology Types (Alharthi et al., 2017). Moreover, this self-determination theory is still being researched and utilized for NLP experiments on Tweets (Stajner et al., 2021).

4.1 Carl Gustav Jung Psychology Types

It is a rare instance that single individuals shape a whole scientific discipline within a field mostly by themselves. Carl Gustav Jung was a student of Sigmund Freud and shaped the by Freud developed paradigm of the unconscious, which in and of itself was neither found by Freud nor Jung (this is attributed to Friedrich Wilhelm Joseph Schelling (Ffytche, 2011, p. 75 ff.)), but experienced their major advancements towards broad acceptance by those two scholars.

4.1.1 Early Years with Sigmund Freud

Sigmund Freud, an Austrian neurologist, was fascinated by severe mental conditions that the established and young field of psychology could not resolve or reasonably treat. At the end of the 19th century, Freud chose a novel approach and employed a broad listening technique (Collin et al., 2012). Participants were asked to describe early memories and experiences. Influenced by the preliminary work of Schelling, Freud constructed the schematic model of the psychological apparatus as consisting of a conscious, an unconscious, and a preconscious part of the mind. The main idea of the unconscious came from various preliminary works, including hypnosis of female *hysteric* (a negatively occupied term nowadays) patients by the scholars Charcot, Breuer, Meynert, Bernheim , and others (Ffytche, 2011, p, 212). Freud's and Joseph Breuer's findings were first published in his 1905 book *Studies on Hysteria* (Freud & Mentzos, 1993), and termed the principals and techniques of psychoanalysis. To Freud, unconsciousness is the state of repressed or forgotten contents (Jung, 1991, p. 3).
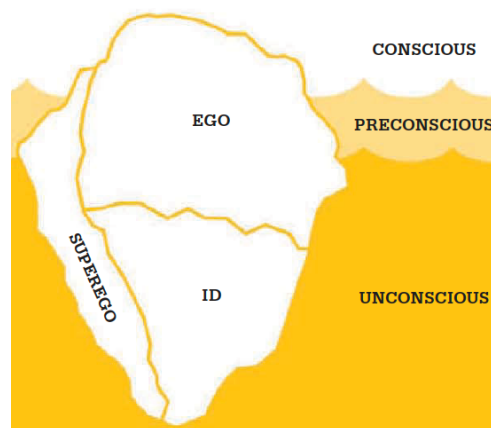
Figure 4.1: Together with Joseph Brot, Sigmund Freud developed a theory to the cognitive psychological apparatus, consisting of a conscious, an unconscious, and a pre-conscious part. An often utilized metaphor is this displayed ice berg, where the largest (i.e. most influential) parts are below the water surface (Collin et al., 2012).

Breuer and Freud successfully cured patients with hysteria, which before was deemed to be unreadable. They claimed that many mental illnesses such as hysteria, anxiety, or compulsions stemmed from traumatic early years experiences, which were oftentimes displaced from the conscious to the unconscious parts of the mind (Ffytche, 2011, p. 245).

4.1.2 Similarities & Difference between Jung and Freud

Carl Gustav Jung was an Austrian psychiatrist. Jung had adopted the teachings and principles of psychoanalysis early on. During his scientific employment in a hospital, Freud reached out to the practitioner. Together, Freud and Jung elaborated upon and researched the principles of the unconscious mind. During Jungs years of practice, Jung's superior psychiatrist Eugen Bleuler introduced Jung to the teachings of Freud and fostered Jung's own research on unconsciousness. During the early 20th century, Jung and Freud first corresponded indirectly but developed an intense professional relationship, disputing over unconsciousness and psychoanalysis (Shamdasani, 2010, p. 47 ff.).
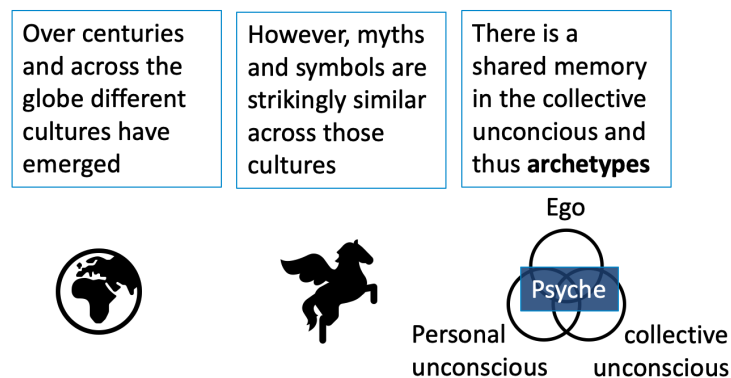
Figure 4.2: Freud and Jung realized that despite the development of diverse cultures there are similar mythologies and symbols across the globe. Both scholars believed unconsciousness to be the source of those similarities. However, whilst Freud viewed unconsciousness as an individual mechanism without external interference, Jung rather believed in a collectively shared unconsciousness (Shamdasani, 2010, p. 78 ff.).

Both, Freud and Jung believed in the unconscious larger part of the psyche, which shapes our thoughts and actions, despite not being accessible by our conscious mind. Freud and Jung both realized that even though diverse and rich cultures emerged among the globe, similar mythologies and symbols emerged, despite the correspondence of those cultures or a shared origin (Collin et al., 2012; Shamdasani, 2010, p. 104, p. 37). To Jung, the unconsciousness thus is not simply the state of repressed or forgotten contents, but furthermore, it has contents and modes of behavior that are – more or loess – the same everywhere and for all individuals (Jung, 1991, p. 4). Both believed that the reason for this shared set of symbols was similar to the idea

of collectively shared shapes by Plaot (Shamdasani, 2010, p. 177) originated from the uncon-sciousness. However, whilst Freud believed the unconsciousness to be unique and individual for each person, Jung rather believed in the existence of a collectively shared unconsciousness across the globe and across all cultures (Shamdasani, 2010, p. 78 ff.). Jung's views, as displayed in Figure 4.2, lead to major disagreements between Freud and Jung in 1912, resulting in the disengagement of both scholars (Shamdasani, 2010, p. 50).

4.1.3 Archetype Theory

During the disengagement with Freud, Jung worked on the first major book about the col-lectively shared unconsciousness and the role of symbolism. His 1912 book *Psychology of the Unconscious: a study of the transformations and symbolisms of the libido* (Jung, 2010) marked the beginning of Jung's own psychological school of thought, which disregarded Freud's ma-jor emphasis of sexual drive or *libido* as the main force behind early childhood traumas and towards the idea of symbols in a collective unconsciousness (Shamdasani, 2010, p. 72).
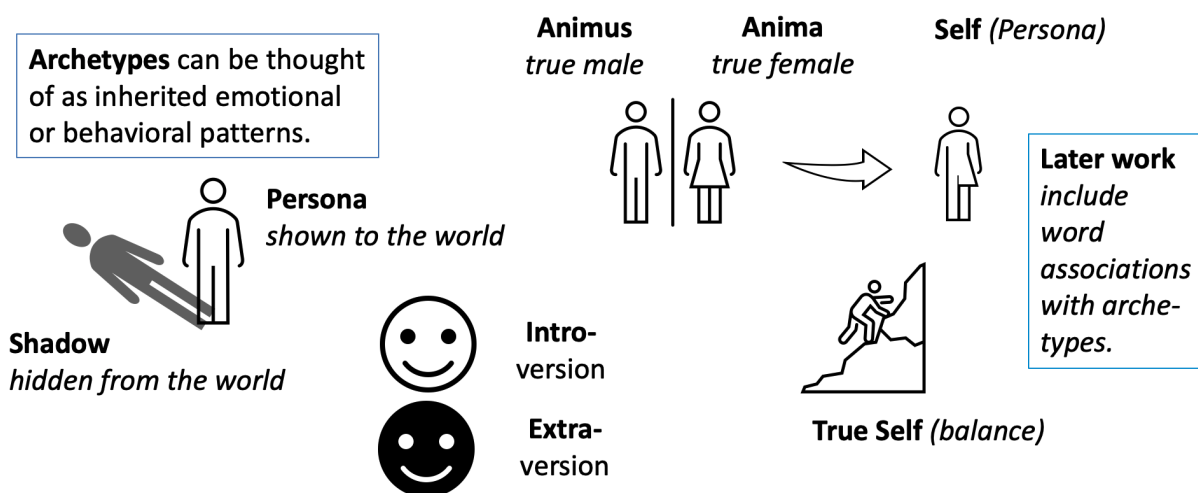
Figure 4.3: Psychological archetypes can be thought of as inherited emotional or behavioral patterns of which there are many (Jung, 1921).

According to Jung, the collective unconsciousness and the similarity between symbols of different mythologies and across the globe means, that humanity shares a hereditary memory, which Jung named *collective memory*. When Jung investigated similar symbols, religions, and cultures, he saw the same type of imagery, myth, or symbol and condensed those cross-cultural similarities to his *Psychological Archetypes*. His major findings were published in the 1934 book *The Archetypes and the Collective Unconscious* (Jung, 1991). Jung adopted the Freudian psychological apparatus (illustrated in Figure 4.1) but interpreted it differently than Freud: to Jung, the psyche is composed of i) the ego, which represents our conscious mind

and thought-out actions, ii) the personal unconscious, which holds the individual memories and iii) the collective unconsciousness of the collectively shared memories across cultures and over human existence (Ffytche, 2011, p. 225). Only the collective unconsciousness contains the Archetypes.

As figure 4.3 illustrates, there are some Archetypes, amongst many, that play a major role, even though Archetypes are mostly non-hierarchical and equivalent in forming the collective unconsciousness. Archetypes are understood as inherited emotional or behavioral patterns. The Anima or Animus, oftentimes referred to as *soul* by Jung when describing both sexual dimensions, is characterized as the inner personality and the attitude towards behaving in relation to one's inner psychic process, which is an inner attitude or inward face. The Anima represents the female inner attitude, whilst the Animus is the same male part. The outward face and *self*, the aspects which we decide to share with the world, are called the Persona (Jung, 1921, p. 428). On the opposite side of this Persona, spectrum lays the Shadow, which we deeply – and thus unconsciously – do not want the world to experience or see (Collin et al., 2012, p. 106).
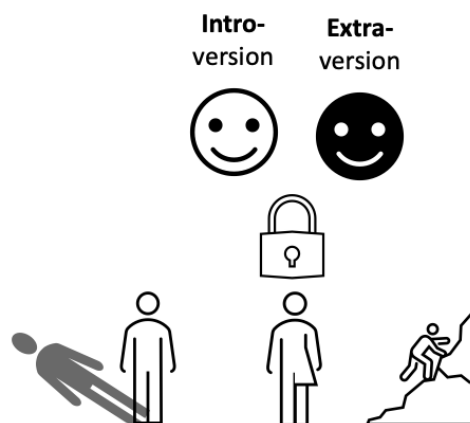
Figure 4.4: Most Archetypes are not hierarchical, but can be thought of as equally influential with the exception of Introversion and Extraverversion, which moderate and thus channel the other Archetypes (Jung, 1921).

The two opposing dimensions of Extraversion and Introversion occupy a niche in the Archetype theory. The Archetypes Extraversion and Introversion determine, how the need and desire for gratification are fulfilled. For extraverts, the feeling of gratification for one's desires stems from the outside, whilst for introverts, this gratification comes from the inside (Collin et al., 2012, p. 107). To Jung, everyone possesses both types, but only the predominance of one over the other determines the manifestation of the type. Extraversion and Introversion moderate all other Archetypes (Jung, 2010, p. 4 ff.), as displayed in Figure 4.4.

The way someone expresses either of the many Archetypes is predominantly determined by the manifestation of the Archetype Extraversion and Introversion, and thus this type can

be considered the most influential besides the omnipotent True Self (Jung, 2010, p. 204). The True Self symbolizes the perfect balance between all of the archetypes and an equilibrium. Since perfect balance is unachievable and thus all individuals suffer from an inner conflict between the extremes of each Archetype. With this, Jung argues closely with the theories of Arthur Schopenhauer and his philosophy of omnipresent desire and suffering (Jung, 2010, p. 88). Jung furthermore connected many mental illnesses, which he had experienced during the correspondence and time together with Freud as an unbalance of the collectively shared unconsciousness and the individual unconsciousness (Jung, 2010, p. 454). Jung was also pioneer for the theorys which subsequently became the PSI theory (Kuhl et al., 2014).

Jung laid the fundamentals of the archtetype theory. Even though many scholars agreed with these ideas and observations, Jung never commercialized it or developed a validated methodology for measuring archetypes. Such an approcha is described in the subsequent section. It is a questionnaire approach for measuring Jungian types.

4.2 Myers-Briggs Type Indicator (MBTI)

Since the release of the book in 1934, the Jungian Psychological Archetypes have laid the foundation for many subsequent metrics, tests, and procedures (e.g. the self-determination theory (Deci & Ryan, 2000; Stajner et al., 2021; Alharthi et al., 2017)). In 1942, the psychologically uneducated Cook Briggs and her daughter Isabel Myers proposed the Myers-Briggs Type Indicator (MBTI, (Briggs Myers et al., 1998)). The MBTI was developed for commercial reasons and resulted in the founding of the Myers-Briggs Company, which distributes the licensed text to most countries. Since the MBTI was never properly validated and shows a severe lack of reliability, it is met with substantial professional criticism from the psychological community (Pittenger, 2005). Despite the criticism, the MBTI is one of the most broadly utilized personality testing procedures with 2 million conducted MBTI questionnaires annually (Quenk, 2009, p. 2).

4.2.1 Development of the MBTI

Mayers and Briggs did not have credentials in the study of Jungian psychological types or psychological diagnostics) first came into contact with said types when *Psychological Types* (Jung, 2010) was translated into English in 1923. Since Jung did not provide any easily adaptable diagnostical procedure for measuring the psychological types despite psychoanalysis, the two women began their own empirical research with the aim to develop an easily applicable and statistically measurable testing procedure (Quenk, 2009, p. 2). Since both Myers and Briggs were interested in the practical utilization rather than scholarly research.

Extraversion/Introversion (EI)
Direct or receive energy based on the inner or outer world

Sensing/Intuition (SN)
Information intake by sense or by interpretations and meaning

Thinking/Feeling (TF)
Decisions by logic or by emotions

Judging/Perceiving (JP)
Organizing the world or staying open for new information
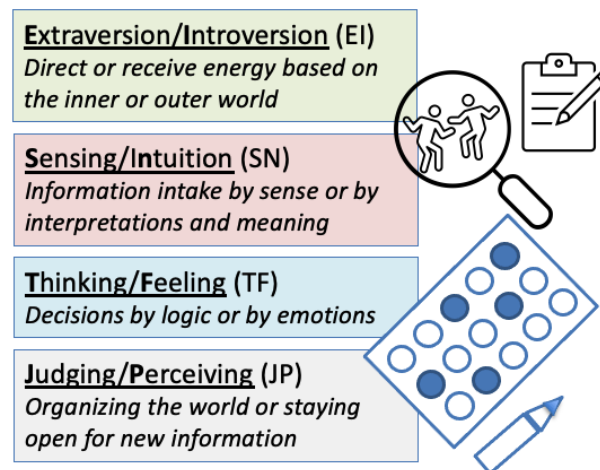
Figure 4.5: Psychological Archetypes can be thought of as inherited emotional or behavioral patterns. They stem from the shared unconsciousness. Many such types exists, but the Animus, Anima, Self, Persona, Shadow, Introversion, Extraversion and the True Self are amongst the most influential Archetypes (Jung, 1921).

Myers and Briggs mainly read and studied the elaborated writings of Jung and identified attitude types of Introversion and Extraversion as impactful Archetypes. Jung also described two opposing perceiving functions Intuition versus Sensing and two opposing functions of judgment, namely Thinking versus Feeling. Observations led Briggs to believe that individuals differ in how they habitually relate to the outside world, which Briggs translated to the two opposing dimensions of Judging versus Perceiving (Quenk, 2009, p. 2).

The MBTI testing procedure was designed explicitly with a commercial interest to support World War II efforts. The developed questionnaire was tested on a few selected individuals, of whom Briggs believed to know their preferences (Quenk, 2009, p. 3). Multiple forms with hundreds of questionnaire items were developed subsequentially.

4.2.2 Functioning

The MBTI utilizes questionnaires to measure and determine opposing dichotomy as momentary preferences rather than permanent personality traits. The testing procedure assigned manifestations of binary dimensions between four so-called functions. Figure 4.5 illustrates the four opposing types.

Table 4.1 displays an excerpt from the MBTI manual with six different questionnaires (i.e. forms) that include from 93 to 290 items (i.e. questions). Two of those forms are self-scorable. The other forms have to be evaluated by trained experts certified by the MBTI company (Quenk, 2009, p. 27).

Form	Number of Items	Approximate Administration Time	Scoring Method for Four-Letter Type	Primary Uses
M	93	15-25 minutes	Templates or computer	In settings where MBTI can be given in advance
M self-scorable	93	15-25 minutes	Hand-scored by respondent or administrator	Where time and subject availability are issues
G	126	30-40 minutes	templates or computer	Preceded Fromt M as standard form
G self-scorable	94	20-30 minutes	Hand-scored by respondent or administrator	Preceded Form M self-scorable
K	131	30-40 minutes	Form G templates or computer	Where EIR subscales will be reviewed later
J	290	75 minutes	Form F templates or computer	Where TDI subscales will be reviewed later

Table 4.1: The MBTI manual displays six forms with a range of 93 to 290 questionnaire items (i.e. questions), two of which are self-scorable (Briggs Myers et al., 1998, p. 107).

After having presented the Jungian theories on archetypes and one approach for measuring them – the MBTI – the following section provides a critical assessment of this diagnostical approach.

4.3 Critical Assessment

A concluding psychological quality assessment of the Jungian psychological types or archetypes is not possible, only diagnostical procedures that utilize the theories of C.G. Jung may be evaluated in terms of their psychological diagnostical quality (Jung, 2010). Broadly and frequently conducted diagnostical approaches do exist, which utilize the archetype theory, e.g. the *Big Five* (see Chapter 6). The connection to those explicit questionnaires is not as pronounced as the connection to psychological types established by implicit or projective approaches (see Chapter 5) (Schultheiss & Brunstein, 2010).

One broadly utilized explicit personality questionnaire on the basis of Jungian types is the MBTI. Even though the test has been monetized from the time it has been first published, nowadays many closely connected free-of-charge tests are available. One of such tests is the *Keirsey Temperament Sorter*. The validity (see Chapter 2) is questionable and its validity evaluation methodologically flawed. Instead of measuring the validity as coherence with other, more established and more clinical psychological procedures or e.g. assessment center observations, the Keirsey Temperament Sorter is validated on its capabilities of capturing the MBTI scales – which in turn have been criticized for its poor validity (Kelly & Jugovic, 2001). This flawed methodology in validating any procedure on the basis of the MBTI by correlating the results to the MBTI and wrongfully claiming this to be valid appears to be systematic. Another procedure, the so-called *Jung Psychology Types Scale* (JPTS) evince the same poor validation methodology (Sato, 2017).

In other words, the validity of most explicit Jungian questionnaire approaches depends on the validity of the oftentimes compared MBTI. However, the validity of the MBTI has mostly been disproven. Whilst some researchers attested the measure's validity (Thompson & Borrello, 1986), others heavily criticize the diagnostic procedure in terms of lacking agreement with known facts and data is poorly testable, and that it shows internal contradictions (Stein & Swan, 2019). Broad surveys find profound methodological flaws in most papers in favor of the metric's validity, such as inadequate sample sizes, inappropriate correlation matrixes (e.g. Pearson), reliability issues in terms of depending on adolescent participants, and being atheoretical statistically driven (Jackson et al., 1996).

This chapter has described and discussed the Jungian psychology types in their theories, their measurement, and their critical evaluation. The same fundamental structure is provided by the subsequent chapter on implicit motives, which are one of the empirically automated metrics described in Part III.

CHAPTER 5

Implicit Projective Procedures

This chapter contains the fundamentals of implicit projective procedures. It consists of first the origins in Section 5.1, therafter the OMT in Section 5.2, multiple hybrid forms, which combine implicit procedures for measuring Jungian types in Section 5.3, and finally provides a position and critical assessment in Section 5.4.

Implicit motives (also called operant motives) are unconscious intrinsic desires, measurable by implicit methods. Those implicit methods, in turn, require participants of this psychometrical testing procedure to use projection and introspection on ambiguous imagery, in order to answer provided emotional and narrational questions (Gawronski & De Houwer, 2014).

During the course of the dissertation project, extensive research on the classification of implicit motives has been performed, resulting in many of the in Part III presented empirical studies.

5.1 Origins

Since implicit projective procedures differ substantially from well-known questionnaires (e.g. the Big Five), standardized tests (e.g. SAPs or IQ), or even open conversations about childhood memories (e.g. psychoanalysis), this section first describes the basic functionality that all implicit methods share (e.g. the TAT, the OMT, or the IPT), and thereafter outlines the origins of this diagnostical paradigm.

Unlike Sigmund Freud, Carl Gustav Jung viewed the unconscious symbolism not as an individual memory of the personal past, but rather as collectively shared memories and thus believed in a collective unconsciousness (see Section 4.1) (Collin et al., 2012; Shamdasani, 2010, p. 104, p. 37).

Not only did Freud and Jung find a new psychological school of thought with the unconsciousness being the most important aspect of it, they furthermore believed the unconsciousness – to Freud the state of suppressed or forgotten contents, but to Jung furthermore with

contents of modes and behavior that are the same everywhere and in all individuals (Jung, 1991, p. 3 f.) – to be the largest part of our psychological apparatus (see Figure 4.1) and determines most of our thoughts and behaviors.

Eventually the theories of shared unconsciousness and Psychological Archetypes (Jung, 1991) became the more accepted theory and Archetypes in general laid the foundations for a central European paradigm shift in psychology towards empiricism (Collin et al., 2012, p. 4, p. 109).

5.1.1 Functionality of Projective Tests

Even though those Jungian Types and the theories related to them were popular, widespread, and accepted amongst diagnostic psychologists, neither Freud, nor Jung could provide validated and reliable measurement procedures for this unconsciousness theory. Measuring the unconsciousness has time and again been proven to be difficult, which already Freud and Jung suspected (Ffytche, 2011, p. 273).

The underlying theory of the Freudian psychic apparatus states, that there are three agents, which interact with each other and the external world – the Id, Ego, and Superego (Ffytche, 2011, p. 255). To Jung, rather than those three agents, the distinction has rather be made between the personal unconscious, the collective unconsciousness, and the ego (Ffytche, 2011, p. 225). Importantly, the ego – a controlling instance that moderates how individuals behave and interact with the external world – only possess comparably small amounts of conscious content. The greatest influential part lies within the unconsciousness and can not be directly accessed. Standardized testing procedures and questionnaires oftentimes rely on accessible content or memories and lack intrinsic unbiased introspection. This introspective shortcoming of so-called explicit psychometrics (i.e. rely on ego access of psychic content) suffers from socio-expectation biases (Schüler et al., 2015).

Implicit tests avoid this socio-expectation bias by circumventing the (later coined super-) ego part of the psychic apparatus and directly relying implicitly on the unconsciousness (Kuhl (2000) cited by Schultheiss & Brunstein (2010, p. 474)).The implicit tests achieve this less bias-prone diagnostic by presenting ambiguous imagery, which participants are asked to interpret. By interpreting those images, individuals project their own emotions and latent desires onto the canvas (Schultheiss & Brunstein, 2010, p. 152). The verbalized reactions and answers thereafter can be analyzed for underlying mental structures and personality traits (Schüler et al., 2015).

5.1.2 Rorschach Test

An early version of projective testing procedures was the well-known but methodologically flawed Rorschach inkblot test. The Rorschach test was developed by a Swiss psychiatrist

and first published in 1921 in his book *Psycho-diagnosis. Method and results of a perception-diagnostic experiment* (Rorschach, 1932). During the testing procedure, participants are shown 10 inkblots printed on separate cards half of which are in color and the others in black and white (see Figure 5.1). The participants are asked to freely associate those inkblots and state what thoughts, feelings, or real-world objects they believe to see. The inkblots are mirrored at the center axis and those show many symmetries.

Figure 5.1: One exemplary card from the Rorschach inkblot test (Lazarevic & Orlic, 2015, p. 91).

The testing procedure lasts about 45 minutes on average and can take up to 2 hours for scoring and interpreting the results. More than 100 characteristics can be scored with three major categories i) content, ii) location, and iii) determinants (e.g. color, movement, or shading) (Lilienfeld et al., 2000, p. 3).

Despite major criticism in terms of a lack of norms, reliability, and validity, the Rorschach test still ranges among the most popular diagnostical personality tests. Reliability coefficients measured during test-retest assertions per criteria ranged from .3 to .9, whilst many scores of the most exhausted system called the Rorschach Comprehensive System (CS) by (Exner, 1974) were never properly tested for test-retest reliability (Lazarevic & Orlic, 2015, p. 91). To Lilienfeld et al. (2000) due to a lack of test-retest assessment the reliability problem of the Rorschach still remains an open one.

5.1.3 Thematic Apperception Test (TAT)

The Thematic Apperception Test (TAT) is another projective test, which took the basic principle described in Subsection 5.1.1 and extended the Rorschach by presenting participants with real-world sceneries and oftentimes with illustrated persons. One example of such an image is displayed in Figure 5.2. Participants were asked to think of a story that might suit the displayed scene. The TAT was first introduced by Morgan & Murray (1935) in their 1935 article *A method for investigating fantasies: the thematic apperception test*.

Annotators (also called raters, scorers, or coders) were poorly trained and mostly relied on their own theoretical perspectives and methods of gaining insight (Schultheiss & Brunstein, 2010, p. 7). The TAT is a construction technique, which means that – different to the Rorschach – participants were asked to construct stories and become active themselves instead of just associating the first thing that comes to mind. Thus, the respondents or participants of the TAT actively interpret their own stimuli instead of having them interpreted by the test conducting expert (Lilienfeld et al., 2000, p. 39). The TAT is performed with 20 out of 31 imagery cards that depict ambiguous situations, whereas most of them are social in nature (e.g. persons are displayed that interact with each other). Even though no gender differences were measured, many different apperception tests for certain demographic groups were developed e.g. for children, for working-class men etc.

Figure 5.2: One exemplary card from the Thematic Apperception Test (TAT) (Carlton & Macdonald, 2003).

When shown an imagery card, the participants are asked to construct a short story of the scene that should describe i) the things that happened before the scene and led to it, ii) events that can be seen on the card, iii) events that will occur after, and iv) what the persons and characters on the card are thinking and feeling with the assumption, that participants would pick on character, which he or she identifies with or views as the main protagonist of the story (Murray (1943) cited by Lilienfeld et al. (2000, p. 39)).

The following section describes one of the modern implicit motive testing procedures, which adopts the fundamentals laid by first the TAT and subsequentially the Picture Story Excercise (PSE), and extends these fundamental works.

5.2 Operant Motive Test (OMT)

The Operant Motive Test (OMT) is one of many predecessors of the TAT and was first introduced by Kuhl & Scheffer (1999). The test – same as most other predecessors of the PSE by McClelland et al. (1989) – avoided some of the main flaws of the TAT. Instead of having participants write lengthy and rather unstructured stories, they now were given four questions, which they answered with mostly few sentences. Furthermore, the annotators (also called raters, scorers, or coders) were trained via a detailled and precise manual.

5.2.1 Testing Procedure

For the OMT testing procedure, participants are confronted with drawn imagery, displaying multiple persons in ambiguous situations. Six examples are displayed in Figure 5.3. Participants are asked to answer four questions: i) what is important for the person in this situation and what is he or she doing? ii) what does the person feel? iii) why does the person feel that way? iv) how does the story end?
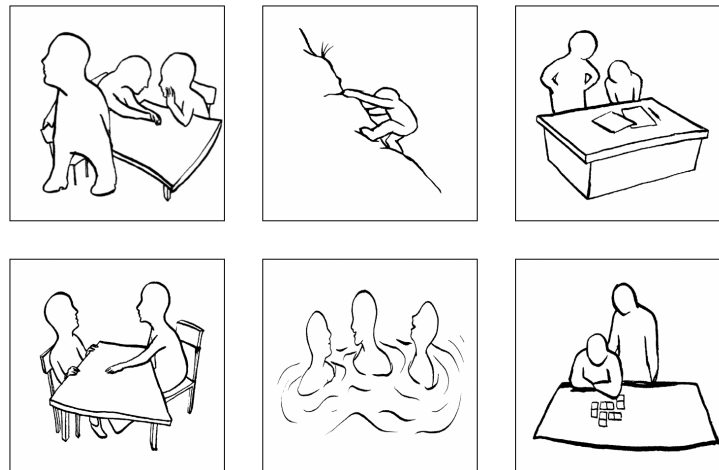
Figure 5.3: Some examples of images to be interpreted by participants utilized for the operant motive test (OMT) (Kuhl & Scheffer, 1999).

The very first observed motive by the expert annotators labels the whole instance, which is the so-called primacy rule (Kuhl & Scheffer, 1999). This primacy rule honors the aspect that the very first impulse and association with an image satisfy the need or desire of the participant. All subsequent identifiable motives would have been less desirable to the participants and thus should not be used as the label. Some exemplary answers with their annotated labels are displayed in Listing 5.1.

```
A sie nimmt am Gespräch nicht teil und
  wendet sich ab. gelangweilt. es
  interessiert sie nicht, worüber die
  anderen beiden reden. schlecht.
M weicht ängstlich zuruück. unterlegen.
  wird zurechtgewiesen. d2
  Gelegenheit den Fehler zu korrigieren
  ------- Translation -----
A she does not take part in the con-
  versation and turns away. bored.
  She does not care what the other
  two are talking about. Bad.
M withdraws anxiously. Inferior.
  is rebuked. Opportunity to
  correct the mistake.
```

Listing 5.1: German text examples of OMT answers with A being Affiliation and M being the power motive. The texts correspond to the first picture of Figure 5.3. Translations into English provided by the authors.

The OMT is a comparably expensive psychometrical procedure. Annotators have to train for 20 hours on average, to become reliable at scoring one of the three motives (Schultheiss & Brunstein, 2010, p. 76) and should undergo at least 12 hours of scoring practice before annotating any real-world data. Annotating the answers given during the PSE of 100 participants takes roughly 20 to 50 hours (Schultheiss & Brunstein, 2010, p. 140 ff.). The intraclass inter-annotator agreement of .85 is achieved by training those experts with a well-defined manual (Impart & Kuhl, 2013).

When inspecting the annotated labels and interviewing annotators on their reasons for choosing one label over the other, at times those annotators refer to their specific manual-based training, and at times they describe it as being intuitive (Johannßen et al., 2019). This intuition exceeds the training and manual but could not be properly translated into reliable rules by the annotators.

According to Schultheiss & Pang (2007), there are three main motives of the operant system: i) affiliation (also referred to as A), which is a desire for establishing positive relationships, ii) achievement (also referred to as L), described as the capacity of mastering challenges and gaining satisfaction from such and iii) power (also referred to as M), which is the desire to have an impact on one's fellows. A so-called zero-motive or zero-level (annotated as 0) are labeled if no clear motive or level can be identified. In addition to the well-established so-called *Big Three* implicit motives, a fourth *freedom* is assumed and currently researched. This freedom motive has barely been researched yet but has a close connection to the power motive. Whilst power-motivated individuals desire control over their fellow humans and their direct

surrounding for the sake of control, freedom motivated individuals seek to express themselves and want to avoid any restraining factors (Baum & Baumann, 2019).

Even though natural languages differ from formal languages, e.g. programming languages, in that they emerged organically and contain many ambiguities (see Section 1.2), those natural languages can still be formalized. NLP allows for the construction of statistical, machine- or deep learning models, which capture linguistic rules or patterns, and can even solve tasks, which require complex reasoning skills (Brown et al., 2020). Interpreting and explaining the decision-making process of deep neural models is still a vastly unsolved research problem. Nonetheless, trained natural language models offer the opportunity of being formally analyzed.

Besides the OMT, other forms of the same testing procedure exists, such as the MIX, employed at the NORDAKADEMIE potential test (see Section 7.5). One example of such testing imagery is displayed in Figure 5.4.

Figure 5.4: Example from the MIX imagery, employed at during the NORDAKADEME aptitude testing procedure (Scheffer & Kuhl, 2013).

5.2.2 Empirical Research

Due to questionable metrics in the past such as the Rorschach test or the MBTI, psychometrics ought to be tested on their test-retest reliability, their validity, and – in the case of personality traits – their stability. In previously published work, empirical validation research was discussed (Johannßen et al., 2019, 2020a; Johannßen & Biemann, 2019).

Modern implicit motive tests, be it the PSE, the OMT, or the IPT (see Subsection 5.3.2) allow for the assertion and prediction of clinically (i.e. under laboratory conditions) measurable non-verbal interpersonal communication such as smiling, laughing or eye contact (McAdams et al., 1984). Lang et al. (2012) performed a broad study with $N \sim 241$ participants on the

connection between motives and job performance. The researchers differentiated between task performance for certain and measurable work-related tasks, and contextual performance, which is behavior that contributes positively to the organization's social and psychological climate. Lang et al. (2012) were able to find significant correlations between the participant's motives and their job performances – both task-related (via the achievement motive) and contextual performance (via the affiliation motive).

In practice, implicit testing procedures not only measure motives, but furthermore other constructs such as Jungian types, or self-regularization. These hybrid metrics are described in the following section and are utilized for the empirical research in Part III.

5.3 Hybrid Forms of Implicit Measures

In addition to the implicit projective procedures, where participants are presented with ambiguous imagery and are asked to answer associating questions on the scenery being displayed, there are additional diagnostic tests, which inherit the implicit paradigm, but surpass natural language reactions or combine them with further testing parts. One of those is the so-called Visual Questionnaire (ViQ) where participants decide on which shapes and colors they prefer and the implicit personality test (IPT), which combines implicit motives with explicit questionnaires.

