KLETT
Hochschulen

WILHELM BÜCHNER
HOCHSCHULE
Mobile University of Technology

Validation of personality survey instruments using vector space representations of natural language

Volker Kempf,
Helge Nuhn

Schriftenreihe der Wilhelm Büchner Hochschule

Band 8 / 2023

Volker Kempf, Helge Nuhn

Schriftenreihe der Wilhelm Büchner Hochschule

Herausgeber Forschungsausschuss der Wilhelm Büchner Hochschule
16.01.2023

Wilhelm Büchner Hochschule

Impressum

ISSN (Online) 2751-0514

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Werden Personenbezeichnungen aus Gründen der besseren Lesbarkeit nur in der männlichen oder weiblichen Form verwendet, so schließt dies das jeweils andere Geschlecht mit ein.

Herausgeber: Forschungsausschuss der Wilhelm Büchner Hochschule
Redaktion: Dr. Marcel Heber
Layout und Satz: Philipp Thißen
Einbandentwurf : Gerhard Kienzle
Projektkoordination: Prof. Dr. Klaus Fischer

Gedruckt auf säurefreiem und chlorfrei gebleichtem Papier.

Validation of personality survey instruments using vector space representations of natural language

Volker Kempf, Helge Nuhn

Zusammenfassung Persönlichkeitstests sind zu einem gängigen Instrument in heutigen Personalauswahlverfahren geworden, da es zahlreiche Belege für den Einfluss von Persönlichkeitsmerkmalen auf die Arbeitsleistung auf individueller Ebene und von Persönlichkeitskonstellationen in Teams auf die Teamleistung gibt. Um einen optimalen Nutzen aus diesen Erkenntnissen zu ziehen, benötigen Unternehmen Persönlichkeitstests, die eine hohe Validität aufweisen, was kurz gesagt bedeutet, dass die Tests qualitativ hochwertige Ergebnisse liefern sollten, die mit der Realität korrelieren. Diese Arbeit befasst sich mit einem neuartigen Ansatz zur automatisierten Ermittlung der Validität solcher Selbsteinschätzungsfragebögen unter Verwendung einer Technik aus der Computerlinguistik, den sogenannten Worteinbettungen. Jedem Wort eines Vokabulars wird ein Vektor in einem hochdimensionalen reellen Vektorraum zugeordnet, seine Einbettung, und die geometrische Beziehung dieser Vektoren trägt semantische Informationen über die Beziehung der Wörter zueinander. Basierend auf diesen Worteinbettungen wird ein Konzept zur Bewertung der Validität von Persönlichkeitsfragebögen entwickelt und an bestehenden Erhebungsinstrumenten getestet, um die Funktionalität von Worteinbettungen für diesen Zweck zu zeigen. Die Ergebnisse der Tests sind vielversprechend und stützen die Annahme, dass Worteinbettungen in diesem Zusammenhang verwendet werden können.

Keywords: Künstliche Intelligenz, Natürliche Sprachverarbeitung, Sozialwissenschaftliche Instrumente, Instrumentvalidierung

Abstract Personality testing has become a common tool in modern personnel selection processes as there is ample evidence for the influence of personality traits on job performance on the individual level and of team personality constellations on team performance. To get an optimal benefit from these insights, organizations require personality tests that show high validity, which in short means that the tests should provide high quality results that correlate with reality. This thesis is concerned with a novel approach to establish the validity of such self-evaluation questionnaires in an automated way by using a technique from natural language processing called word embeddings. Every word of a vocabulary is assigned a vector in a high dimensional real vector space, its embedding, and the geometric relation of these vectors carry semantic information about the relation of the words. Based on these word embeddings, a concept to evaluate the validity of personality questionnaires is developed and tested on existing survey instruments to show the functionality of word embeddings for this purpose. The results of the tests are promising and support the assumption that word embeddings can be used in this context.

Keywords: Natural Language Processing, Social Science Research, Instrument Validation

Table of contents

1 Introduction 1
1.1 Scope and aim of the thesis 1
1.2 Methodical approach and outline of the thesis 3
2 Fundamentals 4
2.1 Natural language processing 4
2.2 Word embeddings 5
2.2.1 Definitions and basics 5
2.2.2 Generation of word embeddings 6
2.2.3 Applications of word embeddings 7
2.2.4 Interpretability of word embeddings 10
2.3 Personality models 12
2.3.1 Big Five personality traits 13
2.3.2 Personality typology 15
2.4 Validity of survey instruments 16
2.4.1 Face validity 17
2.4.2 Content validity 18
2.4.3 Criterion validity 20
2.4.4 Construct validity 21
3 Personality traits and the word embedding space 21
3.1 Personality dimensions in the word embedding space 22
3.2 Clustering of trait descriptive adjectives 24
4 Organizational personality evaluation 29
4.1 Motivation and history of personality testing in organizations 29
4.2 Types and uses of organizational personality tests 30
4.3 Challenges of organizational personality testing 32
4.3.1 Faking in assessment procedures 32
4.3.2 Reactions to personality tests in assessment procedures 34
4.4 Influence of personality trait composition on work teams 36
4.5 Standard inventories for organizational personality tests 38
5 A concept for the validation of personality survey instruments 39
5.1 Observations concerning validity and word embeddings 39
5.2 Term extraction from survey items 40
5.3 Proposed methods for validity analysis 40
5.3.1 Nearest traits 41
5.3.2 Similar items 41
5.3.3 Mean and variance of items 42
5.3.4 Word embedding factor analysis using personality dimensions 43

6 Application to personality survey instruments....44
6.1 Word embedding models and inventories....44
6.2 Application of validity analysis methods to BFI, BFI2 and IPIP items....45
6.2.1 Evaluation of nearest traits....45
6.2.2 Evaluation of similar items....48
6.2.3 Evaluation of mean and variance....49
6.2.4 Word embedding factor analysis....50
7 Conclusion and outlook....52
Limitations of the approach....53
Extensions and future research directions....53
References....54
A Additional tables....64
B Online survey description....67

List of Figures

Figure 2.1: Illustration of the cosine similarity and Euclidean distance functions in two dimensions. 8

Figure 2.2: Illustration of the word analogy finding process for a is to b as c is to x. 9

Figure 2.3: Visualization of the words Easter, Christmas, Actress and King in terms of the tree polar dimensions Spring–Winter, Man– Woman and Weak–Strong. 11

Figure 2.4: Visualization of the projection method for concept dimensions, based on [Roz20, Figure 9, p. 20], adapted.....12

Figure 2.5: Hierarchical system of abstractions of personality description models, based on [Dig90, Figure 1, p. 421], adapted. 14

Figure 2.6: Selection of types and subtypes of validity, based on [Tah16, Figure 1, p. 29], adapted. 17

Figure 2.7: Visualization of the content validity notion, based on [SBG04, Figure 2, p. 386], adapted. 18

Figure 3.1: Positions of the top five trait descriptive adjectives from [Joh21, Table 2.4, p. 50] after performing a three-dimensional principal component analysis. 26

Figure 3.2: Clustering with maximal adjusted mutual information index of the top 8 trait descriptive adjectives from [Joh21, Table 2.4, p. 50]. Stacks represent found clusters, colors the true affiliation of an adjective. 27

Figure 3.3: Clustering with maximal adjusted mutual information index of the top 8 trait descriptive adjectives from [Joh21, Table 2.4, p. 50]. Stacks represent found clusters, colors the true affiliation of an adjective. 28

Figure 4.1: Aspects of job performance, based on [MSA17, Figure 2.1, p. 28], adapted.....30

Figure 4.2: Favorability of selection tools on a scale from 1–5 with standard deviation, blue data from [HDT04, Table 4, p. 659], green data from own online survey, see Figure B.1. 34

Figure 6.1: Confusion matrices of nearest trait classification of BFI2 items for fTCrawl word embeddings, cf. Table 6.3. Item pre-processing stages I1 on the left, I2 in the middle and I3 on the right. Traits are abbreviated by their first letter. 47

Figure 6.2: Confusion matrices of nearest trait classification of IPIP items in I2 stage of pre-processing. Traits are abbreviated by their first letter.....47

Figure 6.3: Average normalized coordinates of positively keyed BFI items of each trait in the personality dimensions. Colorbars indicate values of standard deviation of the averaged data points. 51

Figure 6.4: Average normalized coordinates of positively keyed BFI2 items of each trait in the personality dimensions. Colorbars indicate values of standard deviation of the averaged data points. 51

Figure 6.5: Averages of factor loadings of positively keyed BFI2 items from [SJ17, Table 6, pp. 14f]. Colorbars indicate values of standard deviation of the averaged data points. 51

Figure B.1: Online survey structure and results. 68

List of Tables

Table 2.1: Nearest neighbors to the words ant and crab measured by cosine similarity based on pre-trained word embeddings fTWiki described in Table A.1. 8

Table 2.2: Correct word analogies extracted from the pretrained word embeddings fTWiki described in Table A.1. 10

Table 2.3: CVRcrit values in dependence of panel size, excerpt from [AS14, Table 2, p. 85]. Ncrit is the minimum number of panelists necessary to vote “essential” on an item. 19

Table 3.1: Angles between trait dimensions in pretrained word embeddings set GVWiki described in Table A.1. Traits are abbreviated by their initial letter. 23

Table 3.2: Means and standard deviations of trait angles for different pre-trained word embeddings. 24

Table 3.3: Top 10 trait descriptive adjectives by factor loading, which is given next to each word, for the Big Five domains, from [Joh21, Table 2.4, p. 50]. 24

Table 3.4: Number of misplaced items Nf and adjusted mutual information score RAMI of best achieved clusterings of the top 10 trait descriptive adjectives from [Joh21, Table 2.4, p. 50]. Nm is the number of adjectives that are not in the word embedding vocabulary. 28

Table 4.1: Results of meta study on Big Five trait influence on team performance, data from [DS13, Table 33.1, p. 752]. Shown quantities are weighted mean correlations. n.s. means statistically not significant. – means not measured in cited study 37

Table 5.1: Relevant cosine similarity thresholds for word embeddings of various dimensions, data from [RLH17, Table 1, p. 402]. 42

Table 6.1: Distribution of IPIP items on the Big Five traits 45

Table 6.2: F-scores for classification of BFI items by nearest trait pole in the different stages of pre-processing. I1: items in the original wording; I2: items with removed stop words; I3: descriptive adjectives in place of whole items. Traits are abbreviated by their first letter. 46

Table 6.3: F-scores for classification of BFI2 items in the stages of pre-processing by nearest trait pole. I1: items in the original wording; I2: items with removed stop words; I3: descriptive adjectives in place of whole items. Traits are abbreviated by their first letter. 46

Table 6.4: Similar items found in the BFI and BFI2 survey inventories, cf. Tables A.4 and A.5. w2vNews only contained pairs C4–C5 in BFI and C1–C2 in BFI2, GVWiki only O2–O3 in BFI2. Items that were above the similarity threshold in at least three of the word embedding sets are highlighted in bold. 48

Table 6.5: Average componentwise overlap of intervals of one standard deviation around mean of trait adjectives and inventory items 49

Table A.1: Sets of pre-trained word embeddings used in the thesis.....64

Table A.2: Trait descriptive adjectives for the Big Five domains and their opposites, from [Gol92, Table 3, p. 34]. 64

Table A.3: Trait descriptive adjectives for the Big Five domains with negative factorloadings, from [Joh21, Table 2.4, p. 50]. 65

Table A.4: Items of BFI [JNS08, p. 157], with the extracted key term from each item. (R) indicates reverse scored items. Stop words in italics. Interpreted key terms in bold. “I see myself as someone who . . . ” 65

Table A.5: Items of BFI2 [Joh21, p. 81], with the extracted key term from each item. (R) indicates reverse scored items. Stop words in italics. Interpreted key terms in bold. “I am someone who . . . ” 66

1 Introduction

1.1 Scope and aim of the thesis

In assessment procedures that are conducted to facilitate personnel decisions it is now common practice to generate personality profiles of the candidates in order to gauge how effective a person would be, for example in a newly composed project team [DS13, p. 744]. Two of the most commonly used personality models in this context are the Big Five personality traits, see, e.g., [Joh21], where continuous scales of the five traits extraversion, agreeableness, conscientiousness, neuroticism and openness are used to describe personality, and personality typology approaches as employed for example by the Myers–Briggs Type Indicator, see, e.g., [KM20], which is a popular self-evaluation questionnaire that attempts to categorize a person into one of sixteen distinct and disjunct personality types. Recent research has shown that the personality of employees can have a significant effect on their individual performance and the performance of their work team [DS13; SAM20], thus many companies use some form of personality assessment in hiring and selection procedures. Although there are several methods to assess a candidate's personality, for example during a selection interview [RI13], selfevaluation by questionnaires is most commonly used [CHLS13, p. 477]. A critical role for getting a robust assessment of the personality by this method can be attributed to the suitability of the items on these questionnaires for the intended purpose. This can be described by the concepts of reliability and validity, which are properties that need to be established for every survey instrument in order for it to be accepted by researchers and practitioners. The focus in this thesis is on item and instrument validity, which is the extent to which a test measures what it is supposed to measure [CS10, p. 1].

Contrary to classical ways of evaluating validity, this thesis aims to develop a new technique by using a method from the field of natural language processing, which attaches a high dimensional vector representation to every word of a certain vocabulary. These representations are called word embeddings and can be generated by a variety of methods, for example using a machine learning framework [PSM14a]. The vectors then encode certain characteristics of the words' meanings, whereby the richness of the encoded information depends mainly on two aspects: the base data used to generate the vectors, which is in most cases a vast corpus for texts taken from various sources, and the dimension of the vector itself. The captured information in the word embeddings can be seen, for example, by the fact that embeddings for words with a similar meaning tend to lie close together when measuring the angle between the vectors [RLH17].

Using such a representation of natural language, it may be possible to discover features, characteristics or patterns in the commonly used survey items of personality tests. These insights could be useful to improve or measure the quality, in the sense of validity, of a set of items generated for a certain purpose. Improved validity of an instrument can in the end prove beneficial for personnel decisions that are in part founded on the results of such self-evaluation tests. Furthermore such information could also lead to a novel method to establish the validity of a survey instrument.

The overarching question to be investigated is whether it is possible to validate personality survey instruments using natural language vector space representations. A positive answer to this question could in the future lead to the development of a highly useful tool for the construction of such survey instruments, with the aim of maximizing the gain in knowledge they provide.

As a first step in this direction of research it is necessary to develop a methodology that can be used for the validation of the survey instruments. This concept needs to be versatile, produce consistent and comprehensible results and should have some indicator on the quality of the findings it yields.

In this thesis the focus is laid on the development of a basic concept to validate self-evaluation questionnaires and an application to some existing personality survey instruments to discover the potential for this method and its use in the development of survey instruments for organizational personality testing. However, the basic approach demonstrated here is applicable to different fields as well.

1.2 Methodical approach and outline of the thesis

Chapter 2 contains the results of a literature study on the fundamentals of natural language processing with a focus on vector representations of word semantics, on the most common personality models and on the aspects of validity of survey instruments. Several methods of obtaining word embeddings for a given vocabulary on the basis of a base corpus of text documents are presented and typical use cases are shown and examined. Furthermore an introduction to the Big Five and typology personality models is given in Section 2.3 in order to understand the opportunities and limitations of such frameworks for the assessment of a person's personality.

In Chapter 3 the realization of the Big Five personality traits in the word embedding vector space is discussed. Two concepts to identify the trait structure, personality dimensions and clustering of trait descriptive adjectives, are presented and differences across several sets of pre-trained word embeddings are investigated.

Chapter 4 depicts the current state of personality testing in assessment procedures, and describes the relation of personality traits and job performance, as well as the effects of team personality composition on team performance. An online survey was conducted between 2022-02-08 and 2022-03-10 which gathered 127 respondents and aimed at generating data to support some of the data taken from literature about organizational personality testing.

Since to the best of our knowledge the research direction of validating survey instruments using word embeddings is new, the aim of this thesis is to develop a methodology to tackle this research question. This concept, which is presented in Chapter 5, touches upon the necessary data, pre-processing of survey items and several methods that may be indicators for validity.

To show the functionality of this concept it is demonstrated using an articulate, proof-of-concept-type application in Chapter 6. Throughout, the focus of the thesis is on an application to personality testing survey instruments, so existing self-evaluation questionnaires for the Big Five model from [Joh21; JNS08] are investigated with the methodology described in Chapter 5. However, the concept may be used with some adjustments for other types of survey instruments. Publicly available pre-trained word embeddings are used which were generated with the word2vec, GloVe and fastText algorithms. Since these vector representations differ in the algorithm and base text corpus used to generate them, it is possible to examine differences and make comparisons regarding the suitability of the algorithms when analyzing survey items.

2 Fundamentals

The thesis aims to combine research elements from several fields in order to develop a validation concept for personality survey instruments. Thus, the relevant parts from natural language processing, personality models and instrument validity that are required for the subsequent sections are collected in this chapter.

Throughout the chapter, some examples using word embeddings are presented. The corresponding Python implementation to reproduce the examples can be found at https://gitlab.com/volker.kempf/validation_by_word_ embeddings.

2.1 Natural language processing

The goal of the field of natural language processing is to enable computers to understand and be able to work with human language by converting it to a formal representation [CW08, p. 160]. The first concepts for this were created more than half a century ago, and since then the main focus in research has been machine translation, information retrieval, information extraction, topic modeling and, since the beginning of the social web, opinion mining [CW14, p. 49]. To really achieve these purposes, the syntactic and semantic properties of the individual elements of language have to be known by the machine to an extent that exceeds pure knowledge about key words and word co-occurrences [CW14, p. 53].

When developing a new system for natural language processing, it is often tested and benchmarked on a number of standard tasks that are relevant to a variety of the use cases. They include, see [CW08, p. 161],

- part of speech tagging, i.e., labeling each word in a text by its syntactic role, e.g., “noun”, “verb”,
- named entity recognition, i.e., labeling relevant words in a text by categories, e.g., “PERSON”, “LOCATION”,
- finding semantically related words, i.e., predicting whether words are, e.g., synonyms, hyponyms, antonyms,
- completing word analogies, i.e., finding the missing word x in an analogy “ a is to b as c is to x”.

Modern techniques for natural language processing almost exclusively use machine learning methods, most frequently deep neural networks, to achieve these tasks.

2.2 Word embeddings

Many natural language processing techniques require words to be represented as a mathematical object, so that established mathematical methods can be applied. One of the most common schemes to transfer words to numbers are word embeddings, where every element of a vocabulary is mapped to a vector of real numbers, i.e., elements of $\mathbb{R}^d$, where the dimension of the embedding vector space $d \in \mathbb{N}$ can be chosen arbitrarily. A set of such word embeddings is denoted by W. With such a representation, all the established tools from linear algebra can be used, e.g., orthogonality concepts, geometric aspects, eigenvalue and singular value decomposition of matrices made from sets of words vectors.

The modern processes of generating word embeddings rely on the distributional hypothesis in linguistics, which states that words with similar meaning appear in similar contexts [Sah08, p. 33]. This notion of word similarity led to the concept of a distributional space, where the words are seen as points and the dimensions as linguistic contexts [Len08, p. 11]. Seen from this perspective, the word embedding technique using machine learning methods is the next step to generate an accurate realization of the distributional space.

2.2.1 Definitions and basics

In general, if the goal is only to find a vector representation of the words in a given vocabulary, *one-hot encoded vectors* can be used. With this approach, the vector dimension $d \in \mathbb{N}$ is the number of words in the vocabulary and in the vector for the i -th word of the vocabulary the i-th component is set to 1, while all other components are 0. Consider for example the small vocabulary containing the five arbitrary words *ant, termite, crab, shrimp* and *emu,* then the corresponding one-hot encoded vectors might be

$$\text{ant} \rightarrow \begin{pmatrix}1\\0\\0\\0\\0\end{pmatrix}, \quad \text{termite} \rightarrow \begin{pmatrix}0\\1\\0\\0\\0\end{pmatrix}, \quad \text{crab} \rightarrow \begin{pmatrix}0\\0\\1\\0\\0\end{pmatrix}, \quad \text{shrimp} \rightarrow \begin{pmatrix}0\\0\\0\\1\\0\end{pmatrix}, \quad \text{emu} \rightarrow \begin{pmatrix}0\\0\\0\\0\\1\end{pmatrix}.$$

While these vectors satisfy the above definition of word embeddings, they are not particularly useful. They do not capture any semantic relations of the words they represent since they are all orthogonal. The expectation for word embeddings however would be that the vectors for the words *ant* and *termite* should be more similar to each other than to the vector for the word *emu*. Although this type of vector is not used for word embeddings in the application considered in this thesis, they play a role in some applications, e.g., in information retrieval, and in the generation algorithms for practical word embeddings.

Modern techniques to generate useful embeddings employ machine learning algorithms to find the vectors from a large set of sample texts. These algorithms generate vectors as elements of $\mathbb{R}^d$ with a fixed number of components d, which is independent of the considered vocabulary and usually not larger than a few hundred. For instance, the pre-trained sets of word embeddings of the GloVe project from [PSM14b] have a maximum dimension of d = 300, for a vocabulary of up to 2.2 million words, cf. Table A.1. In this form, each dimension can be seen as one or a combination of several semantic features of the natural language text documents that were used to generate the word embeddings, however it is in general not possible to assign an interpretable meaning to the raw form of those features. An attempt to extract interpretable features from the dense word vectors is presented in Section 2.2.4.

2.2.2 Generation of word embeddings

Word embeddings can be generated by a number of different ways. The most common method currently uses neural net architectures, that are trained on a large corpus of text to capture the semantic meanings of each word in a vector. The three most common algorithms that produce good word embeddings, measured by several common natural language processing benchmark tasks, are *word2vec* [MCCD13], *GloVe* [PSM14a] and *fastText* [BGJM17].

Without going too much into the intricacies of the method, word2vec has two modes when generating the vectors. The first, called *continuous bag of words*, is trained to find the most appropriate word given the surrounding context words, while the second, the *skip-gram* mode, works the other way around, by predicting the most probable context words for a given word [MCCD13, p. 4].