5.3.1 The Visual Questionnaire (ViQ)

Figure 5.5: An example item from the Visual Questionnaire (ViQ), which aims for the participants to decide in accordance to their desires for either structure (left, clear circle) or creative chaos (left, sketched circle) (Sarges & Scheffer, 2008, p. 52 ff.).

The ViQ according to Scheffer & Loerwald (2008) and Scheffer et al. (2016) is an indicator that is primarily measured visually. Subjects of the potential aptitude test of the FH Nordakademie are confronted with the choice between two motives. They decide thereby only, which form appeals to them more. The following can be derived from this according to Scheffer & Loerwald (2008, cf. p. 54) different dimensions of the personality. An example of the test procedure can be found in Figure 5.5. In the following, the dimensions relevant to this dimensions relevant for this work are briefly explained.

ViQ-s (Sensing): This dimension leads to a quantitative, fragmented perception. Simplicity is preferred to a high level of detail. Stimuli of surprise or complexity is avoided. Especially when abstraction is needed, this dimension can bring advantages.

ViQ-n (intuition): The intuition dimension leads to a fast absorption of complex stimuli. Automatic body processes and partially unconscious processes are bypassed. It comes to precise interpretations with minimal information. Thus, this dimension provides for an enormously fast adaptation of highly complex situations and environments.

ViQ-t (Thinking): Fast acquisition of systematic or logical order is determined by the ViQ-t dimension. This means above all the question of truth and untruth as possible answers. Forward-thinking and planning is strived for.

ViQ-f (Feeling): Holistic and emotional ways of looking at things are expressed in the ViQ-f dimension. A strong interest in the social environment and the social social structure is the result. Particularly large amounts of information and impressions can be processed particularly efficiently with a strong ViQ-f dimension. often even in parallel.

ViQ-e (Extraversion): A high score on the ViQ-e dimension is primarily primarily an indicator that a respondent seeks sensations. In addition, there is above all tolerance, but also ambiguity. Decisions are often made on the basis of what external effects they would have on the respondent.

ViQ-j (Judging): Unambiguous decisions, certainty, norm-orientation and avoidance of ambiguity are the focus of people with strong ViQ-j dimensions. Before especially unclear design elements irritate such people.

5.3.2 The WafM Implicit Personality Test (IPT)

It is difficult to measure the psyche or personality directly (Fried & Flake, 2018). The research field of psychology has developed and researched different approaches for measuring manifestations of the underlying mental processes, all of which have advantages and shortcomings. E.g. psychoanalysis tries to assume cognitive mechanisms and past events in dialogues, whilst behaviorism strictly limits statements on empirical and reproducible observations (Mahoney, 1984). Both approaches require controlled environments, extensive manual labor, and time. Testing procedures try to determine personality traits with limited time and budget and thus oftentimes balance reliability (i.e. are results reproducible?), validity (i.e. do results correspond to other observations and measures?), and limited testing resources (Schultheiss & Brunstein, 2010, p. 76f).

Some personality testing procedures utilize questionnaires with high reliability. However, standardized surveys and direct questionnaires at times suffer from socio-expectation bias, i.e. participants rather worry about, what testing personnel might think about them, when answering a question in a certain way, rather than answering freely. This bias can occur if the intentions of questions can be guessed or are assumed (Bogner & Landrock, 2016).
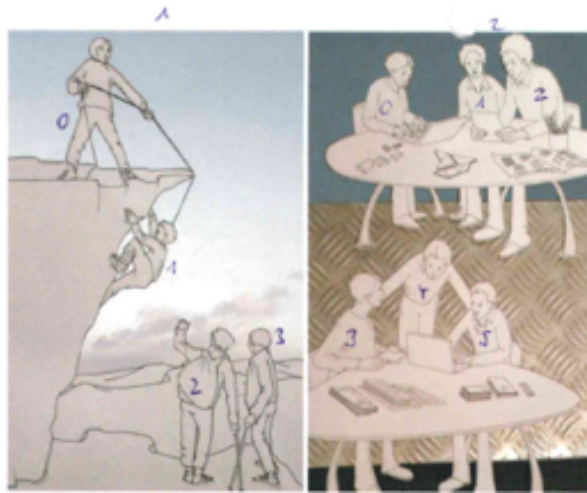
Figure 5.6: During the implicit personality test, participants are presented with projective imagery, to which they answer questions such as who the main person might be and what that person is experiencing. Such projective or implicit tests are designed to reveal intrinsic desires.

Implicit or projective testing procedures overcome this shortcoming by providing participants with ambiguous and situational imagery and asking them to answer questions e.g. who the main character is and what that individual experiences and feels. Those projective methods reveal intrinsic desires. Since there is no socially accepted or wrong answer, the socio-expectation bias is said to be less severe. However, projective methods have been criticized for their reliability (Schultheiss & Brunstein, 2010, p. 119ff).

The IPT is such an implicit test and confronts participants with imagery such as displayed in Figure 5.6. Participants chose the main person and answer questions about what is happening and how that person feels. Some of those answers, manually labeled with either i (introvert) or e (extravert) are displayed in Listing 5.2. The human annotators are psychologists and receive extensive training, which initially is wordlist centered but shifts to narrations over time[27]. The IPT is based on the MBTI and has mainly been utilized for business-oriented aptitude diagnostics.

---

[27]For a closely related testing procedure, please refer to Kuhl & Scheffer (Kuhl & Scheffer, 1999)

```
I          sie  sieht  ihre  schüler.  das  die
           schüler  nachhause  gehen.  genervt
E          Erklährt  jemanden  etwas.  Es
           richtig  zu  machen.  Er  kann  es
 -------- Translation -----
I          she  sees  her  students.  That  the
           students  go  home.  Annoyed
E          Explains  something  to  someone.
           To  do  it  right.  He  can  do  it
```
Listing 5.2: Short examples of answers given during the IPT and corresponding manual labels

5.3.3   Self-regulatory Levels

Children develop Counter-regulation skills during the reciprocal interactions with their parents on self-expressions such as emotions or desires. Those counter-regulation skills are developed not on the basis of positivity or negativity, but rather on the basis of the magnitude of emotional responses. Due to the immediate reactions of parents to self-expressions, children learn to regulate the magnitude of their reactions (Keller (1998) cited by Scheffer & Kuhl (2013))

These self-regulatory levels describe the type of self-regulation when acting out an intrinsic desire, often described by other psychological metrics such as implicit motives or Jungian psychology types. Two steps are described: i) the entrance gradient and ii) the exit gradient. For i), the arousal of positive or negative affect is described, whilst for ii), it is the coping of the affect, e.g. negative or positive coping. The first step i) determines how often and how easily children and persons react to stimuli, whilst the second step ii) determines the magnitude and the own ability to either amplify or dampen those reactions upon stimuli.

For both, metrics and levels, usually, a zero is assigned, if no clear motive or level can be identified. The first level is the ability to self-regulate positive affect, the second level is the sensitivity to positive incentives, the third level is the ability to self-regulate negative affect, the fourth level is the sensitivity to negative incentives and the fifth level is the passive coping of fears (Scheffer & Kuhl, 2013). Table 5.1 summarizes the self-regulatory level reactions.

Finally, after laying the fundamentals, describing the OMT, and presenting hybrid forms, the following section provides a positioning and critical assessment.

5.4   Positioning & Critical Assessment

Scholars and many researchers disregard most projective approaches as not valid (Lilienfeld et al., 2000). However, almost all criticism is directed at the earliest approaches to this im-

Level	Reaction
Level Zero	No self-regulatory level could be identified
Level One	Ability to self-regulate positive affect
Level Two	Sensitivity to positive incentives
Level Three	Ability to self-regulate negative affect
Level Four	Sensitivity to negative incentives
Level Five	Passive coping of fears

Table 5.1: Self-regulatory levels are developed during early childhood as a reaction and conditioning of children, experiencing their parent's reactions to their self-expression. The first step of self-regularization describes the sensitivity towards stimuli, and the second step describes the amplification or dampening of positive or negative reactions (Scheffer & Kuhl, 2013).

plicit diagnostical methodology, which, indeed was methodologically flawed. More recent procedures and tests overcame many of the issues in terms of validity, reliability, and stability. Nonetheless, amongst scholars and researchers, a broad sentiment has been established, that the projective paradigm is invalid altogether. As an example, in the well-established text-book by Schmidt-Atzert et al. (2018), the author discusses the poor validity of the Rorschach test, the TAT, and of a projective test where participants are asked to imagine their family as animals (Brem-Gräser, 1957), all of which were developed in the middle of the 20th century, and concludes, that since those procedures and tests did not show sufficient validation, no implicit projective procedure fulfills the necessary validation criteria (Schmidt-Atzert et al., 2018, p. 308).

However, more recent surveys also include the succeeding projective tests, which display human figure drawings. Even critical assessments conclude, that those human figure drawing projective tests – one of which is the OMT described in Section 5.2 – do achieve substantial higher validation results than previous approaches (Lilienfeld et al., 2000).

Furthermore, one commonly argued criticism of projective diagnostical tests is the lack of correlation with well-established diagnostical procedures. Researchers have compared results from projective tests with other procedures such as an IQ test and criticized the lack of correlation (Schmidt-Atzert et al., 2018, p. 306). Furthermore, it has been criticized that implicit motives do not correlate with other types of motives with target classes power or achievement. According to Schultheiss & Brunstein (2010) this lack of correlation stems from the misconception that explicit motives and implicit motives measure the same construct. However, whilst explicit motives assess conscious attitudes, implicit motives measure intrinsic and unconscious desires. Both measures, despite being addressed with similar names, do not measure the same underlying psychological phenomenons.

The subsequent chapter describes personality questionnaires, which is of importance for the assessment of metric correlations discussed in Part IV.

Personality Questionnaires

This chapter is concerned with the fundamentals of measuring personality in Section 6.1, personality systems as a whole and the methodology of utilizing questionnaires in Section 6.2, whereafter the most broadly utilized personality questionnaire or inventory is described in Section 6.3. Finally, this chapter concludes with a positioning and critical assessment of personality questionnaires in Section 6.4.

In psychology, the term *psycholexical approach* or *lexical hypothesis* refers to the theory that personality traits can be encoded into natural language lexicons and that there are terms for the same personality traits in every natural language. Furthermore, the hypothesis states, that terms describing socially important personal characteristics are more densely distributed, and thus much more frequent in natural language, which in turn leads to basic assumptions leading to the creation of personality questionnaires (Wood, 2015).

6.1 Measuring Personality

The current psychological diagnostic research assumes that there are five main personality categories or traits connected to the lexical hypothesis: openness, conscientiousness, extraversion, agreeableness, and neuroticism (often abbreviated as OCEAN). Those dimensions paired with some attributes are displayed in Figure 6.1. Those five personality categories are called the five-factor model of personality (or also called *Big Five*, (Goldberg, 1993)) and have been (Angleitner, 1991) amongst the most widely utilized personality tests, which rely on the Jungian psychological typologies (see Chapter 4). Those five dimensions can be further decomposed into more detailed facets, e.g. extraversion and warmheartedness, sociability, or enforceability – utilized by the personality test NEO-PI-R. The Big Five has become one of the most broadly utilized and popular personality questionnaire procedures due to multiple reasons: i) the creators of the procedure made the material (questionnaire, manual) freely available, ii) the test has been researched in terms of validity and reliability (different than e.g. the Mayers-Briggs

Type-Indikator, MBTI), iii) the procedure often shows strong correlations with other types of diagnostic procedures and many well-established dimensions of other procedures have proven to describe the same underlying construct, strengthening the validity of the Big Five (Schmidt-Atzert et al., 2018, 239).

Personality traits are mainly measured by either self-assessment (also called self-report), or by behavioral assessment observations. During those observations, participants are asked to perform standardized tasks or solve puzzles, whilst being closely observed by trained psychological experts. In comparison to other personality assessment procedures such as implicit tests (see Chapter 5), questionnaire procedures have many advantages, such as standardization, scalability, and objectivity.

Personality questionnaires are conducted by providing participants with prior defined questions or settings in textual writing. Instructors or written instructions determine the way those questions ought to be answered or settings to be reacted upon. Instructions include the speed of the reactions (e.g. immediately vs. undetermined), the type of answer (e.g. multiple choice, single choice, free answers), or the possibility of alternative answers. Most questionnaires only allow for dichotomous answers (e.g. yes/no), rating scales (e.g. school grades), or forced-choice answers (i.e. a separate answer for each question item). Those restricted answer types are preferable over open answers due to their statistical properties and standardized evaluation methods (Schmidt-Atzert et al., 2018, 240). As described in Subsection 2.1.4, evaluation objectivity is one of the most crucial quality criteria in the field of psychological diagnostics.

After this general description of how to measure personality, the following section briefly discusses an important limitation of questionnaire approaches: the observability and assessability of personality systems.

6.2   Personality Systems and Questionnaires

Personality questionnaires can only assess the personality of expectable and prior defined participants. Especially the most broadly utilized procedures such as the NEO-PI-R rely on participants with the capabilities of objectively self-assessing previous observations. Participants with cognitive, psychiatric, or developmental impairments usually do not get addressed by personality questionnaires. Furthermore, those procedures usually expect a minimal threshold of intelligence as measured by a standardized and representative IQ test (see Subsection 2.1.2 for details on IQ testing). This rather debatable requirement of a minimal IQ score aims to ensure participants are able to fully comprehend the questions and for the participants to be able to self-reflect and self-assess (Schmidt-Atzert et al., 2018, 242).

Even though self-assessed personality questionnaires show good validity and reliability, they suffer from multiple effects, lowering their diagnostic capabilities, such as salience, recency, telescoping, self-deception, effort after meaning, and a social desirability bias.

The most broadly utilized and most influential questionnaire inventory is the Big Five, which is described in the following section. It is furthermore utilized for assessments in Part IV. The MBTI is another broadly utilized but less valid questionnaire, described in Chapter 4.

6.3 Five Factor Personality Test (Big Five) / OCEAN Model

The Big Five model is the by far most broadly utilized model and has become a de-facto standard when it comes to personality testing and questionnaire procedures. In the field of personality diagnostics, any performed research is expected to also include some consideration of a Big Five inventory or the relation between the presented work and the questionnaire. This section describes the development of the Big Five, briefly describes its functioning, and discusses the questionnaire – namely the NEO-PI R.

6.3.1 Development of the Big Five

The Big Five model was developed from lexical observations following the lexical hypothesis. It solved a crisis in personality psychology of not having an established paradigm or shared model across all types of research projects (Goldberg, 1993). Especially the model's capabilities of predicting observable and expectable outcomes in a coherent and – for empirical research – well replicable way led to the broad utilization of the model (Roberts et al., 2007).

For the development of the Big Five model, natural languages were the main source of the material. To utilize natural languages was a rather novel and unconventional idea. Initially, roughly 18,000 personality-related terms were extracted from the well-established Webster's unabridged dictionary by Allport & Odbert (1936). Independent researchers thereafter performed factor analysis on those terms and aggregated a minimal set of relevant dimensions to which each term could be assigned to (Digman, 1990). The first four dimensions were i) neutral terms designating possible personality traits, ii) temporary moods or activity, iii) evaluations, and iv) a 'miscellaneous' category (Raad & Mlacic, 2015). A first initial model was researched and distributed by Tupes & Christal (1992), but could not yet establish a new paradigm. Further research and validation studies by Costa & Mccrae (1992) led to the first well-established version of the Big Five model, the NEO-PI R, and was later popularized by Goldberg (1993), which also added a fifth label, shifting labels from culture to intellect (Raad & Mlacic, 2015). The NEO-PI R was furthermore the first step from the Big Five model toward a questionnaire diagnostic procedure.

6.3.2 Functioning

Openness to Experience
People are open to experience (also called culture or intellect) are often described as being
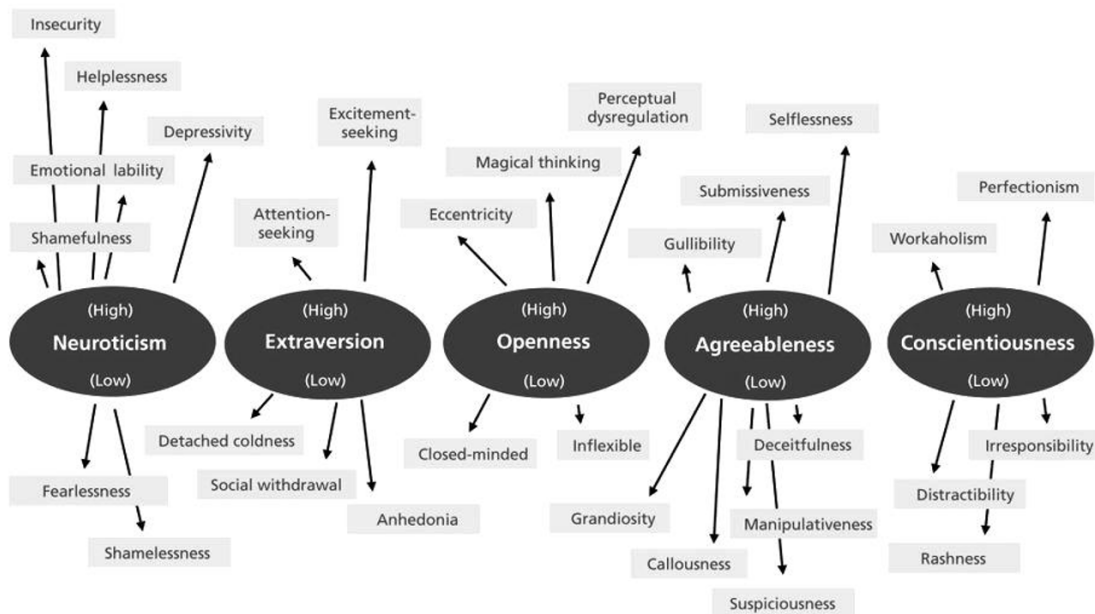
Figure 6.1: An overview of the Big Five dimensions openness, conscientiousness, extraversion, agreeableness, and neuroticism paired with associated behavioral patterns for both, low and high dimension values (Aghdaei & Tabrizi, 2021).

artistic, curious, or imaginative. They are said to have a rich emotional life. This openness to experience describes their endorsement of new things. Besides those positive attributes, people open to experience can appear unfocused or are vulnerable to drug and alcohol abuse. People closed to experience (i.e. low in openness to experience) have fewer interests, are more conventional and experience emotions less intensely (Novikova, 2013; Raad & Mlacic, 2015).

Conscientiousness

conscientiousness is closely connected to self-discipline. People high in this dimension are orderly, responsible, and dutiful. As displayed in Figure 4.1, the superego regulates desires. People high in conscientiousness are able to regulate their impulses and desires and thus possess a strong superego (Roberts et al., 2009). People with low scores in conscientiousness are rather spontaneous, and can be careless and disorganized. Furthermore, they tend to lack clearly planned life goals (Novikova, 2013; Raad & Mlacic, 2015).

Extraversion

People identified as high in extraversion (also called surgency) are assumed to be talkative, assertive, and energetic. They appear cheerful and high in positive affect. Extraverts prefer company and enjoy being around familiar and oftentimes yet unknown people. People low

in extraversion are also referred to as introverts. Introverted people prefer being alone (but do not enjoy the feeling of loneliness), usually have fewer – but at times closer – friends and are rather reserved and serious (Novikova, 2013; Raad & Mlacic, 2015). It is important to note, that the Big Five questionnaire is an explicit procedure, which directly addresses and asks participants instead of utilizing projection. Thus the extraversion of the Big Five does not correlate or relate to the extraversion identified by the IPT (Schultheiss & Brunstein, 2010) – see also Chapter 5 for further details).

Agreeableness

Agreeableness is assumed to be connected to altruistic behavior. People high in agreeableness are cooperative, forgiving, and generous. They furthermore believe in the good in people and their good intentions. Low agreeableness scores are also called disagreeable. People with attributed disagreeableness tend to be skeptical, competitive, and avoid cooperation (Novikova, 2013; Raad & Mlacic, 2015). Oftentimes, sociopaths, which might act as if they were high in agreeableness, show strong disagreeableness and high amounts of neuroticism (Ross et al., 2004).

Neuroticism

People, which were identified as being high in neuroticism are emotionally sensitive. They can quickly be aroused and easily upset, whilst frequently experiencing negative emotions. Oftentimes, neurotic people worry, feel anxious, and are self-conscious. They can be easily stressed and impulsive. On the other hand, people low in neuroticism are emotionally stable (thus this trait is often referred to as emotional stability). They operate calmly and calculated even in high-stress environments and situations. Overall, emotionally stable people experience few negative emotions (Novikova, 2013; Raad & Mlacic, 2015). Due to high requirements of stress resilience in the banking sector, studies have revealed elevated job satisfaction among bank employees with lower scores in neuroticism, i.e. higher emotional stability (Hlatywayo et al., 2013).

6.3.3   NEO-PI R Questionnaire

One of the most broadly and well-researched Big Five questionnaires is the (NEO) Personality Inventory (NEO PI-R), developed by Costa & Mccrae (1992) and revised (*NEO*) by Ostendorf & Angleitner (2004). Whilst the NEO-PI R tests on the five main dimensions, it subdivided them into more detailed facets, resulting in 30 sub-scales. Thus every main dimension is divided into 6 sub-categories with 8 items each, resulting in 30 facets, five main categories in 240 question items in total. An excerpt from the procedure by Costa & Mccrae (1992) is displayed in Figure 6.2.

The items are Likert-scaled, which means that there are five possible answer intensities: i) strong disagree, ii) disagree, iii) neutral, iv) agree, v) strong agree. All items (i.e.) have

to be answered by marking or crossing the most probable reaction by the participants. The procedure of the NEO-PI R takes 30 to 40 minutes and thus is rather fast compared to other personality diagnostical procedures.

The supervised and supervisor-conducted testing procedure of the NEO-PI R is almost similar to the self-reported version. This results in better comparability to other empirical studies and a standardized evaluation of the procedure. Nowadays, the procedure is often conducted at a computer. The results thus can be calculated automatically and exported to statistical evaluation software.

Big five Personality Traits

S#	Items	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
1	I see myself as someone who is talkative	-1-	-2-	-3-	-4-	-5-
2	I see myself as someone who tends to find fault with others (R)	-1-	-2-	-3-	-4-	-5-
3	I see myself as someone who does a thorough job	-1-	-2-	-3-	-4-	-5-
4	I see myself as someone who is depressed	-1-	-2-	-3-	-4-	-5-
5	I see myself as someone who is original, comes up with new ideas	-1-	-2-	-3-	-4-	-5-
6	I see myself as someone who is reserved. (R)	-1-	-2-	-3-	-4-	-5-
7	I see myself as someone who is helpful and unselfish with others	-1-	-2-	-3-	-4-	-5-
8	I see myself as someone who can be somewhat careless (R)	-1-	-2-	-3-	-4-	-5-
9	I see myself as someone who is relaxed, handles stress well (R)	-1-	-2-	-3-	-4-	-5-

Figure 6.2: Excerpt of the NEO-PI R testing procedure question items developed by (Costa & Mccrae, 1992).

The testing procedure has to be standardized depending on the gender and age of participants. The raw values usually have to be documented. Provided T-, stanine- and percentage values provide the capability of standardizing the values to any given norm. For interpreting the results and scores, participants are provided with evaluation manuals. Those manuals have to describe the resulting dimensions and facets as understandable as possible since the NEO-PI R oftentimes is conducted as self-report (Schmidt-Atzert et al., 2018, p. 264).

After having discussed the fundamentals of measuring personality via questionnaires, personality systems, and the most broadly utilized inventory, the following section provides a critical assessment.

6.4   Positioning & Critical Assessment

The NEO-PI R shows a high internal consistency and reliability with a Cronbach's Alpha of $\alpha = .71$. The difference between men and women is marginal. The retest reliability reaches a split-half correlation between $\rho = .82$ and $\rho = .91$ for short periods of times (1 to 2 months) and on a median of $\rho = .75$ for longer periods of five years. As for the validity, the factorial structure of the NEO-PI R scales matches sufficiently well across all 30 scales. However, self-conducted questionnaires and expert observers reported questionnaires to correlate weakly with $\rho = .54$ (Schmidt-Atzert et al., 2018, p. 265).

The NEO-PI R is said to be one of the best types of personality questionnaires, which allows for measuring the five most prominent and well-accepted personality dimensions (OCEAN) in a reliable way. The expert observer questionnaires have been standardized on very large amounts of data, which – in combination with self-reports – allow for a more reliable diagnostic. Even though the scale and sub-scale reliabilities are low and can be criticized, the very broad utilization of the NEO-PI R and the exhaustive validity research, which almost always confirm the practicability and diagnostic capabilities of the NEO-PI R confirm this explicit personality questionnaire approach (Schmidt-Atzert et al., 2018, p. 265).

A key advantage of the NEO-PI R or other Big Five questionnaires in the field of NLP is the conformity of the OCEAN dimensions with natural language and with words. Conversations often relate to behavior and personality traits. Neologisms tend to capture mental processes and even crown psychology and national sentiment can be observed in lexica. Thus, the assertion of the OCEAN dimension and connection to language as the origins of the Big Five promises a deeper understanding of the mind (Wood, 2015).

Some psychological phenomena influence the diagnostic capabilities of explicit questionnaire approaches. One is the socio-desirability bias (see Subsection 2.1.3). This bias leads to participants aiming for leaving a good impression upon answering the NEO-PI R, which results in high correlations between scales, which are known to be independent, e.g. neuroticism and conscientiousness (Ziegler & Buehner (2009) cited by Schmidt-Atzert et al. (2018, p. 245)). Post et al. (2008) observed this effect when comparing statements made by pregnant women on their smoking habits with interviews eleven years after, where their answers differed greatly from the statements made during pregnancy.

A second destructive psychological phenomenon is the so-called *Barnum effect*. This effect states, that participants do not possess the capability of neutrally and objectively assessing own desires and values, which leads to subconscious reflections upon vagueness and self-interpretations and to the acceptance of over-simplified and over-generalized personality descriptions. This Barnum effect can be observed by superstitious individuals, which tend to accept very generalized horoscope descriptions or simple personality types, which could apply to almost anyone (Dickson & Kelly, 1985).

Those biases and effects can be countermeasures with two approaches: i) firstly, the instructions can ask participants to answer as honestly as possible and describe those effects to raise awareness, ii) forced-choice replies with similarly undesirable outcomes have to be taken (Schmidt-Atzert et al., 2018, p. 245)). However, both of those approaches can not guarantee the absence or successful countermeasure of the socio-desirability bias. Alternatives are implicit procedures, which do not suffer from this bias (see Chapter 5).

With the conclusion of this chapter on personality questionnaires, the next Part III presents empirical research conducted during the course, which utilizes many of the introduced psychological metrics. The following chapter describes research on automated aptitude diagnostics and NLPsych for subsequent academic success.

Part III

Empirical Research of NLPsych for Aptitude Diagnostics, Social Unrest, and Pandemical Isolation

CHAPTER 7

NLPsych for Aptitude Diagnostics

This chapter presents empirical research on aptitude diagnostics, more specifically on the selected approach of correlating the achievement motive with grades, as well as investigating the GermEval shared task briefly explained in Chapter 3. For the experiments, the in Chapter 5 presented implicit motives are modeled in Section 7.2. Related work to this niche application domain is presented in Section 7.2.1, followed by the description of the utilized data in Section 7.2.2, the methodology in Section 7.2.3, and finally the results in Section 7.3. A conclusion is drawn and an outlook is provided in Section 7.4. Limitations are discussed in Section 7.6. This chapter ends with the technical description and evaluation of the conducted shared task in Section 7.5.

The goal of our research is to classify psychometric textual data. Furthermore, we aim to investigate algorithmic decision-making and validate automatic annotation by predictions in accordance with the psychometric theory. To pursue this goal, we perform multi-label classification on the Operant Motive Test (OMT, Section 5.2) with four labels. During this OMT, participants textually answer questions on images to the provided questions (see Chapter 5 for details).

Recent advances in artificial neural network architectures have established mechanisms that allow researchers to, in a limited fashion, inspect reasons for algorithmic decisions. One of these mechanisms is called *attention* and was found by Young et al. (2018) to be among the most broadly investigated and adopted elements of deep neural machine learning. We want to investigate access to algorithmic decision-making by employing this attention mechanism (see Subsection 2.2.7 for details).

Lastly, the OMT theory states that some labeled motives allow for predictions of subsequent academic success, which we inspect by counting annotated labels and correlating these counts with participants' academic grades.

Even though there is a high demand for the automation of psychological textual data analysis (NLPsych), comparably little research has been performed on this interdisciplinary task (Johannßen & Biemann, 2018).

In this chapter, we research, whether the attention weights matter and reveal any insights into algorithmic decision-making and whether there are correlations between automatically predicted motives and subsequent academic success.

In the following section, we will present and discuss the approach to utilizing feature engineering (see Section 7.1) and a logistic model tree (LMT, see Section 2.2.3) for classifying the OMT. Since neural approaches oftentimes outperform other machine learning modles, this model is meant for comparison and as a benchmark. The neural approach and the subsequent empirical investigation of predictive capabilities on grades of both models will thereafter be covered from Section 7.2.2 onward.

7.1 Feature Engineering Classification of the LMT

This section describes the modeling of the OMT via an LMT model. Apparent limintations in terms of methodology and results are discussed.

Utilized Data

Data has been collected and hand-labeled by researchers from the IMPART company[28] by having 14,600 anonymized participants textually associate images in German such as the one in Figure 5.4 on the two questions i) who is the main person and what is important for that person? ii) how does that person feel? The participants gave 220,859 answers on 15 different images. After filtering (cf. the pre-processing part of this section), we retain 209,716 text instances.

Each answer was labeled manually with the motives 0, A, L or M and a level ranging from 0 to 5. The annotators were psychologists, trained by the OMT manual by Kuhl & Scheffer (1999). The inter-annotator agreement with previously coded motives using the Winter scale Winter (1994) reached as high as 97% and 95% for the two annotators after the manual training. The pairwise intraclass correlation coefficient is an often utilized agreement measure, developed by Shrout & Fleiss (1979). This coefficient was measured to be .85 on average for the three motives (Schüler et al., 2015), thus showing the difficulty to standardize the labeling process.

The class distributions of motives and levels displayed in Table 7.1 show that the power motive (M) is with 59% nearly three times as frequent as the second largest class of achievement (L) with 19%. Furthermore, levels 4 and 5 together represent more than half of all level-labeled instances. The distribution is displayed in Figure 7.1

---

[28]https://impart.de/?lang=en

Motive distribution
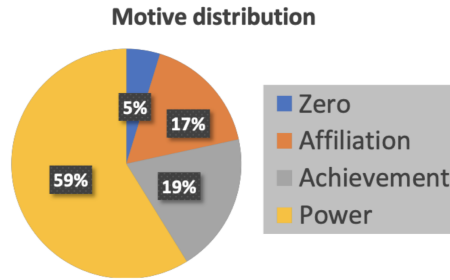
Figure 7.1: Graphical representation of the unevenly distributed motive labels amongst the data set.

In addition to the roughly 220,000 labeled OMT text data instances, a small dataset of related but unlabeled MIX texts from 105 participants is available, which come with the additional information of the bachelor thesis grades of the anonymized participants. We will use this dataset for the extrinsic evaluation below.

Pre-Processing

We pre-processed the data by first removing spam, which mostly contained the same letters repeated, empty answers or a random variation of symbols. Also, we removed entries in different languages other than German. Lastly, texts with encoding problems were either resolved or removed. After this pre-processing, the whole dataset consisted of 209,716 texts. The distribution of filtered questions is uneven.

	0	1	2	3	4	5	Σ
0	7,921	0	2	1	2	6	7,932
A	11	2,888	9,581	1,361	7,617	6,822	28,280
L	6	2,455	12,697	6,405	7,542	3,742	32,847
M	25	11,338	12,353	15,248	36,103	23,610	98,677
Σ	7,963	16,681	34,633	23,015	51,264	34,180	167,736

Table 7.1: The OMT's training classes distribution after filtering and removing a held-out test and development set (10% each).

Feature Engineering

For engineering features, the texts mostly were tokenized and processed per token. Engineered features were the type-token-ratio, the ratio of spelling mistakes and frequencies between 3 and 10 appearances.

Further features are LIWC (see Subsection 2.3.1) and language model perplexities (see Section 2.3.2). The German LIWC allowed for 96 categories to be assigned to each token, ranging from rather syntactic features such as personal pronouns to rather psychometric values such as familiarity, negativity or fear.

Part-of-speech (POS) tags were assigned to each token and thereafter counted and normalized to form a token ratio. We trained a POS tagger via the natural language toolkit (NLTK) on the TIGER corpus, assembled by Brants et al. (2004) and utilizing the Stuttgart-Tübingen-Tagset (STTS), containing 54 individual POS tags.

We trained a bigram language model (see Section 2.3.2) for each class and incorporated Good-Turing smoothing for calculating the perplexity. During training, we tuned parameters (e.g. which smoothing to use) via development set and tested the model with a held-out test set of 20,990 instances.

Model Training

Even though deep learning has shown to be powerful, it often comes with a cost of losing transparency, which is crucial for our task, in which we seek to better understand the connection between psychology and language. Therefore we utilized different classical machine learning algorithms such as Naïve Bayes, LMT or regression and found the logistic model tree (LMT) implementation of Landwehr et al. (2005) to be the best-performing one amongst the tested. An LMT is a decision tree, which performs logistic regressions at its leaves. The root differentiates the language model's perplexities (A, M, and L) and thereafter performs the logistic regressions based on further features.

A qualitative post-hoc analysis by psychologists has resulted in an agreement with the model's predictions, except for too many assigned 0 labels and motives.

Results

Based on the correlation-based *Feature Subset Selection* by Hall (2000), the most influential features are the LIWC categories *I, Anger, Communication, Friends, Down, Motion, Occup, Achieve* and *TV*, as well as the perplexities of the language models affiliation (A), performance (L) and power (M) and attributive possessive pronoun (PPOSAT) POS tag frequency.

		Predicted			
		0	**A**	**L**	**M**	Σ
Actual	**0**	338	92	163	427	1,020
	A	51	2,667	105	708	3,531
	L	115	66	3,151	804	4,136
	M	209	573	556	10,965	12,303
	Σ	713	3,398	3,975	12,904	20,990

Table 7.2: The confusion matrix of the motive classification task (without the levels) on the test set (10% of available data) with filtered values.

The confusion matrices in Table 7.2 illustrate the model's performance for each class. The model scores an F_1 score of 65.4% for classifying the levels and 80.1% for classifying the motives. The resulting LMT is displayed in Figure 2.5 (Section 2.2.3).

Following this feature engineering OMT modeling approach, we also propose a neural model for classifying the OMT in the subsequent section.

7.2 Neural Classificiation of the OMT

This section describes the modeling of the OMT via a bi-LSTM model with attention mechanism. Apparent limintations in terms of methodology and results are discussed.

7.2.1 Related Work

So far, approaches to psychological traits identification from texts often examined the connection between language and mental diseases. Current research mostly focuses on e.g. the detection of dementia (Masrani et al., 2017), crises Demasi et al. (2019), suicide risks (Matero et al., 2019), mental illnesses (Zomick et al., 2019) or anxiety (Shen & Rudzicz, 2017) by the use of some form of natural language processing.

Nonetheless, some findings focus on motivation, success or characteristics. Tomasello (2002) describes the psychology of language as the method of focusing on the way people express themselves rather than to focus on what meaning is conveyed.

So-called closed-class words are by far more informative than open-class words in terms of psychological language research. Closed-class words are words that tend to not change over centuries, which can be e.g. pronouns, prepositions or adverbs. Open-class words, on the other hand, are words that are strongly influenced by the time being, such as historical events or names. Pennebaker et al. (2014) found a link between the usage of closed-class words and academic success. During the study, which used the LIWC tool on written essays of college applicants and connected these to subsequent academic success, the authors showed that the rate of closed-class words are significantly ($p < .01$) positively correlated to subsequent academic success, regardless of the chosen essay topic or sought major.

It is controversial whether the implicit achievement motive (see Section 5.2) is connected with academic success: Scheffer (2004) was able to predict grades with a significant correlation of r = .2, attributed to the intrinsic desire for excellence, whilst McClelland (1988) found that the power motive is rather correlated with academic success if grades are exposed to peers due to the desire to impress fellows.

The utilized data is described in the following section.

7.2.2 Data

The utilized data and pre-processing steps closely assemble the same setup as described for the LMT in Section 7.1.

Our research methodology for assessing aptitude automatically is described in the following section.

7.2.3 Methodology

Our methodology can be divided into two parts: the first is a natural language processing (NLP) task, which counts classified motives per participant and correlating this count to academic grades.

In order to test whether an LSTM with an attention mechanism succeeds in classifying the OMT, we employ the approach by Xu et al. (2015) on an already existing code basis for multiple text classifiers, which is utilized for further benchmarks as well.[29]

As for the word representations, we employed pre-trained fastText word embeddings (see Section 2.3.3) for German (Bojanowski et al., 2017), provided by the developers.[30] In contrast to Word2Vec word embeddings by Mikolov et al. (2013b), fastText has the capability of representing tokens not included in the embedded words on the basis of character n-grams. The OMT data (described in Section 7.2.2) is noisy, has many spelling mistakes and would probably not sufficiently be represented by word-based embeddings.

Benchmarking systems

To our knowledge, psychometrics closely related to the TAT have not been classified with neural methods yet. The only classification on the OMT has been performed by utilizing an LMT model in our previous work (Johannßen et al., 2019), which we compare to our neural approach. In order to put different architectures into perspective and to explore the relationship of our proposed LSTM system with attention mechanism, we performed multiple benchmarking experiments on the task of automatically assigning the four classes of operant motives described in Section 5.2 and thus aim to answer the second question of how well other neural approaches perform in comparison.

For this, we employed the following neural architectures, as reviewed in Section 7.2.1: LSTM, CNN, RNN, RCNN, Bi-LSTM with self-attention, LSTM with attention and Seq2One (a Seq2Seq variant with only one label as output) with attention. Since neural approaches are non-deterministic (Lai et al., 2015), we trained each model three times and averaged the F_1 scores for a stable assessment of results.

Three modifications of the LSTM with attention mechanism are employed: Firstly, we shuffled the attention weights before they got applied to the hidden states. Secondly, we reversed the direction of the input sequence to honor the OMT primacy rule. If this rule is followed and processing order has an influence, processing from right-to-left and classifying on the entire representation could improve results since the most influential signal (the first motive in the text) is accumulated last into the representation. Thirdly, we add comparable

---

[29] https://github.com/prakashpandey9/Text-Classification-Pytorch/tree/master/

[30] Facebook's AI Research, https://fasttext.cc

hand-crafted features as a fully connected input to the final classification softmax layer (e.g. part-of-speech (POS) tags, LIWC categories or the perplexities of trained language models per target motive), following our previous approach (Johannßen et al., 2019) to investigate in how far neural feature induction subsumes these features.

Psychometric predictions

After benchmarking, we utilize the most promising system for predictions in accordance with the OMT theory. 103 participating students answered the questions to 15 images, resulting in 1,545 answer sequences. Further, the data collection includes the grade of their bachelor's thesis, which was completed a few years after the OMT was taken. We employ the experimental design of our previous work (Johannßen et al., 2019) to ensure a fair comparison. For this, we predict the motives of each of the 15 answers given per participant, count the appearances per motive and correlate these to the bachelor's thesis grade.

7.2.4   Model training

All parameters of the models were tuned on a development set. Different fixed input sizes were considered for every architecture: Firstly we considered a fixed input length of 81 since the longest answer contains 81 words. Secondly, the average answer contains 20 words, which we considered as fixed input size in order to take the primacy rule (Section 5.2) into account. Shorter answers than the fixed input length receive the padding token (<pad>), longer ones were truncated. Human annotators are asked to ignore the rest of a sequence after a very first motive could be identified. Terms not observed in the training vocabulary were replaced by an out-of-vocab (OOV) token. Dropouts of .3, .5 and .8 were evaluated, whereas .5 has shown to perform best for the RNNs and has also been suggested by Hinton et al. (2012). The number of iterations was set to 3,600 in 32 batches and two epochs. The models received word embedded fastText inputs with 100 and 300 dimensions, of which the 300-dimensional embeddings reached better results, and had two hidden layers with 256 cells each. Learning rates were set to .0001, .001 and .01 for each model, with .001 performing best. All results are displayed in Table 7.13 and were achieved with these unified best-performing parameters.

As for the LSTM with attention mechanism, which has shown to perform best, the model converged quickly to a loss of approx. .4 and oscillates thereafter.

7.2.5   Attention weights assessment

As shown by Vaswani et al. (2017), the attention mechanism (described in Section 7.2.1) has broadly been believed to contribute to explainable artificial intelligence by shedding light on algorithmic decision making. Many authors have followed the initial idea and e.g. applied heat maps according to attention weights for input sequences and investigated algorithmic decision making. Other studies find contrary evidence that attention weights do not necessarily

reflect true meaning (Jain & Wallace, 2019). Even though we are aware of these controversies and limitations, we follow the critic's suggestion to investigate whether attention weights make a difference in the performance of a system. For this, we measure on which index the most attention weight mass is accumulated. We hypothesized that this might often be the last token since attention weights usually traverse a sequence *in search* (metaphorically speaking) for suiting candidates and mostly does not find any of such, applying the most of the available attention weight to the last possible candidate – the last token. We will further collect sequences that do not show this behavior and thus have the largest attention weight mass assigned to other tokens than the last one. These tokens will be evaluated with the LIWC tool. We would expect the motives to be reflected in the LIWC categories if they meant anything at all. We automatically assembled all classified instances, whose highest attention weight did not assemble on the very last token, exceeded .3 and was classified correctly.

Following this section, we present our results.

7.3   Results

This section presents the results of first the benchmarking of all explored machine learning algorithms and architectures, the best model's performance metrics, the assessment of attention weights, and lastly the empirical correlation with bachelor's thesis grades.

Model	∅ **Accuracy**	∅ **Precision**	∅ **Recall**	∅ F_1 **score**	**F** σ
CNN	63.26	59.34	63.62	61.41	2.36
RNN	68.73	73.10	68.73	70.85	1.59
LSTM	77.84	78.05	77.84	77.92	.65
Sequence to One (Seq2One) with attention	77.34	76.81	77.43	77.12	1.53
LSTM Attn with shuffled attention weights	79.03	78.05	79.03	78.54	.13
RCNN	79.70	79.35	79.81	79.58	.77
Bi-LSTM with self-attention	81.16	80.35	81.16	80.75	.31
LSTM Attn with 129 addit. handcrafted features	80.85	79.86	80.86	80.35	1.23
LSTM Attn with a reversed direction	80.87	80.05	80.87	80.46	.99
LSTM with an attention mechanism (LSTM Attn)	**81.94**	**81.15**	**81.96**	**81.55**	.09
LMT with 129 handcrafted features (baseline)	81.56	80.90	81.60	81.10	0

Table 7.3: The table provides a model benchmark. All models classified with a fixed input size of 20 tokens. The only system overcoming the strong baseline of the feature-based LMT is an LSTM with attention mechanism. We averaged all scores (∅) from three trained models each, and provide the standard deviation across runs (σ).

7.3.1 Model performance

Table 7.13 shows classification performance of the different approaches on the test set. We were able to improve over our previous classifier (Johannßen et al., 2019). Even though neural approaches often perform better than earlier machine learning (Zhang et al., 2018), only the results of the best-performing model, the LSTM with an attention mechanism, outperforms the feature-engineered LMT classification model by an F_1 score of 81.55 (the LMT scored 81.10 and thus only slightly worse) with a fixed input size of 20 tokens. The same model with the fixed size of the longest answer of 81 tokens performed worse with an F_1 score of 80.71 (not shown in Table 7.13). The other approaches, also with a fixed input size of 20 tokens, performed worse, mostly around a 79 F_1 score except for the CNN. Including 129 hand-crafted features, reversing the reading direction and shuffling attention weights did not improve the results, thus indicating that firstly, attention matters, secondly, the direction of classification is not as important and thirdly, the LSTM attention model learns the features (POS, LIWC categories, perplexity) incidentally. The confusion matrix of the best-performing model is displayed in Table 7.4. The same LSTM with attention mechanism enriched by similar hand-crafted features does not improve results further, indicating that the information from these features is subsumed by the induced representations. The inversion of the input sequence resulted in lower scores, indicating that either the model cannot make use of seeing earlier tokens later to account for the primacy rule, or that the primacy rule has not been followed consequently during annotation. Shuffling of the attention weights worsens the results, indicating that these weights matter for the classification task.

		Predicted				
		0	**A**	**L**	**M**	**Σ**
		5%	**17%**	**19%**	**59%**	**100%**
Actual	**0**	283	102	150	478	1,013
	A	29	2,739	112	646	3,526
	L	90	91	3,079	872	4,132
	M	126	657	404	11,102	12,289
	Σ	528	3,589	3,745	13,098	20,960

Table 7.4: The relative motive amounts and confusion matrix of the best performing system (LSTM Attn).

7.3.2 Assessment of the attention weights

Table 7.13 shows that the LSTM with attention mechanism scored significantly lower when its attention weights were shuffled compared to the one with properly trained attention and assigned weights. Jain & Wallace (2019) stated that this case had occurred only rarely in their

experiments, but that if this circumstance holds true, they would assume that attention weights could be considered for interpretation and explanation.

We can observe that on average, 79.85% of the available attention weight mass was assigned to the very last token of each instance. It appears that the mechanism considered one token at a time from left to right and determines whether attention weight mass should be assigned to the token in question. If this is not the case, the attention weight mass is being kept and the successor token is considered. When the mechanism reaches the end of the sequence, it assigns whatever attention weight mass is left to the very last token. The second and third index with the highest following attention weight masses are the second last and third last tokens respectively. According to the OMT theory, the last tokens of a sequence, in general, should not provide the main information for encoding the whole sequence due to the primacy rule, this high attention weight mass on the last token indicates, that for the majority of classified instances, the attention weights do not serve as a widely applicable means to interpret the reasons for classification decisions in this setup.

Besides these last tokens, we aimed to investigate the mechanism further and compare these non-concluding tokens to all tokens by automatically assembling instances and attention weights.

Table 7.5 compares the four most prominent psychologically validated LIWC category memberships in percent per motive of all tokes versus non-final tokens with high attention weight masses. Most of the LIWC category names appear to be representative for the wordlists that they consist of. E.g. *positive emotion* consists of e.g. *love, nice and sweet*.

According to the OMT theory, people with a strong achievement motive desire intrinsic excellence. They tend to analyze problems thoroughly and focus on tasks. This description is reflected by *cognitive mechanism* that is almost twice as present for high attention mass tokens as it is for all tokens (27.39% compared to 14.11%). The categories *occupation* (e.g. observe, conduct, advancing) with 24.66% and *achieve* – already with the same name as the OMT motive – with 23.28% are high in presence as well. Compared to rather low *social, affect and other references*, the OMT theory for the achievement motive appears to be better represented by tokens with high attention. Single words include *intense, concentrated, motivated and capabilities*.

Similarly, the LIWC categories for the affiliation motive are *affect, positive emotion, humans and social* for the left columns and apparently reflect the description of a desire to solve problems cooperatively, whilst avoiding conflicts. However, scores for LIWC categories are rather low at 12.12% and 9.09%. The social LIWC category is strongly present on the right column for all tokens with 19.76%, as well as *affect* with 12.04%. The other two LIWC categories of the right columns *other references* and *cognitive mechanism* do not appear to align well with the affiliation motive.

Even though the desire to influence and alter one's surrounding and fellow beings, the power motive can be identified by positive expressions as well as rather harsh ones. All LIWC

	High attention weight mass			All tokens		
	LIWC	per cent	words	LIWC	per cent	words
Achievement	cognitive mechanism	27.39	intense concentrated motivated capabilities	social	15.17	-
	occupation	24.66		cognitive mechanism	14.11	-
	achieve	23.28		other references	11.44	-
	insight	10.96		affect	10.49	-
Affiliation	affect	12.12	important secure partner interested	social	19.76	-
	positive emotion	12.12		other references	12.04	-
	humans	9.09		affect	10.31	-
	social	9.09		cognitive mechanism	9.48	-
Power	affect	33.95	can feels dominant humiliated	social	18.99	-
	cognitive mechanism	28.91		cognitive mechanism	11.46	-
	positive emotion	24.93		other references	11.25	-
	insight	20.16		affect	9.91	-

Table 7.5: LIWC analysis of tokens that received the most attention weight mass on the left with all tokens on the right separated by predicted labels (left) versus manually annotated labels (right).

categories of these columns on the left appear to align with the power motive, which are *affect (33.95%), cognitive mechanism (28.91%), positive emotion (24.93%) and insight (20.16%)*. The corresponding LIWC categories for all tokens on the right columns correspond with the exception of *other references* but are comparably weaker.

This comparison shows that tokens with high attention mass per motive correspond to the OMT theory e.g. occupation and insight for achievement, whilst all tokens do show some correspondence (e.g. social and affiliation), but in general, do not align well with the OMT theory. Interestingly, when removing the tokens (besides the last ones) that received the most attention weight mass and re-evaluating the answers with the LIWC tool to test the counter-hypothesis that high-attention tokens do not reflect the classes, the categories shift to ones that do not correspond to the OMT theory.

Examples are given in Table 7.6, which displays some tokens highlighted, according to the token's attention weight masses. These examples do not reflect the whole data basis but illustrate a possible aid for understanding the task at hand and might help develop tool support for this task or related psychometrics.

gelangweilt *bored*	weil *because*	sie *she*	jeden *every*	tag *day*	0
geborgen *protected*	weil *because*	die *the*	andere *other*	person *person*	A
gefordert *challenged*	will *wants*	das *the*	ziel *goal*	erreichen *to reach*	L
zu *to*	maßregeln *disciplin*	dominant *dominant*	die *the*	andere *other*	M

Table 7.6: Heatmap according to the attention weights displayed on four example snippets of OMT answers in German with their glossed translations and targets (A for affiliation, M for power and L for achievement).

7.3.3 Correlation with bachelor's thesis grades

As described in Section 7.2.3, in order to analyze the predictive power of motives, we count predicted motives and correlate these counts to academic grades. While we found a weak correlation from the LMT predictions of $r = -.2$ between power motive counts and the bachelor's thesis grade, the experiment with the bi-LSTM revealed a a correlation of $r = -.25$ between the bachelor's thesis grade and the achievement motive in this work, i.e. the higher the achievement motive count, the better the German grade value (1.0 equals *very good*, 5.0 equals *having failed*). The power motive is positively correlated with a small $r = .14$, i.e. the higher the power motive count, the worse the German grade. Figure 7.2 shows scatter plot displaying the counts of the power and achievement motives and the achieved bachelor's thesis grade.

This discrepancy of both model's – the LMT and bi-LSTM – predictions is anomalous. If both models performed comparably well on the same type of data, both models should reveal comparable correlations between counted motives and grades. The investigation of each model's motive predictions per student shows that the LSTM with attention mechanism often assigns the power motive but never zero, whilst the LMT model assigns zero on 17.76% of all cases, indicating that the LMT model often did not predict any motive. Thus, even though the models behave comparably well on test data of the same origin as the training data, they differ in their algorithmic decision making on data from a different origin.

Finally, a conclusion can be drawn in the following section.

7.4 Conclusion and Outlook

We were able classify the OMT by employing an LSTM with an attention mechanism achieving an F_1 score of 81.55. Other architectures such as the RNN, LSTM, Bi-LSTM or the RCNN mostly reached an F_1 score of approx. 79. Attention weights only matter in thus far that the shuffling of these weights worsens the results. The attention weight mass mostly accumulates on the very last token and thus does not allow for insights in the general case. For these cases where
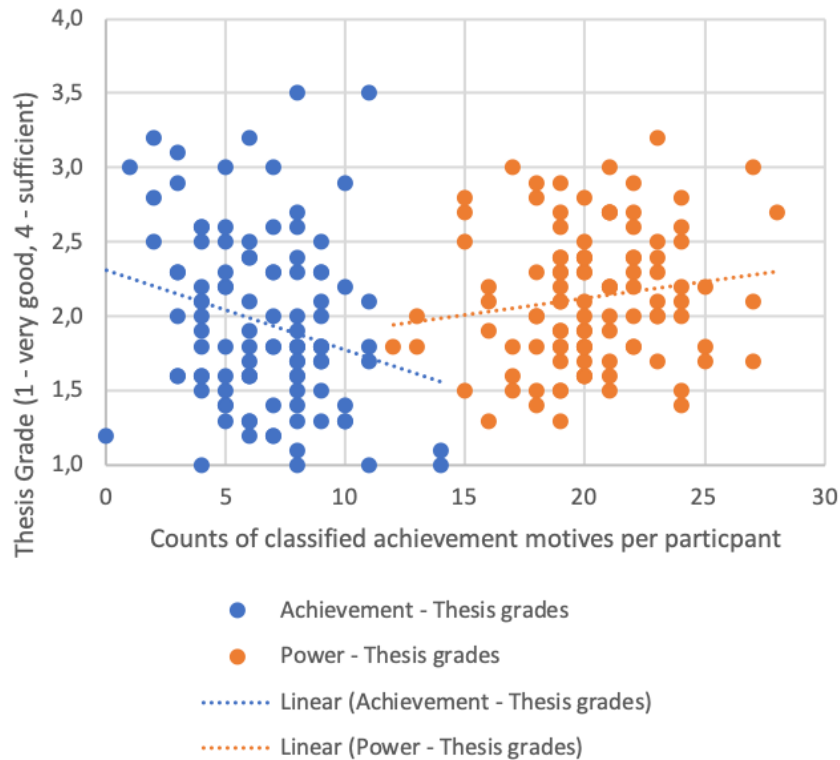
Figure 7.2: After predicting motives, the four motives per participants were counted. The power motive has the highest frequency. By counting predicted motives and correlating them to academic grades, a weak correlation of $r = -.25$ could be observed between the achievement motive (blue dots) and the bachelor's thesis grade (1 being the best, 5 the worst grade). In contrast, the plots shows that the higher the power motive counts (orange dots), the worse the grade with $r = .14$.

the attention weight mass was distributed among other tokens than the last one of a sequence, an analysis with the LIWC tool showed conformity of LIWC categories with the corresponding operant motives compared to these of all words. This indicates an overlap between the memberships per word of both linguistic assessments. This behavior of the highest attention mass on last tokens could be canceled out by employing a Bi-LSTM with attention mechanism and concatenating the attention weights of both systems, which we consider for future experiments. When removing these tokens and re-evaluating the sequence with the LIWC tool, the results shift, which has to be investigated further. A correlation between identified motives and subsequent academic success was assumed. A correlation could be observed with $r = -.25$ between the counted achievement motives and bachelor's thesis grade, which is a weak correlation much different to the alternate predictions of the LMT model that assigned zeros more often than the LSTM model with attention mechanism. Since zero marks indecisiveness, it can be assumed that the LMT model does not generalize as well as the LSTM – though this assumption would have to be further examined by e.g. having trained psychologists assess

the outputs of both models. This effect is displayed in Table 7.7, which shows that the LMT model does not assign the zero motive to any of the instance, whilst the LSTM captures the zero motive comparably better. Furthermore, direct predictions from language to grades could be investigated, hence losing information at the intermediate step of automatically annotated motives.

Achievement	Affiliation	Power	Zero

Logistic Model Tree

Achievement	Affiliation	Power	Zero
7	2	20	0
2	0	27	0
8	3	19	0
6	2	21	0

Bi-LSTM Attention

Achievement	Affiliation	Power	Zero
1	7	14	6
2	2	9	15
6	2	17	3
3	2	16	7

Table 7.7: Comparison of the two approaches for modeling the OMT: an LMT, and a bi-LSTM with an attention mechanism. This Table displays an excerpt from the grade predictions, where the LMT never assignes the zero motive, whilst the bi-LSTM appears much more generalized.

Especially since the application field of aptitude diagnostics renders potentially countless subsequent research objectives, hence it being a large part of psychological diagnostics (Schmidt-Atzert et al., 2018), we crafted a shared task to foster research and provide the community with data. This approach is described in the following section.

7.5 Subsequent Research: GermEval Shared Task 1

This section describes the conducted GermEval shared task 1 on cognitive and motivational style, which aimed to extend the aptitude diagnostic research and aimed to provide the community with the OMT data. For an ethical consideration of the shared task refer to Section 3.3.

Despite the growing interest in NLP and its methods since 2015 (Manning, 2015), application fields of NLP in combination with psychometrics are rather sparse (Johannßen & Biemann, 2018). Aptitude diagnostics can be one of those application fields. To foster research on this particular application domain, we presented the *GermEval-2020 Task 1 on the Classifi-*

cation and Regression of Cognitive and Motivational Style from Text.[31][32] [33] The task contains two subtasks. For Subtask 1, participants are asked to reproduce a ranking of students based on different high school grades and intelligence quotient (IQ) scores solemnly from implicit motive texts. For Subtask 2, participants are asked to classify each motive text into one of 30 classes as a combination of one of five implicit motives and one of six levels. Quantitative details on participation are displayed in Table 7.8.

The validity of high school grades as a predictor of academic development is controversial (Hell et al., 2007; Schleithoff, 2015; Sarges & Scheffer, 2008). Researchers have found indications that linguistic features such as function words used in a prospective student's writing perform better in predicting academic development (Pennebaker et al., 2014) than other methods such as GPA values.

During an aptitude test at a rather small university of applied sciences NORDAKADEMIE in Germany with roughly 500 students enrolling each year, participants take the implicit motive test MIX (see Subsection 5.2).

From a small sample of an aptitude test collected at a college in Germany, the classification and regression of cognitive and motivational style from a German text can be investigated. Such an approach would extend the sole text classification and could reveal insightful psychological traits.

For our task, we provide extensive amounts of textual data from both, the OMT and MIX, paired with IQ and high school grades and labels.

The task is to predict measures of cognitive and motivational style solemnly based on text. For this, z-standardized high school grades and IQ scores of college applicants are summed and globally 'ranked'. This rank is utterly artificial, as no applicant in a real-world-setting is ordered in such fashion but rather there is a certain threshold over the whole of the hour-long aptitude test with multiple different test parts, that may not be undergone by applicants.

Aptitute test and college

Since 2011, the private university of applied sciences NORDAKADEMIE performs an aptitude college application test.

Zimmerhofer & Trost (2008, p. 32ff.) describe the developments of the German Higher Education Act. A so-called Numerus Clausus (NC) Act from 1976 and 1977 ruled that colleges in Germany with a significant amount of applications have to employ a form of selection mechanism. For most colleges, NC was the threshold for many applicants. Even though this value is more complex, it roughly can be understood as a GPA threshold. Since this second

---

[31]GermEval is a series of shared task evaluation campaigns that focus on Natural Language Processing for the German language. The workshop is held as a joint Conference SwissText & KONVENS 2020 in Zürich.

[32]`https://www.inf.uni-hamburg.de/en/inst/ab/lt/resources/data/germeval-2020-cognitive-motive.html`

[33]The data and annotations were provided by Nicola Baumann (Universität Trier) and Gudula Ritz (Impart GmbH).

	Task A	Task B motives	task B levels	task B motives + levels
# Teams	2	3	3	3
# Submissions	6	7	7	7
Best Team	TueOslo	FH Dortmund	FH Dortmund	FH Dortmund
Best pearson r	.3701	–	–	–
Best Macro-F_1	–	70.46	66.50	70.40
Impr. over baseline	.1769	5.52	6.92	5.95

Table 7.8: Quantitative details of submissions.

Higher Education Act, colleges are also free to employ alternate selection forms, as long as they are scientifically sound, transparent, and commonly accepted in Germany.[34]

Even though Hell et al. (2007, p. 46) found the correlation coefficient of high school grades of r = .517 to be the most applicable measure for academic suitability, criticism emerged as well. The authors criticized the measure of grades by just one single institution (i.e. a high school) does not reflect upon the complexity of such a widely questioned concept of intellectual ability. Schleithoff (2015, p. 6) researched the high school grade development of different German federal states on the issue of grade inflation in Germany and found evidence, that supports this claim. Furthermore, in most parts of Germany, the participation grade makes up 60% of the overall given grade and thus is highly subjective.

Since operant motives are said to be less prone to subjectivity, the NORDAKADEMIE decided to employ an assessment center (AC) for research purposes and a closely related aptitude test for the application procedure (Gragert et al., 2018). Rather than filtering the best applicants, the NORDAKADEMIE aims with the test for finding and protecting applicants that they suspect to not match the necessary skills required at the college.[35] Thus, every part of the aptitude test is skill-oriented.

Furthermore, this test contains multiple other parts, e.g. math and an English test, Kahnemann scores, IQ scores, a visual questionnaire, knowledge questions to the applied major or the implicit motives, the MIX.

NORDAKADEMIE Aptitude Data Set

Since 2011, the private university of applied sciences NORDAKADEMIE performs an aptitude college application test, where participants state their high school performance, perform an IQ test, and the implicit psychometrical test MIX. The MIX measures so-called implicit or operant motives by having participants answer questions to those images like the one displayed below such as "who is the main person and what is important for that person?" and "what is that

---

[34]BVerfGE 43, 291 – numerus clausus II
[35]https://idw-online.de/de/news492748

person feeling?". Furthermore, those participants answer the question of what motivated them to apply for the NORDAKADEMIE.

The data consists of a unique ID per entry, one ID per participant, of the applicants' major and high school grades as well as IQ scores with one textual expression attached to each entry. High school grades and IQ scores are z-standardized for privacy protection.

The data is obtained from 2,595 participants, who produced 77,850 unique MIX answers and have agreed to the use of their anonymized data for research purposes.

The shortest textual answers consist of 3 words, the longest of 42 and on average there are roughly 15 words per textual answer with a standard deviation of 8 words. The (for illustrative purposes not z-standardized) average grades and IQ scores are displayed in Table 7.9.

Metric	score	standard deviation
German grade	9.4 points	1.84
English grade	9.5 points	2.15
Math grade	10.1 points	2.2
IQ language	66.8 points	19.0
IQ logic	72.6 points	15.6
IQ averaged	77 points	14.1

Table 7.9: Average scores and standard deviations of data for Subtask 1.

The *IQ language* measures the use of language and intuition such as the comprehension of proverbs. The *IQ logic* tests the relations of objects and an intuitive understanding of mainly verbalized truth systems. The averaged IQ includes IQ language and logic as well as further IQ tests (i.e. language, logic, calculus, technology, and memorization).

OMT Shared Task Data Set

The available data set has been collected and hand-labeled by researchers of the University of Trier. More than 14,600 volunteers participated in answering questions to 15 provided images such as displayed in the figure below.

The pairwise annotator intraclass correlation was $r = .85$ on the Winter scale (Winter, 1994).

The length of the answers ranges from 4 to 79 words with a mean length of 22 words and a standard deviation of roughly 12 words. Table 7.10 shows the class distribution of the motives, the levels, and all the combinations. The number of motives in the available data is unbalanced with power (M) being by far the most frequent with 54.5%. The combined class of M4 is by far more frequent than e.g. the combination F_1. This makes this task more difficult, as unbalanced data sets tend to lead to overfitting. Those percentages were measured on the training set, containing a subset of 167,200 labeled text instances.

		Motives				
	$\sum$	**0**	**A**	**F**	**L**	**M**
$\sum$	**100%**	**4.55%**	**16.83%**	**17.59%**	**19.63%**	**41.02%**
0	**4.6%**	4.55	.01	.00	.00	.01
1	**9.9%**	.00	1.70	1.06	1.43	5.67
2	**20.8%**	.00	5.73	3.33	7.69	4.11
3	**13.6%**	.00	.81	2.57	3.76	6.46
4	**30.7%**	.00	4.51	5.42	4.51	16.25
5	**20.4%**	.00	4.07	5.57	2.24	8.52

Table 7.10: An overview of the Subtask 2 classes distributions (percentages). Values were rounded.

Subtask 1: Regression of artificially ranked cognitive and motivational style

This task had yet never been researched and was open: It was neither certain, whether this task can be achieved, nor how well this might be possible before this task.

The goal of this subtask is to reproduce the artificial 'ranking' of students. Systems are evaluated by the Pearson correlation coefficient between system and gold ranking. An exemplary illustration can be found in Section 8.3. We are especially interested in the analysis of possible connections between text and cognitive and motivational style, which would enhance later submission beyond the mere score reproduction abilities of a submitted system.

A z-standardized example was provided with with a unique ID (consisting of studentID_imageNo_questionNo), a student ID, an image number, an answer number, the German grade points, the English grade points, the math grade points, the language IQ score, the math IQ score, and the average IQ score (all z-standardized). The data is delivered as displayed in Table 7.11.

The data is delivered in two files, one containing participant data, the other containing sample data, each being connected by a student ID. The rank in the sample data reflects the averaged performance relative to all instances within the collection (i.e. within train / test / dev), which is to be reproduced for the task.

The training data set contains 80% of all available data, which is 62,280 expressions and the development and test sets contain roughly 10% each, which are 7,800 expressions for the development set and 7,770 expressions for the test set (this split has been chosen in order to preserve the order and completeness of the 30 answers per participant).

For the final results, participants of this shared task were provided with a MIX_text only and were asked to reproduce the ranking of each student relative to all students in a collection (i.e. within the test set).

System submissions were evaluated on the Pearson rank correlation coefficient (see Subsection 2.1).

Field	Value
student_ID	1034-875791
image_no	2
answer_no	2
UUID	1034-875791_2_2
MIX_text	Die Person fühl sich ein- gebunden in die Unter- hatung. [The person feels involved in the conversation.]

Field	Value
student_ID	1034-875791
german_grade	-.086519991198202
english_grade	.3747985587188588
math_grade	.511555970796778
lang_iq	-.010173719700624
logic_iq	-.136867076187825

Field	Value
student_ID	1034-875791
rank	15

Table 7.11: Subtask 1 asked participants to reconstruct a ranking of cogntiev and motivational style and provided the participants with three separate files: i) with implicit motive data including image numbers and textual answer (topmost table), ii) performance metrics including school grades, math test and IQt scores (middle table). and iii) the rank – all of which share the same *student_ID*.

Subtask 2: Classification of the Operant Motive Test (OMT)

For this task, we provided the participants with a large dataset of labeled textual data, which emerged from an operant motive test (see Chapter 5 for details). The training data set contains 80% of all available data (167,200 instances) and the development and test sets contain 10% each (20,900 instances). The data is delivered as displayed in Table 7.12.

On this task, submissions are evaluated with the macro-averaged F_1 score.

Organizer's baseline systems

For both tasks, the organizers chose rather simple approaches that utilize support vector machines (SVM) paired with frequency-inverse document frequency (tf-idf) document representations.

SVMs are a class of statistical machine-learning algorithms that aim to map data to a higher dimensional feature space that best linearly separates target classes with the largest margin between them, which normally would not be separable linearly (this is called the *kernel trick*) and were first created by Cortes & Vapnik (1995). Tf-idf is a statistical evaluation of how important words are for documents and was first used by Luhn (1957).

Field	Value
UUID	6221323283933528M10
text	Sie wird aus- geschimpft, will jedoch das Gesicht bewahren. [She gets scolded, but wants to save face.]

Field	Value
UUID	6221323283933528M10
motive	F
level	5

Table 7.12: Subtask 2 asked participants to classify the OMT and provided participants with two data files: i) a file containing the textual answers (upper table), and ii) a file containing the corresponding target labels of implicit motives and self-regulatory levels (lower table) – both of which also included a unique ID, the UUID.