The fastText algorithm is a derivative of the word2vec model, which now includes subword information into the word embeddings. This means that incomplete word fragments or misspelled words, even when they are not contained in the word embedding vocabulary, can also be used for analysis. For example, misspelled words are positioned relatively close to the correctly spelled word in the word embedding space [BGJM17].

The third of the widely used algorithms to generate word embeddings is the GloVe algorithm, which is short for Global Vectors [PSM14a]. While the training in the word2vec and fastText algorithms use a local approach that predicts words based on a few of the surrounding words in the sample texts, GloVe combines this approach with the factorization of a global word-word co-occurrence matrix, which results in a global log-bilinear regression model [PSM14a, p. 1532].

For all these algorithms vast text corpora are necessary for the training of the neural networks in order to achieve an optimal model. This means that in order to get good word vectors, considerable time has to be spent on the training process, including several, sometimes manual, steps to find and prepare the raw text data. Common sources for these texts are Wikipedia snapshots, online news websites, Twitter data or just immense amounts of texts taken from the internet by web crawling. Luckily all three projects offer word vectors to download, which are pre-trained on large text corpora containing several billions of words. Throughout the thesis several of these pre-trained embeddings are used, and in Table A.1 an overview of these sets is compiled.

While word embeddings generated from each of the three methods yield good results on various tasks in natural language processing, none clearly outperforms the others in every task. Thus it makes sense to consider comparable pre-trained sets of word embeddings from all three algorithms in the investigations and see if one model is better suited than the others for the specific tasks in this thesis.

2.2.3 Applications of word embeddings

A variety of natural language processing tasks can be handled with word embeddings. One possible use case of these vectors is identifying words that have a similar semantic meaning to a specified target word by finding the nearest neighboring word vectors. Measuring distances or similarities of vectors can be done by different metrics, but mostly the *cosine similarity* and the standard *Euclidean distance* are used. Let $a, b \in \mathbb{R}^d$ be two word embeddings, then their Euclidean distance is defined as the Euclidean norm of their difference, i.e.,

$$\mathrm{dist}_2(a,b) = \|a-b\|_2 = \sqrt{\sum_{i=1}^{d}(a_i-b_i)^2},$$

and their cosine similarity is defined by

$$\mathrm{sim}_{\cos}(a,b) = \frac{a \cdot b}{\|a\|_2 \|b\|_2},$$

which originates from the definition of the scalar product of two vectors $a \cdot b = \|a\|_2 \|b\|_2 \cos\alpha$, where α is the angle between the two vectors. The Euclidean distance of two vectors can be any non-negative real number where low values indicate a higher similarity of the words. The cosine similarity function on the other hand takes values in the interval $[-1,1] \subset \mathbb{R}$, where values close to 1 and −1 indicate semantic similarity and, respectively, anti-similarity, and values closer to 0 mean that the words are not related.

For word embeddings with 300 dimensions, which is the most common size used in this thesis, cosine similarity values of above 0.692 indicate a close semantic relation between words [RLH17, Table 1, p. 402]. Both functions are illustrated in Figure 2.1 for an example in two dimensions.

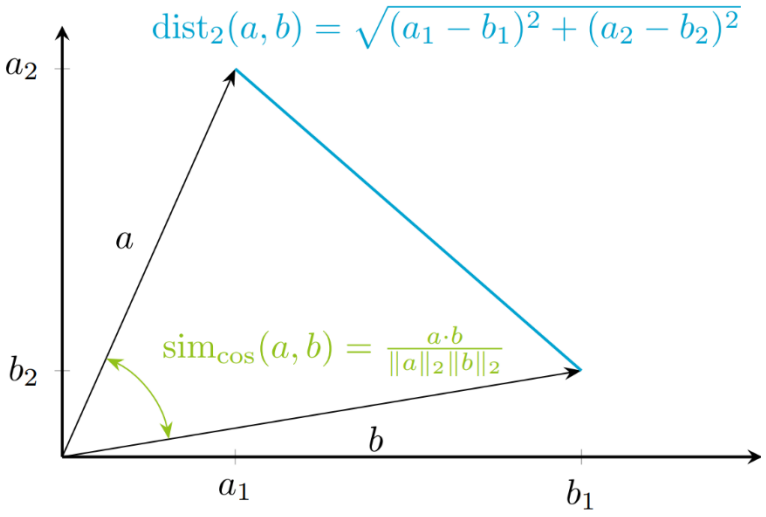

Figure 2.1 Illustration of the cosine similarity and Euclidean distance functions in two dimensions.

Table 2.1 Nearest neighbors to the words ant and crab measured by cosine similarity based on pre-trained word embeddings fTWiki described in Table A.1.

Target word	Neighbors			Vocabulary words	
	No.	Word	$\mathrm{sim}_{\cos}$	Word	$\mathrm{sim}_{\cos}$
ant	1	ants	0.8055	ant	1.0000
	2	insect	0.6960	termite	0.6282
	3	anthill	0.6955	crab	0.4742
	4	wasp	0.6863	shrimp	0.3613
	5	bee	0.6765	emu	0.4079
crab	1	crabs	0.8116	ant	0.4742
	2	lobster	0.7551	termite	0.4794
	3	crabber	0.7358	crab	1.0000
	4	crabbing	0.7204	shrimp	0.7085
	5	crabmeat	0.7199	emu	0.2925

To give an example of the use of this similarity analysis, consider the words ant and crab from before and observe the five nearest word vectors measured by cosine distance. To find those words, the pre-trained word embeddings fTWiki, cf. Table A.1, are used, which are based on a text corpus consisting of Wikipedia 2017, UMBC WebBase and statmt.org news data and comprise a vocabulary of one million words. Table 2.1 shows the nearest neighbors of the two words.

From the data on the right hand side of the table it is also clear that a large similarity of the word pairs *ant–termite* and *crab–shrimp* is encoded in the pre-trained word vectors, as would be expected. The other words from our small vocabulary are less similar to the two target words. In contrast, if the one-hot encoded vectors from Section 2.2.1 would be used, the pairwise cosine similarity of the vocabulary words would be zero, since all those vectors are orthogonal. Another structure that is usually contained in pre-trained word embeddings is that word analogies can be found using the vector difference of a known pair of words. Given three words a, b, and c and an incomplete word analogy of the type a is to b like c is to x, the vector representations can be used to find the word vector x in the word embedding space that is closest, or most similar, to the point y = c + b − a. Written precisely, given a set of pre-trained word vectors W and the word analogy problem, the element x ∈ W for which

$$x = \underset{z \in W \setminus \{a,b,c\}}{\arg\min} \; \mathrm{sim}_{\cos}(z, y)$$

holds needs to be found. An illustration of the word analogy finding task in two dimensions is given in Figure 2.2. As an example, consider again the pre-trained embeddings fTWiki, cf. Table A.1, and the problem to find the missing word x in the analogy *emu is to bird as ant is to x*. As expected, the word that the above formula yields is *insect*. Table 2.2 provides some more examples of word analogies extracted from the pre-trained word embeddings. The results shown in the table demonstrate that the word embeddings used here contain semantic information in various categories.

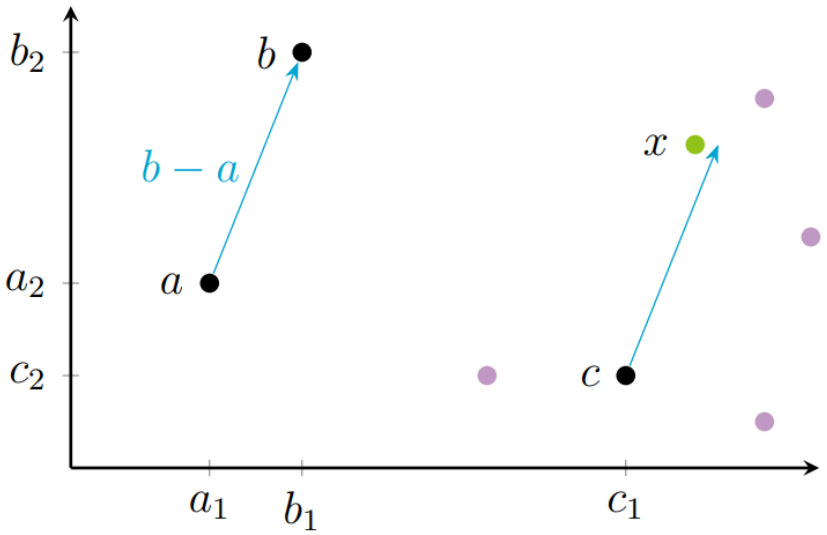

Figure 2.2 Illustration of the word analogy finding process for a is to b as c is to x.

Table 2.2 Correct word analogies extracted from the pretrained word embeddings fTWiki described in Table A.1.

a		b		c		x
actor		actress		waiter		waitress
white		black		up		down
tree	*is to*	leaf	*as*	flower	*is to*	petal
dog		puppy		cat		kitten
pork		pig		beef		cow
cold		colder		small		smaller

2.2.4 Interpretability of word embeddings

The word embedding representation that is generated as dense vectors by the word2vec, GloVe and fastText algorithms has a major drawback that was already briefly mentioned. A vector of a few hundred real numbers is not easily interpretable by a human and the information, which specific semantic feature each component of the vector represents, is not available and cannot accurately be extracted. The semantic information that, as seen in Section 2.2.3, is contained in the embeddings, is not accessible for direct interpretation on the basis of the vectors alone.

The POLAR framework attempts to remedy this shortcoming by introducing artificial interpretable dimensions into the embedding space by means of sets of polar opposites, or antonyms, and rearranging the word embeddings along those axes [MSLS20]. Let $W \subset \mathbb{R}^d$ be a set of normalized word embeddings, i.e., for all word vectors $v \in W$ it holds $\|v\|_2 = 1$. Then a set of n interpretable axes can mathematically be generated by using n pairs of antonyms $(w_i^+, w_i^-) \in W \times W,\ i \in \{1,\dots,n\}$, from the word embedding space, and computing the matrix $A^\pm = (w_1^\pm \cdots w_n^\pm)$ where for each $i \in \{1,\dots,n\}$

$$w_i^\pm = w_i^+ - w_i^-$$

is the column vector pointing from the "negative" word w_i^- to the "positive" word w_i^+ for each antonym pair. Given a word vector $v \in W$, its representation $v^\pm$ in the interpretable antonym space is computed by $v^\pm = (A^\pm)^{-1} v$, where $(A^\pm)^{-1}$ is the Moore-Penrose inverse of the possibly non-square matrix $A^\pm$ [MSLS20, p. 3]. The resulting vector $v^\pm$ holds the coordinates of the word v in terms of the antonym dimensions.

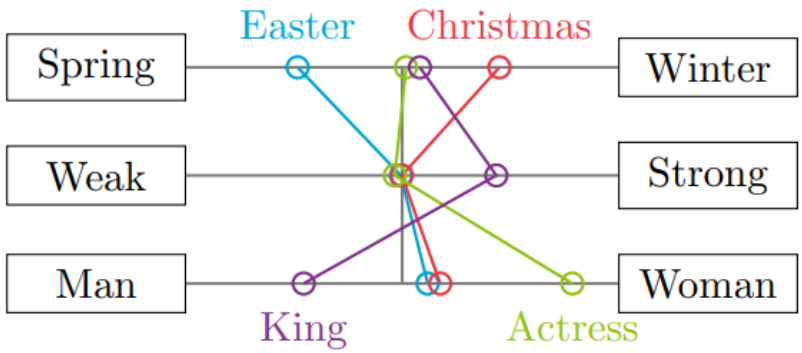

Figure 2.3 Visualization of the words Easter, Christmas, Actress and King in terms of the tree polar dimensions Spring–Winter, Man– Woman and Weak–Strong.

This process can be illustrated with a small example using the pre-trained word embeddings fTCrawl described in Table A.1. Consider the three antonym pairs *Spring–Winter, Man–Woman, Weak–Strong* and the words *Easter, Christmas, Actress* and *King*. Application of the method described above to the pre-trained word embeddings of these antonyms yields the result visualized in Figure 2.3, where the coordinates of the words on the three polar dimensions are shown. While the words *King* and *Actress* are clearly tending to one of the poles on the *Man–Woman* axis, the words *Easter* and *Christmas* are separated on the *Spring–Winter* axis, as could be expected. Among the words in this example only King shows a tendency towards *Strong* in the *Weak–Strong* dimension, the three other words seem to lie almost perfectly orthogonal to that axis.

Instead of using pairs of antonyms to define dimensions, a similar approach, see [Roz20, p. 3], defines concept dimensions, by selecting words that positively and negatively represent the concept. The word vectors of these terms are averaged to a negative and a positive concept pole in the word embedding space. With the positive and negative poles of each concept, the same process as above for the antonym pairs can now be used to generate a transformation matrix onto the concept dimensions.

Another method to use these interpretable dimensions does not require computing the inverse of a matrix. The concept dimension vectors $w_i^{\pm}$ i are instead normalized and the coordinates of a word v on these axes are calculated by a simple projection onto the normalized axis vectors, i.e.,

$$v_i^{\pm} = v \cdot \frac{w_i^{\pm}}{\|w_i^{\pm}\|}.$$

Figure 2.4 illustrates the process of aggregating concept poles w^+ and w^- from two individual words each, generating the concept dimension $w^\pm$ and finally projecting a word vector v onto this axis.

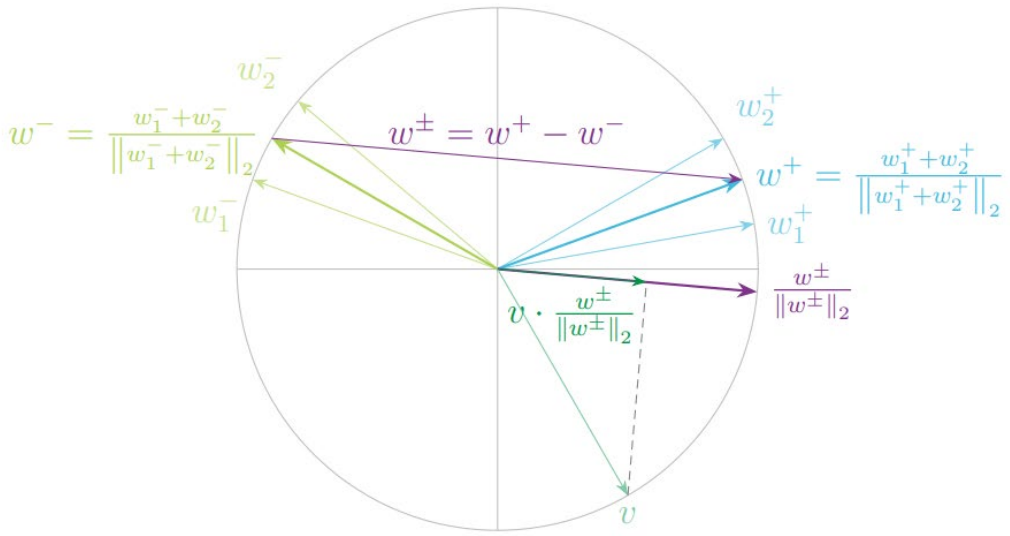

Figure 2.4 Visualization of the projection method for concept dimensions, based on [Roz20, Figure 9, p. 20], adapted.

2.3 Personality models

The origins of personality theory trace back to the works of Allport and Stagner [Cra93, p. 3], with the seminal publications [All37] and [Sta37], which established personality research as a separate subfield in psychology. Personality theory focuses on psychological research concerning the theoretical framework for the study and comprehension of human behavior [Wig88, p. 443]. However, this description of the research aim does not distinguish personality theory from the goal of general behavior theory. The distinction comes mainly from a historical perspective and can be summed up by stating that personality theory has a more humanistic view, while general behavior theory acts in a more abstract, scientific way [Wig88, pp. 444f].

Two general directions in personality research can be seen: the study of individual differences and the study of individual persons as unique wholes. The first approach can be seen as a quantitative, psychometric way to define personality as the sum of all traits of a person, while the second is a qualitative approach that focuses on biographical analyses and case studies. Mostly, personality psychologists focused on the psychometric approach to measure personality differences of populations with the help of personality tests [BW08, pp. 7f].

With the lexical hypothesis, which states that existing and relevant individual differences will over time find their way into the used language [Gol81, pp. 141f], personality research can rely on a semantic foundation, e.g., when designing personality questionnaires. Based on this reasoning, the classification of adjectives describing personalities has been a research focus since the beginnings of the field, see, e.g., [AO36]. The manual and subjective approach of sorting through thousands of personality descriptive adjectives from the beginnings of personality research has since been replaced by more sophisticated methods that observe word co-occurrences [Swi21, p. 14]. These studies ultimately resulted in the by now well established Big Five model of personality traits.

2.3.1 Big Five personality traits

The semantic approach described in the previous paragraph was used to compile word lists containing person descriptors [JNS08, p. 117], which were grouped in different categories, the number of which was largely dependent on the personality psychologist who did the work and mostly ranged from two to 20 [JNS08, p. 114]. Subsequent research lead to the emergence of the Big Five personality traits shown in the following list, where the quoted descriptions are taken from [Joh21, Table 2.2, p. 42]:

extraversion: “Implies an energetic approach toward the social and material world.”

agreeableness: “Contrasts a prosocial and communal orientation toward others with antagonism and hostility.”

conscientiousness: “Describes socially prescribed impulse control that facilitates task- and goal-directed behavior.”

neuroticism: “Contrasts negative emotionality with emotional stability, con tentment, and frustration tolerance.”

openness: “Describes the breadth, depth, originality, and complexity of the per son’s mental and experiential life.”

These traits are considered largely independent [Gol92, p. 26]. The specific names of the traits differ slightly between publications, e.g., openness is often referred to as open-mindedness, but the structure is largely agreed upon [Joh21, p. 49]. Each of these domains is made up of several facets, that combined form the complete trait, and which are in turn based on another layer of more detailed factors.

Figure 2.5 visualizes this concept of a hierarchical system of abstractions. The precise number and definition of the facets is again blurry and many publications have differing views. As an example, the facets of extraversion in [SO99] are defined as *sociability, unrestraint, assertiveness* and *activity–adventurousness*, while in [SJ09] they are called gregariousness, social confidence vs. anxiety and assertiveness. While there is a certain overlap in these facets, they show a different structure. For the purpose of this thesis, the distinction of facets for each trait is not considered further, as the focus is on the more abstracted Big Five traits.

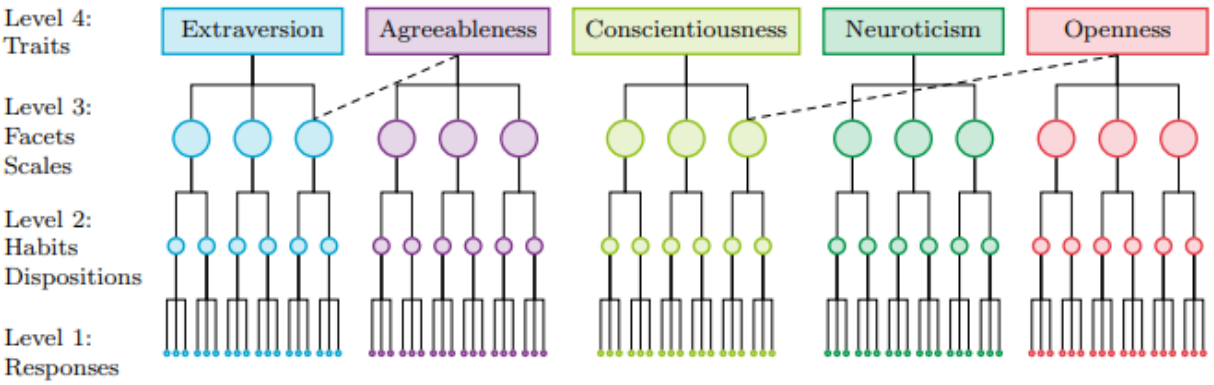

Figure 2.5 Hierarchical system of abstractions of personality description models, based on [Dig90, Figure 1, p. 421], adapted.

Personality self-evaluation questionnaires are commonly used to assess the interrelations between personality traits. After the Big Five structure had been theoretically established, questionnaires were designed to support the model, and the structure of the Big Five, which had been built based on adjectives, could be recovered to a large degree from factor analyses of the survey results [Joh21, pp. 46f].

One of the first survey instruments to measure the Big Five factors was the NEO Personality Inventory, which has been revised several times since its first publication, and which offers tests in different lengths and complexities, however only the long form covers all facets of the Big Five traits [Joh21, p. 47]. Moreover, this instrument is proprietary and not freely accessible. For the later studies of personality survey instruments this thesis resorts to publicly available questionnaires like the Big Five Inventory, see [JNS08, pp. 157], and the Big Five Inventory 2, see [SJ17, pp. 142f].

Though it is currently the most widespread model of personality traits, the Big Five are not without criticism. One point of concern is that the five traits emerged through the method of factor analysis, which identifies factors by shared variance in items and facets. This however only captures what is included in the correlation matrices that are being studied [HOO15, p. 189], which means important information can be overlooked. Another point of criticism is that some constructs of personality are not included in the model.

Examples of such missing traits are honesty, vigilance and graciousness. Though they may be less important than the Big Five, there is sense in observing them, as some may be well suited to draw conclusions for organizational applications of personality theory [HOO15, p. 189].

A modern alternative of the Big Five is the HEXACO model of personality, which extends the Big Five by a sixth trait, the honesty-humility dimension. This additional domain, like the original Big Five, was found as a result from a lexical approach to personality descriptive adjectives [HOO15, p. 196]. Although there is some research in support of the HEXACO model, it will not be investigated in more detail in this thesis, as the Big Five is still the commonly used model in most publications.