For Subtask 1, a Support Vector Regressor (SVR) was utilized. This statistical method tries to find an ideal line that best fits provided training data and thus examines a relationship between two continuous variables. Text is represented via tf-idf and a simple count vectorizer, which tokenizes text and builds vocabulary.

The SVR system achieved a Pearson ρ of .32, which is quite a big signal for data sources produced by human behavior. As there were 260 values to be ranked, we determined a T-value of 5.33 with a degree of freedom of 259, leading to a p-value of 2.096e-07. This means, that the result is highly significant and the null hypothesis can be rejected.

As for the classification task, a linear support vector classifier (SVC) was chosen. 30 (combined motive-level labels) binary SVCs one-vs-all classifiers were trained. The data was centered and C (regularization) was set to the default 1.0 and the chosen loss is the *squared hinge*. It is useful for binary decision or when it is not of importance how certain a classifier is. The loss is either 0 or increases quadratically with the error. The system reached a macro F_1 score of 64.45 on the motive + labels classification task.

Submitted Systems

This section will provide a rough overview of the submitted systems, chosen word representations, some outstanding parameter choices, and some of the most interesting findings. For more details, it is recommended to read the resp. publications. Some details can be found in Table 7.13.

We notice two different approaches from the teams, especially from Subtask 1 to Subtask 2: i) statistical and non-neural word representations and systems and ii) neural approaches and word embeddings.

The team from Tübingen (Çöltekin, 2020) was very successful on the first subtask by using linear models with statistical n-gram features, exceeding the baseline by .1778 points and the

Team	Classifier Approach	Task	Resp. score	Text Features
Tübingen Çöltekin (2020)	Subtask 1	Linear single 2	.3701	n-grams
FH Dortmund Schäfer et al. (2020)	Subtask 1	SVR	.3154	tf-idf
Baseline	Subtask 1	SVR	.1923	tf-idf
FH Dortmund Schäfer et al. (2020)	Subtask 2 motives	BERT ensemble cased	70.46	BERT
Idiap Villatoro-Tello et al. (2020)	Subtask 2 motives	SimpleTransOut BERT LATEST	69.63	BERT
Tübingen Çöltekin (2020)	Subtask 2 motives	SVM adaptive	68.04	n-grams
Baseline	Subtask 2 motives	SVC	64.94	tf-idf
FH Dortmund Schäfer et al. (2020)	Subtask 2 levels	BERT ensemble cased	66.50	BERT
Idiap Villatoro-Tello et al. (2020)	Subtask 2 levels	SimpleTransOut BERT LATEST	65.32	BERT
Tübingen Çöltekin (2020)	Subtask 2 levels	linear-single2	63.35	n-grams
Baseline	Subtask 2 levels	SVC	59.85	tf-idf
FH Dortmund Schäfer et al. (2020)	Subtask 2 motives + levels	DBMDZ uncased	70.40	BERT
Idiap Villatoro-Tello et al. (2020)	Subtask 2 motives + levels	SimpleTransOut BERT LATEST	69.97	BERT
Tübingen Çöltekin (2020)	Subtask 2 motives + levels	SVM adaptive	67.81	n-grams
Baseline	Subtask 2 levels + motives	SVC	64.45	tf-idf

Table 7.13: Overview of the submitted approaches. Only the best submitted systems per team and task were considered. The entries are grouped by the type of task and displayed in descending order. DBMDZ stands for *Digitale Bibliothek Münchener Digitalisierungszentrum* and is a pre-trained German BERT model. SimpleTransOut stands for the Simple Transformer library from pypi.org.

second-placed team FH Dortmund by .0547 points. The authors note in their discussion, that, even though neural approaches nowadays offer broad applicability on all sorts of tasks, for the proposed regression task, their linear approach with n-gram features was sufficient. Even if the authors did not reach the first place on the second subtask with their self-designated *simple* linear and statistical approach, they still surpassed the organizer's baseline system on the second task by 3.36 percent points. Their results showed, that there is a signal in the implicit texts is sufficient for re-creating the ranking above chance.

The team Idiap (Villatoro-Tello et al., 2020) reached the second place for every type of Subtask 2 goal with a *Simple Transformer*, approach, which utilizes the attention mechanism without any recurrent units. Words were represented with pre-trained BERT (Devlin et al., 2019) embeddings. Since the attention mechanism offers the chance of investigating algorithmic decisions made, the authors plan for future work to investigate those, possibly better understanding the OMT and the underlying patterns. During their presentations at the GermEval20 Task 1 session, the authors displayed tokens, which acquired high attention mass and

concluded, that firstly, function words were more influential than content words, secondly, the so-called freedom motive was harder to distinguish from power than e.g. the achievement motives and that finally, negations were influential for classifying the power motive with level 4.

Lastly, the team from the FH Dortmund (Schäfer et al., 2020) utilized BERT word representations, exceeding the baseline-system of the motives + levels approach (30 target classes) by 5.95 percent points and the second-placed team Idiap by .61 percent points. The team experimented with different pre-processing steps but found, that they did not greatly influence the performance of their system, despite the data being mixed with different languages and some noise. For their approaches to solving Subtask 2, the authors experimented with different word representations, namely fasttext (Bojanowski et al., 2017) and BERT. Interestingly, the authors state that it was more useful for solving Subtask 2 to predict all 30 classes with a single model, than to train two classifiers for motives and levels respectively and to combine the predictions.

Discussion

The Organizer's SVM tf-idf systems have shown, that solutions of both subtasks above chance are possible. Subtask 2 with its implicit motives and levels appears to be a bit more trivial, as a macro score of $F_1 = 64.45$ is already strong, considering that the 30 target classes are unevenly distributed.

The submitted systems of the shared task participants revealed some interesting findings, which could be impactful for the implicit motive theory and their practical assessment.

Team Thübingen (Çöltekin, 2020) could re-create the Subtask 1 ranking above chance, even though there were no available manual labels. Since the impacts of identified implicit labels functioned as interim steps for behavioral predictions before (Johannßen & Biemann, 2019), those findings indicate the psychological validity of this implicit psychometric.

Team FH Dortmund (Schäfer et al., 2020) observed that for Subtask 2, excessive pre-processing did not make much of a difference. This, paired with an already strong but simple SVM tf-idf baseline system, indicates that language modeling already could be sufficient for classifying implicit motives and levels. If that were the case, the most impactful utterances per target class should be investigated and compared to the implicit motive theory.

Furthermore, the team found the direct prediction of 30 target classes of the motive + levels combination to be more sufficient than two models separately. This is consistent with the operant motive theory, which states, that the self-regulatory levels are connected (Baumann & Scheffer, 2010).

Lastly, some of the findings by the participants, have shown strong connections to behavioral research made on behalf of the implicit psychometrics theory. Winter (2007) identified so-called activity inhibition (AI) as good behavioral predictors for war and crisis situations by analyzing political speeches. AI is being described as negations in combination with the

power motive. This connection between the power motive and negations was also observed by team Idiap (Villatoro-Tello et al., 2020) and thus reproduces earlier findings in other settings. Those findings could foster implicit psychometrics theory and thus advance aptitude diagnostics, which is the very reason for conducting such shared tasks.

We believe that it is also important to discuss the limitations of our work, in addition to its strengths presented in this chapter. The following section offers clear discussion of limitations.

7.6 Limitations

The most apparent limitation of the work presented in this chapter concerns the validity of the methodology design. Both, the proposed LMT and bi-LSTM models, have shown sufficient performances on classifying the OMT and self-regulatory levels. However, the transistion towards empirical applications of said models oughts to be viewed critically. Especially the behavior of the LMT, to not predict any zero motive on the validation (real world) data underlines, how models can behave differently on other datasets than the ones they were trained on. The most trustfull approach would be to have human experts annotate sufficiently large datasets of any altering data source before applying the models for predictions. However, this comes at the cost of expensive manual labor.

Another point of criticism is the interpretation of attention weights. With the adapted approach of shuffling the attention weights and measuring, whether the performance declines, we were able to demonstrate that these attention weights support the importance for the algorithmic descisions. However, we went even further and aimed to interpret the highest attention weighted tokens with human experts. It appeared that these tokens indeed reflected the modeled task at hand.

Nonetheless, further validation is appropriate due to recent debates upon attention weights as indicators of interpretation. One approach for validation would be to provide trained psychologists for labeling the OMT with tokens that received comparably much attention weight mass and with tokens that did not to measure how many cases would have been identified by said psychologists. Furthermore, we aim to provide annotators with a tool with attention-based highlighting for possibly saving time and expenses during the labeling process. Further numerical improvements could result from using contextualized embeddings, e.g. Bidirectional Encoder Representations from Transformers (BERT, Devlin et al. (2019)).

Lastly, the utilization of psychometric models on data associated with cognitive abilities (and it may be assumed that school und academic grades reflect upon cognitive abilities) raises multiple ethical concerns. These concerns and an assessment of their implications can be found in Section 3.3.

Aptitude diagnostics might be a large part of psychological diagnostics, but it is not sufficient for representing the whole application field. The next chapter therefore describes em-

pirical research on the basis of well-established psychological conflict research and utilizes NLPsych for its automation. The COVID-19 pandemic emerged during the course of the dissertation project and shifted the focus from broader psychological diagnostics towards this black swan event.

NLPsych for Measuring Social Unrest

This chapter describes empirical research on the black swan event of the COVID-19 pandemic. Social unrest was assessed and its detection automated. This research is presented by first presenting the nieche related work in 8.1. The predictors for social unrest as identified by psychologists is presented in Section 8.2. Section 8.4 describes the research methodology. Thereafter, Section 8.5 is concerned with the conducted experiments. The results are presented in the concluding Section 8.6. Limitations are discussed in Section 7.6.

The COVID-19 pandemic and the reactions to it have led to growing social tensions. Gutiérrez-Romero (2020) studied the effects of social distancing and lockdowns on riots, violence against civilians, and food-related conflicts in 24 African countries. The author found that the risk of riots and violence has increased due to lockdowns. Resistance against national health regulations such as the duty to wear masks are partially met with resistance by movements such as anti-maskers or anti-obligation demonstrations.[36] Even anti-democratic alterations of e.g. services offered by the US Postal Service (USPS) to deliver mail-in ballots for the US presidential elections 2020, which are essential for social distancing measures amidst the pandemic, are being utilized amidst this international crisis and foster social unrest and potential outbursts of violence, civil disobedience or uprisings.[37]

Social media has become an important reflection of nationally and internationally discussed topics, and is a predictor of e.g. stock markets, disease outbreaks or political elections (Kalampokis et al., 2013). The majority of human-produced data exists in textual form and broadly in social media and thus, an investigation of social unrest and conflict situations from social media becomes a worthwhile application area for natural language processing (NLP) problems (Gentzkow et al., 2019).

---

[36] https://firstdraftnews.org/latest/coronavirus-how-pro-mask-posts-boost-the-anti-mask-movement/

[37] https://www.businessinsider.com/trump-walks-back-threat-block-covid-relief-over-usps-funding-2020-8?r=DE&IR=T

When speaking about such global phenomena such as a rise in international social unrest and possible occurrences of conflict reflected in the text, the detection of specific keywords or utterances have not been successful in past research. Mueller & Rauh (2017) utilized Laten Dirichlet Allocation (LDA, (Blei et al., 2003)) topic modelling on war-related newspaper items and were not able to improve predictability from other multi-factor models that take into account e.g. gross domestic product (GDP) figures, mountainous terrain or ethnic polarization. Furthermore, Chadefaux (2012) showed that news reports on possible war situations alone did not function as good predictors but identified sharp frequency increases before war emerged, possibly helping with just-in-time safety measures but likely failing to avoid war situations altogether.

Alternatively, the risks of escalation could be determined based on politicians' personalities and the current mood and tone of utterances (Schultheiss & Brunstein, 2010, p. 407). However, intrinsic desires and personality can hardly be measured directly (see Chapter 5). Intrinsic or subconscious desires and motivation would more likely correlate with personalities, tone, and thus possibly social unrest.

8.1   Related Work

Conflict predictions from natural language are rarely encountered applications and have mainly been about content analysis and less about crowd psychology. Kutuzov et al. (2019) used one-to-X analogy reasoning based on word embeddings for predicting previously armed conflict situations from printed news. Johansson et al. (2011) performed named entity recognition (NER) and extracted events via Hidden Markov Models (HMM) and neural networks, which were combined with human intelligence reports to identify current global areas of conflicts, that, in turn, were utilized mainly for world map visualizations.

Investigation of personality traits has mainly been focused on so-called explicit methods. For these, questionnaires are filled out either by interviewers, through observations, or directly by participants. One of the most broadly utilized psychometrics is the Big Five inventory, even though its validity is controversial (Block, 1995). The five-factory theory of personality (later named Big Five) identifies five personality traits, namely *openness to experiences*, *conscientiousness*, *extraversion*, *agreeableness* and *neuroticism* (McCrae & Costa Jr., 1999; Goldberg, 1981). This Big Five inventory was utilized by Tighe & Cheng (2018) for analyzing these five traits of Filipino speakers.

Some studies perform natural language processing (NLP) for investigating personality traits. Lynn et al. (2020) utilized an attention mechanism for deciding upon important parts of an instance when assigning the five-factor inventory classes. The Myers-Briggs Type Indica-

tor (MBTI) is a broadly utilized adaption of the Big Five inventory, which Yamada et al. (2019) employed for asserting the personality traits within tweets.[38]

The research field of psychology has moved further towards automated language assertions during the past years. One standard methodology is the utilization of LIWC (see Subsection 2.3.1). It has been shown that LIWC correlates with the Big Five inventory (McCrae & Costa Jr., 1999). Importantly, the writing style of people can be considered a trait, as it has shown high stability over time, which means that it is not dependent on one's current mood, the time of day, or other external conditions (Pennebaker et al., 1999). Hogenraad (2003) utilized an implicit motive (see Chapter 5) dictionary approach to automatically determine risks of war outbreaks from different novels and historic documents, identifying widening gaps between the so-called power motive and affiliation motive in near-war situations.

Overall, the work on automated classification of implicit motive data or the use of NLP for the assertion of psychological traits in general is rather sparse or relies on rather outdated methods, as this application domain can be considered a niche (Schultheiss & Brunstein, 2010; Johannßen & Biemann, 2019, 2018; Johannßen et al., 2019). One recent event in this area was the GermEval 2020 Task 1 on the Classification and Regression of Cognitive and Motivational Style from Text (Johannßen et al., 2020a). The best participating team reached a macro F_1 score (see Section 2.2.9) of 70.40 on the task of classifying implicit motives combined with self-regulating levels, resulting in 30 target classes. However, behavioral outcomes from automatically classified implicit motives have – to our knowledge – not yet been researched.

Psychological diagnostics is concerned with the detection of social unrest. This can be achieved by measuring patterns or markers, which have shown to be connected to times of social unrest. These predictors are described in the following section.

8.2 Social Unrest Predictors

Times of severe social unrest are reflected by distinct patterns of implicit motives and linguistic features. Winter (2007) surveyed multiple prior studies, identifying three main predictors: *responsibility*, *activity inhibition*, and *integrative complexity*, displayed in Table 8.1. In this study, the author identified and analyzed 8 occurrences of crises and social unrest by examining influential political speeches of this time. Thereafter, the outcomes of these crises – whether they ended peacefully or in conflict – were projected on indicators from earlier research.

Winter & Barenbaum (1985) found that the power motive (M) has a moderating effect on responsibility. In other words, responsibility determines, how vast amounts of power-motivated expressions are behaviorally enacted. If a high responsibility score is measurable,

---

[38]A *tweet* is a short message from the social network microblogging service Twitter (`https://www.twitter.com/`) and consists of up to 240 characters.

power-motivated individuals act pro-social. On the contrary, if the responsibility score is low, aggression and lack of leadership are to be expected.

Activity inhibition is reflected, according to by McClelland & Davis (1972) as the frequency of "not" and "-n't" contradictions in TAT or other verbal texts. Activity inhibition functions as motivational and emotional regulation. The authors identified a negative correlation between activity inhibition and male alcohol consumption. Combined with a high power motive (M) and low affiliation motive (A), subsequent research by McClelland and his colleagues revealed a so-called leadership motive pattern (LMP) (McClelland & Boyatzis, 1982; McClelland, 1988). The higher this LMP, the more responsible leaders act. As for *integrative complexity*, it was observed, that the lower the frequency of utterances in accordance with the 7-point score was (see Table 8.1), the more likely escalations became.

The utilized training and experimental data is described in the following section.

8.3   Data

For testing the proposed hypothesis, we first train a classification model and utilize this model for testing social network textual data. In this section, we will describe the two different data sources for training and the experiments.

8.3.1   Model Training Data

The data utilized for training models is similar to the data set utlized during the GermEval shared task 1 (see Section 7.5). The training set consists of 167,200 unique answers, given by 14,600 participants of the OMT (see Chapter 5. The training data set is imbalanced. The power motive (M) is the most frequent class, covering 41.02% of data points (Johannßen et al., 2020a). The second most frequent class, achievement (L) only accounts for 19.63% and thus is half as frequent as M. The training data was assembled and annotated by the University of Trier, reaching a pairwise annotator intraclass correlation of r = .85. With only 22 words on average per training instance (i.e. a participant's answer) and a standard deviation of 12 words, training a classifier on this data is a short text classification task (Johannßen et al., 2020a).[39]

8.3.2   Experimental Data

The experimental data was collected prior to this work by crawling the Twitter API and fetching 1 percent of the worldwide traffic of this social network (Gerlitz & Rieder, 2013). We sample posts over the time window from March to May of both, 2019 and 2020. There are no apparent

---

[39]The data can be retrieved via `https://www.inf.uni-hamburg.de/en/inst/ab/lt/resources/data/germeval-2020-cognitive-motive.html`

linguistic differences between the two samples. The average word count, part-of-speech (POS) tags, sentence length, etc. are comparable.

Thereafter, we extracted the *text* and *date time* fields of posts marked as German. From those files' hashtags, name references (starting with '@'), corrupted lines, and any posts shorter than three content words were removed. The resulting files for 2019 and 2020 contained more than 1 million instances. Lastly, the instances were randomly shuffled. We drew and persisted 5,000 instances per year for the experiments, as this data set size is large enough for producing statistically significant results. The posts on average consist of 11.97 (2019) and 11.8 (2020) words per sentence, and thus are very short. During the experiments, further pre-processing steps were undertaken, which are described in Section 8.4. By stretching out the data collection time window and by comparing the same periods in two subsequent years, we aim to reduce any bias effect that might impact Twitter user behavior over short periods, e.g. the weather, any sports event, or short-lived political affairs.

The proposed research methodology for measuring social unrest on the basis of the described social unrest predictors is described in the following section.

8.4 Methodology

For constructing a model of sufficient quality to test our hypothesis, we follow our previous work (Johannßen & Biemann, 2019) and train an LSTM combined with an attention mechanism (see Section 2.2).

We decided against additional features such as part of speech (POS) tags or LIWC features like in our previous work (Johannßen & Biemann, 2019), as we did not reach the best results with these additional features. The maximum token length was set to 20, as determined by our preliminary experiments (Johannßen & Biemann, 2019), and reflects the primacy rule of the implicit motive theory. The average answer length of the training data set was 22 tokens (see Section 8.3). With this decision to limit the considered tokens, we aim to closely replicate the implicit motive coding practices manually performed by trained psychologists (Kuhl & Scheffer, 1999). Accordingly, it is preferable to assign the 0 motive (i.e. no clear motive could be identified) than to falsely assign a motive that is not the very first one in the sequence.

The following section describes our conducted NLPsych social unrest experiments.

8.5 Experiments

Similar to previous experiments, this section is divided into the pre-processing and the training phase.

8.5.1 Pre-Processing

Some standard pre-processing steps were applied to reduce noise, which was to remove the Natural Language Toolkit (NLTK) German corpus stop words,[40] to lowercase the text, remove numbers, normalize special German letters (i.e. umlaute). Emojis were removed as well, since Twitter offers a selection of a 3,348 emojis,[41] that in turn mainly do not capture sufficient informational gain per textual answer for the task at hand. To remove stop words has to be an informed choice when it comes to performing NLP on psychological textual data. For example, function words are said to predict academic success (see Section 8.1). However, during our experiments, we saw an increase in model performance when stop words were removed.

After the training, we utilize the model on the two sampled data sets described in Subsection 8.3.2. Following the theories in Section 8.2, we investigate the frequency of the power motive with the self-regulatory level 4, which we expect to be higher. At the same time, we will also analyze the other motives and levels to see which ones are now less frequent and to what extent. Furthermore, we compare different linguistic features and statistics from 2019 to 2020 to see, if any of these show differences that might indicate possible biases in the data.

8.5.2 Training Phase

Our Bi-LSTM model was set to be trained within 3 epochs and with a batch size of 32 instances. The model was constructed having 3 hidden layers and utilized pre-trained fastText embeddings (Bojanowski et al., 2017), as this character-based or word fragment-based language representation has shown to be less prone to noisy data and words that have not been observed yet like e.g. spelling mistakes or slang – both often observable in social media data. The fastText embeddings had 300 dimensions and were trained on a Twitter corpus, ideally matching the task at hand.[42] Explorative experiments with different parameter combinations have shown that a drop-out rate of .3 and step width of .001 produced good results.

The cross-entropy loss was reduced rather quickly and oscillated at 1.1 when we stopped training early during the second epoch. After each epoch, the model was evaluated on a separate development test. After the training was finished, the model was tested once on the GermEval20 Task 1 test data and with the official evaluation script. This provides the chance to compare the achieved results with the best-participating team. Schäfer et al. (2020) achieved a macro F_1 score of 70.40, which our Bi-LSTM model was able to outperform with an F_1 score of 74.08, setting a new state of the art on this dataset.

Finally, we present our results in the following section.

---

[40] The Natural Language Toolkit (NLTK) is a collection of python libraries for NLP `https://www.nltk.org/`.

[41] `https://emojipedia.org/twitter/twemoji-13.0.1/`

[42] The fastText model was obtained from Spinningbytes at `http://spinningbytes.ch/resources/wordembeddings`

8.6 Results

After having trained the Bi-LSTM model and sampled the experimental data, we will describe the results and findings of the conducted Twitter COVID-19 experiments in this section. An overview of all results is displayed in Table 8.2.

To investigate the main predictor for social unrest *activity inhibition* (see Section 8.2), the power motive (M) in combination with level 4 was counted. The self-regulatory level 4 describes the sensitivity for negative incentives (see Chapter 5). These measures are collected for all four data sets. Our Bi-LSTM model assigned power 4 in 33.76% of all cases for the Twitter sample from March to May of 2019, making this the most frequent label. However, for the data sample from 2020, power 4 is as frequent as 37.4%, making this an increase of 10.97%. For calculating the significance of this rise, we perform a t-test on the label confidences for the power motive with self-regulatory level 4 for both, 2019 and 2020 with the 5,000 samples from each year (see Section 8.3).

The two-sample t-test on the confidence levels shows, that the rise in frequency is statistically significant ($p < .05$ with $\bar{x}_1 = .27$, $\bar{x}_2 = .29$, $\sigma_1 = .28$, $\sigma_2 = .28$, $N_1 = 5,000$ and $N_2 = 5,000$).

The affiliation motive (A) is barely classified, covering only 2% (2019) and 1.89% (2020) of all instances. The slight decrease is not statistically significant ($p > .05$). The frequency of self-regulatory level 4 is elevated by 6.7%. The whole of all assigned power motive labels has only risen by 3.64%, both having risen less than the combination of the power motive and level 4 combined. The strongest decline in frequency can be measured for the freedom motive with -12.63%. The other motives of affiliation, achievement, and the null motive have barely changed in comparison to 2019 with 2020. The same holds for the average amounts of words per sentence, verbs, adjectives, and words containing at least 6 letters, all of which have barely changed, not indicating sampling biases. An overview of the class frequencies is provided in Table 8.3.

Since both, responsibility and integrative complexity can only be measured by employing a specific TAT and a questionnaire, which would have to be performed with each Twitter user, we can only investigate activity inhibition as a combination of the power motive with the self-regulatory level 4. However, we will review some psychological LIWC categories, that follow a close description as the five categories of Winter's responsibility scoring system (Winter & Barenbaum, 1985). Relevant LIWC categories for the responsibility is the combination of *family*, which are terms connected to expressions like 'son' or 'brother', and *insight*, which contain expressions such as 'think' or 'know', representing self-aware introspection. Family shows a significant decrease from 2019 (.08) to 2020 (.05) of -37.5%. The frequency of insight terms fell from 2019 (.23) to 2020 (.17) by -26%, all of which are statistically significant changes ($p < .05$ for both categories).

8.6.1   Discussion

We hypothesized that the social unrest predictors by Winter (2007), namely *activity inhibition*, *responsibility*, and *integrative complexity* are automatable and reveal changes in natural language and signs of social unrest observable through the use of social media textual data connected to the COVID-19 pandemic.

The main research objective of this work is to find novel approaches to automatically provide the community with red flags for growing tensions and signs of social unrest via social media textual data. For this, *activity inhibition* is the main predictor. It consists of a distinct shift in implicit motives. It is present when the frequency of the power motive with the self-regulating level 4 (sensitivity for negative incentives, see Chapter 5) is elevated and the affiliation motive is suppressed – even though Winter (2007) did not find clear evidence of the latter. The comparable rise by 10.97% ($p < .01$) is an indicator of the social tension of COVID-19 related social media posts.

Since other linguistic statistics, such as the average amounts of adjectives, verbs, words per sentence, or words containing at least six letters have barely changed, this indicates that the measurable differences in social unrest predictors are content-based and not caused by linguistic biases (see Section 3.2).

It is remarkable, that whilst the power motive has been labeled more frequently, the frequency of the labeled freedom motive has declined by -12.63% from 2019 to 2020. This freedom motive has barely been researched yet but has a close connection to the power motive. Whilst power-motivated individuals desire control over their fellow humans and their direct surrounding for the sake of control, freedom-motivated individuals seek to express themselves and want to avoid any restraining factors. Motives are said to be rather stable but can change over time (Schultheiss & Brunstein, 2010). This change in motive direction could indicate a roughening of verbal textual content and interpersonal communication. Example utterances classified as freedom and power from 2019 compared with 2020 are displayed in Listings 8.1 for 2019 and 8.2 for 2020.

The change of responsibility, as reflected in LIWC categories, retreated by roughly 30% from 2019 to 2020. This responsibility indicates a personal involvement in topics and decisions, that we feel are relevant to our surroundings. If this involvement diminishes, our interest in participating in constructive solutions to problems does as well.

```
M 'RT @FrauLavendel: ist es wahr
  dass schulleitungen den
  schüler*innen drohen
F RT @UteWeber: Nach einem relativ
  unfeierlichen,
  regionalen Offline-Tag aufs Sofa
  sinken, wie von der Tarantel
  gestochen aufspringen und zur...
```

```
A  Weltbestseller  ”P.S.  Ich  liebe
   dich”  bekommt  einen  zweiten  Teil
   https://t.co/9Ifl5CrNAP
   −−−−−−−  Translation  −−−−−
M  ’RT  @FrauLavendel:  is  it  true  that
   principals  threatens  students
F  RT  @UteWeber:  after  a  relatively
   un−celebrational ,  regional  offline
   day ,  as  bitten  by  a  tarantula
   jumping  up
A  world  best−selling  book  ”P.S.  Ich
   liebe  dich”  gets  a  second  part
   https://t.co/9Ifl5CrNAP
```

Listing 8.1: Examples of German tweets with corresponding translations from spring of 2019 (March to May.

```
M  Corona−Regeln  im  Saarland  sind  zum
   Teil  absurd  und  unverhältnismäßig
F  RT  @kattascha:  In  den  USA  bekommen
   viele  Menschen  keine  Lohnfortzahlung
   im  Krankheitsfall .  Das  bedeutet :
   Selbst  bei  Verdacht  auf  #COVID19  w...
A  Wer  einen  Discord−Server  sucht ,
   um  entspannt  mit  seinen  Kollegen
   zu  zocken  oder  gemeinsam  abzuhängen
   ist  hier  genau ...
   −−−−−−−  Translation  −−−−−
M  the  Corona  rules  for  the  Saarland
   are  partially  absurd  and  dis−
   proportionate
F  RT  @kattascha:  in  the  US  a  lot
   of  people  don’t  receive  continued
   pay  in  case  of  illness .  That  means:
   even  in  case  of  suspected  #COVID19
A  Whoever  is  looking  for  a  Discord
   server  for  enjoyably  game  with
   their  colleagues  or  chill  together ,
   is  in  the  right  place ...
```

Listing 8.2: Examples of German tweets with corresponding translations from spring of 2020 (march to may.

8.6.2   Conclusion and Outlook

With this work, we conducted a first attempt at automating psychometrics for investigating social unrest in social media textual data. The Bi-LSTM model combined with an attention mechanism of this work achieved an F_1 score of 74.08 on 30 target classes, making it state of the art on a respective recent shared task dataset. With this model, we measured a statistically significant rise in the power motive with self-regulating level 4, which reflects the social unrest predictor of *activity inhibition* in the direct comparison of the samples from March to May of 2019 vs. 2020.

Furthermore, we investigated *responsibility*, which shows significant reductions during the COVID-19 pandemic, hinting at negative outcomes of interpersonal and verbal communication on the social media platform Twitter.

This first approach most likely does not qualify for a real-world social prediction system. Predictions of such a system can not yet be reliable enough for deriving necessary actions from them. On the upside, implicit motives do not only qualify for examining general socio-economic tensions but can be applied on an individual or small group scale. As an example, detecting tensions within a small group can help to shape the group and guide it into a better fit. Furthermore, we advocate for combining implicit motives with sufficiently many complementary psychometrics and content-based analysis e.g. sentiment analysis, topic modeling, or emotion detection.

Besides those combinations with other information sources for future work, different sampling approaches and larger data set sizes should be utilized for reproducing findings and research correlations with other social unrest predictors and indicators. In this work, we have made the first steps toward understanding the automation of psychological findings. Since only 5,000 samples were drawn from a single social network platform, we advocate for broadening this approach to include many more samples from a wider time window paired with mixing the data sources. In addition to that, deeper investigations into the linguistic variances between times of so-called social unrest and more peaceful times should be performed, as those could reveal patterns and characteristics of time-specific utterances.

The observed correlations and social unrest patterns are in line with an intuitive assumption of how language in social media data changes amid a pandemic. Future work arises in the application of this methodology to other events and crises, eventually providing a quantitative basis for implicit motive research.

We believe that it is also important to discuss the limitations of our work, in addition to its strengths presented in this chapter. The following section offers clear discussion of limitations.

8.7   Limitations

This chapter's methodological approach is only as good as the underlaying prior works by Winter & Barenbaum (1985) and Winter (2007). For these prio works, the authors mainly analyzed political speeches, which differ greatly from tweets. Whilst the authors of the previous works employed many observable categories such as integretive complexity, our work limited itself to only activity inhibition and the leadership motive pattern.

Furthermore, our empirical approach for measuring social unrest during the COVID-19 pandemic has been performed in hindsight. We knew that the panemic struck globally and we also knew that individuals broadly felt agitated. What we did not provide is an assessment of conflict potential before a major event and thus predicting social unrest without the prior knowledge.

In addition, the authors of the prior psychological works did not employ the OMT, but other implicit measures. Futerhermore, the authors investigated the political speeches manually and not automatically. Whether or not our approach reaches the qualities of human experts, is another limitation of this work and should be viewed critially.

The sampling of the validation data is another limitation. It is not trivial to sample data from social media plattforms in a way that it does not introduce any biases or sampling errors. Especially over a rather short period of few months larger topics can have an effect on the sampling contents. Twitter does not offer many detailed information on e.g. demographics. We furthermore did not perform intensive investigations of individuals' tweet history, possibly revealing more information on the trustworthyness. In addition, it is difficult to detect whether or to which extend tweets emerged from so-called bot, meaning programs and not humans.

Lastly, same with the previous empirical work on aptitude diagnostics, the utilization of psychometric models on data associated with mass phenomena observable through social media plattforms raises multiple ethical concerns. These concerns and an assessment of their implications can be found in Section 3.3.

The third and last empirical research conducted during the dissertation project is described in the following chapter. It is concerned with pandemical social isolation, caused by curfues and lockdowns. This research is closely connected to the Jungian psychology types described in Chapter 4 and utilizes the established NLPsych approaches described in Section 2.5, and this chapter.

Category	Measure	Example or Explaination
Responsibility		
i) moral standards	observable, if people, actions, or things are described with either morality or legality	'she wants to do the right thing'
ii) obligation	means, that a character in a story is obliged to act because of a rule or regulation	'he broke a rule'
iii) concern for others	emerges, when a character helps or intends to help others or when sympathy is shown or thought	'the boss will understand the problem and will give the worker a raise'
iv) concerns about consequences	can be identified when a character is anxious or reflects upon negative outcomes	'the captain is hesitant to let the man on board, because of his instructions'
v) self-judgment	scores when a character critically judges his or her value, morals, wisdom, self-control, etc. and has to be intrinsic	'the young man realizes he has done wrong'
Activity inhibition		
linguistic negation	in English terms, the authors describe activity inhibition as the frequency of "not" and "-n't"	responsibility measure, e.g. a variable negatively correlated with male alcohol consumption
leadership motive pattern (LMP)	combined with a high power motive (M) and low affiliation motive (A)	predicts responsible leadership power behaviors instead of profligate impulsive expressions of power
Integrative complexity		
7-point continuum range score from simplicity to	1: no sign of conceptual and differentiation or integration can be observed	only one solution is considered to be legitimate
differentiation and integration	7: overreaching	viewpoints are expressed, involving different relationships between alternate perspectives

Table 8.1: According to Winter (2007), some distinct psychometrics and their combinations predict social unrest – namely responsibility, activity inhibition, and integrative complexity. The table shows their categories, and measurements and offers examples or explanations.

Metric	2019	2020	Percentage delta	Significance
Activity inhibition and responsibility				
Power 4	33.76	37.40	10.97	p<.01***
LIWC Family	.08	.05	-37.60	p<.05*
LIWC insight	.23	.17	-26.09	p<.05*
Implicit motives				
Power motive	65.84	68.24	3.64	p<.01***
Freedom motive	20.28	17.72	-12.63	p<.01***
Achievement motive	6.80	7.00	2.94	p>.05
Affiliation motive	2.00	1.86	-7.00	p>.05
Null motive	5.10	5.10	.00	p>.05
Self-regulatory levels				
Level 1	6.50	6.01	-7.54	p>.05
Level 2	2.76	3.26	18.12	p>.05
Level 3	27.20	25.58	-5.96	p<.01***
Level 4	42.78	45.20	5.67	p<.01***
Level 5	15.86	14.92	-5.93	p<.05*
Linguistic statistics				
Average words	11.97	11.80	-1.42	p>.05
Verbs	1.19	1.22	2.52	p>.05
Adjectives	.43	.43	.00	p>.05
Words >six letters	38.65	38.86	.54	p>.05

Table 8.2: Overview of the different psychometric and statistical results. * represents significant results, *** represents highly significant results. All combinations of motives and levels have been examined.

Implicit motive	Frequency	levels	Frequency
		2019	
Power	3,251	1	492
Affiliation	141	2	193
Achievement	414	3	1,487
Freedom	9622	4	1,872
Zero	232	5	724
		0	232
		2020	
Power	3,433	1	316
Affiliation	90	2	151
Achievement	203	3	1,259
Freedom	923	4	2,233
Zero	761	5	780
		0	261

Table 8.3: Overview of the class frequencies.

Chapter 9

NLPsych for Measuring Signs of Distress

This chapter describes the empirical research on pandemic-related social isolation. As with the previous chapters, it is structured in the Sections on nieche related work in Section 9.1, the utilized data in Section 9.3, the research methodology in Section 9.3, and finally the results in Section 9.4. Limitations are discussed in Section 7.6.

The first cases of individuals reportedly being infected with the SARS-CoV-2 or COVID-19 virus appeared in December of 2019. Ever since, a global pandemic of this highly infectious disease has emerged, which has been met with countermeasures. Those countermeasures include social distancing and temporary lockdowns (Balasa, 2020). Governments stand in the dichotomy of restricting social and public interactions as a measure of safety and risking the mental health of the people affected, as reports of declining mental well-being emerge (Hämmig, 2019).

Even though professional mental consultation and support do exist, it is difficult to identify and contact heavily impacted individuals (Lester & Howe, 2008). The direct approach would not be feasible, as it would tie up the capacities of mental health workers. Broad information campaigns might cause high costs and still not reach individuals in need. Lastly, affected people might not even be aware of their mental health risks and thus not reach out to available mental health consultations. Depression detection systems or even sentiment analyses of e.g. social media posts could potentially support mental health workers (Coppersmith et al., 2018). But those systems often rely on sufficient self-reports or on topics of mental health or loneliness being directly discussed, which require the individuals to already self-reflect and openly discuss their well-being, resp. the decline thereof (Zirikly et al., 2019).

Furthermore, the well-established safety net of e.g. educational facilities, whose staff could identify troubled individuals, can be unavailable due to the lockdown restrictions. Thus, it might be worthwhile to explore alternative and ideally automated approaches.

Mental health detection often focuses on introverts due to their self-inflicted distancing and more frequent occurrence of signs of depression compared with extraverts. Recent empirical

research on the effects of the pandemic confirms those findings (Wei, 2020). Other findings, however, contradict those results and report empirical findings of extraverts' suffering to be comparably worse (Wijngaards et al., 2020).

As with many psychometrics, manual assessment of psychology types can be costly (Johannßen et al., 2019). Furthermore, burdened individuals might not be reachable by broadly conducted surveys amongst a population. Thus, automation of those types with a focus on introverts and extraverts might reveal the additional potential for identifying individuals in need of support.

Therefore, we aim to classify the Jungian psychological types of *extraversion* and *introversion* (see Chapter 4) from German text and to apply such a model to utterances in 2019 compared with 2020 to investigate whether there are noteworthy well-being differences.

In this chapter, we will first discuss related work to automated psychometrics, depression detection, and some psychometrics in Section 9.1. The implicit personality test (IPT) utilized in this work is described in Chapter 5. The description of the dataset for training neural models and for identifying anxious individuals is described in Section 8.3. Section 9.3 discusses the methodology and approach. The results will be presented in Section 9.4 and will be discussed in Section 9.4.2. We conclude our findings in Section 9.4.3 and discuss future outlooks.

9.1 Related Work on Personality Assessment and Pandemical Isolation

The automated assessment of personality or personality traits is a rather recent application domain. Whilst earlier approaches relied more heavily on rule-based systems, themselves mostly divided into wordlist-based versus corpus-induced methods (Johannßen & Biemann, 2018), machine learning has become more widely utilized in recent years (Mehta et al., 2019). Accordingly, the MBTI and the five-factor model of personality (also called *Big Five*, (Goldberg, 1993)) have been (Angleitner, 1991) and are amongst the most widely utilized personality tests, both of which rely on the Jungian psychological typologies (see Chapter 4).

Jungian types have successfully been classified from natural language texts by employing a BERT model by Keh & Cheng (2019). For training their model, the authors scraped data from a self-reporting web forum. The resulting model was utilized for generating personality-induced natural language texts.

The effects of the COVID-19 pandemic have been researched extensively during its outbreak at the end of 2019. Johannßen & Biemann (2020) analyzed social unrest indicators on the application of the pandemic and found that an increase of an implicit motive *power* paired with a self-regulatory passive coping with fears were correlated with signs of crises.

Empirical research on the impacts of the COVID-19 pandemic on introverts and extraverts are somewhat contradictory. Whilst some recent works found extraverts to be more in danger

of mental health degradation (Wijngaards et al., 2020; Gubler et al., 2020), other works come to the opposite conclusion (Wei, 2020).

The next section describes the utilized data.

9.2 Data

Since manually asserting natural language texts on introversion or extraversion is costly and would not be scalable, we will first train a neural model (see Section 9.3) on the data described in this section. We collected German natural language textual data before and from the COVID-19 pandemic and apply said model to this data set. Furthermore, we train in-domain Twitter models.

Model training data

The German natural language textual data utilized for creating the model was collected by a company specialized in aptitude diagnostical testing[43] and is being made public for free use and validation.[44] 2,680 textual answers to provided projection imagery were given by 335 individuals. The population was drawn from the workforce with ages ranging from 18 to 65. Further demographic information was omitted under German data protection laws. The data has been split by separating participants into training (~90%, n=2,360), development, and held-out testing data sets (~5%, n=160 each). Since all 8 answers per participant remained in a data set without being shuffled and separated, we aim to increase the generalization of the model (i.e. rather training to learn the target label and not perform speaker identification). The distribution of answers labeled as extraversion is displayed in Table 9.1. The two labels are distributed unevenly with the vast majority being extraversion (67.4% of all labels with comparable distributions overall data sets). Answers consist of an average of 42 words and thus can be considered short texts. Each answer has been manually labeled with the four typology pairs. Compared to data sources like Twitter, the training data is rather clean without a lot of noise such as spelling mistakes, spam, or unusual characters. The Kohen's Cappa measure for annotator agreement on the task of extraversion and introversion scores $K = .47$ – only *moderate agreement* (McHugh, 2012).

# extra	8	7	6	5	4	3	2	1	0
%	9.7	22.0	21.4	17.3	13.5	8.5	5.9	1.5	.3

Table 9.1: Distribution of answers labeled as extraversion in the training material. The upper row displays the counts of answers labeled as extraversion per participant (8 answers in total), the lower row displays the corresponding percentages.

---

[43]WafM Wirtschaftsakademie GmbH `https://www.wafm.de/`.

[44]`https://www.inf.uni-hamburg.de/en/inst/ab/lt/resources/data/ipt-introextra-2022.html`.

Experimental data

One goal of this work is to research transferability across different data domains, namely from the IPT to tweets. Before utilizing any model for validation purposes on tweets, we first need to measure transferability. For this validation data, we sampled 1,100 tweets from a corpus described hereafter, and had them manually labeled by experts on extraversion and introversion. This costly data is also made available.[45]

Validation data

The experimental data was drawn from Twitter,[46] a micro-messaging service. The service offers an API for downloading 1% of the worldwide traffic of the social network (Gerlitz & Rieder, 2013). Since the goal of this research is to find new ways of identifying individuals in need during the COVID-19 pandemic, we crawled the Twitter API for the period from March to May 2019 and from March to May 2020. Linguistically, the samples are comparably similar (e.g. equal average lengths, equal part-of-speech (POS) tags, sentence lengths, etc).

The crawled instances were filtered by a German flag to only include posts from German individuals. Furthermore, we filtered non-German samples via language detection (Google translate python library[47]). Besides the texts themselves, the field *date time* was included, which functions both as an identifier hence the inclusion of milliseconds, and as an inclusion criterion for the experimental setup. In total, 10,000 instances were sampled, 5,000 per time period (2019, 2020). An answer from 2019 contains 19.77 words on average and 19.76 from 2020, which makes this a short-text classification task. Bias effects have to be assumed when comparing two different time periods. We aimed to reduce this bias by spreading the selection period over three months, hence selective topics like sports, weather, or cultural events should not overshadow the overreaching effects the pandemic might have.

Our proposed research methodology is described in the following section.

9.3 Methodology

In this methodology section, we propose a two-stage approach to asserting domain transferability, describe two employed model architectures, and present the experimental setup.

Two-stage approach

Since there is a considerable difference in labeled data quality and availability between the training data from the IPT and the experimental validation data from Twitter, and since it can be assumed that domain transferability does not produce convincing results, we propose two consecutive experimental stages: i) first, we will train two models from previous experiments (Johannßen et al., 2019; Johannßen & Biemann, 2020) on the IPT data set and validate them

---

[45]https://www.inf.uni-hamburg.de/en/inst/ab/lt/resources/data/ipt-introextra-2022.html.

[46]Twitter https://www.twitter.com

[47]https://pypi.org/project/googletrans/

on the Twitter dataset, and ii) secondly, we will train those models directly on the Twitter validation set. We critically evaluate transferability and validation applicability, as it is often aspired when performing NLP on psychological textual data (Štajner & Yenikent, 2021; Plank & Hovy, 2015).

Bi-LSTM attention Model

Previous work on German natural language textual data with a focus on psychological measures have resulted in a viable model, which has reached state-of-the-art results on a shared task dataset and is being utilized for this work as well (Johannßen et al., 2019; Johannßen & Biemann, 2020).

The first model is an LSTM combined with an attention mechanism (see Section 2.2).

The model is constructed with 5 layers (1 input, 3 hidden, 1 output) and contains 256 units in each hidden layer. Input tokens are represented by 300-dimensional FastText embeddings, pre-trained on *Common Crawl*[48] and *Wikipedia*[49] (Grave et al., 2018). As optimizer we chose Adam (Kingma & Ba, 2015) and the loss was calculated via cross-entropy. Training parameters were set to a step-width of 1e-6, a dropout rate of .5, and mini-batch training of size 32 in 50 epochs.

Logistic Model Tree (LMT) Model

Since previous approaches Johannßen et al. (2019) have shown strong results from trained LMTs (see Subsection 2.2.3) on small datasets, we trained an LMT as a second model to be considered. We performed feature engineering but opted for two different sets of hand-crafted features: one set of features for modeling the IPT and one set of features for modeling the same task on tweets directly.

IPT LMT: As described in our previous work (Johannßen et al., 2019), for firstly engineering the IPT features, the texts mostly were tokenized and processed per token. Engineered features were the type-token ratio, the ratio of spelling mistakes, and frequencies between 3 and 10 appearances. Further features are LIWC and language model perplexities (see Subsection 2.3.2). Part-of-speech (POS) tags were assigned to each token and thereafter counted and normalized to form a token ratio. We trained a POS tagger via the natural language toolkit (NLTK) on the TIGER corpus, assembled by Brants et al. (2004) and utilizing the STTS tagset, containing 54 individual POS tags.

We trained a bigram language model (see Subsection 2.3.2) for each class and incorporated Good-Turing smoothing for calculating the perplexity. During training, we tuned parameters (e.g. which smoothing to use) via development set and tested the model with the held-out test set.

Twitter model: Secondly, we engineered features for the same task on the labeled Twitter data directly. For the class extraversion, the most influential tasks reflected upon stimulus *from the outside*, such as many add symbols (@) and hashtags (#), plural forms, and plural pronouns.

---

[48]Common Crawl, `https://commoncrawl.org/`.

[49]Wikipedia, `https://www.wikipedia.org/`.

Furthermore, multiple exclamation marks (often used by German speakers to emphasize and *shout*), instances written in all caps, and emojis indicate extraversion in tweets. As for introversion, mostly the opposite features indicate the class: only a few emojis, exclamation marks, hashtags, or add symbols. Singular forms and singular pronouns indicate introversion, as well as lowercased tokens (unusual in German, since common and proper nouns are spelled with an initial uppercase).

Pre-processing

Since additional features did not enhance the model's performance metrics in preliminary experiments, we decided against adding any (e.g. POS tags, spelling mistakes, or LIWC). We follow the pre-processing steps by Johannßen & Biemann (2020) by removing stop-words, numbers, emojis, or Twitter-typical special characters, as well as auto-correcting spelling mistakes. 1,000 remaining pre-processed tweets were drawn.

Experimental Setup

There are contradictory empirical findings on whether introverts or extraverts are more mentally challenged during the pandemic. To investigate this contradiction, we collected data from 2019 and 2020, as described in Section 9.2. The proposed models (see Section 9.3) will be trained on the task of classifying extraverts and introverts by their use of natural textual language and will thereafter be utilized for classifying labels to the tweets from 2019 and 2020. Finally, we will divide extraverts and introverts of both years and investigate their linguistic tone and mood. This investigation will be performed by the use of LIWC. From those LIWC category word percentages, we will investigate, whether the tone of extraverts and introverts has significantly changed and in which way.

The following section presents the experimental and analytical results.

9.4 Results

The results are divided into model benchmarks, the IPT model validation via Twitter data, an in-domain validation, and finally an error analysis.

Model benchmarks

Firstly, we performed benchmarks to confirm our model choices. The Benchmarks displayed in Table 9.2 have shown that the proposed Bi-LSTM model with attention mechanism achieves the best results on this classification task, even outperforming a BERT base model (see Subsection 2.3.4). It can be assumed that BERT base fails to capture the task due to little training data and diverging content meanings compared with everyday use of language (Ezen-Can, 2020).

IPT model performances and Twitter validation

The confusion matrix of the IPT Bi-LSTM is displayed in Table 9.4. The current state-of-the-art (SOTA) approach for classifying English introversion and extraversion by Plank & HovyPlank

Model	Accuracy	Precision	Recall	$F1$ Score
BERT base	.70	.49	.70	.58
CNN	.72	.70	.72	64
LMT + features	.66	.65	.66	.65
RNN	.66	.64	.66	.65
Self attention	.68	.71	.68	.69
LSTM	**.73**	.70	**.73**	.69
Bi-LSTM attn.	.71	**.73**	.71	**.72**

Table 9.2: Benchmark performances of different model architectures. The proposed Bi-LSTM model with attention mechanism achieves the highest $F1$ score. Whilst oftentimes BERT outperforms other architectures, the employed BERT base might fail to capture the signals.

& Hovy (2015) scores F_1 = .72. Even though those scores are not comparable due to the different languages and corpora, the proposed model nonetheless achieves comparable results with F_1 = .72 on the task with German textual data. The performance of the IPT LMT model is slightly worse than the performance of the Bi-LSTM attention model with F_1 = .69 with perplexity (and thus introversion/extraversion bigram language models) being the discriminating feature on its root node.

Model	Bi-LSTM att.	LMT
Precision	.736	.693
Recall	.7125	.685
F-Measure	.7203	.689

Table 9.3: Displayed are the Bi-LSTM attention model and LMT model performance measures of precision, recall, and the F-measure for the task of classifying the Jungian psychology types of extraversion and introversion.

		Predicted		
		Extra	**Intro**	Σ
Actual	**Extra**	83	29	112
	Intro	17	31	48
	Σ	100	60	160

Table 9.4: The confusion matrix of the Bi-LSTM attention model on the IPT classification task test set.

Despite the proposed Bi-LSTM model scoring well on the held-out test IPT dataset, does not validate well on the experimental Twitter dataset. When utilizing this model on a held-out test set (n = 160) of the 1,000 hand-labeled tweets and measuring its performance, the model

scores $F_1 = .5$, indicating uninformed decisions based on chance. The same can be observed for the proposed IPT LMT model, which scores an even worse $F_1 = .3$, rendering it unapplicable for cross-domain tweet classification.

In-domain Twitter model and validation

The proposed Bi-LSTM model with attention mechanism fails to capture the aspects of introversion and extraversion from the small Twitter dataset. The model scores a mere $F_1 = .4$ on the Twitter held-out test set and thus is not applicable for being utilized for any further predictions.

In contrast to the Bi-LSTM model, the feature engineered and in-domain trained LMT Twitter model achieves good results on the held-out Twitter test set with $F_1 = .69$. The LMT model's confusion matrix is displayed in Table 9.5, showing that the model performs sufficiently well on both classes and especially introversion, which seems to be harder to model in general (Štajner & Yenikent, 2021). Influential features include the POS tags KOUI, PPOSAT, VAPP, and pronouns, as well as LIWC categories Other, Past, School, and Physical. Lastly, frequencies of exclamation marks, hashtags, emojis, and add tags.

From those results, we can conclude that the out-of-domain transferability between IPT models and tweets does not validate. The Bi-LSTM model performs well on the IPT but fails when being trained directly on the Twitter dataset. The LMT IPT model performs slightly worse. When training a feature-engineered LMT directly on tweets, it performs sufficiently. Hereafter, we will only discuss the IPT Bi-LSTM and Twitter LMT. Additionally, we will utilize the Twitter LMT for further validation studies on the Covid-19 validation dataset described in Section 8.3.

| | Predicted | | |
	Extra	**Intro**	Σ
Extra	37	21	58
Intro	13	37	50
Σ	50	58	108

(row labels grouped under *Actual*)

Table 9.5: The confusion matrix of the LMT model on the Twitter data test set.

Error Analysis

The employed attention mechanism at least partially allows for the investigation of the algorithmic importance of single input tokens for the IPT Bi-LSTM classification task at hand. As Jain & Wallace (2019) point out, the distribution of attention weight mass does not necessarily correspond to the underlying theories of the task at hand. However, in earlier work, we have explored the attention weights of the proposed model in more depth and found them to be in line with implicit test theory (anonymous reference). With the limitations and the possibility of some explainability in mind, we present the attention weight mass during the training phase

in Table 9.6. Those tokens with higher mass indeed appear to correspond with the psychological theory of introversion and extraversion. In those examples, calmness is rather associated with introversion and togetherness rather than extraversion.

verwenden	erschaffen	ruhe	arbeit	vertieft	
use	create	calm	work	being absorbed	intro
gemeinsam	ideen	nachbar	vertrauen	gedicht	
together	ideas	neighbour	trust	poem	extra

Table 9.6: Visualization of the attention weight mass per German token with corresponding translations during the training phase. The tokens that received the highest mass do correspond with the psychological theory of extroversion vs. introversion.

The errors made by the IPT Bi-LSTM attention model are displayed in Table 9.7. Very short and uncontextualized answers were more often mistaken by the model and classified incorrectly. Furthermore, instances that require broader world knowledge (e.g. holding a rope being equivalent to team mountaineering) were misclassified.

Label	Text	Pred.
E	King kills; kills; drill in his hand	I
E	Hears his volleagues; to understand everything	I
I	Persons climbing; secures rope; in focus; reaction	E
I	sees landscape; holds rope; feels responsible	E

Table 9.7: Errors made by the Bi-LSTM attention model. Apparently, short answers and those that require broader world knowledge were difficult to model. The labels read E for Extraversion and I for Introversion.

The LMT Twitter model made similar mistakes as the IPT Bi-LSTM model, which indicates, that despite the data sources being different (IPT vs. tweets), there are overreaching linguistic challenges when attempting to model the task of classifying Jungian introversion and extraversion. Once again, short and noisy instances are prone to being misclassified, as well as those instances, which require world knowledge. This is in line with the findings from Štajner & Yenikent (2021) on why the MBTI (including introversion and extraversion) is difficult to model.

9.4.1 Validation study: Twitter LMT Model & LIWC categories

The most precise method of identifying individuals in need of support would either be self-reports or medical diagnoses made by trained physicians. Both information are sparse and those individuals with the most severe threat of mental suffering oftentimes do not self-report their struggling or visit facilities. With limited information, we aim to determine whether classifications of introversion and extraversion differentiate the observed tweets not only into

those two psychological types, but also into groups that are challenged by the pandemic at different levels.

As described in Section 9.3, we utilize the psychological dictionary tool LIWC. Table 9.8 displays those results. Six LIWC categories were investigated that correspond to mental health and the social background (Pennebaker et al., 2007). Those are *inhibition positive feeling*, *insight*, *anxiety*, *sad*, *sex* and *eat*.

Table 9.8 is divided into three table paragraphs. The first displays tweets classified as introversion from 2019 compared with 2020. The second table paragraph displays tweets classified as extraversion, and the third table paragraph compares the whole instance data set without this introversion/extraversion differentiation in order to provide a comparison point (whether those changes are specific for either of the two psychological types or are present in the entire data set).

Even though we investigated the changes from 2019 compared with 2020 a confounding analysis showed differences in LIWC categories between extraversion and introversion in multiple categories, including those in Table 9.8, indicating an unrecognized explanatory variable.

		Inhibition	Positive feeling	Insight	Anxiety	Sad	Sex	Eat
Introversion	'19	.27	.20	1.35	.12	.34	.33	.13
	'20	.31	.21	1.71	.20	.28	.25	.09
	Δ	.04	.24	.36	.08	-.06	-.09	-.04
	%	12.4	3.7	22.1	40.3	-21.8	-35.0	-56.7
Extraversion	'19	.29	.21	1.54	.13	.31	.24	.13
	'20	.27	.27	1.57	.12	.37	.35	.17
	Δ	-.02	.06	-.03	-.01	-.07	.10	.04
	%	-7.1	24.2	3.0	-9.8	18.9	30.1	26.7
Control	'19	.28	.21	1.42	.12	.33	.30	.13
	'20	.29	.23	1.65	.16	.32	.29	.12
	Δ	.01	.04	.23	.04	.01	-.01	.01
	%	5.2	13.3	14.9	26.2	-2.7	-3.3	-8.6

Table 9.8: The first table paragraph displays psychological LIWC categories per instance with noticeable fluctuations from 2019 compared with 2020. The second table paragraph displays the corresponding LIWC categories for extraversion predictions.

Table 9.8 shows some fundamental differences between the groups of tweets classified as introverted and extraverted. Accordingly, *inhibition* declined for rose by 12%, whilst having

increased by 7% for extraverts. While *positive feelings* barely changed for introverts, they increased by 24% for extraverted. Insight was greatly increased for introverts (+ 22%). The big difference occurs for anxiety, which sharply increased by 40% for individuals classified as introverts, whilst having declined roughly 10% for extraverted instances.

Noteworthy, *sad* did increase for extraverts (+19%), whilst having decreased for introverts (-22%). The category includes utterances such as crying, grief, or sadness. Instance examinations showed that instances high in *sadness* mostly read 'i miss you' or missing someone or something.

The social factors of *sex* and *eat* (being physical closeness and topics such as restaurants, dining, etc.) further differentiate those two groups by having decreased for introverts (-35% and -57%), whilst being increased in its frequency for instances classified as extraversion (+30% and +27%).

Needless to say, neither the attention weights, the binary classifications, nor the LIWC psychological categories can assert the individual's state of mind for certain. Nonetheless, they can serve as indicators. Following, we will discuss those findings, put them into relation to the pandemic, and will discuss the current research on this topic from Section 9.1 with regard to those findings.

9.4.2   Discussion

As shown in Section 9.4, the proposed IPT Bi-LSTM model reaches comparably strong performances on the binary classification task between introversion and extraversion. The attention weights during training as displayed in Table 9.6 appear to be aligned with the theory of Jungian psychology types. For tweets, an in-domain LMT was trained.

The results in Table 9.8 add novel findings to the current discussion. Whilst introverts expressed fewer optimistic utterances, those worries did not increase for extraverts. Rather than that, negative emotions rose sharply for introverts, which can be interpreted as clear signs of worry. Anxiety generally increased but slightly more for introverts. Noteworthy, sadness increased for extraverts. But as single instance observations reveal, instances high in *sadness* mostly miss persons or e.g. restaurants. This direction of energy towards the outside suits extraversion and would explain this rather negative emotion being increased for extraverts. The last two observed LIWC categories with remarkable changes from 2019 compared with 2020 are of social relevance (*sex* and *eat*). Firstly, utterances associated with physical closeness are less frequent for introverts, whilst being by far more frequent for extraverts. Utterances associated with dining, eating, or visiting restaurants decreased for introverts, whilst being increased for extraverts. This, again, suits the understanding of Jungian extraversion (see Chapter 4).

Extraversion has been interpreted as sensitivity to positive affect and optimism, and introversion, on the other hand, as lacking sensitivity to positive affect and pessimism (Watson

& Clark, 1997; Watson & Tellegen, 1985). Positive affect (i.e. extraversion) is crucial in times of crisis to see the broader picture, cope with depressive thoughts and ruminations, and stay action-oriented. Introverts, which lack this disposition to experience positive affect tend to be "state-oriented" and even depressed, especially in times of crisis (Kuhl & Kazén, 1999). This could explain the higher frequencies of negative emotions in the tweets.

All of those characteristics are unfavorable during lockdowns or other inclined types of isolations and social distancing. Those findings are supported by current empirical research, such as conducted by Wei (2020), who also found introverts to be rather inclined to suffer during the pandemic.

9.4.3 Conclusion & Outlook

The Corona or COVID-19 pandemic can be described as an event of a century. Many governments have resorted to measurements of social distancing or lockdowns. Even though those measurements save lives and help to fight this menacing disease, it also burdens individuals. The aim of this work to build an NLP binary classifier of the Jungian psychology types of introverts and extraverts and investigate whether they react differently to those methods has been reached with comparably strong results. Even though the model showed strong results on the held-out test set, the Bi-LSTM model was not applicable for out-of-domain data from Twitter. Therefore, we crafted a second model on hand-labeled tweets. All data was made public.

Experiments on Twitter data from 2019 compared with 2020 differentiated by introverts and extraverts revealed that the mental suffering of introverts during the pandemic is comparably more severe, adding novel findings to the current and contradictory debate. Introverts show a higher frequency of utterances associated with isolation, showed less optimism, spoke less about social interactions, and showed more frequent anxiety utterances. Meanwhile, extraverts showed less frequent utterances of isolation and more frequent friendships. With our approach, we offer an approach to identify individuals, that show elevated signs of worry. With those findings, those individuals could be supported by mental health services. Furthermore, it underlines the necessity as a society to look out for those individuals, that have become especially retracted or express themselves with isolating language.

For future outlooks, some indicators such as the confounding analysis, some already infrequent LIWC counting measures, and the rather weak introversion classification capabilities of the model should be taken into account for further critical analyzations. The findings in this paper should be viewed critically and examined with complementary experiments. Furthermore, we aim to deepen those findings and provide systems for automated personality detections, which then could help society to better overall mental health.

9.4.4   Author's Position on the Ethical Consideration

The aspects closely connected to the ethical evaluation of NLPsych approaches such as the one described in this chapter are described in Chapter 3 and more precisely in Section 3.3.

Even though this research is intended to foster psychological diagnostic research and mental health, such work poses the problem of an ethical dilemma between risks and promises (Johannßen et al., 2020b). NLPsych systems can be misused (dual use Williams-Jones et al. (2014)), misunderstood (Luhmann system theory (Görke & Scholl, 2006)), and will contain severe biases, which are hard to detect due to data protection laws (Diehl et al., 2015).

The proposed classification approach can neither replace clinical examinations nor should it be used for anything else than the performed validation study: mass observations with in-domain data for research purposes and without the intention of diagnosing individuals. This, however, is not what this work intends to provide. Rather, we aimed to support psychologists with additional and evaluation objectivity tools and shed validating light on the effects of the pandemic. We believe this work to add insights into human well-being during the COVID-19 pandemic and hope to foster research for increased mental health, which is a result of a wide range of research findings.

We believe that it is also important to discuss the limitations of our work, in addition to its strengths presented in this chapter. The following section offers clear discussion of limitations.

9.5   Limitations

The most apparent limition of this chapter is the methodology. Sampling tweets without the in-depth knowledge of the speaker does not guarantee that our assumptions on whether an individual is rather introverted or extraverted are correct.

Furthermore, these predictions and classifications are based on very short texts. Jungian types are said to be rather stable constructs (see Chapter 4). However, the way an indual expresses him- or herself in tweets can depend on the current mood. Many wordings furthermore do not allow for pscyhology diagnostic statements. This data was not collected clinically.

Since we utilized the same dataset as during our empirical work on social unrest prediction in Chapter 8, the same data sampling limitations apply, described in Section 8.7.

In addition, we only had very few tweets annotated by human experts. The utilization of our crafted models on these data instances revealed a lack of transferability of these models. The same has to assumed for the LMT model trained on tweets directly. Whether they can be applied to any other type of text, has to be doubted.

Lastly, same with the previous empirical works on aptitude diagnostics and social unrest prediction, the utilization of psychometric models on data associated with mass phenomena observable through social media plattforms raises multiple ethical concerns. These concerns and an assessment of their implications can be found in Section 3.3.

This chapter concludes the Part on empirical research. Thus far, this dissertation project automated and researched three psychometrics: i) the OMT, ii) self-regulatory levels, and iii) the Jungian types of introversion and extraversion. The available implicit data was extended to also include some Big Five assessments. The next Chapter 10 and Part IV presents analyses on the assumption, that the three automated metrics might correlate with each other or with Big Five dimensions. Such correlations could indicate underlying shared constructs captured by the metrics and would thus provide steps towards a solution of an yet open research problem.

Part IV

Psychological Pragmatics of NLPsych

CHAPTER 10

Correlation between NLPsych Psychometrics

This chapter presents the analyses of assumed correlations between the automated metrics theoretically described in Part II, and empirically researched in Part III. It is structured similarly to the previous chapters, containing first the research objectives in Section 10.1, utilized data in Section 10.2, the methodology in Section 10.3, the experiments in Section 10.4, and our results in Section 10.5. Finally, we discuss the results and draw a conclusion in Section 10.6. Limitations are discussed in Section 10.7.

Ever since psychology has become the study of one's mind, cognitive processes, behavior, and the connection between natural language and psychology has been established (Collin et al., 2012, p. 37). However, as Wittgenstein famously noted, it could be that "the limits of my language mean the limits of my world" (Wittgenstein & Schulte, 1963). The term *psychological pragmatics* has been established to describe the phenomenon that natural language speakers can communicate beyond what is explicitly said (Neale, 1992).

Psychological metrics, as they have been introduced in this dissertation, aim to reveal underlying cognitive processes, which can not be directly observed or are too complex to be fully understandable by observational processes (e.g. electroencephalography, EEG) (Schmidt-Atzert et al., 2018, p. 2). These psychometrics thus additionally aim for reducing the complexity of one's mind as much as possible, without losing the ability to explain cognitive processes. However, time and again, researchers identify unexplainable variance, criticize the validity of psychological diagnostical tests, or assume yet undiscovered cognitive processes not yet diagnosted (Schmidt-Atzert et al., 2018, p. 286), as it is also happening with the not-yet affirmed freedom motive (see Chapter 5).

Since sufficiently large annotated data can not be easily acquired for modeling the complexity of language, it appears that psychological metrics are necessary for complexity reduction. Nonetheless, in this chapter, we analyze the assumption that correlations between the three

researched metrics of implicit motives, self-regulatory levels, and Jungian psychology types exist.

10.1 Research Objectives (RO)

This section describes the research objective (RO) as part of the third research question (RQ3) proposed in Section 1.3. This RO can be divided into the goals of measuring metric inter-correlations (RO I) and of measuring diagnostical capabilities to find a metric consensus (RO II).

10.1.1 RO I: Metric Correlations

A first, rather basic assumption and goal is the identification of correlations between the previously researched diagnostical metrics. The naming of some dimensions is equal, such as *extraversion*, which is one dimension of the Big Five questionnaire inventory, as well as one Jungian psychology type. It has been largely researched that the Big Five dimensions are observable in a multitude of other diagnostical tests and procedures, despite being named differently. However, as Schultheiss & Brunstein (2010, p. XV) points out, these similarly named dimensions or metrics do not necessarily represent the same psychological constructs. Accordingly, the explicit extraversion from the Big Five usually does not correlate with the implicit extraversion described as the Jungian psychological type.

This first goal, to explore correlations, thus emerges from this phenomenon of similarly named metrics, which seemingly describe different constructs. Since undiscovered variables for explaining the varriance cognitive processes have time and again been suspected (Schmidt-Atzert et al., 2018, p. 2), correlations can be assumed.

10.1.2 RO II: Capabilities of Reaching Metric Consensus

All of the researched metrics have in common, that they are utilized for – amongst others – personality diagnostics (Schmidt-Atzert et al., 2018). Despite the complexity of natural languages and the complexity of cognitive processes, personality diagnostical tests and procedures mostly reduce the human mind to only a few target classes. If personality diagnostics does capture the proclaimed personality characteristics and if there are only a few temperaments or body fluids (Flaskerud, 2012), as assumed centuries ago, then similar psychological metrics should arrive at similar personality descriptions.

The second goal thus is to explore whether the automatically classified texts from multiple metrics display linguistic psychological content that is coherent with what these multiple psychometrics ought to describe.

Since we analyze textual data, this utilized data is described in the following section before moving on to describe the research methodology thereafter.