2.3.2 Personality typology

Another approach to personality analysis tries to sort the personality of people in several dichotomous type categories. An instrument that employs this idea is the Myers–Briggs Type Indicator, which is also one of the best known and most used tools to determine personality factors in organizational settings [Mur90, p. 1187]. It is built around the theory of Carl Jung which states that the personality of people can be captured in type categories [Jun71]. Thus in this theory, the variances in personalities are abstracted by sorting them into distinct groups which supposedly have little variance internally while the difference to the other type groups is large [Pit05, p. 211]. The Myers–Briggs Type Indicator is a tool that tries to operationalize Jung's theory by using different versions of self-evaluation inventories to determine the position of an individual on the four bipolar scales [Car77, p. 461]

- Extraversion – Introversion
- Sensation – Intuition (S-N),
- Thinking – Feeling (T-F) and
- Judgment – Perception (J-P).

The first forms of the inventory were introduced in the 1940s, and were since then redeveloped and reintroduced several times [Mur90, p. 1188]. After taking the test, every individual is categorized to one of the ends on each of the four scales, giving a total of 16 combinations and thus 16 different personality types, which are usually abbreviated by the letters seen in the above list. As an example, a possible result is ESTJ, which indicates that the tested candidate scored higher on the extraversion, sensation, thinking and judgment sides of the scales. Each of the types has a specific personality description associated with it [BMQH98, Chapter 4].

Though the Myers–Briggs Type Indicator is, due to the simplicity of the 16 types, a widely popular instrument for a variety of psychometric applications, it has several drawbacks that make its results unusable in professional settings, which is why it is not further considered in this thesis. A main criticism of this instrument is that it neglects the continuous nature of the personality traits and focuses on the classification into the 16 types, which not only restricts the statistical analysis [Boy95, p. 73], but also leads to changing results in re-tests, especially when the actual score of a test candidate is near the middle of the scale [Pit05, p. 214]. Several evaluations showed that only around 50 percent of those tested retain their type when re-tested [DB91, pp. 96f].

The last aspect is particularly concerning when thinking of the further application of the test results, e.g., in the context of personnel assessment. In the proposed theoretical foundation of the instrument the type of a person is supposed to be set at birth [Pit05, p. 212]. If the resulting type can easily change for a large part of the test audience, but each type offers a distinctly different interpretation [McC00, p. 118], the implications drawn from the results must be unreliable.

A large part of the publications in support of the theoretical foundation, reliability and validity of the Myers–Briggs Type Indicator is published in journals of the Center for the Applications of Psychological Type [CMHE04, p. 50], which also sells licenses and trainings for the Myers–Briggs Type Indicator, indicating a potential conflict of interest. Moreover, the supporting studies are inconsistent and scientifically weak [GM96, p. 78].

Another reason against further considering the Myers–Briggs Type Indicator and the typology approach to personality as a whole in this thesis it that the inventories are not publicly available and licenses have to be bought for every test that is to be administered [Mye22].

2.4 Validity of survey instruments

In order to develop a concept that can be used to validate personality questionnaires, it is necessary to get a clear picture of what validity means in the context of survey instruments. In a very general sense, validity can be defined as how accurately a test, or in our case a survey instrument, measures what it is supposed to measure [CS10, p. 1]. Other references describe validity as “the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests” [AAN14, p. 11]. However, such abstract descriptions can hardly be of use for practical validation purposes. Thus, a large variety of different types and subtypes of validity were developed and can be found in the scientific literature, as evidenced by the summary list of more than 150 types of validity in [NS14, Table 1.3, p. 8].

The immense amount of separate and intersecting concepts shows that it is hard to define a conclusive notion for validity [NS14, p. 183], so for the purpose of this text it is necessary to define the types of validity that are to be investigated in more detail.

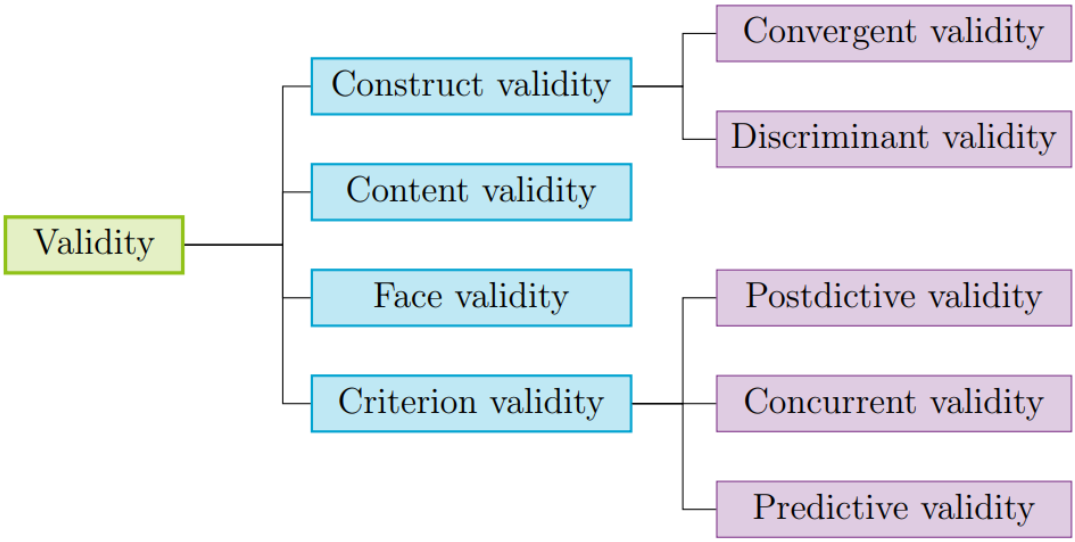

Figure 2.6 Selection of types and subtypes of validity, based on [Tah16, Figure 1, p. 29], adapted.

Figure 2.6 shows the types of validity that are discussed in the following subsections and which are focused on for the concept development later on. The four categories face, content, criterion and construct validity are the most prominent types in the relevant literature. For some time, content, criterion and construct validity were seen as the three most important types of validity and they were treated as separate kinds of validity. Research in this area has by now agreed that they are merely three aspects of validity, which each describe some part of the general notion of validity [Gui98, p. 236].

By whichever way the validity of any test is attempted to be established, it can only be measured compared to different tests, as there is no absolute authority on the question of validity [CS10, p. 1].

2.4.1 Face validity

The first type of validity, called face validity, is less technical than those in the subsequent sections. It is a relatively subjective aspect of a given test, survey instrument or item [Hol10], and can be described as the degree to which a nonprofessional considers the test relevant to the subject and the surrounding situation in which it is taken. If a clear relation can be seen, the test can be classified as face valid [HJ79, pp. 460f]. It is important to note that for face validity it is not the researcher's judgment that is important, but the test respondents' opinions. As such, specific knowledge of a subject or test environment is not a prerequisite for the evaluation of face validity, rather the contrary [Cro90, pp. 216f].

Face validity is a weak form of validity, and should not be taken as only measure of the validity of a survey instrument [Cro90, pp. 216f]. Some authors argue that it is not even a real form of validity, but can still be useful for test takers as it can be motivational for them to recognize from the questions which subject they are tested for [KS17, p. 136]. On the other hand, it is harder for test takers to manipulate the result if they do not know what a test item is supposed to measure, i.e., if the item lacks face validity [CS10, p. 1]. So missing face validity can also be an advantage.

2.4.2 Content validity

A deeper analysis than simply assessing the appearance of the survey items is necessary if content validity is in question, which is a measure of the relevance of the survey instrument [Cro90, p. 170], and which gauges whether the questionnaire items are representative and comprehensive with respect to the extent of the theoretical construct that the instrument is supposed to measure [Jac04, p. 9].

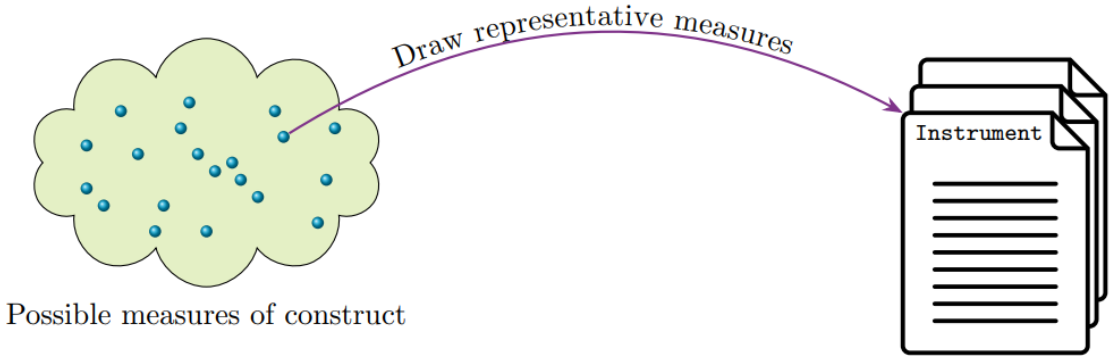

Figure 2.7 Visualization of the content validity notion, based on [SBG04, Figure 2, p. 386], adapted.

Figure 2.7 shows a visual representation of the concept of content validity. While many options are available for creating measurement tools for a specific construct, a choice has to be made while constructing a survey instrument or test. The question of content validity now basically boils down to how good these choices were during the design process of the test items. An instrument thus either contains measurement errors when chosen items do not represent the construct, or excludes construct facets when important items are excluded [SBG04, p. 386].

Different ways to establish content validity are possible, the simplest is judging the instrument items on the basis of the findings of an extensive literature review. Additionally a panel of experts can pass a judgment on the content validity of an instrument [SBG04, p. 387].

To get a better and more convincing result from the content validation procedure, an additional empirical validation step can be performed, which computes a metric of content validity [Law75]. It has been extensively used for content validation purposes in various research areas [WPS12, p. 199].

A panel of experts judges the content validity by opining whether each individual item is essential in order to test for the specific construct the item is designed for. Each panelist answers this question on their own on a trichotomous scale, i.e., by determining whether they find the item

- essential,
- useful but not essential, or
- not necessary.

From the data obtained in this way, the content validity ratio

$$\mathrm{CVR} = \frac{n_e - \frac{N}{2}}{\frac{N}{2}}$$

as a measure for the agreement of the experts is computed, where ne is the number of "essential" votes and N is the total number of experts on the panel [Law75, p. 567]. While this coefficient is only a rescaling of the percentage of agreeing panelists, the resulting scale for the CVR coefficient lies between unity and negative unity, which makes the extent of the consensus among the judges more tangible [Law75, p. 567]. When all judges vote "essential", the CVR of this item is 1, while it is 0 when only half of them vote "essential" and the other half sees the item as not essential or not necessary.

In the original publication that introduced the content validity ratio CVR, a table provided suggestions for the critical CVR value CVRcrit, above which the item could be considered to have content validity, depending on the panel size N [Law75, Table 1, p. 568]. However, a subsequent article doubted whether the computation of these values was correct and developed their own table, based on similar or slightly different assumptions, see [WPS12, Table 2]. These values were doubted and recalculated again, and in Table 2.3 some of the newest CVRcrit values from [AS14, Table 2, p. 85] are given to get an impression of the intra-panel agreement necessary to validate the item content.

Table 2.3: CVRcrit values in dependence of panel size, excerpt from [AS14, Table 2, p. 85]. Ncrit is the minimum number of panelists necessary to vote "essential" on an item.

N	N_{crit}	$\mathrm{CVR}_{\mathrm{crit}}$
5	5	1.000
10	9	0.800
20	15	0.500
30	20	0.333
40	26	0.300

Apart from the mentioned methods to validate items, the proportion of items in an instrument related to each of the constructs being measured should reflect the importance of the construct as determined by a literature review or expert testimony [CS10, p. 1]. It may be difficult to determine the exact number of items needed for each construct, however developing too few items is deemed disadvantageous, as eliminating inadequate items at a later stage is easier that introducing new items [CZ79, p. 21].

A major problem when evaluating the content validity of an instrument is that the extent of the domain from which the items are chosen, pictured as the green cloud in Figure 2.7, in itself is unknown, which means that a proper coverage by the items is hard to verify [SBG04, p. 387].

2.4.3 Criterion validity

This type of validity concerns the accuracy of an instrument, and can be assessed when comparing the scores with those of an instrument whose validity has already been established. Both instruments need to be related, which can be assessed by determining their correlation, in order for this method to yield viable results [CS10, p. 1].

In general, criterion validity states that the measuring instrument accurately estimates the criterion it is supposed to measure [CZ79, p. 17]. As an example, a university entry exam is criterion valid, if it highly correlates with the actual study success of the individual students. This is at the same time an example for a subtype of criterion validity, which is called predictive validity. It can be present if the criterion the instrument is supposed to correlate with does not yet exist when the instrument is applied, as is the case for the study success of students taking an entry exam. Concurrent validity on the other hand can be measured by correlating the test result with an existing criterion, at the same time [CZ79, p. 18]. The third and lesser used form of criterion validity is postdictive validity, which has a hindsight perspective [Tah16, p. 33].

A commonly used method to verify the criterion validity of a test is calculating a validation coefficient, which is the correlation coefficient between a score of a test result and a criterion variable, which may be a dummy variable [Fra02, p. 37].

In addition to the necessary verification of the survey instrument or test, it is also imperative to put thought into the measurement of the actual criterion, that the instrument is designed for. If there is no real way of measuring the real world criterion, the criterion validity of a test can not be established properly [CZ79, p. 19].

2.4.4 Construct validity

In contrast to the types of validity described in the preceding sections, which each only encompass an aspect of validity, construct validity can be seen as a scientific concept for validity in general [Loe57, p. 636]. Thus, as content and criterion validity only cover a small part of what validity is concerned with, construct validity should always be considered when assessing instruments [Mes87, p. 11].

In general, construct validity can be seen as a measure of how well instrument items capture the relations between constructs. The two categories of construct validity are thus convergent validity, which describes whether items measure the construct they are supposed to measure even if the instrument also considers other constructs, and discriminant validity, which describes how well items can distinguish between constructs [SBG04, p. 388]. As an example, consider a test that investigates the two constructs A and B. When the test scores show that an item has a high factor loading for A, i.e., it is a good indicator for construct A, then the convergent validity of this item for construct A is established. If the same item also shows a high factor loading for B, then the discriminant validity of this item can not be established.

When determining the construct validity of a test, a factor analysis employing principal component analysis in combination with the varimax rotation method can be used on sample test scores, see, e.g., [HBBA18, pp. 121ff]. Items with a factor loading of above 0.4 on the desired construct can be considered as convergent, and convergent items with no cross loadings on other constructs above 0.4 are discriminant [SBG04, p. 410]. Other thresholds may be applicable, see [HBBA18, p. 152].

3 Personality traits and the word embedding space

Due to the mentioned weaknesses of the Myers–Briggs Type Indicator, only the Big Five model with the domains extraversion, agreeableness, conscientiousness, neuroticism and openness is considered further. The question to be answered now is how these five domains are represented in a word embedding vector space. The Python implementation used to perform the tests with the word embedding spaces in this chapter can be found at https://gitlab.com/volker.kempf/ validation_by_word_embeddings.

3.1 Personality dimensions in the word embedding space

The dimensions of the vector space can not easily be identified with individual features of the word semantics [MSLS20, p. 1548], however they hold certain information about this aspect of natural language, as shown in the examples from Section 2.2. Following a concept from [Roz20], interpretability is attempted to be added to the word embedding space by identifying *trait dimensions*.

This method identifies for each trait one direction in the vector space that represents this trait. The steps are explained with the trait *extraversion*, the other directions can be found analogously. First, a pole $w_e \in \mathbb{R}^d$ d for extraversion is identified in the vector space by choosing several words that represent this trait and subsequent averaging of the corresponding word vectors, and the same is done for the polar opposite, i.e., introversion, resulting in the pole $w_i \in \mathbb{R}^d$. To find these poles, this example uses the lists of adjectives given in [Gol92, Table 3, p. 34], which have high factor loadings for the Big Five traits, and which are given in Table A.2. Thus the word set W_e with the word vectors for the words in the extraversion column in Table A.2, and the set W_i with the embeddings of the words in the introversion column of the table is generated. To compute the poles w_e, w_i, the vectors for the words in the two sets are normalized, summed and again normalized, i.e.,

$$w_e = \frac{\sum_{w \in W_e} \frac{w}{\|w\|_2}}{\left\| \sum_{w \in W_e} \frac{w}{\|w\|_2} \right\|_2}, \qquad w_i = \frac{\sum_{w \in W_i} \frac{w}{\|w\|_2}}{\left\| \sum_{w \in W_i} \frac{w}{\|w\|_2} \right\|_2}.$$

From these two poles representing the two extreme ends of the extraversion– introversion scale, the extraversion dimension v e can be computed by

$$v^e = \frac{w_e - w_i}{\|w_e - w_i\|},$$

i.e., subtracting the two poles and normalizing the result.

The process is the same as pictured in Figure 2.4. The result is a vector of unit length in the direction of the extraversion–introversion axis. Repeating the process for the other personality domains gives the five trait dimension vectors v^e, v^a, v^c, v^n and v^o.

In Big Five related literature from the psychology field, the five domains are said to be orthogonal factors contributing to the description of personality [Gol90, p. 1216], where orthogonality means that a factor analysis of test results with subsequent orthogonal varimax rotation of the found factors yields the Big Five traits. With the tool of the word embedding space where two vectors are orthogonal if the angle between them is exactly 90 degrees, this orthogonality relation can be inspected mathematically. Having defined a direction in the word embedding space for each trait, it is possible to check whether the pre-trained word embeddings correctly reproduce the pairwise orthogonality relation between the traits. The expectation for such an experiment is that the angles are close to $90°$, but since the word embeddings are experimental and based on a fixed amount of texts some deviance from the optimal value is acceptable.

Table 3.1 Angles between trait dimensions in pretrained word embeddings set GVWiki described in Table A.1. Traits are abbreviated by their initial letter.

	E	A	C	N	O
E	0°	82°	74°	91°	72°
A	82°	0°	65°	98°	71°
C	74°	65°	0°	88°	67°
N	91°	98°	88°	0°	76°
O	72°	71°	67°	76°	0°

To this end, the previously defined cosine similarity function from Equation (2.1) is used, and the angle between the trait directions is recovered after applying the inverse cosine function arccos. To get, e.g., the angle α_e^a between the extraversion and agreeableness dimensions it is necessary to compute

$$\alpha_e^a = \arccos\big(\mathrm{sim}_{\cos}(v^e, v^a)\big) = \arccos(v^e \cdot v^a).$$

The subsequent computations are based on the 300-dimensional embeddings GVWiki, cf. Table A.1, which are pre-trained on a Wikipedia text corpus, and the results are given in Table 3.1. Trait adjectives that are not in the vocabulary of the pre-trained set are discarded and not used for the averaging process during the computation of the trait poles. As expected, the trait dimensions are not perfectly orthogonal. All angles lie between 65° and 98°, with an average difference of only 13.4° from the perfect 90°.

This means that while there is a slight correlation of the trait dimensions in the word embedding space, the highest between the agreeableness and conscientiousness traits with an angle of 65°, they can be considered independent enough for applications. From a practical point of view, the trait directions could be orthogonalized, e.g., using the Gram–Schmidt process, but this would probably alter the underlying semantic structure which needs to be as accurate as possible.

Table 3.2 Means and standard deviations of trait angles for different pre-trained word embeddings.

	mean	SD
fTCrawl	80.0°	17.3°
fTWiki	81.4°	16.8°
GVCrawl	74.5°	14.8°
GVTwitter	81.8°	10.5°
GVWiki	78.4°	10.4°
w2vNews	81.6°	14.9°

In addition, this computation was performed for all pre-trained word embeddings listed in Table A.1, and the mean and standard deviation of the observed angles of the trait dimensions are given in Table 3.2. The results in the table show that the word embeddings GVTwitter which were generated from Twitter data are in fact closest to the orthogonal structure of the trait dimensions.

3.2 Clustering of trait descriptive adjectives

As a second way to show that the inherent semantic structure of the word embedding space contains enough information about the personality domains to be usable for the intended purpose, a clustering approach similar to the method employed in [Swi21, Section 4] is presented in this section. For this study, instead of the trait descriptive adjectives from [Gol92], now those from [Joh21, Table 2.4, p. 50] are used, which are given in Table 3.3. Using principal component analysis as a dimensionality reduction technique for the word embeddings of these adjectives, it is possible to distinguish the regions of the Big Five traits. Figure 3.1 shows a three dimensional representation of the top 5 words from each trait measured by factor loading. Even with this qualitative view of the word embedding space, it is clear that the adjectives belonging to one trait seem to cluster in certain regions.

Table 3.3 Top 10 trait descriptive adjectives by factor loading, which is given next to each word, for the Big Five domains, from [Joh21, Table 2.4, p. 50].

Extraversion		Agreeableness		Conscientiousness		Neuroticism		Open-mindedness	
talkative	.85	sympathetic	.87	organized	.80	tense	.73	imaginative	.76
assertive	.83	kind	.85	thorough	.80	anxious	.72	original	.73
active	.82	appreciative	.85	planful	.78	nervous	.72	intelligent	.72
energetic	.82	affectionate	.84	efficient	.78	moody	.71	insightful	.68
outgoing	.82	soft-hearted	.84	responsible	.73	worrying	.71	curious	.64
outspoken	.80	warm	.82	reliable	.72	touchy	.68	sophisticated	.59
dominant	.79	generous	.81	dependable	.70	fearful	.64	artistic	.59
forceful	.73	trusting	.78	conscientious	.68	high-strung	.63	clever	.59
enthusiastic	.73	helpful	.77	precise	.66	self-pitying	.63	inventive	.58
show-off	.68	forgiving	.77	practical	.66	temperamental	.60	sharp-witted	.56

To get an even clearer result, the k-means clustering algorithm is used on the word embeddings corresponding to these adjectives. This algorithm gets as input a discrete set $D \subset \mathbb{R}^d$ with n elements, and a predefined number $m < n$ of clusters, into which these points are to be divided. The algorithm then completes the following steps:

1. Initialize the iterative process by choosing m of the n elements from the set D as centroids c_i , $i \in \{1, \ldots, m\}$, at random.