10.2 Data

The data utilized in this chapter has been collected by a company specialized in aptitude diagnostical testing[50]. For the utilized data set, the IPT (see Subsection 5.3.2) has been conducted mainly amongst prospective German managers from mid-sized enterprises. The average age was 42 and the participants were mostly male (without further demographic information in accordance with German data protection laws). Therefore, the data itself is expected to be biased and imbalanced towards middle-aged, well-educated, caucasian men (on biases, refer to Subsection 3.2.1). The collected IPT data is available in sufficient quantities ($n = 2,680$). However, those instances, which also contain Big Five dimension scores, are few in number with only $n = 863$. During the preliminary experiments (see Subsection 10.3.1), it was not possible to build classification models for the Big Five inventory of acceptable quality from this small available corpus. Thus, the Big Five data can only be utilized for correlation and interconnection analyses described in Section 10.3.

Table 10.1 displays some data characteristics. The data instances contain 52 words on average with roughly 17 words per sentence. Thus the instances from the IPT are significantly longer than the average answer lengths during the OMT testing procedure (roughly 22 words per instance). Furthermore, roughly 27 percent of all words are longer than 6 letters. Filler utterances, quotations or exclamation marks are rare and the majority of sentences end with a period. On average, one data instance contains 6 sentences. The data appears to be rather clean without many apparent grammar or spelling mistakes.

Characteristic	Value
Average number of words	52
Average number of words per sentence	17
Average number of sentences	6
Percentage of words longer than 6 letters	27

Table 10.1: The table displays the characteristics of the IPT dataset. Text instances are significantly longer than data from the OMT with 52 words, 17 words per sentence, and 6 sentences per instance on average.

The proposed research methodology for measuring correlations between the empirically researched psychometrics is described in the following section.

---

[50]WafM Wirtschaftsakademie GmbH `https://www.wafm.de/`.

10.3 Methodology

This section describes the experimental methodology and research design. First, we discuss preliminary research and thereafter provide a brief discussion of the utilized models, which are presented in more details in Chapters 7, 8, and 9. Lastly, the research objectives are described.

10.3.1 Preliminary Research

Our preliminary research, which tried to omit the intermediate and moderating psychological metrics and e.g., predicting academic success directly from language, has remained unsuccessful. Furthermore, preliminary research, which aimed to classify the Big Five dimensions from text (see Chapter 6), has also remained unsuccessful.

For the first preliminary goal, we utilized the best models from the empirical research presented in Chapter 7, Chapter 8, and Chapter 9 on the dataset described in Section 7. The goal of this preliminary research was closely related to the empirical research presented in Chapter 7 with the difference that we aimed to directly predict academic success in the form of grades solemnly from textual answers given during the OMT part of the aptitude test without classifying implicit motives first and correlating the motive counts.

For the second preliminary goal, we tried to model the Big Five questionnaire (see Chapter 6) from a small part of the IPT dataset described in Section 10.2, which was annotated with both, implicit motives and the OCEAN dimensions.

Both resulting models of this preliminary research did not surpass the minimal baseline of chance, scoring an F_1 of .5 and thus were not able to model academic success or the Big Five from natural language without moderating psychometrics.

10.3.2 Experimental Models Overview

The models originate from the presented research and experiments of previous chapters. As the benchmarks displayed in Tables 7.13 and 9.2 have demonstrated, the bi-LTSM model with attention mechanism (bi-LSTM attn, see Subsection 2.2.6) has outperformed other neural and non-neural architectures, including BERT (see Subsection 2.3.4). This has been confirmed by the GermEval shared task 1 on the prediction of cognitive style and motivation from text (see Section 3.3) and subsequent research from Johannßen & Biemann (2020), where said bi-LSTM attn was able to subsequentially outperform the participant's systems on modeling self-regulatory levels and implicit motives, making it the SOTA system for similar tasks.

Even though the best-established implicit motive theories assume the existence of the three main motives (affiliation, achievement, power) with the non-motive null (see Chapter 5), experiments have shown that from a data perspective, the assumed fourth freedom motive is differentiable from the other *Big Three* motives and thus it can be assumed that this freedom

motive does matter in the field of personality diagnostics. Therefore, the resulting bi-LSTM attn presented in Section 8.5 on social unrest prediction is utilized, which already included the freedom motive. This model reached an $F_1 = 74.08$ on the task of classifying the five target motives (see Subsection 2.2.9 for details on evaluation measures).

The bi-LSTM attn model for predicting self-regulatory levels has been created during the prediction of aptitude in Chapter 7. On the 6 target self-regulatory levels (one to five for self-regulation and a null motive if the self-regulation is neutral or not observable), the model achieves an $F_1 = 65.4$ score.

Lastly, the model of the Jungian psychological types of extraversion and introversion was first introduced during the research on pandemic isolation in Chapter 9. This model was able to achieve an $F_1 = 72.0$ on the IPT data. Even though this model was not applicable to tweets, it still renders the best approach for the following experiments, since the IPT itself is an implicit test, was conducted by the same company by the same standards, and the data basis of the experiments in this chapter emerged from an IPT testing procedure.

The methodological approach of this analysis thus is to automatically predict labels and to correlate them. Since these models achieved sufficient F_1 scores, at least signals of correlations are to be expected, even though the limitations of this approach (see Section 10.7) are likely to weaken and distort these signals.

10.3.3 Identification of Correlations between Psychometrics

The first step for achieving the research goal of identifying trait similarities is to perform a correlation analysis. Since the Big Five personality questionnaire (see Chapter 6) is considered one of the most influential and best-validated personality diagnostical tests, the correlations between the three researched metrics i) implicit motives, ii) self-regulatory levels, and iii) Jungian psychology types with the Big Five are of high interest.

First, all textual instances as described in Section 10.2 are classified by the respective psychometric model, two at a time for the subsequent comparison. Thereafter, each metric prediction will be standardized into their dichotomized values, i.e. with the binary differentiation of either being classified as a metric dimension (e.g. self-regulatory level five) or not. Only the textual instances classified as scoring high in a metric are kept, the rest is not considered.

Thereafter, for the combination of both metrics, the Pearson point-biserial correlation will be calculated and reported (see Section 2.1 for details on statistical measures). As stated in Subsection 10.1.1, it can be expected that implicit motives and Jungian psychology types do not correlate with the Big Five questionnaires, as the first two metrics are implicit, whilst the latter is explicit. However, correlations between self-regulatory levels could occur, since these are connected to volitional processes and to the implicit motives (Baumann & Kuhl, 2020). This methodology is displayed in Figure 10.1.
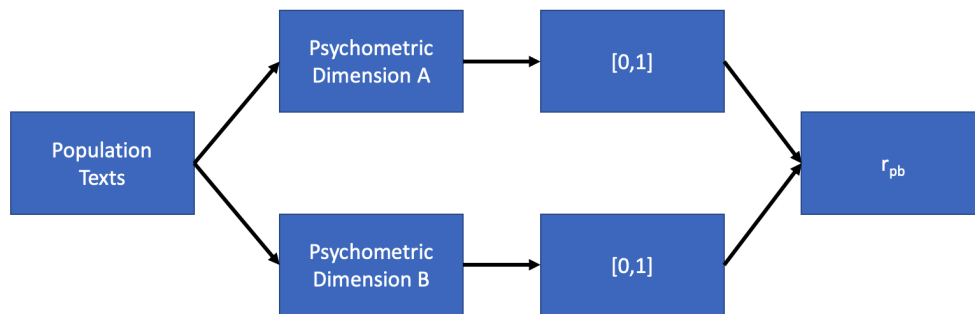
Figure 10.1: For the identification of correlations between different psychometrics, first two metrics are applied on the population, whereafter the predicted labels are standardized into dichotomized values and a point-biseral correlation is calculated.

10.3.4 Psychometric Consens

The goal described as *consensus* aims to combine two metrics and analyze, whether those texts, which have been classified as being both, high in metric A and high in metric B, score high in LIWC categories associative with both metrics – that is, whether these texts are categorized coherently with their respective theory (e.g. Jungian introversion describes individuals that gain gratification intrinsically). This comparison is made between the implicit metrics and the Big Five, since the Big Five is the most broadly utilized questionnaire approach.

For measuring the psychometric consensus, we first apply the first step from the correlation evaluation, which is to utilize two metrics or metric models each for label predictions from the population – an implicit one and a Big Five dimension. Only those classified as this metric dimension is kept, and the other texts are discarded. Different from the correlation approach, we will not standardize the values onto the integral between $[0, 1]$ but keep the texts as they are. Instead, the unification of both metrics is compared with the population (e.g. Big Five openness with self-regulatory level one vs the population).

Even though NLP researchers are aware of the limitations of word-list-based or dictionary approaches (see Section 2.4), the field of psychological diagnostics has successfully utilized the psychological dictionary tool LIWC (see Subsection 2.3.1) for validation studies. Therefore, for a second step we apply LIWC on the first the unification of both metrics (i.e. Jungian introversion and Big Five openness) and separately on the population.

Lastly, we will calculate first the arithmetic mean of both LIWC results (population vs. metrics). These mean values will subsequentially be compared and the absolute delta (Δ), as well as the percentage increase is calculated, as well as the percentage increases. We only consider increases, since the absence of utterances belonging to a LIWC category does not mean, that any psychological desire has diminished – such an absence could be obversable

due to different topics of interests. An increase, however, does indicate an increased desire or focus on these topics.

LIWC is a tool, which counts words belonging to a dictionary category. Thus, LIWC calculates absolute numbers. We only considered those categories that score at least an average of one word belonging to the category on average. Otherwise, if only very few appearances on average are considered (e.g. .07), the slightest change from the metrics would increase the percentage change by a large extent without being relevant in absolute terms. As an example: the change from .07 to .14 represents a 100% increase, whilst only having increased in absolute terms by .07 words on average, which is a very minor change. Furthermore, we only considered increases of at least 20%, as the methodological limitation of having very few instances per assessment (see Section 10.7) lessen the relevance of these changes.

Figure 10.2 visualizes the approach.
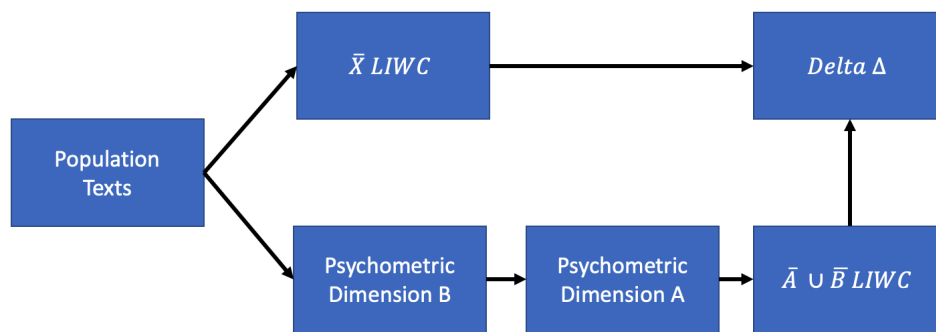
Figure 10.2: For analyzing the psychometric consensus, we first predict labels from the population and separate those texts into two metrics, which were classified as the metric at hand. Thereafter, we apply LIWC, calculate the mean and compare these mean values between the two metrics.

The following section describes our conducted experiments for performing correlation analyses.

10.4 Experiments

The experiments closely assemble the approaches from the respective research chapters, mentioned in Subsection 10.3.2. Most importantly, the pre-processing of each model has to be applied identically to the experimental data, as it had been to the training data. The cleaned data mostly assembles the original data. The performed experiments did not result in novel models, but since the already crafted models were utilized on implicit textual data, the experiments are rather concerned with data analysis than with performance assessments. Furthermore, for a

valid performance assessment, human expert annotators would be necessary for creating gold labels (see the limitations presented and discussed in Section 10.7).

Since the data described in Section 10.2 was collected and matched with the Big Five questionnaire recently, it does not align with the IPT data from previous research. We present our results in the following section.

10.5   Results

In this section, the resulting analyses presented in Section 10.3 are described, namely the identification of correlations and psychometric consensus.

10.5.1   Metric Correlations

The results of the metric correlations are displayed in Table 10.2. The table displays the Big Five dimensions, implicit motives, self-regulatory levels, and the Jungian psychology types of extraversion and introversion. First of all, the measured correlations are weak overall with mostly ranging between $r_{pb} = -.13$ and $r_{pb} = .15$ (see Subsection 2.1) and thus there are no strong connections between the metrics. This generally indicates that there are no underlying explaining metrics or variables and that each metric in and of itself is necessary. The self-regulatory levels *one, null*, and *three* correlate weakly with the explicit Big Five dimensions *openness, conscientiousness*, and *agreeableness*, whilst levels *two, four*, and *five* correlate with the implicit metrics of motives and Jungian psychological types. This could slightly indicate, that the self-regulatory levels function as a bridging variable between implicit and explicit metrics. The other metrics – Big Five dimensions, implicit motives, and Jungian Types – do not correlate with each other.

10.5.2   Metric Consensus

The results from the consensus analysis are separated by the type of psychometric. Table 10.3 displays the combination of implicit motives and the Big Five questionnaire. The table is ordered ascending from the lowest percentage increase from the population texts to the metric combination to the highest percentage increase (last column). $n(A \cup B)$ denotes the absolute number of instances, which score high in both Metric A and Metric B. *LIWC cat.* denotes the LIWC category, *example category words* present some words belonging to this category, $\overline{X}$ denotes the mean of this LIWC category for the population, $\overline{A} \cup \overline{B}$ denotes the mean for this LIWC category for the combination of metrics A and B, Δ denotes the absolute increase of counted LIWC category words from the popaulation $\overline{X}$ to the metrics $\overline{A} \cup \overline{B}$, and % denotes the percentage increase.

Metric X	Metric Y	r_{pb} **Correlation**

Introversion		
Big Five Openness	Level One	-.12
Big Five Conscientiousness	Level Null	.13
Big Five Agreeableness	Level Null	.13
Big Five Agreeableness	Level Three	-.11

Extraversion		
Level Two	Motive Affiliation	.15
Level Four	Motive Affiliation	.1
Level Two	Motive Power	.11
Level Five	Intro / Extra	-.09

Table 10.2: This table displays the correlations measured between all metrics: Big Five dimensions, self-regulatory levels, implicit motives, and the Jungian psychology types of extraversion and introversion.the measured correlations are weak overall with mostly ranging from $r_{pb} = -.13$ to $r_{pb} = .15$.

Interestingly, the highest increase comes from Big Five agreeableness and the affiliation motive for the LIWC category 'friend' (words associated are e.g. friend, boyfriend, or dude). In absolute terms, the metric combination contained 1.7 words more on average in the 'friend' category than the population. However, the number of instances is with $n(A \cup B) = 27$ sparse. Neuroticism paired with the power motive contained 50% more utterances associated with negative emotions (e.g. bad, hate, hurt). Big Five openness combined with the affiliation motive contained .61 more words on average categorized as 'leisure', e.g. game, fun, or party.

Metric A	Metric B	$n(A \cup B)$	LIWC cat.	Example category words	$\overline{X}$	$\overline{A \cup B}$	Δ	%
Big Five O	Achievement	272	risk	secur*, protect*, pain, risk	1.25	1.52	+.27	+22.02%
Big Five O	Affiliation	43	leisure	game*, fun, play, party*	1.94	2.55	+.61	+31.14%
Big Five N	Power	10	negmo	bad, hate, hurt, tired	1.66	2.51	+.85	+50.91%
Big Five A	Affiliation	27	friend	friend*, girlfriend*, dude	1.48	3.17	+1.7	+114.88%

Table 10.3: This table shows the results from the metric combinations of Big Five with implicit motives. The average LIWC frequencies per category of the population are compared with the combined metrics. The changes are rather minor and inconclusive.

As for the combination of Big Five and self-regulatory levels, there were barely any changes observable. Only Big five extraversion combined with self-regulatory level five was associated with more frequent 'negative emotion' utterances (.74 more words on average and an increase by 44%, but with $n = 16$) and Big Five openness combined with level one was associated with more frequent utterances of 'see', e.g. view, saw, seen (an absolute increase of 1.08 word on average and relatively +60%, $n = 29$).

Metric A	Metric B	$n(A \cup B)$	LIWC cat.	Example category words	$\overline{X}$	$\overline{A \cup B}$	Δ	%
Big Five E	Level Five	16	negemo	bad, hate, hurt, tired	1.65	2.39	+.74	+44.76%
Big Five O	Level One	29	see	view, saw, seen	1.78	2.86	+1.08	+60.42%

Table 10.4: This table shows the results from the metric combinations of Big Five with self-regulatory. The average LIWC frequencies per category of the population are compared with the combined metrics. The changes are rather minor and inconclusive.

As for the combination of Big Five with Jungian extraversion or introversion, no apparent signals could be detected.

Finally, a conclusion is drawn in the following section.

10.6   Discussion & Conclusion

In this chapter, the correlation and consensus between all three proposed implicit metrics of motives, self-regulatory levels, and Jungian psychology types was assessed. These three metrics were combined with the Big Five personality questionnaire in form of novel IPT implicit test data. Since personality diagnostics rely on comparable and sometimes identical underlying cognitive mechanisms and traits, one assumption was that there are correlations between all personality diagnostical metrics. Furthermore, it was questioned, whether there is a metric consensus.

For this, two experiments and analyses were conducted. Firstly, all metrics were classified by the use of previously crafted models. The limiting factor of this approach is the imperfection of the models' F_1 scores, ranging from $F_1 = 65.6$ to $F_1 = 74.08$. For the correlation assessment, the resulting predicted labels were then correlated and utilized for the comparison with LIWC categories. The consensus was assessed by classifying combined metrics and comparing the instances with the population in terms of LIWC category increases.

As for the correlations between all metrics, there were only very few to be identified. These correlations furthermore were quite weak with a Pearson point-biserial correlation coefficient of mostly $r_{pb} = .1$, which renders a noticeable correlation for human-produced textual data but would be by far too weak for substitutional purposes or the assumption of a further, undiscovered explaining variable. One can only speculate on the reasons for this weak correlation. One major reason could very well be the in terms of classification perfect insufficient utilized models on out-of-domain data. If a model already misclassifies roughly one fourth of all instances and is correlated with another model performing similarly, even existing correlations will most likely be masked by noise. This does not mean, that there can not be correlations, but rather, that this multi-model correlation approach is insufficient in discovering them.

It was expected that there would be no useful correlation between explicit and implicit metrics. The Big Five is an explicit metric and implicit motives, as well as the Jungian psychology types extraversion and introversion are implicit. Despite the Big Five having a dimension also called *extraversion*, the extraversion of the Big Five and the Jungian types do not represent the same underlying traits.

However, one noticeable correlation result is the correlation between self-regulatory levels and Big Five dimensions, as well as self-regulatory levels with implicit motives and Jungian types. These self-regulatory levels do not describe a personality trait itself, but rather the type of desire, with which these traits are satisfied. This connection ought to be researched in subsequent experiments, but could reveal a novel connection, which the psychological field of diagnostics did not address yet.

The broader combination between Big Five and the psychological metrics of self-regulatory levels, Jungian types, and implicit motives was rather inconclusive. At most, texts associated with two combined metrics (one of the three empirically researched psychometrics with the Big Five) contained 1.7 more words associated with a LIWC category compared to the population. As described, the textual basis is rather long. 1.7 words do not represent a large change and most of the categories for both, implicit motives and self-regulatory levels combined with the Big Five questionnaire, were increased by less than one word. However, a total number of $n = 27$ text instances relevant for this assessment is not sufficient, which leads to the next section that discusses limitations.

Especially since the metric consensus or combination was inconclusive, we believe that it is also important to discuss the limitations of our work, in addition to its strengths presented in this chapter. The following section offers a clear discussion of limitations.

10.7   Limitations

The very few signals identified and displayed in Tables 10.3 and 10.4 are weak and could have occurred by pure chance, considering the fast variety of combinations analyzed. The Jungian types did not show any noticeable increase.

Reasons for this lack of signals could be the very small data basis. The population contains $n = 863$ instances, which are already few in number. But when classified and then dichotomously filtered, this data basis becomes too small to be statistically relevant. When two model predictions are applied to the $n = 863$ instances, the only noteworthy changes are observed for as little as $n = 10$ instances. This sparse data basis is not sufficient for making any claims from this analysis.

Furthermore, the lack of correlations between the explicit Big Five questionnaire and the implicit Jungian types or implicit motives already indicate that a combination would not describe any psychological trait or construct, but results in random textual selections.

The most severe limitation stems from the approach to automatically assign labels and correlate them in multiple steps as displayed in Figures 10.1 and 10.2. The employed models, even if they achieve SOTA performance scores, do not score perfectly. If one model achieves an accuracy of 70% correctly classified instances and a second model is applied to the same textual data, then – under the assumption of independence – the overall accuracy drops to 49% correctly classified instances. For three subsequent testing procedures, this accuracy scores as low as 34%. This, in combination with the next limitation of out-of-domain-data, severely distorts correlation signals.

Lastly, the previous empirical works on e.g. pandemic isolation (see Chapter 9) has demonstrated, that these proposed models do not generalize well on other data sources than the training material. Whether it these models are applicable on e.g. the IPT data set of this chapter would have to be measured by having human experts annotate large chunks of the instances and by comparing the model's predictions with these gold labels.

However, all of these limitations do not disproof that there might be interesting signals to be found. Nor do they mean, that the signals identified do not matter, but rather, that exerpiments without these limitations would have to calrify, whether they occured by chance. This would be a worthwile future endeavour.

The following chapter concludes the dissertation project as a whole. It summarizes the empirical research, discusses the conducted shared task, and the empirical pandemic research. This correlation chapter will be critically assessed and finally, the research questions from the introductory Chapter 1 are answered. Since a dissertation project can only provide some steps towards a larger picture, future outlooks of the possible subsequent research are provided.

Part V

Conclusion and Outlook

Conclusion and Outlook

In this chapter, a conclusion of the dissertation project is drawn. Even though excessive research has been performed on the interdisciplinary field of NLP for psychometric textual data, the presented work can only provide first aspects and results in a large field of research. Thus, an outlook on future directions and research are provided.

First, empirical evidence of psychometric modeling capabilities of personality traits, aptitude, and behavior from Chapter 7 is summarized, followed by the automated assessment of social unrest and pandemic isolation indicators, originally described in Chapter 6. Lastly, the similarities between all utilized psychometrics is discussed, which were presented in Chapter 10.

This chapter will be concluded with the answering of the three research questions presented in Chapter 1 and the future outlook of NLPsych research and future expectable research in the domain of psychological diagnostics.

11.1 Empirical Evidences of Psychometric Modelling Capabilities of Personality Traits, Aptitude, and Behavior

Especially for psychological diagnostics, it tends to be difficult to produce approaches and tools, which allow for stable, reliable, and consistent results.

Furthermore, those psychological diagnostical tests available suffer from an array of effects and biases. One is the halo effect, which is the tendency to attribute attractive people with further positive characteristics and traits, regardless of their age, e.g. assuming that a visually attractive student also possesses above-average intelligence. The Barnum effect describes the tendencies of self-attributing characteristic descriptions, which are unspecific and generalized, leading to an easy belief in unspecific personality diagnostics. Explicit personality diagnostic tests, such as personality questionnaires, oftentimes suffer from a socio-desirability bias,

leading to participants rather answering in a socially desirable way instead of a more neutral introspection (see e.g. Section 6.4).

This thesis proposes steps towards a solution to all of the above challenges, which is the utilization of NLP methods for the processing of psychometric textual data. When modeling diagnostical metrics, one could i) enable perfectly replicable results, ii) challenge the halo effect, since those models do not honor visual attractiveness, iii) the Barnum effect since models are only acceptable when target classes are clearly separatable and thus specific enough, and iv) the socio-desirability bias, that can be challenged by implicit methods (see Chapter 5), which had been too expensive to manually analyze but could now be modeled by the use of NLPsych.

During the course of the dissertation project, state-of-the-art (SOTA) models have been crafted to classify the OMT with its implicit motives, self-regulatory levels, and the Jungian psychological types of introversion and extraversion. The models achieved high F_1 score, were determined to be reliable, showed human-annotator-like stability, and revealed behavioral predictions during validation studies.

A validation study, which utilizes a proposed LSTM model with an attention mechanism on implicit motivational texts for aptitude diagnostics achieved an $F_1 = 81.55$, outperforming the LMT approach. Since *attention is not explanation*, the attention weights of the model were shuffled, which resulted in worse performance metrics, indicating algorithmic importance. Observed attention weights confirmed the implicit motive theory. The predicted implicit motive *achievement* correlated with $r = -.25$ with bachelor's thesis grades (since 1.0 represents the best results, the correlation is negative, reading: the higher the count of predicted achievement motive per student, the better the grade), whilst the *power motive* showed a weak correlation with $r = .14$, indicating that a higher frequency of power-motivated utterances predict a worse bachelor's thesis grade (see Chapter 7).

The conducted shared task aimed to foster apditutde diagnostical NLPsych research and provide the community with annotated data. However, it also sparked an intense international ethical debate upon NLPsych as a whole, which is discussed in the following section.

11.2 Shared Task on the Prediction of Cognitive and Motivational Style from Text

After the first findings of predictive power through the moderating variable of implicit motives, the aim of a conducted GermEval shared task on the prediction of cognitive and motivational style from the text was to further investigate, whether criticized diagnostic approaches such as IQ scores, school grades, or standardized testing could be compensated with a more neutral and less biased measure of implicit motives. In addition to this research objective, the shared

task aimed at the distribution of implicit motive data with useful additional data points (see Section 7.5).

The task provided participants with implicit motive texts from college students paired with a rank. The rank represented a hierarchical order of all students, calculated as the harmonic mean of different school grades (English, German, and math) and the IQ test dimensions *language* and *logic*. Three teams participated in the task and crafted different systems for surpassing the baseline systems (tf-idf SVMs), including SVMs, linear n-gram models, and BERT transformer networks. All systems were able to model the subtasks, i) recreating the rank, and ii) performing classification on the 30 target classes combination of implicit motives and self-regulatory levels. Since all systems were able to clearly outperform the baseline systems, it has been shown that aptitude diagnostics does not need to rely on explicit diagnostic procedures or standardized testing but could rather utilize projective procedures. To achieve a board and differentiated assessment, a mixture of standardized, explicit, and implicit diagnostical tests can be advised.

The shared task had caused an extensive discussion regarding the IQ testing parts associated with the provided data. The main criticism focussed on dual use, surveillance, and dangers associated with trying to predict IQ scores from textual data. This justified and necessary critical assessment of research on NLP methodology for the assessment of psychological textual data has to lead to a publication concerned with the ethical assessment of said GermEval shared task. The raised concerns could be refuted in that i) the conducted IQ test during the college's potential assessment honors all forms of standardization and established quality criteria in psychological diagnostics, ii) in Germany, socio-economical biases and differentiating factors emerge early on, most prominently during the high school years, iii) the ground population of college students are relatively homogeneous due to the dual study program offered by the private college, iv) the goal of the shared task is mainly to reduce standardized tests – which are trainable–, not increase them, v) the utilization of resulting models on any non-implicit texts would be methodologically flawed, vi) the task reflects common practices for aptitude diagnostics in the private industry sectors of Europe, vii) it is better to research possibly harmful effects of such practices, rather than forbidding the research, viii) the marketplace of ideas will self-regulate those research approaches, which are valid and share community consensus, ix) the system's theory by Luhmann states and acknowledges, that neither psychologists, nor NLP researchers, but only researchers from interdisciplinary fields may be able to determine the methodological soundness of such a task, and x) knowledge can not and should not be restrained (see Section 7.5). In summary, the ethical consideration of the conducted shared task concluded, that the task was methodologically sound, ethically correct, and worthwhile being researched.

The COVID-19 pandemic struk during the course of the dissertation project and shifted the focus towards researching effects on individuals. These effects and the conducted research is evaluated in the following section.

11.3 Automated Assessment of Social Unrest and Pandemic Isolation Indicators

As a consequence from the GermEval shared task, a bi-LSTM model with attention mechanism was trained on the publicly available data from said shared task and surpassed the participant's systems – including the BERT approach –, resulting in a state-of-the-art model for implicit motives and self-regulatory levels.

Amongst multiple conclusions of the ethical considerations of the GermEval shared task was the recommendation to strongly focus on validation studies when crafting NLP models for classifying and predicting psychological effects. Since diagnostical tests and metrics aim to reveal underlying traits and mental processes, and since the field of psychological diagnostic has shifted towards behaviorism as established mainly in the USA in the 19th century, NLPsych systems should not only be evaluated on their reliability and in terms of the established machine learning metrics (e.g. precision, recall, F_1), but also be tested in their observable validity by behavioristic empirical experiments (see Section 2.1).

During the course of this dissertation project, the COVID-19 pandemic greatly affected the well-being and social structures globally. The research activity thereafter was reoriented towards researching the effects of this pandemic on people and societies. The crafted SOTA implicit motive model was applied to a pandemic corpus (later published) drawn from the 1% Twitter stream. The corpus consists of randomly sampled Tweets from spring (March to May) of 2019 and spring of 2020. David G. Winter identified a pattern of implicit motives and self-regulatory levels, indicating social unrest. By applying the model on the Twitter pandemic corpus, the same patterns were observable to be more pronounced and elevated in 2020 compared with 2019, indicating growing signs of social unrest during the pandemic (see Chapter 8).

The same neural architecture was applied to a novel dataset consisting of texts from the implicit personality test (IPT) and hand-labels of introversion or extraversion as defined by C.G. Jung. The model achieved an $F_1 = 72.03$, scoring comparably well as the SOTA model for the English language. However, when applying this model to prior hand-labeled Tweets, it was not able to classify introversion and extraversion beyond mere chance. Therefore, a second logistic model tree (LMT) model was trained directly from Tweets, achieving an $F_1 = 68.9$. This second model was applied to the same pandemic dataset. Current research assumes, that individuals classified or identified as introverts experienced more mental suffering during the lockdown phases of the pandemic compared to individuals identified as extraverts. The con-

ducted research for this dissertation project confirmed those findings. Individuals classified as introverts produced more utterances associated with anxiety, loneliness, and isolation compared to extraverts. To ensure reliable results, a base population was also compared to those two groups. All results were statistically significant (see Chapter 9).

The following section critically assesses the analyses on assumed correlations between these empirically researched psychometrics. Since the automation approaches and models replicated psychological findings, it is assumed that they did model these metrics. Furthermore, the following section evaluates the assumption of correlations between these psychometrics, indicating an underlying shared construct.

11.4 Correlations between Psychometrics

The assumption of correlations between all personality diagnostical metrics was researched in Chapter 10. In preliminary experiments, it was questioned, whether metrics are necessary in the first place or whether it would be more valuable, to disregard those intermediate moderators and e.g. predict academic success directly from the text. These preliminary experiments were unsuccessful. Predicted psychometric labels were thus correlated and utilized for the comparison with LIWC categories and the consensus of multiple merics with the Big Five questionnaire was researched.

Correlations between the metrics were quite weak with a Pearson point-biserial correlation coefficient of r_{pb} = .1. Self-regulatory levels and Big Five dimensions correlated weakly, as well as implicit motives and Jungian types and could reveal a novel connection, which the psychological field of diagnostics did not address yet.

The assumed consensus between multiple metrics, namely the three empirically researched metrics of Jungian types, self-regulatory levels, and implicit motives with the Big Five questionnaire was inconclusive.

With all the above results, we can finally answer the proposed research questions.

11.5 Answering of the Research Questions

In this section, the three main research questions as defined in Chapter 1 are answered.

RQ1: Can NLP systems model psychological metrics?

The first research question can be affirmed. Previous research approaches had unsuccessfully attempted to automizing psychological metrics such as motives (Schultheiss & Brunstein, 2010; Schultheiss, 2013). Previous approaches oftentimes aimed for modeling the metrics by

solemnly relying on wordlists and dictionary approaches. Correlations did not surpass $\rho = .37$ (Schultheiss, 2013). Computer-based modeling of implicit motives had previously not been achieved. Schultheiss & Brunstein (2010, p. 186 ff.) concluded in 2010, that the modelling of implicit motives with computer models would greatly facilitate research.

During this dissertation project, multiple psychological diagnostic metrics based on textual inputs were successfully modeled. A proposed bi-LSTM neural network with an attention mechanism achieved SOTA performances of $F_1 = 81.55$ for the five motive classes (affiliation, power, achievement, freedom, and zero). Out of an array of architectures, this model was identified to be the most suitable for the task, even surpassing the BERT approaches.

The combination of implicit motives and self-regulatory levels was modeled with an macro $F_1 = 70.40$ by Schütze (2020). The subsequentially crafted bi-LSTM model with attention mechanism surpassed the model's performance with an $F_1 = 74.08$ for all 30 target classes as a combination of the five motives and six self-regulating levels.

Lastly, the implicitly measured Jungian psychology types of extraversion and introversion could be modeled with the self bi-LSTM attention model with an $F_1 = 72.03$, achieving as high of results as the SOTA model of the English language by Plank & Hovy (2015). This model is the first of its kind. Especially since modeling the Jungian psychology types is said to be difficult, this model offers novel subsequent research objectives to be investigated (Stajner et al., 2021).

RQ2: Do modeled psychometrics predict behavioral observations?

The second research question can be affirmed. The RQ1 described models might have achieved high F_1 scores. Since the classification of psychological diagnostical metrics is not a goal in itself, those resulting models had to be externally validated (see Subsection 2.1.4). This can be achieved by utilizing the models for classifying target labels – in this case psychological categories – and combining those assigned labels with observable behavior. This observed behavior would have to align with the expected behavior in accordance with the psychological metric.

The implicit motive models are the most versatile validated in this dissertation project. First, both, the logistic model tree (LMT) and LSTM model with attention mechanism, were validated by classifying motives of college students and thereafter aligning those motive labels with their achieved bachelor's thesis grades. As expected, the achievement motive correlates with $\rho = -.25$ weakly, but significantly, with better thesis grades (see Chapter 7).

The self-regulatory levels were utilized in combination with the implicit motives to measure the social unrest patterns identified by Winter (2007). A Twitter corpus containing tweets before the pandemic and during was utilized to assess the motives and self-regulatory levels, revealing that the power motive in combination with the self-regulatory 4th level – the sensitivity for negative incentives – increased in its frequency by 10.97% in 2020 as compared to 2019 (see Chapter 8).

Lastly, the model capable of classifying introversion and extraversion from short implicit texts was considered for validating related works' findings on experiences of loneliness for introverts during the COVID-19 pandemic. However, the bi-LSTM attention model failed to classify introversion and extraversion on a hand-labeled Twitter dataset beyond chance. A second LMT model was applicable and revealed an increase in anxiety, inhibition, and a decrease in positive feelings, as well as social topics such as sexual intercourse and dining (see Chapter 9).

RQ3: Do automated psychomerics correlate in their assessment on similar texts?

Even though some promising novel findings could be made by the research presented in Chapter 10, possibly leading to a new understanding of self-regulatory levels and providing practitioners with a more handleable set of fewer important tokens to be investigated, the overall research question 3 on correlations between the researched metrics have to be negated.

It appears that for personality diagnostic developed and utilized metrics are necessary for identifying traits, which in turn allow for behavioral predictions. It was not possible to neglect those intermediate metrics. However, such a result – that well validated and crafted metrics are neglectable – would have been of the uttermost of surprise.

The consensus combination between the empirically researched metrics and the Big Five questionnaire was inconclusive. This is to say, the limitations of the proposed experiments were so vast, that identified signals and changes might be of importance, but they could also have been the result of pure chance. The data set was too small, the labels were too uncertain, and the transferability too questionnable to reliably arrive at an answer to the question, whether consensus exists. As for the state of research, this has to be negated. Nonetheless, more sophisticated experimental research should be performed on this consensus assumption.

Since NLPsych is a fragmented, rather novel applicaiton field and since a dissertation project can only result in so many publications and solved research problems, the implicaitons towards future research approaches by far exceed the taken steps towards the larger picture, which is why the following section provides some of these future directions and outlooks.

11.6 Future Direction of NLP Research in the Domain of Psychological Diagnostics

The research field of combining NLP methods with psychological diagnostic is still a niche. Dedicated workshops such as CLPsych mostly publish few proceedings and are rather unique in their kind. It is still difficult to combine those fields, since for the NLP community, the need for interpretable models in psychology lead to less innovative models and architectures and

for the psychology community, those methods utilized by the NLP community are oftentimes too intransparent.

However, NLP research has shifted from ever larger transformer models towards explainability, applications, and even down-sizing of models. This trend is likely to continue. The field of empirical or clinical psychology has become more open to novel methods as well. It is likely, that the – so far still niche – the field of NLPsych will continue to grow and be more widely accepted amongst both communities.

To direct research activities toward those needs, future NLPsych research will continue to model established psychological metrics and support decision-making in both, psychology and applications (e.g. aptitude diagnostics). Since most of the proposed psychometrics are utilized in very sensitive application areas, it remains advisable to rather view the resulting models as case studies, as decision supporting systems, and as work in progress. Even though systems achieve human-like performance, error analysis has shown that in cases of misclassifications, the false classes can appear further from the true labels than those assigned by human expert annotators. Whether or not trained models should be carefully kept up to date, has to be researched. On the one hand, language does change fast and events like the pandemic with estimated 2,000 novel words in the German vocabulary can heavily influence the use of language. On the other hand, related work and in this dissertation project conducted experiments have shown that rather grammar and function words hold psychological importance.

Explainable artificial intelligence (XAI) has become one of the main trends in NLP research and psychology for providing replicable, explainable, and validated studies. Hence, XAI will most likely be one main research objective for the near and further future of NLPsych research. This need for XAI will most likely be combined with NLPsych tool development. In the area of psychological diagnostics, the importance of natural languages is noticeable in almost all aspects: IQ tests always include linguistic tasks, the Big Five was developed on the basis of lexical analyses, and LIWC – as simple as this dictionary approach might be – remains one of the most broadly utilized text analysis tools for psychologists and empirical psychological research. The field of psychology is in need of more modern and more context-sensitive tools, which have to be interpretable. Thus far, neural approaches or word embeddings could not convince the psychology community due to their poor explainability. Even embedding visualizations and distance metrics could not satisfy this empirical research. However, if future tools provide an array of NLP analyzations such as heat map attention weights, part-of-speech tags, dependency trees, or confidences. First simple tools, which were developed during this dissertation project (see Section 10.4) have convinced most of the researchers, that utilized this tool and those models for empirical psychological research. A first prototype for providing researchers with a tool for utilizing computer-aided psychometrics with NLP, developed by the author, is displayed in Figure 11.1.

Yet to be researched is the open question of whether psychological diagnostical metrics, tests, and procedures are necessary. According to the classical test theory (CTT, see Subsection
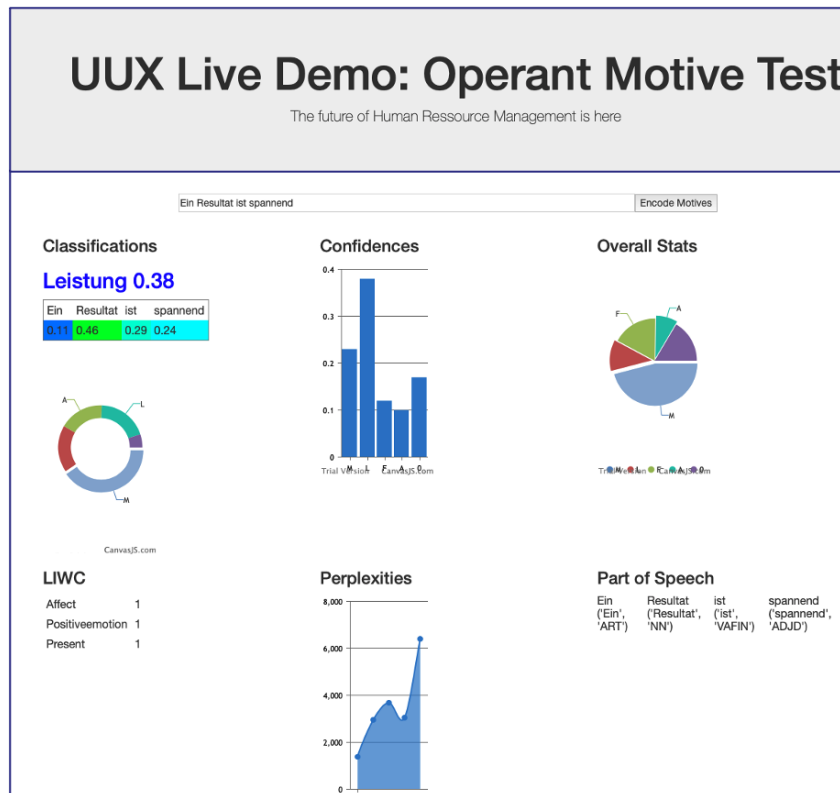
Figure 11.1: During the course of this dissertation project, one proposed implicit motive model has been crafted into an usable tool, which has been utilized empirically by economy and psychology scientists. For an outlook, it is intended to further provide scientific communities with NLPsych tools.

2.1.1), with increasing numbers of test executions, the testing error approaches 0. However, this only implies that a test is working as intended. It does not state that the metric fully explains mental processes. Psychological diagnostic has been struggling with fully accessing mental processes and explicit metrics can only make statements upon observable behavior. Implicit metrics, Carl Gustav Jung, and Sigmund Freud all acknowledge unconsciousness to be of great importance, which is hardly observable. It is possible that all those metrics (e.g. IQ scores, aptitude tests, implicit motives) only shed light on selected skills and mental processes, but fail to capture larger underlying processes, which influence all metrics. With NLP and novel, giant models such as GPT3, it could be possible to directly combine the use of language with observable behavior to assess underlying mental processes. It could be promising to dismiss moderating psychological diagnostic metrics altogether since this reduction of complex language onto very few target classes omits many signals. However, no evidence and no experiments conducted as part of this dissertation project were yet able to achieve such a direct approach. Nonetheless, this dissertation project contributed a few pieces to the puzzle of combining psychology with natural language processing.

Bibliography

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mane, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viegas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., & Zheng, X. (2016). TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv:1603.04467 [cs]*.
URL `http://arxiv.org/abs/1603.04467`

Aghdaei, R., & Tabrizi, A. (2021). A Review Study of How and Why People Are Different. *Journal of Social Sciences and Humanities*, *9*, 1–11.

Alharthi, R., Guthier, B., Guertin, C., & El Saddik, A. (2017). A Dataset for Psychological Human Needs Detection From Social Networks. *IEEE Access*, *5*, 9109–9117.

Allport, G. W., & Odbert, H. S. (1936). Trait-names: A psycho-lexical study. *Psychological Monographs*, *47*(1), i–171.

Altman, M. C. (2011). *Kant and Applied Ethics: The Uses and Limits of Kant's Practical Philosophy*. Malden, MA: Wiley-Blackwell, 1 ed.

Angleitner, A. (1991). Personality psychology: trends and developments. *European journal of personality*, *5*(3).
URL `https://pub.uni-bielefeld.de/record/1779606`

Archer, R., Buffington-Vollum, J., Vauter, R., & Handel, R. (2006). A Survey of Psychological Test Use Patterns Among Forensic Psychologists. *Journal of personality assessment*, *87*, 84–94.

Ashford, E., & Mulgan, T. (2018). Contractualism. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2018 ed.
URL `https://plato.stanford.edu/archives/sum2018/entries/contractualism/`

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv:1409.0473 [cs, stat]*, *1409*.
URL `http://arxiv.org/abs/1409.0473`

Balasa, A. (2020). COVID – 19 on Lockdown, Social Distancing and Flattening the Curve – A Review. *European Journal of Business and Management Research*, *5*.

Basile, A., Pérez-Torró, G., & Franco-Salvador, M. (2021). Probabilistic Ensembles of Zero- and Few-Shot Learning Models for Emotion Classification. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, (pp. 128–137). Held Online: INCOMA Ltd.
URL `https://aclanthology.org/2021.ranlp-1.16`

Baum, I., & Baumann, N. (2019). Autonomous creativity: The implicit autonomy motive fosters creative production and innovative behavior at school. *Gifted and Talented International*, *33*, 1–11.

Baumann, N., & Kuhl, J. (2020). Nurturing your self: Measuring and changing how people strive for what they need. *The Journal of Positive Psychology*, (pp. No Pagination Specified–No Pagination Specified). Place: United Kingdom Publisher: Taylor & Francis.

Baumann, N., & Scheffer, D. (2010). Seeing and Mastering Difficulty: The role of affective change in achievement flow. *Cognition and Emotion*, *24*, 1304 – 1328.

Benikova, D., Wojatzki, M., & Zesch, T. (2018). What Does This Imply? Examining the Impact of Implicitness on the Perception of Hate Speech. In *International Conference of the German Society for Computational Linguistics and Language Technology: Language Technologies for the Challenges of the Digital Age*, (pp. 171–179).

Benson, E. (2003). Intelligent intelligence testing. *Monitor of Psychology*, *34*(2), 48–49.
URL `https://www.apa.org/monitor/feb03/intelligent`

Biemann, C., Heyer, G., & Quasthoff, U. (2022). *Wissensrohstoff Text: Eine Einführung in das Text Mining*. Wiesbaden: Springer Vieweg, 2nd ed. 2022 ed.

Binckebanck, L. (2019). Digitale Unterstützung für Personaler – Mitarbeitende finden mithilfe von Künstlicher Intelligenz.
URL `https://www.nordakademie.de/news/digitale-unterstuetzung-fuer-personaler-mitarbeitende-finden-mithilfe-von-kuenstlicher`

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993–1022.

Block, J. (1995). A contrarian view of the five-factor approach to personality description. *Psychological Bulletin*, *117*(2), 187–215.

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, (pp. 5454–5476). Online: Association for Computational Linguistics.
URL `https://aclanthology.org/2020.acl-main.485`

Bogner, K., & Landrock, U. (2016). *Response Biases in Standardised Surveys (Version 2.0)*. GESIS - Leibniz-Institut für Sozialwissenschaften.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146.
URL `https://www.aclweb.org/anthology/Q17-1010`

Bornmann, L. (2013). What is social impact of research and how can it be assessed? A literature survey. *Journal of the American Society for Information Science and Technology*, *64*, 217–233.

Boyd, R. L., & Schwartz, H. A. (2021). Natural Language Analysis and the Psychology of Verbal Behavior: The Past, Present, and Future States of the Field. *Journal of Language and Social Psychology*, *40*(1), 21–41.
URL `https://doi.org/10.1177/0261927X20967028`

Brants, S., Dipper, S., Hansen, S., Lezius, W., & Smith, G. (2004). The TIGER treebank. *Journal of Language and Computation*, *2*, 597–620.

Braunack-Mayer, A. J. (2001). What makes a problem an ethical problem? An empirical perspective on the nature of ethical problems in general practice. *Journal of Medical Ethics*, *27*(2), 98–103.

Brem-Gräser, L. (1957). *Familie in Tieren: Die Familiensituation im Spiegel der Kinderzeichnung. Entwicklung eines Testverfahrens*. Reinhardt.

Brey, P., Reijers, W., Rangi, S., Toljan, D., Romare, J., & Collste, G. (2015). International differences in ethical standards and in the interpretation of legal frameworks. *Rapport Technique SATORI.[Google Scholar]*.

Briggs Myers, I., McCaulley, M. H., Quenk, N. L., & Hammer, A. L. (1998). *MBTI Manual: A Guide to the Development and Use of the Myers - Briggs Type Indicator*. Palo Alto, Calif: Consulting Psychologists Press, inc., 3 ed.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language Models are Few-Shot Learners. *arXiv:2005.14165 [cs]*.
URL `http://arxiv.org/abs/2005.14165`

Brunstein, J. C. (2008). Implicit and explicit motives. *Motivation and Action*, (pp. 227–246).

Carlton, L., & Macdonald, R. A. R. (2003). An Investigation of the Effects of Music on Thematic Apperception Test (TAT) Interpretations. *Musicae Scientiae*, *7*(1_suppl), 9–30.
URL `https://doi.org/10.1177/10298649040070S101`

Carroll, J. (2010). *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*. Cambridge ; New York: Cambridge University Press, illustrated edition ed.

Chadefaux, T. (2012). Early Warning Signals for War in the News. *Journal of Peace Research*, *51*(1).

Charter, R. (2003). A Breakdown of Reliability Coefficients by Test Type and Reliability Method, and the Clinical Implications of Low Reliability. *The Journal of general psychology*, *130*, 290–304.

Clark, R. (2004). The classical origins of Pavlov's conditioning. *Integrative physiological and behavioral science : the official journal of the Pavlovian Society*, *39*, 279–94.

Collin, C., Grand, V., Benson, N., Lazyan, M., Ginsburg, J., & Weeks, M. (2012). *The Psychology Book*. New York: DK.

Coppersmith, G., Dredze, M., Harman, C., & Hollingshead, K. (2015). From ADHD to SAD: Analyzing the Language of Mental Health on Twitter through Self-Reported Diagnoses. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, (pp. 1–10). Denver, CO, USA: Association for Computational Linguistics.
URL `https://aclanthology.org/W15-1201`

Coppersmith, G., Leary, R., Crutchley, P., & Fine, A. (2018). Natural Language Processing of Social Media as Screening for Suicide Risk. *Biomedical Informatics Insights*, *10*.
URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6111391/`

Coppersmith, G., Ngo, K., Leary, R., & Wood, A. (2016). Exploratory Analysis of Social Media Prior to a Suicide Attempt. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, (pp. 106–117).
URL `http://www.aclweb.org/anthology/W16-0311`

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Mach. Learn.*, *20*(3), 273–297.
URL `https://doi.org/10.1023/A:1022627411411`

Costa, P., & Mccrae, R. (1992). Neo PI-R professional manual. *Psychological Assessment Resources*, *396*.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334.
URL `https://doi.org/10.1007/BF02310555`

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302.

Danilevsky, M., Qian, K., Aharonov, R., Katsis, Y., Kawas, B., & Sen, P. (2020). A Survey of the State of Explainable AI for Natural Language Processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, (pp. 447–459). Suzhou, China: Association for Computational Linguistics.
URL `https://aclanthology.org/2020.aacl-main.46`

Darwin, C. (1859). *On the origin of species by means of natural selection, or, The preservation of favoured races in the struggle for life*, vol. 1859. London: John Murray.
URL `https://www.biodiversitylibrary.org/item/135954`

Deci, E. L., & Ryan, R. M. (2000). The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry*, *11*(4), 227–268.

Demasi, O., Hearst, M. A., & Recht, B. (2019). Towards augmenting crisis counselor training by improving message retrieval. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, (pp. 1–11). Minneapolis, MN, USA.
URL `https://www.aclweb.org/anthology/W19-3001`

Descartes, R. (1956). *Discourse on method*. Library of liberal arts (Macmillan Publishing Company). New York, London: Macmillan ; Collier Macmillan.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, (pp. 4171–4186). Minneapolis, MN, USA: Association for Computational Linguistics.
URL `https://www.aclweb.org/anthology/N19-1423`

DeYoung, C. G. (2011). Intelligence and personality. In *The Cambridge handbook of intelligence*, Cambridge handbooks in psychology, (pp. 711–737). New York, NY, US: Cambridge University Press.

Dickson, D., & Kelly, I. (1985). The 'Barnum Effect' in Personality Assessment: A Review of the Literature. *Psychological Reports*, *57*, 367–382.

Diehl, C., Hunkler, C., & Kristen, C. (2015). *Ethnische Ungleichheiten im Bildungsverlauf: Mechanismen, Befunde, Debatten.* Springer VS, 1 ed.

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology*, *41*, 417–440.

Ehrlich, H. (2009). Poe in Cyberspace: Google's First Trillion Pages: Web 2.0 and Beyond. *The Edgar Allan Poe Review*, *10*(1), 87–91.
URL http://www.jstor.org/stable/41507866

Ethayarajh, K. (2019). How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, (pp. 55–65). Hong Kong, China: Association for Computational Linguistics.
URL https://aclanthology.org/D19-1006

Exner, J. E. (1974). *The Rorschach: A comprehensive system.* The Rorschach: A comprehensive system. Oxford, England: John Wiley & Sons.

Ezen-Can, A. (2020). A Comparison of LSTM and BERT for Small Corpus. *CoRR, abs/2009.05451*.
URL https://arxiv.org/abs/2009.05451

Fernandez-Kelly, P. (2015). The Unequal Structure of the German Education System: Structural Reasons for Educational Failures of Turkish Youth in Germany. *Spaces & flows : an international journal of urban and extraurban studies*, *2*, 93–112.

Ffytche, M. (2011). *The foundation of the unconscious: Schelling, Freud and the birth of the modern psyche.* Cambridge University Press.

Fine, A., Frank, A. F., Jaeger, T. F., & Van Durme, B. (2014). Biases in Predicting the Human Language Model. In *52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference*, vol. 2, (pp. 7–12). Baltimore, MD, USA.

Flaskerud, J. (2012). Temperament and Personality: From Galen to DSM 5. *Issues in mental health nursing*, *33*, 631–4.

Folger, R., Cropanzano, R., & Goldman, B. (2005). What is the relationship between justice and morality. In J. Greenberg, & J. Colquitt (Eds.) *Handbook of organizational justice*, (pp. 215–245). Mahwah, NJ: Erlbaum.

Freud, S., & Mentzos, S. (1993). *Bruchstück einer Hysterie-Analyse: Nachw. v. Stavros Mentzos.* Frankfurt am Main: FISCHER Taschenbuch, 2 ed.

Fried, E. (2017). What are psychological constructs? On the nature and statistical modeling of emotions, intelligence, personality traits and mental disorders. *Health Psychology Review*, *11*, 1–11.

Fried, E. I., & Flake, J. K. (2018). Measurement Matters. *APS Observer*, *31*(3).
URL https://www.psychologicalscience.org/observer/measurement-matters

Fujita, H., Selamat, A., & Omatu, S. (2017). *New Trends in Intelligent Software Methodologies, Tools and Techniques: Proceedings of the 16th International Conference Somet 2017*. Washington, DC, USA: IOS Press.

Galati, G. (2015). A Simultaneous Invention – The Former Developments. In *100 Years of Radar*, (pp. 55 – 77). New York, NY, USA: Springer, 1 ed.

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635–E3644.
URL https://www.pnas.org/doi/10.1073/pnas.1720347115

Garimella, A., Banea, C., Hovy, D., & Mihalcea, R. (2019). Women's Syntactic Resilience and Men's Grammatical Luck: Gender-Bias in Part-of-Speech Tagging and Dependency Parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, (pp. 3493–3498). Florence, Italy: Association for Computational Linguistics.
URL https://aclanthology.org/P19-1339

Gawronski, B., & De Houwer, J. (2014). Implicit measures in social and personality psychology. *Handbook of research methods in social and personality psychology*, *2*, 283–310.

Gensler, H. J. (2011). *Ethics: A Contemporary Introduction.* New York: Routledge, 2nd ed ed.

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as Data. *Journal of Economic Literature*, *57*(3), 535–574.

Gerlitz, C., & Rieder, B. (2013). Mining One Percent of Twitter: Collections, Baselines, Sampling. *M/C Journal*, *16(2)*(620).

Gershon, E. S. (1983). Should science be stopped? The case of recombinant DNA research. *The Public interest*, *71*, 3–16.

Goldberg, L. R. (1981). Language and individual differences: The search for universals in personality lexicons. *Review of personality and social psychology*, *2*(1), 141–165.

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *The American Psychologist*, *48*(1), 26–34.

Goodman, J. C., Dale, P. S., & Li, P. (2008). Does frequency count? Parental input and the acquisition of vocabulary. *Journal of Child Language*, *35*(3), 515–531.

Gordon, J. (1997). John Stuart Mill and the "Marketplace of Ideas". *Social Theory and Practice*, *23*(2), 235–249.

Gottenbarn, D., Brinkman, B., Flick, C., Kirkpatrick, M. S., Miller, K., Varansky, K., & Wolf, M. J. (2018). ACM Code of Ethics and Professional Conduct.
URL `https://www.acm.org/code-of-ethics`

Gragert, E., Meier, J., & Fülscher, C. (2018). Campus Forum. Tech. Rep. 66, NORDAKADEMIE, Elmshorn, Germany.
URL `https://www.nordakademie.de/sites/default/files/2019-08/CF_66_final.pdf`

Graham, G. (2010). *Theories of Ethics: An Introduction to Moral Philosophy with a Selection of Classic Readings*. New York: Taylor & Francis Ltd, 1st edition ed.

Grankvist, G., Kajonius, P., & Persson, B. (2016). The Relationship between Mind-Body Dualism and Personal Values. *International Journal of Psychological Studies*, *8*, 126–132.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning Word Vectors for 157 Languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, (pp. 3483–3487). Miyazaki, Japan.

Gubler, D. A., Makowski, L. M., Troche, S. J., & Schlegel, K. (2020). Loneliness and Well-Being During the Covid-19 Pandemic: Associations with Personality and Emotion Regulation. *Journal of Happiness Studies*.
URL `https://doi.org/10.1007/s10902-020-00326-5`

Gulliksen, H. (1950). *Theory of mental tests*. Theory of mental tests. Hoboken, NJ, US: John Wiley & Sons Inc.

Gupta, A., Agrawal, D., Chauhan, H., Dolz, J., & Pedersoli, M. (2018). An Attention Model for Group-Level Emotion Recognition. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, ICMI '18, (pp. 611–615). New York, NY, USA.

Gutiérrez-Romero, R. (2020). Conflict in Africa during COVID-19: social distancing, food vulnerability and welfare response. *SSRN Electronic Journal*.

Görke, A., & Scholl, A. (2006). Niklas Luhmann'S theory of social systems and journalism research. *Journalism Studies*, *7*, 644–655.

Görz, G., Schmid, U., & Braun, T. (2020). *Handbuch der Künstlichen Intelligenz*. Berlin ; Boston: De Gruyter Oldenbourg, 6 ed.

Hall, M. A. (2000). *Correlation-Based Feature Selection for Machine Learning*. dissertation, University of Auckland, New Zealand.

Hansson, S. O. (2013). *Defining pseudoscience and science*. Chicago, IL, USA: University of Chicago Press.

Hansson, S. O. (2017). Science and Pseudo-Science. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2017 ed.
URL `https://plato.stanford.edu/archives/sum2017/entries/pseudo-science/`

Harris, Z. S. (1954). Distributional structure. *Word*, *10*, 146–162.

Hausknecht, J., Halpert, J., Paolo, N., & Gerrard, M. (2007). Retesting in selection: A meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology*, *92*, 373–385.

Hawkins, R., & Boyd, R. (2017). Such Stuff as Dreams Are Made On: Dream Language, LIWC Norms, & Personality Correlates. *Dreaming*, *27*.

Hell, B., Trapmann, S., & Schuler, H. (2007). Eine Metaanalyse der Validität von fachspezifischen Studierfähigkeitstests im deutschsprachigen Raum. *Empirische Pädagogik*, *21*(3), 251–270.

Hermann, K. M., Kočiský, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., & Blunsom, P. (2015). Teaching Machines to Read and Comprehend. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, (pp. 1693–1701). Cambridge, MA, USA.
URL `http://dl.acm.org/citation.cfm?id=2969239.2969428`

Heyer, G., Quasthoff, U., & Wittig, T. (2006). *Text Mining: Wissensrohstoff Text: Konzepte, Algorithmen, Ergebnisse*. Herdecke: W3L GmbH, 1 ed.

Hildebrandt, B., & Köhler, S. (2022). *Statistik I*. Ruhr-Universität Bochum, 1 ed.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

Hlatywayo, C., Mhlanga, T., & Zingwe, T. (2013). Neuroticism as a Determinant of Job Satisfaction among Bank Employees. *Mediterranean Journal of Social Sciences*, *4*, 549–554.

Hochreiter, S. (1998). The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *6*(02), 107–116.

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-term Memory. *Neural computation*, *9*, 1735–1780.

Hogenraad, R. (2003). The Words that Predict the Outbreak of Wars. *Empirical Studies of the Arts*, *21*, 5–20.

Homan, C., Johar, R., Liu, T., Lytle, M., Silenzio, V., & Alm, C. O. (2014). Toward macro-insights for suicide prevention: Analyzing fine-grained distress at scale. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, (pp. 107–117). Association for Computational Linguistics.
URL `http://www.aclweb.org/anthology/W14-3213`

Hovy, D., & Prabhumoye, S. (2021). Five sources of bias in natural language processing. *Language and Linguistics Compass*, *15*(8), e12432.
URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/lnc3.12432`

Hovy, D., & Søgaard, A. (2015). Tagging Performance Correlates with Author Age. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, (pp. 483–488). Beijing, China: Association for Computational Linguistics.
URL `https://aclanthology.org/P15-2079`

Hursthouse, R., & Pettigrove, G. (2018). Virtue Ethics. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2018 ed.
URL `https://plato.stanford.edu/archives/win2018/entries/ethics-virtue/`

Hämmig, O. (2019). Health risks associated with social isolation in general and in young, middle and old age. *PLoS ONE*, *14*(7), e0219663.
URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6638933/`

Impart, I., & Kuhl, J. (2013). *Auswertungsmanual für den Operanten Multi-Motiv-Test OMT*. Münster: sonderpunkt Verlag, 1 ed.

Jackson, S., Parker, C., & Dipboye, R. (1996). A Comparison of Competing Models Underlying Responses to the Myers-Briggs Type Indicator. *Journal of Career Assessment*, *4*, 99–111.

Jain, S., & Wallace, B. C. (2019). Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, (pp. 3543–3556). Minneapolis, MN, USA.
URL `https://www.aclweb.org/anthology/N19-1357`

Johannßen, D., & Biemann, C. (2018). Between the Lines: Machine Learning for Prediction of Psychological Traits - A Survey. In *Proceedings of the International Cross-Domain Conference*, (pp. 192–211). Hamburg, Germany: Springer.
URL `https://www.researchgate.net/publication/327180648_Between_the_Lines_Machine_Learning_for_Prediction_of_Psychological_Traits_-_A_Survey_Second_IFIP_TC_5_TC_8WG_84_89_TC_12WG_129_International_Cross-Domain_Conference_CD-MAKE_2018_Hamburg_Germany_August`

Johannßen, D., & Biemann, C. (2019). Neural classification with attention assessment of the implicit-association test OMT and prediction of subsequent academic success. In *Proceedings of the 15th Conference on Natural Language Processing, KONVENS*, (pp. 68–78). Erlangen, Germany: German Society for Computational Linguistics & Language Technology.
URL `https://corpora.linguistik.uni-erlangen.de/data/konvens/proceedings/papers/KONVENS2019_paper_14.pdf`

Johannßen, D., & Biemann, C. (2020). Social Media Unrest Prediction during the COVID-19 Pandemic: Neural Implicit Motive Pattern Recognition as Psychometric Signs of Severe Crises. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, (pp. 74–86). Barcelona, Spain (Online).
URL `https://www.aclweb.org/anthology/2020.peoples-1.8`

Johannßen, D., Biemann, C., Remus, S., Baumann, T., & Scheffer, D. (2020a). GermEval 2020 Task 1 on the Classification and Regression of Cognitive and Motivational style from Text. In *Proceedings of the GermEval 2020 Task 1 Workshop in conjunction with the 5th SwissText & 16th KONVENS Joint Conference 2020*, (pp. 1–10). Zurich, Switzerland (online): German Society for Computational Linguistics & Language Technology.

Johannßen, D., Biemann, C., & Scheffer, D. (2019). Reviving a psychometric measure: Classification of the Operant Motive Test. In *Proceedings of the Sixth Annual Workshop on Computational Linguistics and Clinical Psychology (CLPsych)*, (pp. 121–125). Minneapolis, MN, USA.
URL `https://www.aclweb.org/anthology/W19-3014`

Johannßen, D., Biemann, C., & Scheffer, D. (2020b). Ethical considerations of the GermEval20 Task 1. IQ assessment with natural language processing: Forbidden research or gain of knowledge? In *Proceedings of the GermEval 2020 Task 1 Workshop in conjunction with the 5th SwissText & 16th KONVENS Joint Conference 2020*, (pp. 30–44). Zurich, Switzerland (online): German Society for Computational Linguistics & Language Technology.

Johannßen, D., Biemann, C., & Scheffer, D. (2022). Classification of German Jungian Extraversion and Introversion Texts with Assessment of Changes during the COVID-19 Pandemic. In *LREC 2022 Workshop RaPID-4: Resources and Processing of linguistic, para-linguistic and extra-linguistic Data from people with various forms of cognitive, psychiatric and/or developmental impairments*, (pp. 31–40). Marseille, France: European Language Resources Association (ELRA).
URL `https://www.aclweb.org/anthology/2020.peoples-1.8`

Johansson, F., Brynielsson, J., Hörling, P., Malm, M., Mårtenson, C., Truvé, S., & Rosell, M. (2011). Detecting Emergent Conflicts through Web Mining and Visualization. In *Proceedings - 2011 European Intelligence and Security Informatics Conference, EISIC 2011*, (pp. 346 – 353). Athens, Greece: Springer.

Jung, C. G. (1921). *Psychologische Typen.* Zürich, Rascher,.

Jung, C. G. (1991). *The Archetypes and the Collective Unconscious.* London: Taylor & Francis Ltd., 2 ed.

Jung, C. G. (2010). *Psychology of the Unconscious: A Study of the Transformations and Symbolisms of the Libido A Contribution to the History of the Evolution of Thought.* Nabu Press.

Jurafsky, D., & Martin, J. H. (2008). *Jurafsky, D: Speech and Language Processing: International Edition.* Upper Saddle River, NJ: Prentice Hall, 2 ed.

Kalampokis, E., Tambouris, E., & Tarabanis, K. (2013). Understanding the Predictive Power of Social Media. *Internet Research*, *23*(5).

Kaptein, M., & Wempe, J. (2002). Three General Theories of Ethics and the Integrative Role of Integrity Theory. *SSRN Electronic Journal*.

Keh, S. S., & Cheng, I.-T. (2019). Myers-Briggs Personality Classification and Personality-Specific Language Generation Using Pre-trained Language Models. *ArXiv*, *abs/1907.06333*.

Keller, H. (1998). *Lehrbuch Entwicklungspsychologie.* Bern Göttingen: Hogrefe AG.

Kelly, K., & Jugovic, H. (2001). Concurrent Validity of the Online Version of the Keirsey Temperament Sorter II. *Journal of Career Assessment*, *9*, 49–59.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In Y. Bengio, & Y. LeCun (Eds.) *3rd International Conference on Learning Representations, ICLR*. San Diego, CA, USA.

Kozlowski, A. C., Taddy, M., & Evans, J. A. (2019). The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings. *American Sociological Review*, *84*(5), 905–949.
URL `https://doi.org/10.1177/0003122419877135`

Kramer, R.-T., Helsper, W., Thiersch, S., & Ziems, C. (2009). *Selektion und Schulkarriere: Kindliche Orientierungsrahmen beim Übergang in die Sekundarstufe I*. Studien zur Schul- und Bildungsforschung. VS Verlag für Sozialwissenschaften.
URL `https://www.springer.com/de/book/9783531162096`

Kshirsagar, R., Morris, R., & Bowman, S. (2017). Detecting and Explaining Crisis. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*, (pp. 66–73). Vancouver, BC, Canada.

Kuhl, J. (2000). A Functional-Design Approach to Motivation and Self-Regulation: The Dynamics of Personality Systems Interactions. *Handbook of Self-regulation*.

Kuhl, J., & Kazén, M. (1999). Volitional facilitation of difficult intentions: Joint activation of intention memory and positive affect removes Stroop interference. *Journal of Experimental Psychology: General*, *128*(3), 382–399.