2. Sort every element into one of the m temporary cluster sets D_{c_i}, depending on which of the centroids it lies closest to, measured in the Euclidean distance.

3. Update the centroid positions, by setting c_i to

$$c_i = \frac{1}{|D_{c_i}|} \sum_{a \in D_{c_i}} a,$$

for all $i \in \{1, \ldots, m\}$.

4. Check whether the updated centroid positions differ from the old positions. If all positions are unchanged return the found clusters D_{c_i}, else go to step 2.

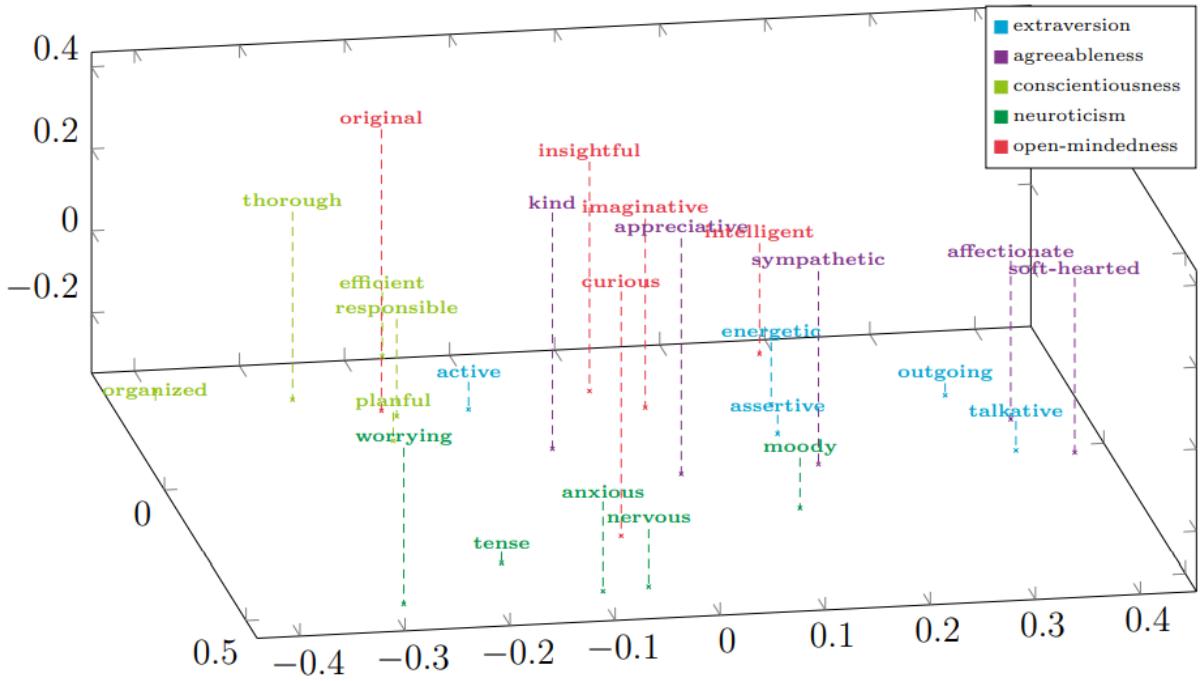

Figure 3.1 Positions of the top five trait descriptive adjectives from [Joh21, Table 2.4, p. 50] after performing a three-dimensional principal component analysis.

Since there is a random initialization involved in the algorithm, it is nondeterministic and yields possibly different results for each run. To get around this problem, a quality metric for the resulting clusterings is used to determine the best clustering after a fixed number of runs of the algorithm. The ground truth clustering is known in this case, cf. Table 3.3, so the experiment uses the adjusted mutual information index to measure and maximize the clustering quality over the separate runs. For a computed clustering and the given true clustering, the adjusted mutual information index returns a number $R_{AMI} \leq 1$ which is 1 for a perfect match and 0 if the found clustering resembles a random labeling [VEB10, p. 2845].

Since the random initialization chooses 5 elements out of the given 50 trait descriptive adjectives as initial centroids for the clusters, there are a total of $\binom{50}{5} = 2118760$ possible initial settings for the algorithm, which possibly lead to different clusters. Luckily due to the computing power of current tabletop computers, it is possible to find a not necessarily unique clustering that maximizes the adjusted mutual information score in reasonable time by trying every possible starting configuration of centroids.

Using the word embeddings fTCrawl, the best clustering result for the adjectives from Table 3.3 has an adjusted mutual information index of RAMI ≈ 0.832 and is visualized in Figure 3.2. It is apparent that all clusters are largely recovered by the algorithm, which speaks for the capability of the word embeddings' inherent semantic structure to represent the Big Five traits.

The conscientiousness, neuroticism and open-mindedness adjectives were all assigned to the same respective cluster, only two words each from the extraversion and agreeableness trait were put into wrong clusters. This result is significantly better than the one shown in [Swi21, Figure 8b, p. 127], where the trait descriptive adjectives from [Gol92], cf. Table A.2, were clustered and a total of seven words were put in wrong clusters. A possible reason may be the differences in the used adjectives or pre-trained word embeddings, as the reference used a fastText model which was trained on another corpus. Another difference is that all possible combinations of starting centroids were tested to get the result with a maximal RAMI-score seen in Figure 3.2, while only 1000 random initializations were used in [Swi21].

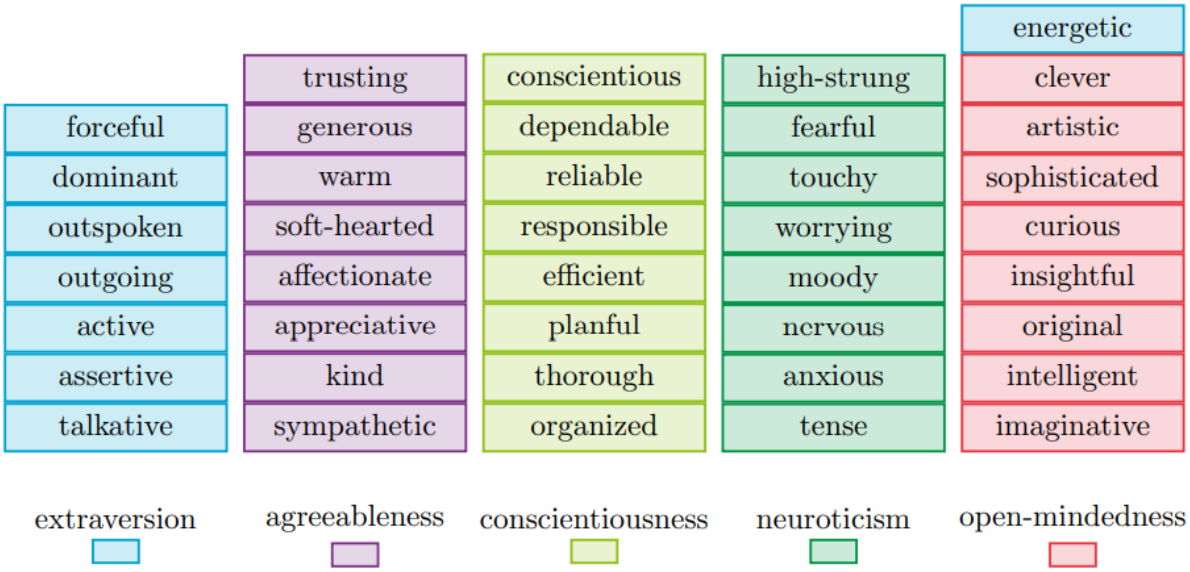

Figure 3.2 Clustering with maximal adjusted mutual information index of the top 8 trait descriptive adjectives from [Joh21, Table 2.4, p. 50]. Stacks represent found clusters, colors the true affiliation of an adjective.

The clustering result is already very good and supports the hypothesis, that the Big Five trait structure is present in the word embedding model. However, by reducing the number of clustered vectors in this experiment to the top 8 adjectives of each trait, it is possible to obtain an almost perfect clustering, where only the word energetic, which should be in the extraversion cluster is placed in the wrong cluster. This clustering reaches an almost perfect adjusted mutual information score of RAMI ≈ 0.943 and is visualized in Figure 3.3.

The clustering analysis with all ten adjectives from each trait was performed with all the different pre-trained word embeddings from Table A.1, with the results given in Table 3.4. The three sets of pre-trained word embeddings GVCrawl, fTWiki and fTCrawl contain more semantic information to separate the Big Five into distinct regions in the word embedding space as evidenced by the high adjusted mutual information score and low number of wrongly clustered adjectives.

These sets also contain all of the trait descriptive adjectives in their vocabulary, or in the case of the fastText models can place them appropriately due to the contained subword information. The word2vec and smaller GloVe models on the other hand do not contain all the words, e.g., in the word2vec case all adjectives that contain a hyphen like show-off are not contained in the vocabulary, i.e., no word vector for them exists in the model. Furthermore, the GloVe model that was trained on text data from Twitter performs by far worst in this task, which means that Twitter text data contains not enough semantic details to build a proper model of the Big Five in the word embedding space.

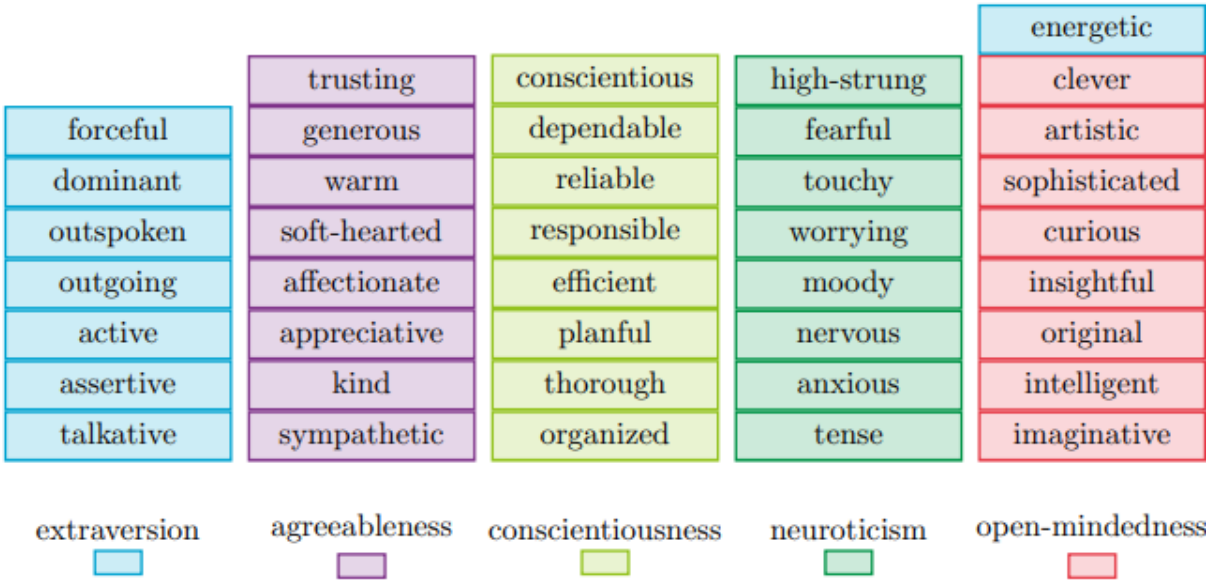

Figure 3.3 Clustering with maximal adjusted mutual information index of the top 8 trait descriptive adjectives from [Joh21, Table 2.4, p. 50]. Stacks represent found clusters, colors the true affiliation of an adjective.

Table 3.4 Number of misplaced items Nf and adjusted mutual information score RAMI of best achieved clusterings of the top 10 trait descriptive adjectives from [Joh21, Table 2.4, p. 50]. Nm is the number of adjectives that are not in the word embedding vocabulary.

	R_{AMI}	N_f	N_m
fTCrawl	0.832	4	0
fTWiki	0.835	4	0
GVCrawl	0.879	3	0
GVTwitter	0.575	18	2
GVWiki	0.717	7	3
w2vNews	0.762	5	5

Due to these findings it can be hypothesized that the word embeddings GVWiki, GVTwitter and w2vNews will perform worse than the embeddings GVCrawl, fTWiki and fTCrawl in applications that rely on the Big Five structure.

4 Organizational personality evaluation

Knowledge about the personality of employees or job candidates is a vital part of today's business world. Depending on the job requirements, it can make a significant difference in job performance if the worker has a personality with high conscientiousness or low extroversion scores. This chapter discusses the motivation and practical aspects behind personality testing in professional settings.

4.1 Motivation and history of personality testing in organizations

The earliest systematic efforts of personality and trade testing can be pinpointed to the build-up of the United Stated army in preparation of their involvement in World War One, where a large number of people had to be assessed for their mental and leadership capabilities as well as technical skills [ZK13, p. 175]. During that time, in order to identify recruits in advance that would likely show extreme anxiety when first exposed to battle, a condition then called shell-shock and which would now be considered a type of post traumatic stress disorder, the Psychoneurotic Tendencies scale was created, a first inventory for assessment. When the war was won by the Allies, the developer released the test in a revised version called Woodworth Personal Data Sheet and marketed it as a tool to sort out unsuitable workers [ZK13, p. 176].

This was the beginning of organizational personality testing, which was at that time mainly used to filter out unwanted applicants even though the methods were largely unscientific [ZK13, p. 176]. A prominent example of this misguided practice that was promoted mostly for prospective financial gain or fame is the case of Katherine Blackford, who claimed to have developed a method to determine the personality of persons by visual cues, even by simply examining photographs of people [Bla18].

Taking the mentioned Woodworth Personal Data Sheet as a starting point, several personality tests were created which were largely used for testing the adjustment trait of workers, since inadequate emotional adjustment was seen as the main factor leading to poor work performance. So in the decades leading up to the Second World War, personality inventories were developed specifically for the use of identifying potential labor and union activists [Zic01, pp. 153f]. Some tests in the United States were specifically marketed and used to circumvent a law from 1935 that made it illegal to ask job candidates about their view on unions [ZK13, p. 178]. Such uses of personality tests are now widely seen as unethical.

Instead of focusing on just one construct domain, the next generation of personality tests were multi-factorial instruments that evaluated several personality aspects. The first of these new tests was the Bernreuter Personality Inventory that was developed in 1931.

It had 125 items, measured neurotic tendency, self-sufficiency, introversion–extroversion and dominance–submission scales, and was used in several professional settings, e.g., selection of traveling sales representatives and engineers [GZ08, pp. 171f]. Other instruments of this type include the Minnesota Multiphasic Personality Inventory, which saw use in organizational selection processes until the 1990s, and the Myers–Briggs Type Indicator [ZK13, pp. 178f], which was already described in Section 2.3.2.

After a period of low popularity beginning in the 1960s due to seemingly low validity of the existing test instruments [ZK13, p. 182], organizational personality testing came back into general use in the 1990s. Several publications using a new meta-analysis approach had reviewed existing studies and found that personality tests could provide information that was relevant for work situations, specifically if a conscientiousness scale was included [ZK13, p. 183]. In the same timeperiod, the Big Five trait model, cf. Section 2.3.1, was developed and came into widespread use. The field of organizational psychology largely also switched over and accepted the Big Five model [ZK13, p. 184]. As mentioned in Section 2.3.1, there are approaches using the HEXACO six factor model, however those are not considered here. Since interest in organizational personality testing had started to grow again, it has gotten very popular in the business world, with reports of 20–40% of companies in the United States using some form of personality test in their application process [RG06, p. 156]. The widespread use of personality evaluation in hiring procedures is also supported by the data of an online survey conducted for this thesis, see Appendix B, where 27% of the respondents who had gone through an application process stated that a personality questionnaire or conversation with a psychologist was part of the process.

4.2 Types and uses of organizational personality tests

A main goal of organizational psychology research is to determine to which extent personality predicts how well a worker performs in his job. The overall job performance comprises the three aspects task performance, organizational citizenship performance and counterproductive work behavior [SAM20, p. 429], see Figure 4.1.

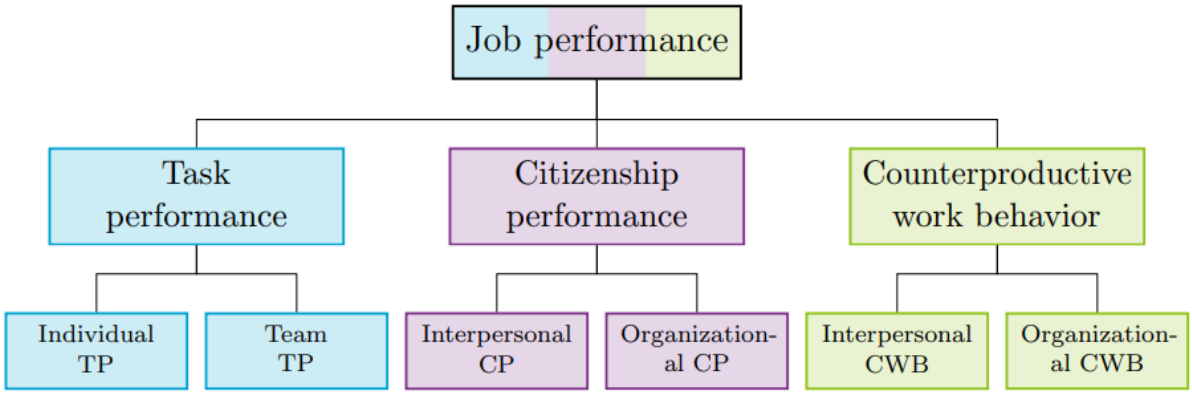

Figure 8 Aspects of job performance, based on [MSA17, Figure 2.1, p. 28], adapted.

Task performance is divided into individual and team task performance and describes how well the employees can perform the main functions of their job. Citizenship performance is a measure for how well employees have internalized the goals of the organization and are working to accomplish them, by acting outside of their usual and required work function. Counterproductive work behavior is defined as purposeful behavior of an employee against the interests of the organization [MSA17, pp. 28f]. Counterproductive work behavior can be directed against other employees, e.g., in the form of bullying or aggressiveness, or against the organization itself, e.g., as theft or voluntary absenteeism [HD12, p. 544].

Getting the data that is required to evaluate job performance is not always straight-forward, especially in occupations where no product is generated as is the case for, e.g., professors, service and public security personnel [MSA17, p. 29]. When objectively measurable performance related data, e.g., production output, is not available, the only way to assess job performance is through subjective ratings from supervisors, co-workers or subordinates. However, this method is prone to biases and other inaccuracies, which reduces its reliability [VO00, p. 222].

Concerning the relation of job performance and personality, an initial study on the predictive capability of the Big Five traits for job performance found that conscientiousness and emotional stability, the opposite of neuroticism, are valid predictors for general job performance [BM91]. Additionally in occupations that involve frequent contacts with people, agreeableness and extroversion are good predictors [MSA17, p. 39]. Subsequent meta-studies could confirm these findings [SAM20, p. 430].

Organizational citizenship performance, which includes, e.g., volunteering for additional duties and taking on more responsibility [SAM20, p. 431], is found to be predicted by conscientiousness, extraversion, agreeableness and emotional stability, with conscientiousness being the best predictor [Jud+13, Table 7, p. 890].

For organizations, counterproductive behavior of employees towards the company or co-workers must be avoided. Here also conscientiousness and agreeableness are valid to predict the lack of deviant behavior [Sal02, pp. 121f], which is a subtype of counterproductive behavior, and which includes, e.g., theft, property damage, substance abuse [Sal02, Table 1, p. 118]. With neuroticism replaced by its opposite, emotional stability, all Big Five traits were predictors of a lack of personnel turnover [Sal02, pp. 121f].

Another interesting topic are *dark side* personality factors, i.e., traits that are socially undesirable but can have negative as well as positive consequences for organizations and individuals [JL07, p. 334]. Examples for these kinds of traits are narcissism, dominance and Machiavellianism [Fur18, p. 548].

People with high scores on neuroticism have been found to think they are more effective leaders, have higher task and citizenship performance than they really do [JL07, p. 337]. These traits also seem to predict short-term success in the workplace, but long term failure [Fur18, p. 547].

Overall, the Big Five have been shown to be useful predictors of job performance as a whole and its subdomains, with conscientiousness being the most useful. As such, it is safe to assume that a benefit for companies can be drawn from systematic personality testing in personnel selection and assessment procedures both for new applicants and internal staff.

4.3 Challenges of organizational personality testing

In addition to the mentioned problems concerning uncertainties and biases in the measuring process of job performance, which blurs the validity of personalitybased selection procedures, there are other aspects of personality theory and personality testing that can have a negative impact. Although there are more challenges in existence, only the issues of faking during personality tests and the response of applicants to personality tests are discussed in detail.

4.3.1 Faking in assessment procedures

While self-evaluation questionnaires are practical and convenient to use, they can be susceptible to self-deceptive biases of the candidates, but also to intentional faking of answers, which means that the applicant chooses answers that from their perspective look beneficial for a successful outcome of the application. This issue complicates the interpretation of such tests and reduces their benefit regarding the decisions they are supposed to support [Tet13, pp. 855f].

The severity of this problem has been analyzed with a scheme of seven nested questions [GR13, p. 254], the first of which asks whether faking behavior is even an identifiable construct, which is basically a question for a proper definition of the phenomenon. As a brief answer, faking seems to be an intentional behavior with differentiated motivations, and seems to have multiple facets and a set of correlated behavioral patterns, with the purpose of scoring higher in assessment procedures. Thus it can be considered a construct in the common sense, but also as a “multifaceted goal-oriented behavior” [GR13, p. 258].

The second question, whether people can be expected to fake, can also be answered positively. The nature of assessment and selection procedures incentivizes faking. Applicants do in general not have a strong connection and sense of loyalty to a company in the corporate sector, and tend to view corporations as purely professional workplaces [GR13, p. 259]. In addition, applicants in assessment settings do not have to fear severe consequences for faking in tests [GM06, p. 7]. This kind of mindset facilitates faking behavior.

The next question is whether people have the capability to fake in test settings. The intuitive answer to this question has been supported by some studies, which showed that people can fake by about one standard deviation when instructed to do so [Tet13, p. 856]. This can be combined with the fourth question, whether people actually engage in faking behavior during assessments. Research showed that a substantial percentage of applicants do fake in personality tests, averaging at around 30 % of candidates who intentionally try to increase their scores [GR13, p. 263].

On the question whether everybody has the same ability to fake, research shows that persons with higher mental ability tend to fake less often, but when they do, their faking is more effective compared to faking of people with lower mental abilities [LMC09, p. 278]. In addition, applicants whose real score on job relevant traits is low have more room and also more motivation for faking and inflating results. Thus the combination of high mental abilities and a true low score on job relevant traits in an applicant is the most challenging for the development of robust selection procedures [Tet13, p. 856].

The sixth question is whether faking is important to consider, i.e., if the effect of faking is large enough to alter the end results. Here it has been found that faking does negatively affect the decisions of the selection process, meaning that faking applicants increased their chances to be hired [GR13, p. 267]. In addition, those who require faking on self-evaluation tests to improve scores on job-relevant traits are likely to be unqualified for the job requiring those traits. If a faking applicant is hired, it is likely that a better fitting candidate does not get hired, which hurts the hiring organization [Tet13, p. 856]. Those are just a few of the reasons why faking is a problematic aspect in self-evaluation tests during assessments and selection processes.

Lastly, what can be done to reduce or prevent faking? The approaches that can be taken to mitigate the effects of faking can be put into two categories, preventive or remedial methods. The focus of preventive methods is to reduce the amount of faking that is happening by reducing motives, opportunities and abilities of applicants to fake. In contrast, remedial methods try to factor out the influence of faking after a test has been taken [Tet+06, p. 62]. The remedial methods rely heavily on sophisticated statistical methods, which is beyond the scope of the thesis to explain, so only some preventive methods are briefly described.

The method that is probably easiest to implement also seems to have the best effect in preventing faking. It consists of warning against faking in advance of the test application. Different kinds of warnings can be issued:

- Warning against faking,
- Laying out penalties should faking be identified, e.g., exclusion from the selection process,
- 1Laying out other consequences, e.g., if applicants do not answer honestly and get the job due to the faked test, they would not be happy in the job, since it has requirements they do not fit.

Several studies found that such warnings can reduce the extent of faking [GR13, p. 270]. Another possible preventive method is speeding, i.e., setting a tight time limit for the test, however this practice does not seem to reduce faking [GR13, p. 271].

4.3.2 Reactions to personality tests in assessment procedures

Even though personality self-evaluations are a tool often used in personnel selection procedures, usually no thought is given to how applicants react to taking such personal and intimate tests as a requirement for being hired. People tend to react more negatively on personality tests as part of a selection process than most other assessment tools like interviews, work samples and cognitive ability tests [HDT04, p. 659]. The favorability scores of several selection tools are shown in Figure 4.2. The green data point in the figure represents part of the results from a small scale online survey, see Appendix B, where the favorability of personality tests in hiring processes was asked from the view of an applicant. The results of the survey clearly support the data from the reference.

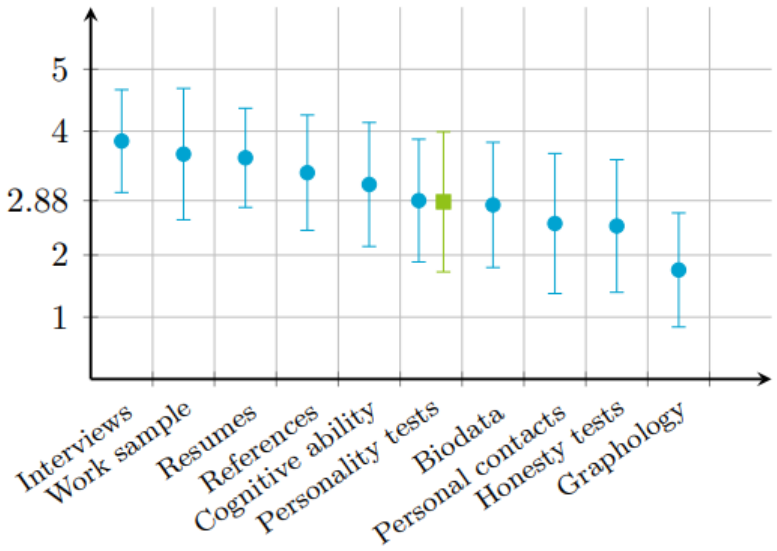

Figure 9 Favorability of selection tools on a scale from 1–5 with standard deviation, blue data from [HDT04, Table 4, p. 659], green data from own online survey, see Figure B.1.

The contrast to interviews which rank highest is especially interesting, since questions in interviews also tend to assess personality traits. However, they do this embedded in a context of work related questions and personal interaction between interviewer and applicant, while personality self-evaluation questionnaires on the other hand ask for personal and private information in a general setting that is not directly related to the job a candidate has applied for [McF13, pp. 283f]. Also in contrast to an interview, personality tests tend to pose very general questions and do not leave room for the candidates to explain their answers [McF13, p. 284]. Consider for example the item “I see myself as someone who does things efficiently” from the Big Five Inventory [JNS08, p. 157], cf. Table A.4, which is supposed to be rated on a five point scale from “strongly agree” to “strongly disagree”. Then without the opportunity to explain, a candidate, who is efficient at work but likes to be unbound elsewhere, may be in a conflict about how to answer, while a similar question in an interview could be answered in a nuanced way.

Applicants may also feel negatively about personality tests because it is mostly easy to see what the items are aiming to assess [MR00, p. 817]. With this knowledge, it is easy to deduce which and how item responses can be faked to achieve a more desirable score. So even applicants who do not intend to fake can see the weakness in the personality measure and feel cheated when they think some other job candidates may be faking in their answers [McF13, p. 284].

Negative reactions of applicants towards the selection process may hurt the organization in the long run, as dismissed candidates who feel that the process was unfair or just bad in general may not re-apply or deter people they know from applying, which reduces the future pool of candidates for the company. Thus it makes sense from the perspective of an organization to improve the applicants’ reception of personality tests, for which several strategies are possible.

The first approach goes in the direction of improving face validity of the test items. The rationale behind this is that when the construct that is being assessed is clearly noticeable from the items, the candidates can more easily see the relevance for the job they apply to. At the same time, greater face validity means that item answers can more easily be faked to achieve higher scores, which means that a certain balance must be found [McF13, p. 285]. One relatively simple and moderately effective remedy against faking was already mentioned in the previous section: Warning applicants against faking in the personality tests has been shown to reduce the amount of faking [McF03, p. 273]. Depending on the type and how it is delivered, a warning can positively or negatively affect the reactions towards the test, however not a lot of research is available for this aspect [McF13, p. 292].

Other approaches include offering an explanation for the test use, which was shown to improve the perception of a personality test [McF13, p. 287], or assessing personality by different means, for example during an interview.

4.4 Influence of personality trait composition on work teams

As described in the previous sections, aspects of personality have a clear relation to job performance for an individual worker. However, in most modern professional environments work is a team effort, and this trend is growing over time [Hal+05, p. 84]. Several employees have to work together to successfully achieve the team's tasks. These work teams may be long lived, consisting of the same employees working together for years, but in a lot of business branches it is common practice to assemble teams for a specific project, bundling competencies from different areas, and to disband the team when the project is finished. Software engineering is a prime example where work teams are usually assembled and managed with the help of some form of agile project management system. In such short lived project teams the members do not know each other beforehand, so in order to be effective a certain compatibility of the personalities seems intuitively necessary [DS13, p. 744].

Work teams can differ in more ways than just the already mentioned time of operation, for example in the number of team members, the type of work task or whether in-person meetings are common. This variety of work teams means that insights concerning the influence of personality on team performance should be broken down by team type. General guidelines for team composition may not be appropriate for a particular practical setting [Bel07, p. 608].

The results of a meta study concerning the influence of the team members' personality traits on team performance are presented in Table 4.1 [DS13]. For each of the Big Five traits it was considered whether the mean score of the team, the minimum and maximum scores in the team and the variance of trait scores had an influence on the team's performance. Considering the mean score is interesting for the hypothesis that the effectiveness of a team increases with increased scores of the trait in the team. On the other hand, minimum and maximum scores of an individual in a team matter when one member can have a singular effect on team performance. Variance in scores matters when it is expected that either similarity or heterogeneity among team members boosts team performance [DS13, p. 747]. The results across all studies shown in Table 4.1 confirm the expectation from the results concerning individual job performance that a higher mean conscientiousness score of a team has a large positive influence on team performance. The same is true for team mean agreeableness and to a smaller extent some of the studies showed the same for the other Big Five, when neuroticism is replaced by the positive version of this trait, emotional stability. Concerning minimum and maximum scores of traits in teams, the results are clear for minimum conscientiousness and agreeableness, where a higher minimum in the team positively correlates with team performance.

For intra-team trait variance, there is little statistically significant support except again for conscientiousness and agreeableness, where trait variance is negatively correlated with team performance, i.e., teams with more variance tend to perform worse.

Table 4.1 Results of meta study on Big Five trait influence on team performance, data from [DS13, Table 33.1, p. 752]. Shown quantities are weighted mean correlations. n.s. means statistically not significant. – means not measured in cited study

Trait criteria		Study [Bel07]	[PTRR06]	[Pre+09]
Extraversion	Mean	0.14	n.s.	0.09
	Maximum	n.s.	–	0.12
	Minimum	n.s.	–	n.s.
	Variance	n.s.	n.s.	0.06
Agreeableness	Mean	0.28	0.17	0.10
	Maximum	n.s.	–	n.s.
	Minimum	0.30	–	0.10
	Variance	n.s.	−0.09	−0.07
Conscientiousness	Mean	0.28	0.15	0.13
	Maximum	n.s.	–	0.09
	Minimum	0.22	–	0.13
	Variance	n.s.	−0.17	n.s.
Emotional stability	Mean	0.18	n.s.	0.08
	Maximum	n.s.	–	0.11
	Minimum	0.09	–	n.s.
	Variance	n.s.	n.s.	n.s.
Openness	Mean	0.20	n.s.	–
	Maximum	0.14	–	–
	Minimum	n.s.	–	–
	Variance	n.s.	n.s.	–

In contrast to these general findings, there can be some negative effects of generally desirable traits. Conscientiousness tends to be a trait that positively affects team performance, see Table 4.1, however when adaption of the team's behavior after an unforeseen event is necessary, the decisions of persons with high conscientiousness scores are worse than those of persons with low scores [LeP03, p. 29]. It seems that individuals with high dependability, which is a facet of conscientiousness, adapt worse to new situations [JL07, p. 342]. In addition, team mean conscientiousness, while positively correlated to the quality of results, seems to be negatively related to result quantity in creative tasks. This was shown in a study where groups were given the task to find as many different uses for some objects [WB98, p. 626]. The high conscientiousness groups seemed to emphasize quality of results over the explicitly stated objective to produce as many uses as possible [WB98, p. 631].

Normally agreeableness is also a desirable trait for groups as agreeable people tend to be cooperative and helpful. However, in team situations where it is necessary to voice an opinion and challenge established routines, agreeable persons may not speak out, which in such settings can be detrimental to group performance [JL07, p. 343].

4.5 Standard inventories for organizational personality tests

Since the resurgence of interest in organizational personality testing in the 1990s, many test instruments were developed and older ones revised. Many of them aim for the Big Five model, or a subset of it, while others follow different approaches, e.g., the Myers–Briggs Type Indicator. Although most of these tests are already in use for decades, none has clearly turned out to work best for all circumstances. Each test is rather marketed with some specialty of application. To get a better overview, an extensive study compared 12 of the most popular tests [PTC13]. Without going into the details of each of the questionnaires, the main results of the comparison are interesting and briefly presented below.

The study concluded that the Hogan Personality Inventory (HPI), the Occupational Personality Questionnaire-32n and -32i (OPQ-32), the Personality Research Form (PRF) and the Wonderlic 5 tests performed best in the comparison, mainly since the methodology of these instruments is appropriate for organizational use and since the criterion validity evidence is convincing [PTC13, pp. 218f].

Tests in the middle of the comparison are the 16 Personality Factor Questionnaire (16PF), the Caliper Profile, the Global Personality Inventory–Adaptive (GPI-A) an the Neuroticism-Extraversion-Openness Personality Inventory 3 (NEO-PI-3). Although these tests have certain strengths and can have some uses in human resource management, they are lacking with respect to normative data and validity for job related criteria compared to the more favorably rated instruments [PTC13, p. 219].

The tests on the lower end of the comparison are the California Psychological Inventory (CPI), the Myers–Briggs Type Indicator (MBTI) and the Minnesota Multiphasic Personality Inventory 2 (MMPI-2). Main criticism of the CPI is its lack of validity for job-related criteria, of the MBTI its reliance on the binary categories, see Section 2.3.2, and of the MMPI-2 its lack of validity evidence for a normal working population, as it was developed for the use in law enforcement [PTC13, p. 219].

Due to the limited scope of this thesis, none of these tests can be considered further. The focus will be on the Big Five Inventory (BFI), see [JNS08], the Big Five Inventory 2 (BFI2), see [SJ17], which are general inventories for the Big Five traits. In order to have another large set of test items to work with, the items from the International Personality Item Pool (IPIP), see [Gol22; Gol+06], aimed at the Big Five traits will be used for some evaluations.

5 A concept for the validation of personality survey instruments

The motivation to have good survey inventories for the assessment of personality in organizational settings is evident after considering the research on this topic that was laid out in the previous chapter. Validity is a good indicator of the quality of a questionnaire, however, the process of validation is a complex and time consuming matter. The aim of this chapter is to develop a concept with which aspects of the validity of personality survey instruments can be assessed by investigating the questionnaire items on the basis of word embeddings.

5.1 Observations concerning validity and word embeddings

Using a tool like word embeddings that is based purely on semantic information found in a large text corpus, it is not possible to evaluate the validity of a survey instrument in the same way it is classically done, some methods were described in Section 2.4. Typical evaluations of convergent and discriminant validity for example rely heavily on the data generated from applying the tests to sample populations and statistically analyzing the given answers, e.g., with principle component analysis. Working only with the vector representation of the words in the items, such a type of examination is of course not possible. However in this chapter some methods are developed on the basis of word embeddings that can help with the evaluation of some aspects of validity of personality tests.

Another limitation of word embeddings is that they can only integrate the semantic information that is present in the base corpus of text from which they are built. So in order to get results that are optimally fitted to the type of investigated survey instruments, it would make sense to specifically train word embedding models on texts of the relevant research fields. Unfortunately there are only very few topic specific text corpora available, and within the review of literature and online data sources conducted for this thesis no personality psychology specific corpus was found. So in order to train a set of word embeddings specifically designed for applications with a psychological background, a lot of work is needed to curate a custom text corpus, which is beyond the scope of this thesis. Additionally, there are also some approaches to integrate topics and generate topic-specific word embeddings [ZHZ20]. In this thesis for the case of personality survey instruments, the standard pre-trained word embedding models that are freely available to download, cf. Table A.1, are used.

5.2 Term extraction from survey items

In order to work effectively with word embeddings in connection with items from a questionnaire, common words in the items that do not carry any semantic value need to be stripped from the items. In natural language processing these so called stop words are regularly dismissed from the processed texts beforehand. Examples for stop words are is, a and the.

In the tests of the proposed techniques for item validation in the following chapter, the items are used in three stages of pre-processing in order to see how robust the proposed methods are, and which stage is most practical:

I1: The raw items are used with minimal pre-processing like removing punctuation to facilitate the automated natural language processing.

I2: Stop words are removed from the items.

I3: Each item is reduced to one adjective that tries to capture the meaning of the item.

The word embeddings of all the remaining words of an item at each stage are then averaged into a single vector for the further steps. Averaging word vectors as an attempt to capture the meaning of sentences is a known method that works [LM14, p. 1189], but it has its limitations and the more words are averaged the less accurate it becomes. The last stage of pre-processing has to be performed manually and while for most items one adjective can be chosen from the words of the item, some items require interpretation in order to find appropriate adjectives. This reduces the value of the results for those items that required interpretation, as not the original wording is investigated. However, it can help to test the performance of the proposed methods from the next section under more ideal conditions, i.e., when the meaning of an item is more clearly defined.

There is an extension to the word2vec algorithm, that generates vector representations for sentences, paragraphs or whole documents called doc2vec [LM14]. Those might lead to better results in this use case than the averaging procedure proposed here, however since no pre-trained models are available using this for the practical tests is beyond the scope of this thesis.

5.3 Proposed methods for validity analysis

For the purpose of validating a survey instrument, some methods that use word embeddings are developed and described in the following subsections. Each of these methods tries to exploit a certain aspect of the word embedding space in combination with the inspected survey items, in order to see whether some insights can be drawn from the results concerning the validity of items.

5.3.1 Nearest traits

In Chapter 3 it was shown that the Big Five personality traits are encoded in the structure of the word embedding space. Using the trait descriptive adjectives seen in Table 3.3, a trait pole can be generated by averaging all the adjectives belonging to each trait, which gives the set of the five trait poles $V^{BF} = \{v^e, v^a, v^c, v^n, v^o\} \subset \mathbb{R}^d$. Consider an item of a survey instrument, which has the corresponding averaged word vector v_i. Finding the closest trait v_i^x to this item in the word embedding space measured by the cosine similarity then corresponds to computing

$$v_i^x = \arg\min_{v \in V^{BF}} \{\mathrm{sim}_{\cos}(v, v_i)\}.$$

The reasoning behind this method is that an item probably belongs to the trait it lies closest to. By this rationale, this method may be seen as a way to evaluate the face validity of a questionnaire item.

Since for the testing purposes in the next chapter the correct trait of the items in the investigated survey instruments is known, it is possible to evaluate whether the classification of the item via the nearest trait in the word embedding space is correct. From this data, the $F\text{-}score$ for each trait can be computed by

$$F = \frac{2PR}{P + R},$$

where P is called $precision$, which is the number of correct item assignments to the trait divided by the total number of item assignments to the trait, and R is the $recall$, which is the number of correct assignments to the trait divided by the number of items that should have been assigned to the trait. Taken for the whole survey instrument, the computed F-scores are a measure for the correctness of the found nearest traits. The F-score takes values between 0 and 1, where 1 means everything is labeled correctly. The F-scores for the five traits can be averaged to get a combined F-score for the inventory.

5.3.2 Similar items

When there are items in an inventory that are very similar in meaning, they probably are measures for almost the same detail of a construct, so it may make sense to switch out items so that only one of the similar items remains. This could contribute to content validity, as it more evenly distributes the items throughout the semantic region of the construct in the word embedding space. The word vector space seems perfectly outfitted for identifying similar items, as it is straightforward to compute the pairwise cosine similarity of all items. The only question for this approach is where the threshold should be set for too similar items.

Table 5.1 Relevant cosine similarity thresholds for word embeddings of various dimensions, data from [RLH17, Table 1, p. 402].

	Thresholds		
Dimensions	Lower	Main	Upper
200	0.737	0.756	0.767
300	0.692	0.708	0.726

A recent publication used frequency analytic methods to derive such thresholds for the cosine similarity in word embedding spaces in dependence of the space dimension [RLH17]. Table 5.1 shows the relevant thresholds from [RLH17, Table 1, p. 402] for the word embeddings used in this thesis. For the evaluations in the next chapter, the upper thresholds are used, in order to allow for more similarity of items that belong to the same trait. So for the 200-dimensional word vectors from GVTwitter the threshold 0.767 and for the other word embeddings the threshold of 0.726 will be used. When evaluating an existing survey instrument, it makes sense to only check the similarity between items that are supposed to measure the same construct.

The results then provide sets of similar items, that need to be reevaluated and may have to be redesigned by the developers of the survey instrument.

5.3.3 Mean and variance of items

Content validity was described in Section 2.4.2 as a means to measure how well an instrument samples the construct domain it is supposed to analyze. In the context of the word embedding space this notion may be investigated in a rather literal way. For this purpose, the construct domain in the vector space must first be properly defined, and subsequently the position and distribution of the inventory items relative to this domain can be evaluated.

As seen in Section 3.2, the Big Five traits form relatively well separated clusters in the word embedding space that can be identified by the trait descriptive adjectives from Table 3.3. The proposed method computes the mean and standard deviation of the trait descriptive adjectives of each of the Big Five domains, to define the cluster positions and sizes in the word embedding space. Then the items of the inventory are grouped, placed in the word embedding space and also mean and standard deviation are computed.

From this data, d-dimensional hypercubes, where d is the dimension of the word embeddings, are constructed for the sets of trait descriptive adjectives and items from each trait by using in each dimension an interval of one standard deviation around the mean value. With this in mind it is now possible to compute the overlap of the two regions in the following sense:

Let μ^j and s^j be the mean and component-wise standard deviation vectors of the trait descriptive adjectives for trait j, and v^j, t^j the analogous vectors computed from the inventory items for trait j. Then the component-wise overlaps of trait j are computed by

$$C_i^j = \frac{\left|\left[\mu_i^j - s_i^j, \mu_i^j + s_i^j\right] \cap \left[v_i^j - t_i^j, v_i^j + t_i^j\right]\right|}{\left|\left[\mu_i^j - s_i^j, \mu_i^j + s_i^j\right] \cup \left[v_i^j - t_i^j, v_i^j + t_i^j\right]\right|},$$

and the overall overlap comes from averaging the component-wise overlaps, i.e.,

$$C^j = \frac{1}{d}\sum_{i=1}^{d} C_i^j.$$

This should give an indication of the sampling quality of the inventory in order to estimate the content validity.

5.3.4 Word embedding factor analysis using personality dimensions

In Section 3.1 the personality trait dimensions in the word embedding space were introduced. The trait dimensions can be generated, e.g., using the trait descriptive adjectives from Table 3.3. They can be used as a tool to further check the validity of the items of an inventory, by transforming the item vectors to the five dimensional personality subspace with the methods described in Section 2.2.4. This yields for each item a reduced vector with five components that represents the item's position on the personality trait axes.