Kuhl, J., Kazén, M., & Quirin, M. (2014). The Theory of Personality Systems Interactions (PSI). *Revista Mexicana de Psicología*, *31*, 90–99.

Kuhl, J., & Scheffer, D. (1999). *Der operante Multi-Motiv-Test (OMT): Manual [The operant multi-motive-test (OMT): Manual]*. Osnabrück, Germany: University of Osnabrück: Impart.

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. New York: Springer, 1st ed. 2013, corr. 2nd printing 2018 edition ed.

Kunze, L. (2019). Can We Stop the Academic AI Brain Drain? *KI - Künstliche Intelligenz*, *33*(1), 1–3.
URL `https://doi.org/10.1007/s13218-019-00577-2`

Kutuzov, A., Velldal, E., & Øvrelid, L. (2019). One-to-X Analogical Reasoning on Word Embeddings: a Case for Diachronic Armed Conflict Prediction from News Texts. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, (pp. 196–201). Florence, Italy: Association for Computational Linguistics.
URL `https://www.aclweb.org/anthology/W19-4724`

Lai, S., Xu, L., Liu, K., & Zhao, J. (2015). Recurrent convolutional neural networks for text classification. In *Twenty-ninth AAAI conference on artificial intelligence*, (pp. 2267–2273). Austin, TX, USA.

Lancashire, I., & Hirst, G. (2009). Vocabulary Changes in Agatha Christie's Mysteries as an Indication of Dementia: A Case Study. In *Cognitive Aging: Research and Practice*, Cognitive Aging: Research and Practice, (pp. 8–10).

Landwehr, N., Hall, M. A., & Frank, E. (2005). Logistic Model Trees. *Machine Learning*, *59*(1), 161–205.

Lang, J. W. B., Zettler, I., Ewen, C., & Hülsheger, U. R. (2012). Implicit motives, explicit traits, and task and contextual performance at work. *The Journal of Applied Psychology*, *97*(6), 1201–1217.

Lazarevic, L., & Orlic, A. (2015). *Implicit Assessment: Paradigm of Implicit Measurement in the Field of Individual Differences*. Institute of Psychology, Faculty of Philosophy and Faculty of Sport and Physical Education, University of Belgrade, Serbia.

Lester, H., & Howe, A. (2008). Depression in primary care: three key challenges. *Postgraduate Medical Journal*, *84*(996), 545–548.

Liepmann, D., Beauducel, A., Brocke, B., & Amthauer, R. (2007). *Intelligenz-Struktur-Test 2000 R*. Hogrefe Verlag, Göttingen, Germany.

Lilienfeld, S., Wood, J. M., & Garb, H. (2000). The Scientific Status of Projective Techniques. *Psychological science in the public interest : a journal of the American Psychological Society*.

Lindberg, V. (2008). *Intellectual Property and Open Source: A Practical Guide to Protecting Code*. Sebastopol, CA, USA: O'Reilly & Associates.

Lombardo, P. (2010). Three Generations, No Imbeciles: Eugenics, the Supreme Court, and Buck v. Bell. *Three Generations, No Imbeciles: Eugenics, the Supreme Court, and Buck v. Bell*, *101*, 1–365.

Lord, F., Novick, M., & Birnbaum, A. (1968). *Statistical theories of mental test scores*. Statistical theories of mental test scores. Oxford, England: Addison-Wesley.

Luhn, H. P. (1957). A Statistical Approach to Mechanized Encoding and Searching of Literary Information. *IBM J. Res. Dev.*, *1*(4), 309–317.
URL https://doi.org/10.1147/rd.14.0309

Lynn, V., Balasubramanian, N., & Schwartz, H. A. (2020). Hierarchical Modeling for User Personality Prediction: The Role of Message-Level Attention. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, (pp. 5306–5316). Online: Association for Computational Linguistics.
URL `https://www.aclweb.org/anthology/2020.acl-main.472`

Mahoney, M. J. (1984). Psychoanalysis and Behaviorism. In H. Arkowitz, & S. B. Messer (Eds.) *Psychoanalytic Therapy and Behavior Therapy: Is Integration Possible?*, (pp. 303–325). Boston, MA: Springer US.
URL `https://doi.org/10.1007/978-1-4613-2733-2_23`

Manning, C., & Schütze, H. (1999). *Foundations of Statistical Natural Language Proc*. Cambridge, Mass: MIT Press.

Manning, C. D. (2015). Computational Linguistics and Deep Learning. *Computational Linguistics*, *41*, 699–705.

Martin, B. (1978). Can scientific development be stopped? *Australian Science Teachers Journal*, *24*(1), 65–70.

Masrani, V., Murray, G., Field, T., & Carenini, G. (2017). Detecting Dementia through Retrospective Analysis of Routine Blog Posts by Bloggers with Dementia. In *BioNLP 2017*, (pp. 232–237). Vancouver, BC, Canada: Association for Computational Linguistics.

Matero, M., Idnani, A., Son, Y., Giorgi, S., Vu, H., Zamani, M., Limbachiya, P., Guntuku, S. C., & Schwartz, H. A. (2019). Suicide Risk Assessment with Multi-level Dual-Context Language and BERT. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, (pp. 39–44). Minneapolis, MN, USA.
URL `https://www.aclweb.org/anthology/W19-3005`

McAdams, D. P., Jackson, R. J., & \mboxKirshnit, C. (1984). Looking, laughing, and smiling in dyads as a function of intimacy motivation and reciprocity. *Journal of Personality*, *52*(3), 261–273.

McClelland, D. C. (1988). *Human Motivation*. Cambridge University Press.

McClelland, D. C., & Boyatzis, R. (1982). Leadership Motive Pattern and Long-Term Success in Management. *Journal of Applied Psychology*, *67*, 737–743.

McClelland, D. C., & Davis, W. N. (1972). *The Drinking Man: Alcohol and Human Motivation*. New York: Free Press.

McClelland, D. C., Koestner, R., & Weinberger, J. (1989). How do self-attributed and implicit motives differ? *Psychological Review*, *96*(4), 690–702.

McCrae, R. R., & Costa Jr., P. T. (1999). A Five-Factor theory of personality. In *Handbook of personality: Theory and research, 2nd ed*, (pp. 139–153). New York, NY, US: Guilford Press.

McHugh, M. (2012). Interrater reliability: The kappa statistic. *Biochemia medica: časopis Hrvatskoga društva medicinskih biokemičara / HDMB*, *22*, 276–82.

Mehta, Y., Majumder, N., Gelbukh, A., & Cambria, E. (2019). Recent trends in deep learning based personality detection. *Artificial Intelligence Review*, *53*(4), 2313–2339.
URL `http://dx.doi.org/10.1007/s10462-019-09770-z`

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). Efficient Estimation of Word Representations in Vector Space. In *Proceedings of the International Conference on Learning Representations, ICLR*. Scottsdale, AZ, USA.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013b). Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26*, (pp. 3111–3119). Lake Tahoe, NV, USA: Neural Information Processing Systems Conference.

Mill, J. S. (1871). *Utilitarianism*. Longmans, Green, Reader, and Dyer.

Mill, J. S. (2011). *On Liberty*. Cambridge Library Collection - Philosophy. Cambridge University Press.

Mohammady, E., & Culotta, A. (2014). Using County Demographics to Infer Attributes of Twitter Users. In *Proceedings of the Joint Workshop on Social Dynamics and Personal Attributes in Social Media*, (pp. 7–16). Baltimore, MD, USA: Association for Computational Linguistics.
URL `https://aclanthology.org/W14-2702`

Morales, M., Scherer, S., & Levitan, R. (2017). A Cross-modal Review of Indicators for Depression Detection Systems. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*, (pp. 1–12). Association for Computational Linguistics.
URL `http://www.aclweb.org/anthology/W17-3101`

Morgan, C. D., & Murray, H. A. (1935). A method for investigating fantasies: the thematic apperception test. *Archives of Neurology & Psychiatry*, *34*, 289–306.

Mueller, H., & Rauh, C. (2017). Reading Between the Lines: Prediction of Political Violence Using Newspaper Text. *American Political Science Review*, *112*, 1–18.

Murray, H. A. (1943). *Thematic apperception test*. Thematic apperception test. Cambridge, MA, US: Harvard University Press.

Nachtwei, J., & Schermuly, C. C. (2009). Acht Mythen über Eignungstests. *Harvard Business Manager*, (04/2009), 6–10.
URL `https://www.harvardbusinessmanager.de/heft/d-64530881.html`

Neale, S. (1992). Paul Grice and the Philosophy of Language. *Linguistics and Philosophy*, *15*, 509–559.

Niederhoffer, K., Schler, J., Crutchley, P., Loveys, K., & Coppersmith, G. (2017). In your wildest dreams: the language and psychological features of dreams. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*, (pp. 13–25). Association for Computational Linguistics.
URL `http://www.aclweb.org/anthology/W17-3102`

Novikova, I. A. (2013). Big Five (The Five-Factor Model and The Five-Factor Theory). In *The Encyclopedia of Cross-Cultural Psychology*, (pp. 136–138). John Wiley & Sons, Ltd.
URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118339893.wbeccp054`

Oak, M., Behera, A., Thomas, T., Alm, C. O., Prud'hommeaux, E., Homan, C., & Ptucha, R. (2016). Generating Clinically Relevant Texts: A Case Study on Life-Changing Events. *Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology*, (pp. 85–94).
URL `https://aclanthology.info/papers/W16-0309/w16-0309`

O'Leary, D. (2019). GOOGLE'S Duplex: Pretending to be human. *Intelligent Systems in Accounting, Finance and Management*, *26*, 46–53.

O'Neill, J. (1995). The role of ARPA in the development of the ARPANET, 1961-1972. *Annals of the History of Computing, IEEE*, *17*, 76 – 81.

Ones, D. S., Dilchert, S., Viswesvaran, C., & Salgado, J. F. (2017). Cognitive ability: Measurement and validity for employee selection. In *Handbook of Employee Selection, Second Edition*, (pp. 251–276). Taylor and Francis.

Ostendorf, F., & Angleitner, A. (2004). *NEO-Persönlichkeitsinventar nach Costa und McCrae: NEO-PI-R ; Manual*. Hogrefe.
URL `https://pub.uni-bielefeld.de/record/1878577`

Owsnicki-Klewe, B. (2002). *Algorithmen und Datenstrukturen. Inf & Ing, Bd. 5*. Augsburg: Wißner-Verlag, 5. aufl. edition ed.

Pang, J., & Ring, H. (2020). Automated coding of implicit motives: A machine-learning approach. *Motivation and Emotion*, *44*.

Paulhus, D. (1984). Two-Component Models of Socially Desirable Responding. *Journal of Personality and Social Psychology*, *46*, 598–609.

Pavlov, I. P. (1906). The Scientific Investigation of the Psychical Faculties or Processes in the Higher Animals. *Science*, *24*(620), 613–619.
  URL `https://www.science.org/doi/abs/10.1126/science.24.620.613`

Pece, C. (2020). Federal R&D Obligations Increase 8.8% in FY 2018; Preliminary FY 2019 R&D Obligations Increase 9.3% Over FY 2018. Tech. Rep. 20-308, National Science Foundation.
  URL `https://www.nsf.gov/statistics/2020/nsf20308/`

Pennebaker, J., Chung, C., Ireland, M., Gonzales, A., & Booth, R. (2007). The Development and Psychometric Properties of LIWC2007. *Software manual. http://liwc.wpengine.com.*

Pennebaker, J. W., Chung, C. K., Frazee, J., Lavergne, G. M., & Beaver, D. I. (2014). When Small Words Foretell Academic Success: The Case of College Admissions Essays. *PLOS ONE*, *9*(12), e115844.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (1999). Linguistic inquiry and word count (LIWC). *Software manual..*

Pennebaker, J. W., & King, L. (2000). Linguistic styles: Language use as an individual difference. *Journal of personality and social psychology*, *77*, 1296–312.

Pflege, L., & Menche, N. (2014). *Pflege Heute: mit Zugang zu pflegeheute.de: Lehrbuch für Pflegeberufe. + pflegeheute.de. Inkl. Download.* München: Urban & Fischer Verlag/Elsevier GmbH, 6 ed.

Pisarov, J., & Mester, G. (2020). The Future of Autonomous Vehicles. *FME Transactions*, *49*, 29–35.

Pittenger, D. (2005). Cautionary comments regarding the Myers-Briggs Type Indicator. *Consulting Psychology Journal: Practice and Research*, *57*, 210–221.

Plank, B., & Hovy, D. (2015). Personality Traits on Twitter—or—How to Get 1,500 Personality Tests in a Week. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, (pp. 92–98). Lisboa, Portugal: Association for Computational Linguistics.
  URL `https://www.aclweb.org/anthology/W15-2913`

Plate, G. (2016). Häufig gestellte Fragen (FAQs) - NORDAKADEMIE - Hochschule der Wirtschaft.
  URL `https://www.nordakademie.de/bewerber/kontakt/haeufig-gestellte-fragen-faqs/`

Pool, C., & Nissim, M. (2016). Distant supervision for emotion detection using Facebook reactions. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, (pp. 30–39).
URL `https://aclanthology.info/papers/W16-4304/w16-4304`

Popper, K. R. (2002). *Logik der Forschung*. Tübingen: Mohr Siebeck.

Post, A., Gilljam, H., Bremberg, S., & Galanti, M. R. (2008). Maternal smoking during pregnancy: a comparison between concurrent and retrospective self-reports. *Paediatric and Perinatal Epidemiology*, *22*(2), 155–161.

Prost, F., Thain, N., & Bolukbasi, T. (2019). Debiasing Embeddings for Reduced Gender Bias in Text Classification. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, (pp. 69–75). Florence, Italy: Association for Computational Linguistics.
URL `https://aclanthology.org/W19-3810`

Pérez-Rosas, V., Mihalcea, R., Resnicow, K., Singh, S., & An, L. (2016). Building a Motivational Interviewing Dataset. *Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology*, (pp. 42–51).
URL `https://aclanthology.info/papers/W16-0305/w16-0305`

Quenk, N. L. (2009). *Essentials Myers-Briggs Type Indicator Assessment*. Hoboken, N.J: Wiley, 2 ed.

Raad, B., & Mlacic, B. (2015). Big Five Factor Model, Theory and Structure. In *International Encyclopedia of the Social & Behavioral Sciences*, (pp. 559–566).

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. *OpenAI blog*.
URL `https://openai.com/blog/language-unsupervised/`

Ragheb, M., & Dickinson, M. (2013). Inter-annotator Agreement for Dependency Annotation of Learner Language. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, (pp. 169–179). Atlanta, GA, USA: Association for Computational Linguistics.
URL `https://aclanthology.org/W13-1723`

Rainey, S., McGillivray, K., Akintoye, S., Fothergill, T., Bublitz, C., & Stahl, B. (2020). Is the European Data Protection Regulation sufficient to deal with emerging data concerns relating to neurotechnology? *Journal of Law and the Biosciences*, *7*.

Ramet, G., Garner, P. N., Baeriswyl, M., & Lazaridis, A. (2018). Context-Aware Attention Mechanism for Speech Emotion Recognition. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, (pp. 126–131). Athens, Greece.

Rammstedt, B., Lechner, C. M., & Danner, D. (2018). Relationships between Personality and Cognitive Ability: A Facet-Level Analysis. *Journal of Intelligence*, *6*(2), 28.
URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6480763/`

Rao, D., Mohandas, G., & McMahan, B. (2019). *Natural Language Processing with PyTorch: Build Intelligent Language Applications Using Deep Learning*. Sebastopol, CA: O'Reilly UK Ltd., 1st edition ed.

Reece, A. G., Reagan, A. J., Lix, K. L. M., Dodds, P. S., Danforth, C. M., & Langer, E. J. (2017). Forecasting the onset and course of mental illness with Twitter data. *Nature Scientific Reports*, *7*(1), 13006.
URL `https://www.nature.com/articles/s41598-017-12961-9`

Roberts, B., Kuncel, N., Shiner, R., Caspi, A., & Goldberg, L. (2007). The Power of Personality The Comparative Validity of Personality Traits, Socioeconomic Status, and Cognitive Ability for Predicting Important Life Outcomes. *Perspectives on Psychological Science*, *2*.

Roberts, B. W., Jackson, J. J., Fayard, J. V., Edmonds, G., & Meints, J. (2009). Conscientiousness. In *Handbook of individual differences in social behavior*, (pp. 369–381). New York, NY, US: The Guilford Press.

Roberts, J. (2014). Freddie Lee Hall, Petitioner v. Florida.
URL `https://www.apa.org/about/offices/ogc/amicus/hall`

Rorschach, H. (1932). *Psychodiagnostik. Methodik und Ergebnisse eines wahrnehmungsdiagnostischen Experiments. [Psycho-diagnosis. Method and results of a perception-diagnostic experiment.].* Psychodiagnostik. Methodik und Ergebnisse eines wahrnehmungsdiagnostischen Experiments. Oxford, England: Huber.

Ross, S., Lutz, C., & Bailley, S. (2004). Psychopathy and the Five Factor Model in a Noninstitutionalized Sample: A Domain and Facet Level Analysis. *Journal of Psychopathology and Behavioral Assessment*, *26*, 213–223.

Roth, M., & Herzberg, P. Y. (2008). Psychodiagnostik in der Praxis: State of the Art? *Klinische Diagnostik und Evaluation*, *1*(1), 5–18.

Runge, M., Lang, J., Engeser, S., Schüler, J., Hartog, S., & Zettler, I. (2016). Modeling motive activation in the Operant Motive Test: A psychometric analysis using dynamic Thurstonian item response theory. *Motivation Science*, *2*, 268–286.

Rushton, J. P., & Jensen, A. R. (2005). Thirty years of research on race differences in cognitive ability. *Psychology, Public Policy, and Law*, *11*(2), 235–294.

Russell, S., & Norvig, P. (2012). *Künstliche Intelligenz*. München u.a.: Pearson Studium, 3 ed.

Russell, S., & Norvig, P. (2016). *Artificial Intelligence: A Modern Approach, Global Edition*. Boston Columbus Indianapolis New York San Francisco Upper Saddle River Amsterdam, Cape Town Dubai London Madrid Milan Munich Paris Montreal Toronto Delhi Mexico City Sao Paulo Sydney Hong Kong Seoul Singapore Taipei Tokyo: Addison Wesley, 3. edition ed.

Sahakian, W. S., & Sahakian, M. L. (1993). *Ideas of the Great Philosophers*. New York, NY, USA: Barnes & Noble.

Sahlgren, M. (2008). The distributional hypothesis. *Italian Journal of Linguistics*, *20*.

Sanger, R. M. (2015). IQ, Intelligence Tests, 'Ethnic Adjustments' and Atkins. SSRN Scholarly Paper ID 2706800, Social Science Research Network, Rochester, NY, USA.
URL https://papers.ssrn.com/abstract=2706800

Sarges, W., & Scheffer, D. (2008). *Innovative Ansätze für die Eignungsdiagnostik [Innovative Approaches for Aptitute Diagnostics]*. Göttingen, Germany: Hogrefe Verlag, 1st ed.

Sato, J. (2017). Additional report about the validity of the Jung Psychological Types Scale. *Online Journal of Japanese Clinical Psychology*, *4*, 1–7.

Savoldi, B., Gaido, M., Bentivogli, L., Negri, M., & Turchi, M. (2021). Gender Bias in Machine Translation. *Transactions of the Association for Computational Linguistics*, *9*, 845–874.
URL https://aclanthology.org/2021.tacl-1.51

Scheffer, D. (2004). *Implizite Motive: Entwicklung, Struktur und Messung [Implicit Motives: Development, Structure and Measurement]*. Göttingen, Germany: Hogrefe Verlag, 1st ed.

Scheffer, D., Eisermann, J., Tissen-Diabaté, T., Gemar, C., & Boecker, D. (2016). An implicit Approach in Measuring Personality Traits by the Visual Questionnaire (ViQ®): Psychometric Properties, Validation and Scope of Application. Tech. rep., ViQ Academy.
URL http://viq-academy.de/wp-content/uploads/2016/06/ViQ-White-Paper.pdf

Scheffer, D., & Kuhl, J. (2013). *Auswertungsmanual für den Operanten Multi-Motiv-Test OMT*. Münster, Germany: sonderpunkt Verlag.

Scheffer, D., & Loerwald, D. (2008). Messung von Persönlichkeitseigenschaften mit dem Visual Questionnaire (ViQ). In *Innovative Ansätze für die Eignungsdiagnostik*, (pp. 51–63). Göttingen, Niedersachs: Hogrefe Verlag, 1 ed.

Schira, J. (2012). *Statistische Methoden der VWL und BWL: Theorie und Praxis*. Pearson Studium, 4 ed.

Schlehahn, E., Aichroth, P., Mann, S., Schreiner, R., Lang, U., Shepherd, I., & Wong, B. (2015). *Benefits and Pitfalls of Predictive Policing*. Springer.

Schleithoff, F. (2015). Noteninflation im deutschen Schulsystem — Macht das Abitur hochschulreif? *ORDO — Jahrbuch für die Ordnung von Wirtschaft und Gesellschaft*, *66*, 3–26.

Schmidt, A., & Wiegand, M. (2017). A Survey on Hate Speech Detection using Natural Language Processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, (pp. 1–10). Association for Computational Linguistics.
URL http://aclweb.org/anthology/W17-1101

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, *124*(2), 262–274.

Schmidt-Atzert, L., Amelang, M., Fydrich, T., & Moosbrugger, H. (2018). *Psychologische Diagnostik: Extras Online*. Berlin Heidelberg: Springer, 5 ed.

Schultheiss, O. (2013). Are implicit motives revealed in mere words? Testing the marker-word hypothesis with computer-based text analysis. *Frontiers in Psychology*, *4*.
URL https://www.frontiersin.org/articles/10.3389/fpsyg.2013.00748

Schultheiss, O., & Pang, J. (2007). Measuring implicit motives. In *Handbook of research methods in personality psychology*, (pp. 322–344). New York, NY, US: Guilford Press.

Schultheiss, O. C., & Brunstein, J. C. (2010). *Implicit Motives*. Oxford ; New York: Oxford Univ Pr, 1 ed.

Schäfer, H., Idrissi-Yaghir, A., Schimanowski, A., Bujotzek, M. R., Damm, H., Nagel, J., & Friedrich, C. M. (2020). Predicting Cognitive and Motivational Style from German Text using Multilingual Transformer Architectures. In *Proceedings of the GermEval 2020 Task 1 Workshop in conjunction with the 5th SwissText & 16th KONVENS Joint Conference 2020*, (pp. 17–22). Zurich, Switzerland (online): German Society for Computational Linguistics & Language Technology.

Schüler, J., Brandstätter, V., Wegner, M., & Baumann, N. (2015). Testing the convergent and discriminant validity of three implicit motive measures: PSE, OMT, and MMG. *Motivation and Emotion*, *39*(6), 839–857.
URL https://doi.org/10.1007/s11031-015-9502-1

Schütze, H. (2020). ACL Code of Ethics.
URL https://www.aclweb.org/portal/content/acl-code-ethics

Scutari, M., & Denis, J.-B. (2014). *Bayesian Networks: With Examples in R*. s.l.: Chapman & Hall/Crc Texts in Statistical Science.

Segal, D., & Coolidge, F. (2001). Diagnosis and Classification. In *Journal of Personality Assessment*, (pp. 5–22). PubMed.

Serrà, J., Leontiadis, I., Spathis, D., Stringhini, G., Blackburn, J., & Vakali, A. (2017). Class-based Prediction Errors to Detect Hate Speech with Out-of-vocabulary Words. In *Proceedings of the First Workshop on Abusive Language Online*, (pp. 36–40). Association for Computational Linguistics.
URL `http://www.aclweb.org/anthology/W17-3005`

Shah, D. S., Schwartz, H. A., & Hovy, D. (2020). Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, (pp. 5248–5264). Online: Association for Computational Linguistics.
URL `https://aclanthology.org/2020.acl-main.468`

Shamdasani, S. (2010). *Jung and the Making of Modern Psychology: The Dream of a Science*. Cambridge, UK ; New York: Cambridge University Press.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technology Journal*, *27*(3), 379–423.
URL `http://dblp.uni-trier.de/db/journals/bstj/bstj27.html#Shannon48`

Shen, J. H., & Rudzicz, F. (2017). Detecting Anxiety on Reddit. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*, (pp. 58–65). Vancouver, BC, Canada.
URL `http://www.aclweb.org/anthology/W17-3107`

Shrout, P., & Fleiss, J. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428.

Silver, D., Huang, A., Maddison, C., Guez, A., Sifre, L., Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*, 484–489.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, *15*(1), 72–101.

Stajner, S., Yenikent, S., Ghanem, B., & Franco-Salvador, M. (2021). What Motivates You? Benchmarking Automatic Detection of Basic Needs from Short Posts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, (pp.

803–810). Online: Association for Computational Linguistics.
URL https://aclanthology.org/2021.acl-short.101

Stein, R., & Swan, A. B. (2019). Evaluating the validity of Myers-Briggs Type Indicator theory: A teaching tool and window into intuitive psychology. *Social and Personality Psychology Compass*, *13*(2), e12434.
URL https://onlinelibrary.wiley.com/doi/abs/10.1111/spc3.12434

Steinberg, L., & Thissen, D. (2013). Item response theory. In *The Oxford handbook of research strategies for clinical psychology*, Oxford library of psychology, (pp. 336–373). New York, NY, US: Oxford University Press.

Tan, Z., Wang, M., Xie, J., Chen, Y., & Shi, X. (2018). Deep Semantic Role Labeling with Self-Attention. In *Thirty-Second AAAI Conference on Artificial Intelligence*, (pp. 4929–4936). New Orleans, LA, USA.

Thompson, B., & Borrello, G. M. (1986). Construct Validity of the Myers-Briggs Type Indicator. *Educational and Psychological Measurement*, *46*(3), 745–752.
URL https://doi.org/10.1177/0013164486463032

Tighe, E., & Cheng, C. (2018). Modeling Personality Traits of Filipino Twitter Users. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, (pp. 112–122). Louisiana, LA, USA: Association for Computational Linguistics.
URL https://www.aclweb.org/anthology/W18-1115

Tomasello, M. (2002). *The New Psychology of Language: Cognitive and Functional Approaches to Language Structure*. NJ, USA: Psychology Press, 2nd ed.

Tucker, J. B. (2012). *Innovation dual use and security: Managing the risks of emerging biological and chemical technologies*. MIT Press.

Tupes, E. C., & Christal, R. E. (1992). Recurrent personality factors based on trait ratings. *Journal of Personality*, *60*(2), 225–251.

Turkheimer, E., Haley, A., Waldron, M., D'Onofrio, B., & Gottesman, I. (2003). Socioeconomic Status Modifies Heritability of IQ in Young Children. *Psychological science*, *14*, 623–8.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is All you Need. In *Advances in Neural Information Processing Systems 30*, (pp. 5998–6008). Curran Associates, Inc.

Villatoro-Tello, E., Parida, S., Kumar, S., Motlicek, P., & Zhan, Q. (2020). Idiap & UAM participation at GermEval 2020: Classification and Regression of Cognitive and Motivational Style from Text. In *Proceedings of the GermEval 2020 Task 1 Workshop in conjunction with the 5th SwissText & 16th KONVENS Joint Conference 2020*, (pp. 11–16). Zurich, Switzerland (online): German Society for Computational Linguistics & Language Technology.

Vinyals, O., Kaiser, L., Koo, T., Petrov, S., Sutskever, I., & Hinton, G. (2015). Grammar As a Foreign Language. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, (pp. 2773–2781). Cambridge, MA, USA.
URL http://dl.acm.org/citation.cfm?id=2969442.2969550

Štajner, S., & Yenikent, S. (2021). Why Is MBTI Personality Detection from Texts a Difficult Task? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, (pp. 3580–3589). Online: Association for Computational Linguistics.
URL https://aclanthology.org/2021.eacl-main.312

Waluchow, W. J. (2003). *The Dimensions of Ethics: An Introduction to Ethical Theory*. Broadview Press.

Warner, W., & Hirschberg, J. (2012). Detecting hate speech on the World Wide Web. In *Proceedings of the Second Workshop on Language in Social Media*, (pp. 19–26). Montreal, QC, Canada: Association for Computational Linguistics.
URL http://dl.acm.org/citation.cfm?id=2390374.2390377

Watson, D., & Clark, L. A. (1997). Chapter 29 - Extraversion and Its Positive Emotional Core. In R. Hogan, J. Johnson, & S. Briggs (Eds.) *Handbook of Personality Psychology*, (pp. 767–793). San Diego, CA, USA: Academic Press.
URL https://www.sciencedirect.com/science/article/pii/B9780121346454500305

Watson, D., & Tellegen, A. (1985). Toward a consensual structure of mood. *Psychological Bulletin*, *98*(2), 219–235.

Wei, M. (2020). Social Distancing and Lockdown – An Introvert's Paradise? An Empirical Investigation on the Association Between Introversion and the Psychological Impact of COVID19-Related Circumstantial Changes. *Frontiers in Psychology*, *11*.

Weischedel, W. (2005). *Die philosophische Hintertreppe: 34 großen Philosophen in Alltag und Denken*. München: dtv Verlagsgesellschaft.

Werner, M. H. (2020). *Einführung in die Ethik*. Berlin Heidelberg: J.B. Metzler, 1. Aufl. 2021, zweifarbig Edition ed.

Westerlund, M. (2019). The Emergence of Deepfake Technology: A Review. *Technology Innovation Management Review*, *9*, 39–52.

Wijngaards, I., Sisouw de Zilwa, S. C. M., & Burger, M. J. (2020). Extraversion Moderates the Relationship Between the Stringency of COVID-19 Protective Measures and Depressive Symptoms. *Frontiers in Psychology*, *11*.

Williams-Jones, B., Olivier, C., & Smith, E. (2014). Governing 'Dual-Use' Research in Canada: A Policy Review. *Science and Public Policy*, *41*, 76–93.

Winter, D. (1994). *Manual for scoring motive imagery in running text*. Dept. of Psychology, University of Michigan (unpublished).

Winter, D. G. (2007). The Role of Motivation, Responsibility, and Integrative Complexity in Crisis Escalation: Comparative Studies of War and Peace Crises. *Journal of personality and social psychology*, *92*, 920–37.

Winter, D. G., & Barenbaum, N. B. (1985). Responsibility and the power motive in women and men. *Journal of Personality*, *53*(2), 335–355.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques. With download-code*. Burlington, MA: Morgan Kaufmann, 3 ed.

Wittgenstein, L., & Schulte, J. (1963). *Tractatus logico-philosophicus: Logisch-philosophische Abhandlung*. Frankfurt am Main: Suhrkamp Verlag.

Wolf, M., Gotterbarn, D., & Kirkpatrick, M. (2019). *ACM Code of Ethics: Looking Back and Forging Ahead*. Association for Computing Machinery.

Wolf, M., Horn, A. B., Mehl, M. R., Haug, S., Pennebaker, J. W., & Kordy, H. (2008). Computergestützte quantitative Textanalyse: Äquivalenz und Robustheit der deutschen Version des Linguistic Inquiry and Word Count. *Diagnostica*, *54*(2), 85–98.

Wood, D. (2015). Testing the Lexical Hypothesis: Are Socially Important Traits More Densely Reflected in the English Lexicon? *Journal of personality and social psychology*, *108*, 317–35.

Xu, K., Ba, J. L., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R. S., & Bengio, Y. (2015). Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, (pp. 2048–2057). Lille, France.
URL http://dl.acm.org/citation.cfm?id=3045118.3045336

Yamada, K., Sasano, R., & Takeda, K. (2019). Incorporating Textual Information on User Behavior for Personality Prediction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, (pp. 177–182). Florence, Italy: Association for Computational Linguistics.
URL `https://www.aclweb.org/anthology/P19-2024`

Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent Trends in Deep Learning Based Natural Language Processing [Review Article]. *IEEE Computational Intelligence Magazine*, *13*, 55–75.

Zhang, X., Wang, N., Ji, S., Shen, H., & Wang, T. (2018). Interpretable Deep Learning under Fire. *CoRR*, *abs/1812.00891*.

Zhou, Q., & Wu, H. (2018). NLP at IEST 2018: BiLSTM-Attention and LSTM-Attention via Soft Voting in Emotion Classification. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, (pp. 189–194). Brussels, Belgium.
URL `https://www.aclweb.org/anthology/W18-6226`

Ziegler, M., & Buehner, M. (2009). Modeling Socially Desirable Responding and Its Effects. *Educational and Psychological Measurement*, *69*, 548–565.

Zimmerhofer, A., & Trost, G. (2008). Auswahl- und Feststellungsverfahren in Deutschland - Vergangeheit, Gegenwart und Zukunft. In *Studierendenauswahl und Studienentscheidung*, (pp. 32 – 42). Germany: Hogrefe Verlag, 1., aufl. ed.

Zirikly, A., Resnik, P., Uzuner, O., & Hollingshead, K. (2019). CLPsych 2019 Shared Task: Predicting the Degree of Suicide Risk in Reddit Posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, (pp. 24–33). Minneapolis, MN, USA.
URL `https://www.aclweb.org/anthology/W19-3003`

Zomick, J., Levitan, S. I., & Serper, M. (2019). Linguistic Analysis of Schizophrenia in Reddit Posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, (pp. 74–83). Minneapolis, MN, USA.
URL `https://www.aclweb.org/anthology/W19-3009`

Çöltekin, Ç. (2020). Predicting Educational Achievement Using Linear Models. In *Proceedings of the GermEval 2020 Task 1 Workshop in conjunction with the 5th SwissText & 16th KONVENS Joint Conference 2020*, (pp. 23–29). Zurich, Switzerland (online): German Society for Computational Linguistics & Language Technology.

Eidesstattliche Versicherung[51]

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Hamburg, den 12. September 2022   (Dirk Johannßen)

---

[51]Gemäß § 7 Abs. 4 der Promotionsordnung der Universität Hamburg.