The hypothesis is that the component entries of the reduced vectors can be interpreted in terms of the convergent and discriminant validity concepts, similar to factor loadings in a classical factor analysis. For example, an item for the extraversion trait should have a relatively high value in the component that corresponds to the extraversion trait axis, which could be seen as an aspect of convergent validity. In order to satisfy the requirements for discriminant validity on the other hand, the same item should have entries close to zero in the components corresponding to the other four trait dimensions. To evaluate this in terms of the whole survey instrument, the results for all items belonging to one trait can be averaged.

6 Application to personality survey instruments

The concept from the preceding section provides a new approach to validate survey instruments. Using some well known personality self-evaluation inventories, this chapter shows that the concept is more than just a theoretical construct and can be used to achieve results in practice. For this demonstration, an automated Python-script was written, which can be found in a git repository at https://gitlab.com/volker.kempf/validation_by_word_embeddings.

6.1 Word embedding models and inventories

For the proof of concept application the word embeddings from Table A.1 that were already used throughout the thesis are taken as a basis for the analysis of the instruments. Two of these embeddings are based on Wikipedia data, another two on general text data from the internet, one on news texts and one on Twitter data. The different origins of the semantic data for the embeddings may cause starkly differing results, as already seen in Section 3.2. In addition to the selected sets of word embeddings, several other pre-trained embeddings are available. However those listed in Table A.1 are roughly comparable with respect to the number of tokens in the base corpora, the vocabulary size and number of dimensions of the vectors.

For the comparison studies the personality inventories Big Five Inventory (BFI) from [JNS08, pp. 157f] and a revised version, the Big Five Inventory 2 (BFI2) from [SJ17, pp. 142f] are used. Additionally for some of the tests, items aimed at the Big Five traits from the International Personality Item Pool (IPIP), see [Gol22; Gol+06], are evaluated statistically.

The BFI contains a total of 44 items, ten for openness, nine each for agreeableness and conscientiousness and eight each for extraversion and neuroticism. Its revised form, the BFI2 contains 60 items, that are uniformly assigned to the five traits. In both tests, the answers to each item are given on a five point Likert-type scale with the choices $disagree\ strongly$, $disagree\ a\ little$, $neither\ agree\ nor\ disagree$, $agree\ a\ little$ and $agree\ strongly$, however this is not really important for the analyses conducted in this chapter.

Some of the items of the BFI, and 50% of the BFI2 items are reverse scored items, which means the answers to these items need to be inverted for the evaluation of the tests. For the intended analysis in this thesis these items also require special attention. For example for the analysis of the nearest trait of an item, it can of course not be expected that a reverse scored item is closest to the positive pole of the trait it belongs to, but rather closest to the negative pole.

The items of the IPIP on the other hand are not considered as part of a fixed inventory, but used to generate statistical data to compare the performance of the different word embeddings. The item pool offers a list of 3805 personality items that are assigned to various traits and facets, including the Big Five traits. In an initial step, out of these items those are selected that can be assigned to one of the Big Five traits by comparing the facet description in the list with the various facets of the Big Five, see, e.g., [SJ17, Table 1, p. 121]. After extraction of the relevant items by their facet label, assigning them to one of the Big Five traits and dismissing double entries, 672 items are left that are distributed over the Big Five traits as shown in Table 6.1.

Table 6.1 Distribution of IPIP items on the Big Five traits

Trait	No. items
Extraversion	147
Agreeableness	117
Conscientiousness	168
Neuroticism	101
Openness	139
Total	672

The next step is to get the inventory items to a format that is suitable for working with word embeddings, see Section 5.2. For the items from the BFI and BFI2, Tables A.4 and A.5 contain the forms of the items that are being used for the analysis. In the tables, the stop words that are removed in the first stage of pre-processing are written in italics. For the manual step of pre-processing, where each item is replaced by one key adjective, most of the items already contain an appropriate word that can be chosen, but for some items the chosen adjective is not contained in the item text itself.

6.2 Application of validity analysis methods to BFI, BFI2 and IPIP items

The several different approaches from Section 5.3 to confirm some type of validity of the items are now tested with the BFI and BFI2 survey instruments. Due to the well chosen items from these established questionnaires, good results for most of the methods can be expected.

6.2.1 Evaluation of nearest traits

The method from Section 5.3.1 to find the nearest traits was executed for the BFI, BFI2 and IPIP items on the basis of all the sets of pre-trained word embeddings described in Table A.1. The trait poles for the Big Five traits that are required for this approach were generated by averaging the word vectors from the ten trait descriptive adjectives seen in Table 3.3. For this first investigation the reverse-keyed items were neglected, so that, e.g., from the 60 items of the BFI2 only 30 are considered.

The resulting F-scores for the classification of the items are presented in Tables 6.2 and 6.3. From the data in the table it is clear that compared to the original wording, columns I1, removing the stop words in pre-processing, columns I2, results in a large improvement for almost all word embeddings except for the word2vec data set, where no difference in the average F-score is visible. On the other hand, the manual extraction of one descriptive adjective from the items, columns I3, improved the score in some instances and lowered the score in others, which means this step does not generate a significant benefit for this task.

Table 6.2 F-scores for classification of BFI items by nearest trait pole in the different stages of pre-processing. I1: items in the original wording; I2: items with removed stop words; I3: descriptive adjectives in place of whole items. Traits are abbreviated by their first letter.

	fTCrawl			fTWiki			GVCrawl			GVTwitter			GVWiki			w2vNews		
	I1	I2	I3	I1	I2	I3	I1	I2	I3	I1	I2	I3	I1	I2	I3	I1	I2	I3
E	0.73	0.83	0.91	0.60	0.91	1.00	0.75	0.89	1.00	0.75	0.91	0.83	0.67	0.75	0.91	0.67	0.60	1.00
A	0.83	1.00	0.80	0.91	1.00	0.80	0.83	0.91	0.80	0.83	0.91	0.89	0.67	0.91	0.80	0.83	0.80	0.91
C	0.75	1.00	0.89	0.83	1.00	0.80	0.80	0.89	0.80	0.89	0.89	0.89	0.77	0.91	0.67	0.80	0.91	0.89
N	0.83	0.91	1.00	0.91	0.91	1.00	1.00	1.00	1.00	0.77	0.91	1.00	1.00	1.00	1.00	0.81	0.91	1.00
O	0.77	0.77	1.00	0.67	0.86	1.00	1.00	0.94	1.00	0.86	0.86	1.00	0.93	1.00	1.00	0.86	0.86	1.00
Avg.	0.78	0.89	**0.93**	0.77	**0.93**	**0.93**	0.89	**0.93**	**0.93**	0.82	0.89	**0.93**	0.82	**0.92**	0.89	0.82	0.82	**0.96**

Table 6.3 F-scores for classification of BFI2 items in the stages of pre-processing by nearest trait pole. I1: items in the original wording; I2: items with removed stop words; I3: descriptive adjectives in place of whole items. Traits are abbreviated by their first letter.

	fTCrawl			fTWiki			GVCrawl			GVTwitter			GVWiki			w2vNews		
	I1	I2	I3	I1	I2	I3	I1	I2	I3	I1	I2	I3	I1	I2	I3	I1	I2	I3
E	1.00	1.00	1.00	0.80	1.00	0.92	0.80	0.91	1.00	0.80	0.80	0.92	0.80	0.80	1.00	0.80	0.80	1.00
A	0.91	1.00	0.86	0.83	0.91	0.92	0.92	0.86	0.86	0.63	0.92	0.91	0.83	0.77	0.86	0.83	0.83	0.86
C	0.91	1.00	0.80	0.75	0.92	0.91	0.83	0.91	0.80	0.29	0.91	0.71	0.71	0.77	0.80	0.77	0.77	0.80
N	0.80	0.92	1.00	0.92	0.92	1.00	1.00	1.00	1.00	0.75	0.80	0.83	1.00	1.00	1.00	0.92	0.92	1.00
O	0.91	0.91	0.83	0.67	0.91	0.91	0.92	1.00	0.83	0.91	0.91	0.60	0.83	1.00	0.83	0.83	0.83	0.83
Avg.	0.91	**0.97**	0.90	0.79	**0.93**	**0.93**	0.90	**0.94**	0.90	0.67	**0.87**	0.80	0.84	0.87	**0.90**	0.83	0.83	**0.90**

In order to get a better idea of what these F-score values mean, Figure 6.1 shows the confusion matrices for the results of this experiment with the fTCrawl word embeddings and the BFI2. From this visualization it is easy to see where the wrong classifications occurred. For example in the left matrix, where the result for the I1 stage of item pre-processing is shown, the column labeled “N” shows that all items belonging to the neuroticism trait were correctly classified as such. However the “N”-labeled row shows that three items in total were wrongly classified as belonging to neuroticism, while in truth they belong to the agreeableness, conscientiousness and openness traits.

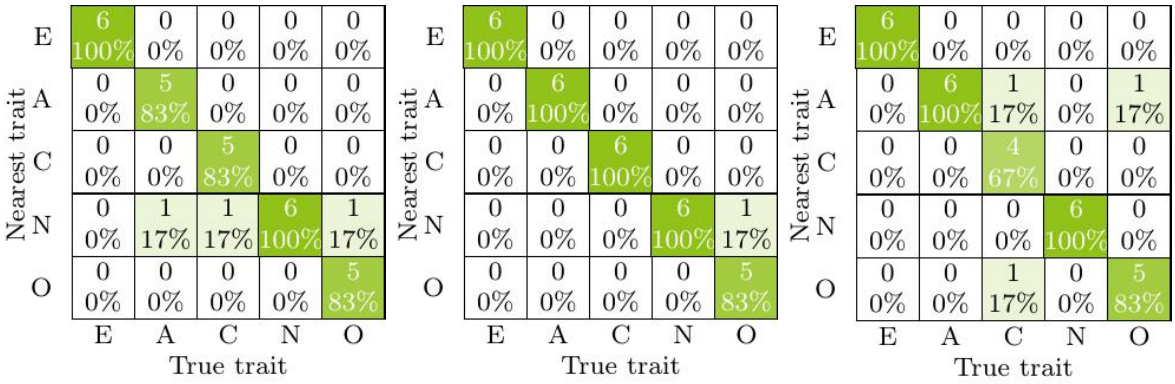

Figure 6.1 Confusion matrices of nearest trait classification of BFI2 items for fTCrawl word embeddings, cf. Table 6.3. Item pre-processing stages I1 on the left, I2 in the middle and I3 on the right. Traits are abbreviated by their first letter.

Overall, the results are very encouraging and show that the items from the BFI and BFI2 can to a large degree be mapped to their trait by this technique using word embeddings. For the BFI2 items in particular, very good results were achieved with the three sets of word embeddings, fTCrawl, fTWiki and GVCrawl, that showed the best cluster structure in Section 3.2.

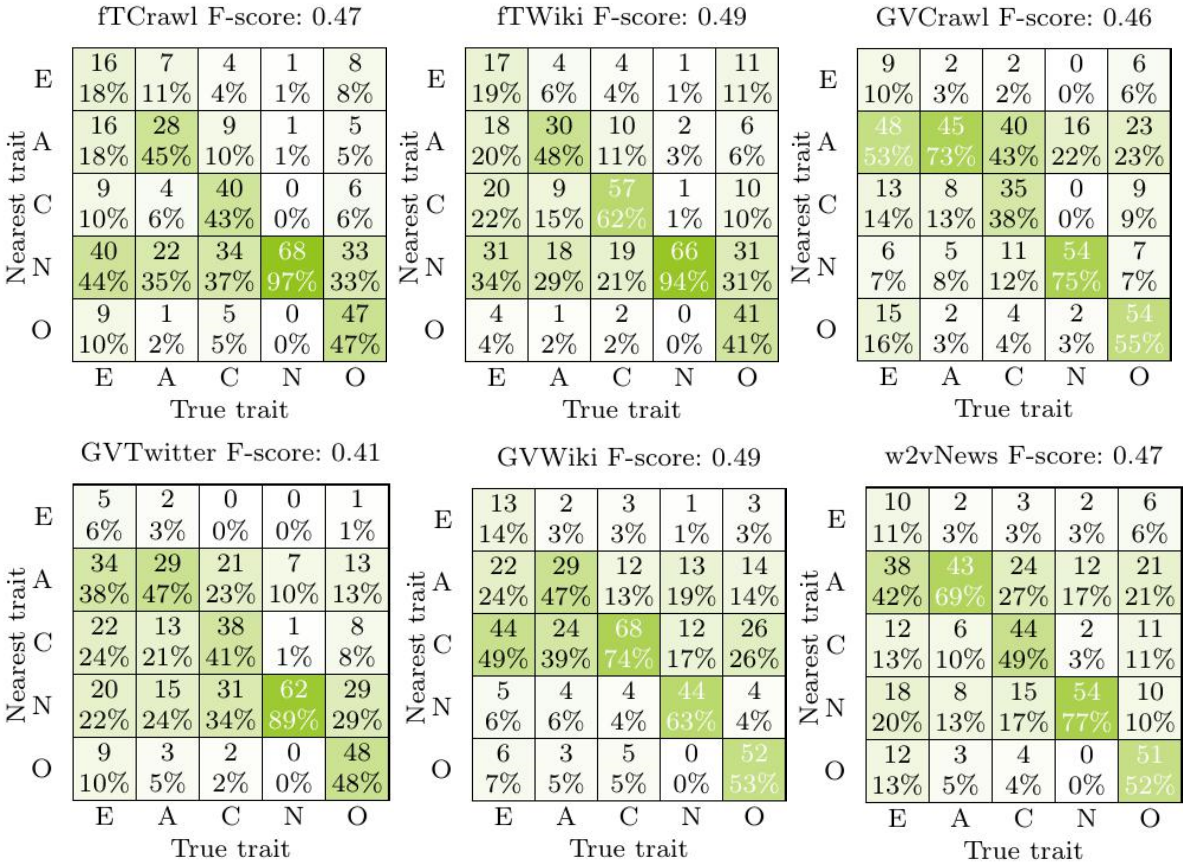

Figure 6.2 Confusion matrices of nearest trait classification of IPIP items in I2 stage of pre-processing. Traits are abbreviated by their first letter.

With the positively keyed items from the IPIP the analogous study was conducted for items in stage I2 of pre-processing. The resulting confusion matrices are presented in Figure 6.2. The results for these items are less clear than those for the BFI and BFI2.

Especially the items from the extraversion trait are not close to the corresponding trait pole, for all of the tested word embeddings. On the other hand the neuroticism items are classified very well in most embeddings.

6.2.2 Evaluation of similar items

For the evaluation of item similarity, the pairwise cosine similarity values of BFI and BFI2 items were computed individually for each trait and within the traits positive- and reverse-keyed items were considered separately. Thus in total the pairwise similarity was investigated within ten sets of items for each of the inventories and for each of the different word embeddings. While conducting the tests, it became obvious that the pre-processing stage I1 of the items, see Section 5.2, can not be used for this task, as the high number of identical stop words in the items in combination with the averaging of all word vectors of an item leads to the identification of too many similar item pairs. Thus the following results were produced with the pre-processing stage I2 of the items. Table 6.4 contains the results of these investigations and in order to keep the table as clear as possible the shorthand for the items givenin the “No.” column of Tables A.4 and A.5 is used.

Table 6.4 Similar items found in the BFI and BFI2 survey inventories, cf. Tables A.4 and A.5. w2vNews only contained pairs C4–C5 in BFI and C1–C2 in BFI2, GVWiki only O2–O3 in BFI2. Items that were above the similarity threshold in at least three of the word embedding sets are highlighted in bold.

	fTCrawl	fTWiki	GVCrawl	GVTwitter
BFI	E1–E8	–	–	–
	–	A2–A7	–	–
	C4–C5	**C4–C5**	**C4–C5**	–
	–	N2–N7	–	–
	–	O1–O2	–	–
	–	O1–O6	–	–
	O1–O8	**O1–O8**	–	**O1–O8**
	–	O2–O8	O2–O8	–
	–	O6–O10	–	–
BFI2	E1–E10	–	–	–
	–	E3–E4	–	–
	E6–E8	E6–E8	–	–
	C1–C2	**C1–C2**	**C1–C2**	–
	C3–C9	–	–	–
	–	C4–C7	–	–
	C4–C8	**C4–C8**	–	**C4–C8**
	–	C4–C11	–	–
	C7–C8	C7–C8	C7–C8	–
	–	N5–N6	–	–
	–	N5–N10	–	N5–N10
	N6–N10	**N6–N10**	**N6–N10**	–
	N8–N11	N8-N11	N8–N11	N8–N11
	O2–O3	**O2–O3**	**O2–O3**	**O2–O3**
	–	O2–O12	–	–
	O3–O12	O3–O12	–	–
	O4–O7	O4–O7	–	–
	–	O5–O11	–	–
	O6–O11	**O6–O11**	**O6–O11**	–

Of all the similar items, the three pairs C4–C5 from the BFI and C1–C2, O2–O3 from the BFI2 were identified as similar in at least four of the six word embedding sets. It can be noted that C4–C5 in the BFI and C1–C2 in the BFI2 are identical item pairs. As an example of the similarity, the items O2–O3 from the BFI2 read “is curious about many different things” and “is inventive, finds clever ways to do things”, where the stop words are written in italics. These items indeed seem to have a similar meaning.

In summary, this test showed that the approach of identifying similar items with word embeddings is promising. However, the different sets of pre-trained embeddings yield very different results, from identifying only one pair of similar items, see the results for GVWiki, to identifying maybe even too many, see the results for fTWiki. Since it is difficult to ascertain beforehand which model works best, it may be best to use the approach from this section to run the evaluation with different sets of word embeddings and only look at those pairs of similar items in detail that were identified by several models. This way the method can be a tool to find similar items, which can then be reevaluated and redesigned by the developers of the survey instruments.

6.2.3 Evaluation of mean and variance

For this method, at first the mean and standard deviation vectors of the trait descriptive adjectives from Tables 3.3 and A.3 and those from the inventory items are computed. Then the component-wise overlaps of the intervals defined by these vectors are computed and averaged. This again requires for the items of the BFI and BFI2 of each trait to be separated into positively and negatively keyed items, which gives a total of ten overlap values for each inventory. Subsequently, the values of the positive and negative items are averaged for each trait resulting in one value for each trait. In Table 6.5 the resulting values are presented. The values for the BFI2 are higher than those for the BFI with all the tested word embeddings, which means that the revision was an improvement measured in this metric.

Table 6.5 Average componentwise overlap of intervals of one standard deviation around mean of trait adjectives and inventory items

		fTCrawl	fTWiki	GVCrawl	GVTwitter	GVWiki	w2vNews
BFI	E	0.56	0.55	0.55	0.55	0.55	0.53
	A	0.55	0.55	0.54	0.57	0.55	0.55
	C	0.44	0.45	0.47	0.48	0.50	0.49
	N	0.48	0.48	0.48	0.49	0.49	0.50
	O	0.41	0.42	0.43	0.42	0.43	0.44
BFI2	E	0.56	0.56	0.57	0.57	0.57	0.55
	A	0.59	0.58	0.59	0.59	0.61	0.59
	C	0.50	0.50	0.51	0.54	0.54	0.53
	N	0.50	0.49	0.49	0.50	0.50	0.51
	O	0.44	0.45	0.46	0.44	0.47	0.47

On the other hand, it seems that the differences between the sets of word embeddings are only minimal, which is surprising when looking at the large differences in the results from the preceding tests. This may be an artifact of the averaging process, and may be an indication that the method could be refined and improved.

6.2.4 Word embedding factor analysis

For this analysis the required trait dimensions are constructed as described in Section 3.1. The trait descriptive adjectives seen in Table 3.3 are averaged to create the positive trait poles in the word embedding space, and similarly the adjectives from Table A.3 indicating low scores of the five traits are used to construct the negative poles. The following evaluation uses the technique presented in Section 2.2.4: The vectors representing the trait dimensions are written in a matrix that is then inverted, which gives the transformation matrix with which the vectors of the BFI and BFI2 items can be transformed into the five dimensional personality space spanned by the trait dimensions. All transformed item vectors of items belonging to a trait are then averaged in order to get an overview of the results. For the experiments only the positively keyed items in pre-processing stage I2 of the instruments are considered, since averaging with the negatively keyed items is counter-productive as they are expected to cancel each other out. Another post-processing step is taken by normalizing each axis. To get the normalizing factors, the positive and negative trait poles are mapped to the personality space using the transformation matrix. This gives a value for each pole in the corresponding trait dimension, which is used to set up a linear transformation which maps the value of the positive pole to the value 1 in the trait dimension and the value of the negative pole to −1. This is done due to the hypothesis that the poles are the extreme ends of the trait dimensions, and the other vectors are situated somewhere in between. The normalization also allows for a certain comparability between the five trait dimensions. The results of this analysis are presented in Figures 6.3 and 6.4, where the averaged values of the items belonging to a trait have the same color as the poles of this trait. These results show that in all word embeddings the expected trend can be seen to a certain extent: The items belonging to a trait indeed have the highest value on the corresponding trait dimension, which means that the convergent validity for most cases might reasonably be seen as satisfied. On the other hand, some of the curves show a substantial semantic cross-loading on the other personality dimensions, which could speak against discriminant validity. While the general trend is similar in all the tested word embeddings, there is some variance in the details. The w2vNews word embeddings seems to show the best result with respect to the discriminant validity criterion, i.e., the cross-loading is least pronounced here. In all tested word embeddings, most of the data curves reach the value of 0.4 on the associated personality dimension, which is seen in classical factor analysis as the required value for convergent validity.

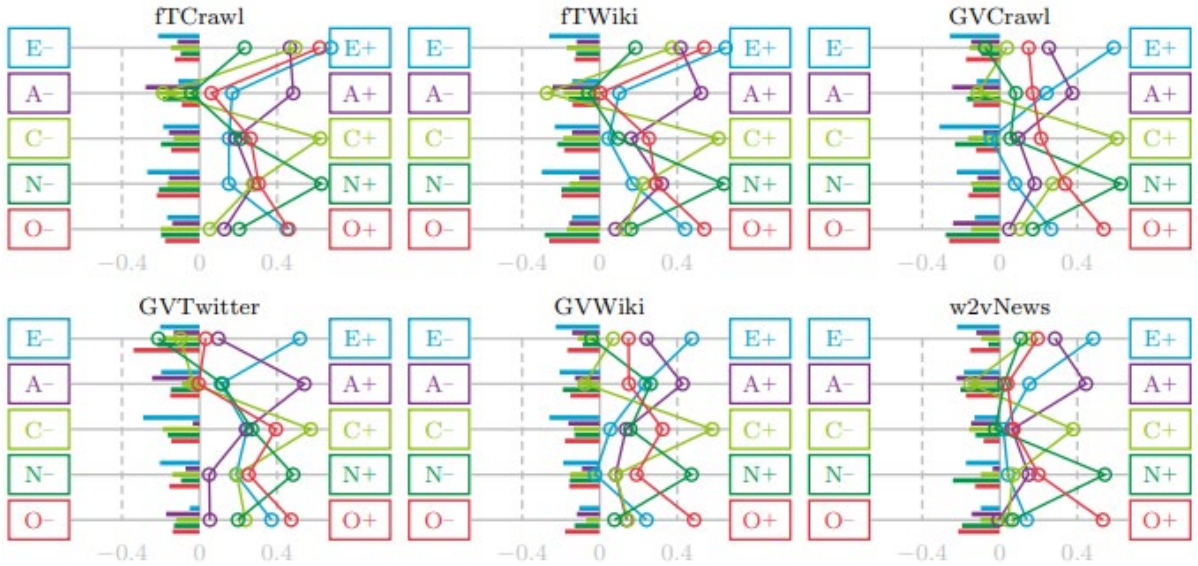

Figure 6.3: Average normalized coordinates of positively keyed BFI items of each trait in the personality dimensions. Colorbars indicate values of standard deviation of the averaged data points.

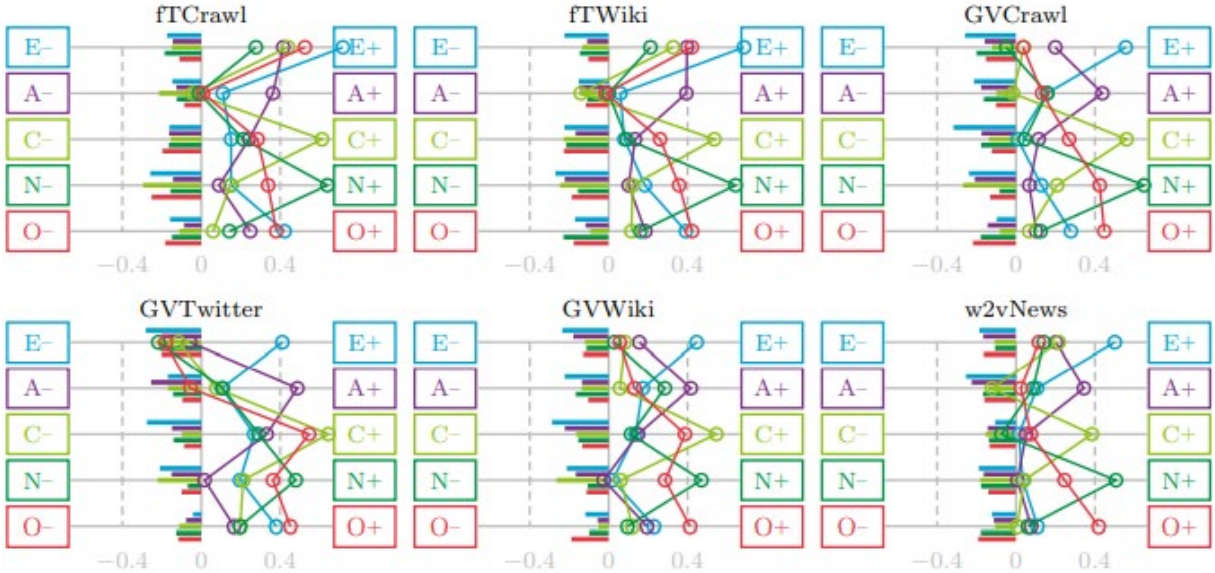

Figure 10.4 Average normalized coordinates of positively keyed BFI2 items of each trait in the personality dimensions. Colorbars indicate values of standard deviation of the averaged data points.

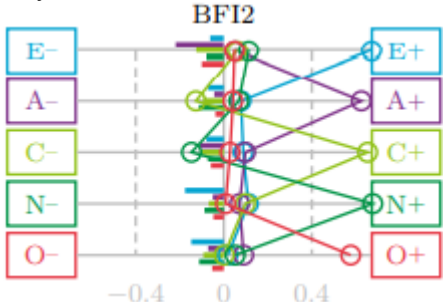

Figure 6.5 Averages of factor loadings of positively keyed BFI2 items from [SJ17, Table 6, pp. 14f]. Colorbars indicate values of standard deviation of the averaged data points.

For comparison, the factor loading averages of the positively keyed items from the BFI2 from a classical factor analysis with varimax rotation are given in Figure 6.5 [SJ17, Table 6, pp. 14f]. Clearly this data contains a lot less noise and the discriminant and convergent validity is easily recognizable. However, the approach using only generic word embeddings without any adaption to the specific field of personality psychology can reproduce the general trend, which is encouraging for further investigations.

7 Conclusion and outlook

Personnel selection procedures for hiring as well as promotion purposes and internal selection often employ some sort of personality evaluation of the candidates. The presented research supports that this practice is common and justified, as personality aspects have an evident influence on job performance on the individual level, but also in the context of work teams, where team personality composition, like, e.g., the mean score of a trait, can impact the team performance significantly. Thus, organizations benefit from the use of accurate and reliable personality tests for these evaluations, where accuracy can be seen as the validity of the survey instrument.

The investigation of several word embedding spaces from different sources using a clustering approach has shown that these spaces contain the structure of the Big Five personality traits to an extent that makes them a useful tool for the analysis of questionnaire items. Based on this nice finding, a concept with several approaches to validate a personality survey instrument was developed. The first step of the concept is a pre-processing stage for the items to facilitate the subsequent natural language processing, where stop words are removed from the items' texts. Then the concept proposes four methods to evaluate the face, content and construct validity of individual items or the instrument as a total.

Using two established personality survey instruments for the Big Five traits, the concept was tested to ascertain its applicability. Since these questionnaires were already validated by standard methods, good results could be expected. Indeed, in all of the experiments the methods were able to generate results that are in line with their proposed use concerning the validity evidence.

Summed up, this thesis provides a first promising indication that the research question, whether it is possible to validate survey instruments on the basis of word embeddings, can be answered positively. The most promising approach of the four presented methods seems to be the semantic factor analysis, which can be seen as an analogy to the classical statistical factor analysis using sample applications of a newly developed survey instrument.

Limitations of the approach

Despite this promising insight, there are some limitations of the conducted studies that need to be addressed. Firstly, word embeddings can only reflect the word semantics that are present in the text corpus on which they were trained. Consider for example the GVTwitter word embeddings, which are based on Twitter data that consists only of phrases with a limited number of characters that also often have wrong grammar and spelling. Intuitively those embeddings internalize less semantic information than word embeddings based on scientific literature, and the test computations in the thesis showed that they indeed perform worse in most of the tasks concerning personality traits. The GVTwitter embeddings strongly show this inadequacy, however none of the used word embeddings in this thesis is trained on a text corpus specifically curated for psychological evaluations, so the results presented in Chapter 6 might be improved when a set of word embeddings specialized for psychological application is used.

Furthermore, the standard model of word embeddings can not capture polysemy of words, i.e., when a word has different meanings for the same spelling. The training process of the word embedding algorithms can not differentiate the separate meanings, and thus produces a word vector that can be seen as a sort of weighted average of the different meanings. This can lead to inaccuracies when the word vectors of words with multiple meanings are used.

Extensions and future research directions

Directly linked to the first mentioned limitation of this thesis, a first extension that may improve the results from Chapter 6 significantly is the use of a custom set of word embeddings that is trained on psychological literature. To this end, a large set of psychological texts needs to be compiled to train word embedding models on. From the experiences with the pre-trained models in this thesis, this collection of texts should comprise a few billion words at least, which means that an automated process to gather these texts must be developed, possible sources are scientific journals in the psychology field.

Another possible extension of the presented methods is additional and custom natural language pre-processing of the survey items. For now only stop words were removed, but other often used pre-processing techniques like lemmatization or stemming could be applied additionally. These methods remove conjugation and declension from words and bring them to a standard, non-inflected form. By reducing words in this way, ambiguity might be removed and words are reduced to their pure form, which may improve the performance of methods based on word embeddings.

References

[All37] G. W. Allport. Personality: a psychological interpretation. New York: Holt, 1937. url: https://lccn.loc.gov/37025297.

[AO36] G. W. Allport and H. S. Odbert. “Trait-names: A psycho-lexical study.” Psychol. Monogr. 47.1 (1936), i–171. doi: 10.1037/h0093360.

[AAN14] American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, eds. Standards for educational and psychological testing. Washington, DC: American Educational Research Association, 2014. url: https://lccn.loc.gov/2014009333.

[AS14] C. Ayre and A. J. Scally. “Critical Values for Lawshe’s Content Validity Ratio”. Meas. Eval. Couns. Dev. 47.1 (2014), 79–86. doi: 10.1177/0748175613513808.

[BW08] N. B. Barenbaum and D. G. Winter. “History of Modern Personality theory and research”. In: Handbook of Personality: Theory and Research. Ed. by O. P. John, R. W. Robins, and L. A. Pervin. 3rd ed. Guilford Publications, 2008, 3–26. url: https://lccn.loc.gov/2008006659.

[BM91] M. R. Barrick and M. K. Mount. “The Big Five personality dimensions and job performance: A meta-analysis.” Pers. Psychol. 44.1 (1991), 1–26. doi: 10.1111/j.1744-6570.1991.tb00688.x.

[Bel07] S. T. Bell. “Deep-level composition variables as predictors of team performance: A meta-analysis.” J. Appl. Psychol. 92.3 (2007), 595–615. doi: 10.1037/0021-9010.92.3.595.

[Bla18] K. M. Blackford. Reading character at sight. New York: Independent corporation, 1918. url: https://lccn.loc.gov/19011868.

[BGJM17] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. “Enriching Word Vectors with Subword Information”. Transactions of the Association for Computational Linguistics 5 (2017), 135–146. doi: 10.1162/tacl_a_00051.

[Boy95] G. J. Boyle. “Myers–Briggs Type Indicator (MBTI): Some Psychometric Limi ta-tions”. Aust. Psychol. 30.1 (1995), 71–74. doi: 10 . 1111 / j . 1742 - 9544.1995.tb01750.x .

[BMQH98]I. Briggs Myers, M. H. McCaulley, N. L. Quenk, and A. L. Hammer. MBTI manual : a guide to the development and use of the Myers–Briggs Type Indicator. Palo Alto, CA: Consulting Psychologists Press, 1998.

[CW14] E. Cambria and B. White. “Jumping NLP Curves: A Review of Natural Language Processing Research”. IEEE Comput. Intell. Mag. 9.2 (2014), 48–57. doi: 10.1109/mci.2014.2307227.

[Car77] M. Carlyn. “An Assessment of the Myers-Briggs Type Indicator”. J. Pers. Assess. 41.5 (1977), 461–473. doi: 10.1207/s15327752jpa4105_2.

[CZ79] E. Carmines and R. Zeller. Reliability and Validity Assessment. Thousand Oaks, CA: SAGE Publications, 1979. doi: 10.4135/9781412985642.

[CHLS13] N. D. Christiansen, B. J. Hoffman, F. Lievens, and A. B. Speer. “Assessment Centers and the Measurement of Personality”. In: Handbook of Personality at Work. Ed. by N. D. Christiansen and R. P. Tett. Routledge, 2013, 477–497. doi: 10.4324/9780203526910.ch21.

[CMHE04] F. Coffield, D. Moseley, E. Hall, and K. Ecclestone. Learning styles and pedagogy in post-16 learning: a systematic and critical review. London: Learning and Skills Research Centre, 2004.

[CW08] R. Collobert and J. Weston. “A unified architecture for natural language processing”. In: Proceedings of the 25th international conference on Machine learning - ICML ’08. ACM Press, 2008. doi: 10 . 1145 / 1390156 . 1390177.

[CS10] F. L. Coolidge and D. L. Segal. “Validity”. In: The Corsini encyclopedia of psychology. Ed. by I. B. Weiner and W. E. Craighead. John Wiley & Sons, 2010. doi: 10.1002/9780470479216.corpsy1019.

[Cra93] K. H. Craik. “The 1937 Allport and Stagner Texts in Personality Psychology”. In: Fifty Years of Personality Psychology. Springer, 1993, 3–20. doi: 10.1007/ 978-1-4899-2311-0_1.

[Cro90] L. J. Cronbach. Essentials of psychological testing. 5th ed. New York, NJ: Harper & Row, 1990. url: https://lccn.loc.gov/89027772.

[Dig90] J. M. Digman. “Personality Structure: Emergence of the Five-Factor Model”. Annu. Rev. Psychol. 41.1 (1990), 417–440. doi: 10.1146/annurev.ps.41. 020190.002221.

[DS13] J. E. Driskell and E. Salas. “Personality and Work Teams”. In: Handbook of Personality at Work. Ed. by N. D. Christiansen and R. P. Tett. Routledge, 2013, 744–771. doi: 10.4324/9780203526910.ch33.

[DB91] D. Druckman and R. A. Bjork, eds. In the mind's eye: enhancing human performance. Washington, DC: National Academy Press, 1991. url: https: //lccn.loc.gov/91023941.

[Fac20] Facebook Inc. fastText. 2020. url: https://fasttext.cc/ (visited on 2022- 01-06).

[Fra02] M. D. Franzen. Reliability and Validity in Neuropsychological Assessment. 2nd ed. New York: Springer, 2002. doi: 10.1007/978-1-4757-3224-5.

[Fur18] A. Furnham. "Personality and Occupational Success". In: The SAGE Handbook of Personality and Individual Differences: Volume III: Applications of Personality and Individual Differences. SAGE Publications, 2018. doi: 10. 4135/9781526451248.n23.

[GM96] W. L. Gardner and M. J. Martinko. "Using the Myers–Briggs Type Indicator to Study Managers: A Literature Review and Research Agenda". J. Manage. 22.1 (1996), 45–83. doi: 10.1177/014920639602200103.

[GZ08] R. E. Gibby and M. J. Zickar. "A history of the early days of personality testing in American industry: An obsession with adjustment." Hist. Psychol. 11.3 (2008), 164–184. doi: 10.1037/a0013041.

[Gol22] L. Goldberg. International Personality Item Pool: A Scientific Collaboratory for the Development of Advanced Measures of Personality Traits and Other Individual Differences. 2022. url: https://ipip.ori.org/ (visited on 2022- 02-21).

[Gol81] L. R. Goldberg. "Language and Individual Differences: The Search for Universals in Personality Lexicons". In: Review of Personality and Social Psychology. Ed. by L. Wheeler. Vol. 2. SAGE, 1981, 151–165.

[Gol90] L. R. Goldberg. "An alternative "description of personality": The Big-Five factor structure." J. Pers. Soc. Psychol. 59.6 (1990), 1216–1229. doi: 10.1037/ 0022-3514.59.6.1216.

[Gol92] L. R. Goldberg. "The development of markers for the Big-Five factor structure". Psychol. Assessment 4.1 (1992), 26–42. doi: 10.1037/1040- 3590.4.1.26.

[Gol+06] L. R. Goldberg, J. A. Johnson, H. W. Eber, R. Hogan, M. C. Ashton, C. R. Cloninger, and H. G. Gough. "The international personality item pool and the future of public-domain personality measures". J. Res. Pers. 40.1 (2006), 84–96. doi: 10.1016/j.jrp.2005.08.007.

[Goo13] Google Inc. word2vec. 2013. url: https://code.google.com/archive/p/ word2vec/ (visited on 2022-01-17).

[GM06] R. L. Griffith and M. McDaniel. “The nature of deception and applicant faking behavior”. In: A closer examination of applicant faking behavior. Ed. by R. L. Griffith and M. H. Peterson. Information Age Publishing, 2006, 1–19. url: https://lccn.loc.gov/2006007218.

[GR13] R. L. Griffith and C. Robie. “Personality Testing and the “F-Word””. In: Handbook of Personality at Work. Ed. by N. D. Christiansen and R. P. Tett. Routledge, 2013, 253–280. doi: 10.4324/9780203526910.ch12.

[Gui98] R. M. Guion. Assessment, Measurement, and Prediction for Personnel Decisions. London: Lawrence Erlbaum Associates, 1998. url: https://lccn.loc. gov/96036765.

[HBBA18] J. Hair, W. Black, B. Babin, and R. Anderson. Multivariate Data Analysis. 8th ed. Andover: Cengage, 2018. url: https://lccn.loc.gov/2019301366.

[Hal+05] T. Halfhill, E. Sundstrom, J. Lahner, W. Calderone, and T. M. Nielsen. “Group Personality Composition and Group Effectiveness”. Small. Gr. Res. 36.1 (2005), 83–105. doi: 10.1177/1046496404268538.

[HDT04] J. P. Hausknecht, D. V. Day, and S. C. Thomas. “Applicant Reactions to Selection Procedures: An Updated Model and Meta-Analysis”. Pers. Psychol. 57.3 (2004), 639–683. doi: 10.1111/j.1744- 6570.2004.00003.x.

[HD12] B. J. Hoffman and S. Dilchert. “A Review of Citizenship and Counterproductive Behaviors in Organizational Decision-Making”. In: The Oxford Handbook of Personnel Assessment and Selection. Ed. by N. Schmitt. Oxford University Press, 2012, 543–569. doi: 10.1093/oxfordhb/9780199732579.013.0024.

[Hol10] R. R. Holden. “Face Validity”. In: The Corsini encyclopedia of psychology. Ed. by I. B. Weiner and W. E. Craighead. John Wiley & Sons, 2010. doi: 10.1002/9780470479216.corpsy0341.

[HJ79] R. R. Holden and D. N. Jackson. “Item subtlety and face validity in personality assessment”. J. Consult. Clin. Psychol. 47.3 (1979), 459–468. doi: 10.1037/ 0022-006x.47.3.459.

[HOO15] L. M. Hough, F. L. Oswald, and J. Ock. “Beyond the Big Five: New Directions for Personality Research and Practice in Organizations”. Annual Review of Organizational Psychology and Organizational Behavior 2.1 (2015), 183–209. doi: 10.1146/annurev-orgpsych-032414-111441.

[Jac04] S. F. Jacobson. “Evaluating Instruments for Use in Clinical Nursing Research”. In: Instruments for clinical health-care research. Ed. by M. Frank-Stromborg and S. J. Olsen. 3rd ed. Jones & Bartlett Learning, 2004, 3–19. url: https: //lccn.loc.gov/2003015429.

[Joh21] O. P. John. “History, Measurement, and Conceptual Elaboration of the Big-Five Trait Taxonomy”. In: Handbook of Personality: Theory and Research. Ed. by O. P. John and R. W. Robins. 4th ed. Guilford Publications, 2021, 35–82. url: https://lccn.loc.gov/2020042618.

[JNS08] O. P. John, L. P. Naumann, and C. J. Soto. “Paradigm shift to the integrative big five trait taxonomy”. In: Handbook of Personality: Theory and Research. Ed. by O. P. John, R. W. Robins, and L. A. Pervin. 3rd ed. Guilford Publications, 2008, 114–158. url: https://lccn.loc.gov/2008006659.

[JL07] T. A. Judge and J. A. LePine. “The Bright and Dark Sides of Personality: Implications for Personnel Selection in Individual and Team Contexts”. In: Research Companion to the Dysfunctional Workplace. Edward Elgar Publishing, 2007. doi: 10.4337/9781847207081.00028.

[Jud+13] T. A. Judge, J. B. Rodell, R. L. Klinger, L. S. Simon, and E. R. Crawford. “Hierarchical representations of the five-factor model of personality in predicting job performance: Integrating three organizing frameworks with two theoretical perspectives.” J. Appl. Psychol. 98.6 (2013), 875–925. doi: 10.1037/a0033901.

[Jun71] C. G. Jung. Collected Works of C.G. Jung, Volume 6: Psychological Types. Ed. by G. Adler and R. F. Hull. Princeton University Press, 1971. doi: 10. 1515/9781400850860.

[KS17] R. M. Kaplan and D. P. Saccuzzo. Psychological Testing. Boston, MA: Cengage Learning, 2017. url: https://lccn.loc.gov/2016948139.

[KM20] S. P. King and B. A. Mason. “Myers-Briggs Type Indicator”. In: The Wiley Encyclopedia of Personality and Individual Differences: Measurement and Assessment. Ed. by B. J. Carducci, C. S. Nave, J. S. Mio, and R. E. Riggio. Wiley, 2020, 315–319. doi: 10.1002/9781119547167.ch123.

[Law75] C. H. Lawshe. “A quantitative approach to content validity”. Pers. Psychol. 28.4 (1975), 563–575. doi: 10.1111/j.1744-6570.1975.tb01393.x.

[LM14] Q. Le and T. Mikolov. “Distributed Representations of Sentences and Documents”. In: Proceedings of the 31st International Conference on Machine Learning. Ed. by E. P. Xing

and T. Jebara. Vol. 32. Proceedings of Machine Learning Research 2. Bejing, China: PMLR, 2014, 1188–1196. url: https: //proceedings.mlr.press/v32/le14.html.

[Len08] A. Lenci. “Distributional semantics in linguistic and cognitive research”. Rivista di Linguistica 20.1 (2008), 1–31. url: https : / / www . italian - journal - linguistics.com/app/uploads/2021/05/1_Lenci.pdf.

[LeP03] J. A. LePine. “Team adaptation and postchange performance: Effects of team composition in terms of members’ cognitive ability and personality”. J. Appl. Psychol. 88.1 (2003), 27–39. doi: 10.1037/0021-9010.88.1.27.

[LMC09] J. Levashina, F. P. Morgeson, and M. A. Campion. “They Don’t Do It Often, But They Do It Well: Exploring the relationship between applicant mental abilities and faking”. Int. J. Select. Assess. 17.3 (2009), 271–281. doi: 10. 1111/j.1468-2389.2009.00469.x.

[Loe57] J. Loevinger. “Objective Tests as Instruments of Psychological Theory”. Psychol. Rep. 3.3 (1957), 635–694. doi: 10.2466/pr0.1957.3.3.635.

[MSLS20] B. Mathew, S. Sikdar, F. Lemmerich, and M. Strohmaier. “The POLAR Framework: Polar Opposites Enable Interpretability of Pre-Trained Word Embeddings”. In: Proceedings of The Web Conference 2020. 2020, 1548–1558. doi: 10.1145/3366423.3380227.

[McC00] M. H. McCaulley. “Myers-Briggs Type Indicator: A bridge between counseling and consulting.” Consulting Psychology Journal: Practice and Research 52.2 (2000), 117–132. doi: 10.1037/1061-4087.52.2.117.

[McF03] L. A. McFarland. “Warning Against Faking on a Personality Test: Effects on Applicant Reactions and Personality Test Scores”. Int. J. Select. Assess. 11.4 (2003), 265–276. doi: 10.1111/j.0965-075x.2003.00250.x.

[McF13] L. A. McFarland. “Applicant Reactions to Personality Tests”. In: Handbook of Personality at Work. Ed. by N. D. Christiansen and R. P. Tett. Routledge, 2013, 281–298. doi: 10.4324/9780203526910.ch13.

[MR00] L. A. McFarland and A. M. Ryan. “Variance in faking across noncognitive measures.” J. Appl. Psychol. 85.5 (2000), 812–821. doi: 10 . 1037 / 0021 - 9010.85.5.812.

[Mes87] S. Messick. “Validity”. ETS Research Report Series 1987.2 (1987), i–208. doi: 10.1002/j.2330-8516.1987.tb00244.x.

[MCCD13] T. Mikolov, K. Chen, G. Corrado, and J. Dean. “Efficient Estimation of Word Representations in Vector Space” (2013).

[MSA17] S. Moscoso, J. F. Salgado, and N. Anderson. “How Do I Get a Job, What Are They Looking For? Personnel Selection and Assessment”. In: An Introduction to Work and Organizational Psychology. John Wiley & Sons, Ltd, 2017, 25–47. doi: 10.1002/9781119168058.ch2.

[Mur90] J. B. Murray. “Review of Research on the Myers-Briggs Type Indicator”. Percept. Mot. Skills 70.3 (1990), 1187–1202. doi: 10.2466/pms.1990.70.3c. 1187 (cit. on p. 17). [Mye22] Myers–Briggs Company. The Myers–Briggs Company. 2022. url: https : //eu.themyersbriggs.com/ (visited on 2022-01-28).

[NS14] P. E. Newton and S. D. Shaw. Validity in Educational & Psychological Assessment. London: SAGE Publications, 2014. 280 pp. url: https://lccn.loc. gov/2013946019.

[PTRR06] M. A. G. Peeters, H. F. J. M. van Tuijl, C. G. Rutte, and I. M. M. J. Reymen. “Personality and team performance: a meta-analysis”. Eur. J. Personality 20.5 (2006), 377–396. doi: 10.1002/per.588.

[PSM14a] J. Pennington, R. Socher, and C. Manning. “Glove: Global Vectors for Word Representation”. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, 2014, 1532–1543. doi: 10.3115/v1/d14-1162.

[PSM14b] J. Pennington, R. Socher, and C. D. Manning. GloVe: Global Vectors for Word Representation. 2014. url: https://nlp.stanford.edu/projects/glove/ (visited on 2021-12-20).

[Pit05] D. J. Pittenger. “Cautionary comments regarding the Myers-Briggs Type Indicator.” Consulting Psychology Journal: Practice and Research 57.3 (2005), 210–221. doi: 10.1037/1065-9293.57.3.210.

[PTC13] M. S. Prewett, R. P. Tett, and N. D. Christiansen. “A Review and Comparison of 12 Personality Inventories on Key Psychometric Characteristics”. In: Handbook of Personality at Work. Ed. by N. D. Christiansen and R. P. Tett. Routledge, 2013, 191–225. doi: 10.4324/9780203526910.ch10.

[Pre+09] M. S. Prewett, A. A. G. Walvoord, F. R. B. Stilson, M. E. Rossi, and M. T. Brannick. “The Team Personality–Team Performance Relationship Revisited: The Impact of Criterion Choice, Pattern of Workflow, and Method of Aggregation”. Hum. Perform. 22.4 (2009), 273–296. doi: 10.1080/08959280903120253 .

[RI13] P. H. Raymark and C. H. V. Iddekinge. “Assessing Personality in Selection Interviews”. In: Handbook of Personality at Work. Ed. by N. D. Christiansen and R. P. Tett. Routledge, 2013, 419–438. doi: 10.4324/9780203526910.ch18.

[RLH17] N. Rekabsaz, M. Lupu, and A. Hanbury. “Exploration of a Threshold for Similarity Based on Uncertainty in Word Embedding”. In: Advances in Information Retrieval (ECIR2017). Ed. by J. M. Jose, C. Hauff, I. S. Altıngovde, D. Song, D. Albakour, S. Watt, and J. Tait. Springer International Publishing, 2017, 396–409. doi: 10.1007/978-3-319-56608-5_31.

[RG06] M. G. Rothstein and R. D. Goffin. “The use of personality measures in personnel selection: What does current research support?” Hum. Resour. Manage. R. 16.2 (2006), 155–180. doi: 10.1016/j.hrmr.2006.03.004.

[Roz20] D. Rozado. “Wide range screening of algorithmic bias in word embedding models using large sentiment lexicons reveals underreported bias types”. PLOS One 15.4 (2020). Ed. by C. Schwieren, e0231189. doi: 10.1371/journal.pone. 0231189.

[Sah08] M. Sahlgren. “The distributional hypothesis”. Rivista di Linguistica 20.1 (2008), 33–53. url: https://www.italian-journal-linguistics.com/app/ uploads/2021/05/2_Sahlgren-1.pdf.

[Sal02] J. F. Salgado. “The Big Five Personality Dimensions and Counterproductive Behaviors”. Int. J. Select. Assess. 10.1&2 (2002), 117–125. doi: 10.1111/1468- 2389.00198.

[SAM20] J. F. Salgado, N. Anderson, and S. Moscoso. “Personality at Work”. In: The Cambridge Handbook of Personality Psychology. Cambridge University Press, 2020, 427–438. doi: 10.1017/9781108264822.040.

[SO99] G. Saucier and F. Ostendorf. “Hierarchical subcomponents of the Big Five personality factors: A cross-language replication.” J. Pers. Soc. Psychol. 76.4 (1999), 613–627. doi: 10.1037/0022-3514.76.4.613.

[SJ09] C. J. Soto and O. P. John. “Using the California Psychological Inventory to assess the Big Five personality domains: A hierarchical approach”. J. Res. Pers. 43.1 (2009), 25–38. doi: 10.1016/j.jrp.2008.10.005.

[SJ17] C. J. Soto and O. P. John. “The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power.” J. Pers. Soc. Psychol. 113.1 (2017), 117–143. doi: 10.1037/pspp0000096.

[Sta37] R. Stagner. Psychology of personality. New York: McGraw-Hill, 1937. url: https://lccn.loc.gov/37020553.

[SBG04] D. Straub, M.-C. Boudreau, and D. Gefen. “Validation Guidelines for IS Positivist Research”. Communications of the Association for Information Systems 13 (2004), Article 24. doi: 10.17705/1cais.01324.

[Swi21] V. Swift. “Validating Word Embedding as a Tool for the Psychological Sciences”. PhD thesis. University of Toronto, 2021. url: https://hdl.handle.net/ 1807/104927.

[Tah16] H. Taherdoost. “Validity and Reliability of the Research Instrument; How to Test the Validation of a Questionnaire/Survey in a Research”. SSRN Electronic Journal (2016). doi: 10.2139/ssrn.3205040.

[Tet+06] R. P. Tett, M. G. Anderson, C. Ho, T. S. Yang, L. Huang, and A. Hanvongse. “Seven nested questions about faking on personality tests: An overview and interactionist model of item-level response distortion.” In: A closer examination of applicant faking behavior. Ed. by R. L. Griffith and M. H. Peterson. Greenwich, CT: Information Age Publishing, 2006, 43–83. url: https: //lccn.loc.gov/2006007218.

[Tet13] R. P. Tett. “Personality Psychology in the Workplace”. In: Handbook of Personality at Work. Ed. by N. D. Christiansen and R. P. Tett. Routledge, 2013, 849–862. doi: 10.4324/9780203526910.ch38.

[VEB10] N. X. Vinh, J. Epps, and J. Bailey. “Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance”. J. Mach. Learn. Res. 11.95 (2010), 2837–2854. url: https : //jmlr.org/papers/v11/vinh10a.html.

[VO00] C. Viswesvaran and D. S. Ones. “Perspectives on Models of Job Performance”. Int. J. Select. Assess. 8.4 (2000), 216–226. doi: 10.1111/1468-2389.00151.

[WB98] M. Waung and T. S. Brice. “The Effects of Conscientiousness and Opportunity to Caucus on Group Performance”. Small. Gr. Res. 29.5 (1998), 624–634. doi: 10.1177/1046496498295005.

[Wig88] J. S. Wiggins. Personality and prediction: principles of personality assessment. Malabar, FL: Krieger Pub. Co., 1988. url: https://lccn.loc.gov/87017348.

[WPS12] F. R. Wilson, W. Pan, and D. A. Schumsky. “Recalculation of the Critical Values for Lawshe’s Content Validity Ratio”. Meas. Eval. Couns. Dev. 45.3 (2012), 197–210. doi: 10.1177/0748175612440286.

[ZHZ20] L. Zhu, Y. He, and D. Zhou. “A Neural Generative Model for Joint Learning Topics and Topic-Specific Word Embeddings”. Transactions of the Association for Computational Linguistics 8 (2020), 471–485. doi: 10.1162/tacl_a_00326.

[Zic01] M. J. Zickar. “Using Personality Inventories to Identify Thugs and Agitators: Applied Psychology’s Contribution to the War against Labor”. J. Vocat. Behav. 59.1 (2001), 149–164. doi: 10.1006/jvbe.2000.1775.

[ZK13] M. J. Zickar and J. A. Kostek. “History of Personality Testing Within Organizations”. In: Handbook of Personality at Work. Ed. by N. D. Christiansen and R. P. Tett. Routledge, 2013, 173–190. doi: 10.4324/9780203526910.ch9.

A Additional tables

Table A.1: Sets of pre-trained word embeddings used in the thesis.

ID	Algorithm	Base data source	Tokens	Vocab. size	Dim.	Source
fTCrawl	fastText	Common crawl	$6.0 \cdot 10^{11}$	$2.0 \cdot 10^{6}$	300	[Fac20]
fTWiki	fastText	Wikipedia, news	$1.6 \cdot 10^{10}$	$1.0 \cdot 10^{6}$	300	[Fac20]
GVCrawl	GloVe	Common crawl	$8.4 \cdot 10^{11}$	$2.2 \cdot 10^{6}$	300	[PSM14b]
GVTwitter	GloVe	Twitter	$2.7 \cdot 10^{10}$	$1.2 \cdot 10^{6}$	200	[PSM14b]
GVWiki	GloVe	Wikipedia, news	$0.6 \cdot 10^{10}$	$0.4 \cdot 10^{6}$	300	[PSM14b]
w2vNews	word2vec	Google news	$1.0 \cdot 10^{11}$	$3.0 \cdot 10^{6}$	300	[Goo13]

Table A.2: Trait descriptive adjectives for the Big Five domains and their opposites, from [Gol92, Table 3, p. 34].

Extraversion	Agreeableness	Conscientiousness	Neuroticism	Open-mindedness
extraverted	kind	organized	anxious	intellectual
talkative	cooperative	systematic	moody	creative
assertive	sympathetic	thorough	temperamental	complex
verbal	warm	practical	envious	imaginative
energetic	trustful	neat	emotional	bright
bold	considerate	efficient	irritable	philosophical
active	pleasant	careful	fretful	artistic
daring	agreeable	steady	jealous	deep
vigorous	helpful	conscientious	touchy	innovative
unrestrained	generous	prompt	nervous	introspective
			insecure	
			fearful	
			self-pitying	
			high-strung	
Introversion	**Disagreeableness**	**Heedlessness**	**Emotional stability**	**Close-mindedness**
introverted	cold	disorganized	unenvious	unintellectual
shy	unkind	careless	unemotional	unintelligent
quiet	unsympathetic	unsystematic	relaxed	unimaginative
reserved	distrustful	inefficient	imperturbable	uncreative
untalkative	harsh	undefendable	unexcitable	simple
inhibited	demanding	impractical	undemanding	unsophisticated
withdrawn	rude	negligent		unreflective
timid	selfish	inconsistent		imperceptive
bashful	uncooperative	haphazard		uninquisitive
unadventurous	uncharitable	sloppy		shallow

Table A.3: Trait descriptive adjectives for the Big Five domains with negative factorloadings, from [Joh21, Table 2.4, p. 50].

Extraversion		Agreeableness		Conscientiousness		Neuroticism		Open-mindedness	
quiet	-.83	fault-finding	-.52	careless	-.58	stable	-.39	commonplace	-.74
reserved	-.80	cold	-.48	disorderly	-.53	calm	-.35	simple	-.67
shy	-.75	unfriendly	-.45	frivolous	-.50	contended	-.21	shallow	-.55
silent	-.71	quarrelsome	-.45	irresponsible	-.49			unintelligent	-.47
withdrawn	-.67	hard-hearted	-.45	slipshod	-.40				
retiring	-.66	unkind	-.38	undependable	-.39				
		cruel	-.33	forgetful	-.37				
		stern	-.31						
		thankless	-.28						
		stingy	-.24						

Table A.4: Items of BFI [JNS08, p. 157], with the extracted key term from each item. (R) indicates reverse scored items. Stop words in italics. Interpreted key terms in bold. “I see myself as someone who . . . ”

	No.	Item		Key term
Extraversion (8 items)	E1	. . . *is* talkative		talkative
	E2	. . . *is* reserved	(R)	reserved
	E3	. . . *is* full *of* energy		energetic
	E4	. . . generates *a* lot *of* enthusiasm		enthusiastic
	E5	. . . tends *to be* quiet	(R)	quiet
	E6	. . . *has an* assertive personality		assertive
	E7	. . . *is sometimes* shy, inhibited	(R)	shy
	E8	. . . *is* outgoing, sociable		outgoing
Agreeableness (9 items)	A1	. . . tends *to find* fault *with others*	(R)	fault-finding
	A2	. . . *is* helpful *and* unselfish *with others*		helpful
	A3	. . . starts quarrels *with others*	(R)	quarrelsome
	A4	. . . *has a* forgiving nature		forgiving
	A5	. . . *is* generally trusting		trusting
	A6	. . . *can be* cold *and* aloof	(R)	cold
	A7	. . . *is* considerate *and* kind *to almost everyone*		kind
	A8	. . . *is sometimes* rude *to others*	(R)	rude
	A9	. . . likes *to* cooperate *with others*		cooperative
Conscientiousness (9 items)	C1	. . . *does a* thorough job		thorough
	C2	. . . *can be* somewhat careless	(R)	careless
	C3	. . . *is a* reliable worker		reliable
	C4	. . . tends *to be* disorganized	(R)	disorganized
	C5	. . . tends *to be* lazy	(R)	lazy
	C6	. . . perseveres *until the* task *is* finished		persevering
	C7	. . . *does* things efficiently		efficient
	C8	. . . makes plans *and* follows *through with them*		planful
	C9	. . . *is* easily distracted	(R)	distractable
Neuroticism (8 items)	N1	. . . *is* depressed, blue		depressed
	N2	. . . *is* relaxed, handles stress *well*	(R)	relaxed
	N3	. . . *can be* tense		tense
	N4	. . . worries *a* lot		worrying
	N5	. . . *is* emotionally stable, *not* easily upset	(R)	stable
	N6	. . . *can be* moody		moody
	N7	. . . remains calm *in* tense situations	(R)	calm
	N8	. . . gets nervous easily		nervous
Openness (10 items)	O1	. . . *is* original, comes *up with* new ideas		original
	O2	. . . *is* curious *about many* different things		curious
	O3	. . . *is* ingenious, *a* deep thinker		ingenious
	O4	. . . *has an* active imagination		imaginative
	O5	. . . *is* inventive		inventive
	O6	. . . values artistic, aesthetic experiences		artistic
	O7	. . . prefers work *that is* routine	(R)	**commonplace**
	O8	. . . likes *to* reflect, play *with* ideas		**insightful**
	O9	. . . *has few* artistic interests	(R)	unartistic
	O10	. . . is sophisticated in art, music or literature		sophisticated

Table A.5: Items of BFI2 [Joh21, p. 81], with the extracted key term from each item. (R) indicates reverse scored items. Stop words in italics. Interpreted key terms in bold. “I am someone who . . . ”

		Item		Key term
Extraversion (12 items)	E1	. . . *is* outgoing, sociable.		outgoing
	E2	. . . *has an* assertive personality.		assertive
	E3	. . . rarely feels excited *or* eager.	(R)	unexcitable
	E4	. . . tends *to be* quiet.	(R)	quiet
	E5	. . . *is* dominant, acts *as a* leader.		dominant
	E6	. . . *is less* active *than other* people.	(R)	inactive
	E7	. . . *is sometimes* shy, introverted.	(R)	shy
	E8	. . . finds *it* hard *to* influence people.	(R)	**withdrawn**
	E9	. . . *is full of* energy.		energetic
	E10	. . . *is* talkative.		talkative
	E11	. . . prefers *to have others take* charge.	(R)	**inhibited**
	E12	. . . shows *a* lot *of* enthusiasm.		enthusiastic
Agreeableness (12 items)	A1	. . . *is* compassionate, *has a* soft heart.		compassionate
	A2	. . . *is* respectful, treats *others with* respect.		respectful
	A3	. . . tends *to find* fault *with* others.	(R)	fault-finding
	A4	. . . feels little sympathy *for* others.	(R)	unsympathetic
	A5	. . . starts arguments *with* others.	(R)	**quarrelsome**
	A6	. . . *has a* forgiving nature.		forgiving
	A7	. . . *is* helpful *and* unselfish *with* others.		helpful
	A8	. . . *is sometimes* rude *to* others.	(R)	rude
	A9	. . . *is* suspicious *of* others’ intentions.	(R)	**distrustful**
	A10	. . . *can be* cold *and* uncaring.	(R)	cold
	A11	. . . *is* polite, courteous *to* others.		polite
	A12	. . . assumes *the* best *about* people.		**trustful**
Conscientiousness (12 items)	C1	. . . tends *to be* disorganized.	(R)	disorganized
	C2	. . . tends *to be* lazy.	(R)	lazy
	C3	. . . *is* dependable, steady.		dependable
	C4	. . . *is* systematic, likes *to keep* things *in* order.		systematic
	C5	. . . *has* difficulty getting started *on* tasks.	(R)	**negligent**
	C6	. . . *can be* somewhat careless.	(R)	careless
	C7	. . . keeps things neat *and* tidy.		neat
	C8	. . . *is* efficient, gets things done.		efficient
	C9	. . . *is* reliable, *can always be* counted on.		reliable
	C10	. . . leaves *a* mess, *doesn’t* clean up.	(R)	messy
	C11	. . . *is* persistent, works *until the* task *is* finished.		persevering
	C12	. . . *sometimes* behaves irresponsibly.	(R)	irresponsible
Neuroticism (12 items)	N1	. . . *is* relaxed, handles stress well.	(R)	relaxed
	N2	. . . stays optimistic *after* experiencing *a* setback.	(R)	optimistic
	N3	. . . *is* moody, *has up and down* mood swings.		moody
	N4	. . . *can be* tense.		tense
	N5	. . . feels secure, comfortable *with* self.	(R)	secure
	N6	. . . *is* emotionally stable, *not* easily upset.	(R)	stable
	N7	. . . worries *a* lot.		worrying
	N8	. . . *often* feels sad.		sad
	N9	. . . keeps *their* emotions *under* control.	(R)	unemotional
	N10	. . . rarely feels anxious *or* afraid.	(R)	**calm**
	N11	. . . tends *to* feel depressed, blue.		depressed
	N12	. . . *is* temperamental, gets emotional easily.		temperamental
Open-mindedness (12 items)	O1	. . . *has few* artistic interests.	(R)	unartistic
	O2	. . . *is* curious *about many* different things.		curious
	O3	. . . *is* inventive, finds clever ways *to do* things.		inventive
	O4	. . . *is* fascinated *by* art, music, *or* literature.		**cultivated**
	O5	. . . avoids intellectual, philosophical discussions.	(R)	unintellectual
	O6	. . . *has* little creativity.	(R)	uncreative
	O7	. . . values art *and* beauty.		**artistic**
	O8	. . . *is* complex, *a* deep thinker.		complex
	O9	. . . *has* difficulty imagining things.	(R)	unimaginative
	O10	. . . thinks poetry *and* plays *are* boring.	(R)	**shallow**
	O11	. . . *has* little interest *in* abstract ideas.	(R)	**uninquisitive**
	O12	. . . *is* original, comes *up with* new ideas.		original

B Online survey description

For this thesis an online survey at https://www.empirio.de was conducted from 2022-02-08 to 2022-03-10 which attracted 127 respondents. The aim of the survey was to gather some data concerning the use of personality survey instruments and the experience and opinion applicants have of them in professional settings. Figure B.1 details the survey questions and answers given by the respondents. The original survey was conducted in German with German respondents, the items given in the figure are translated to fit this thesis.

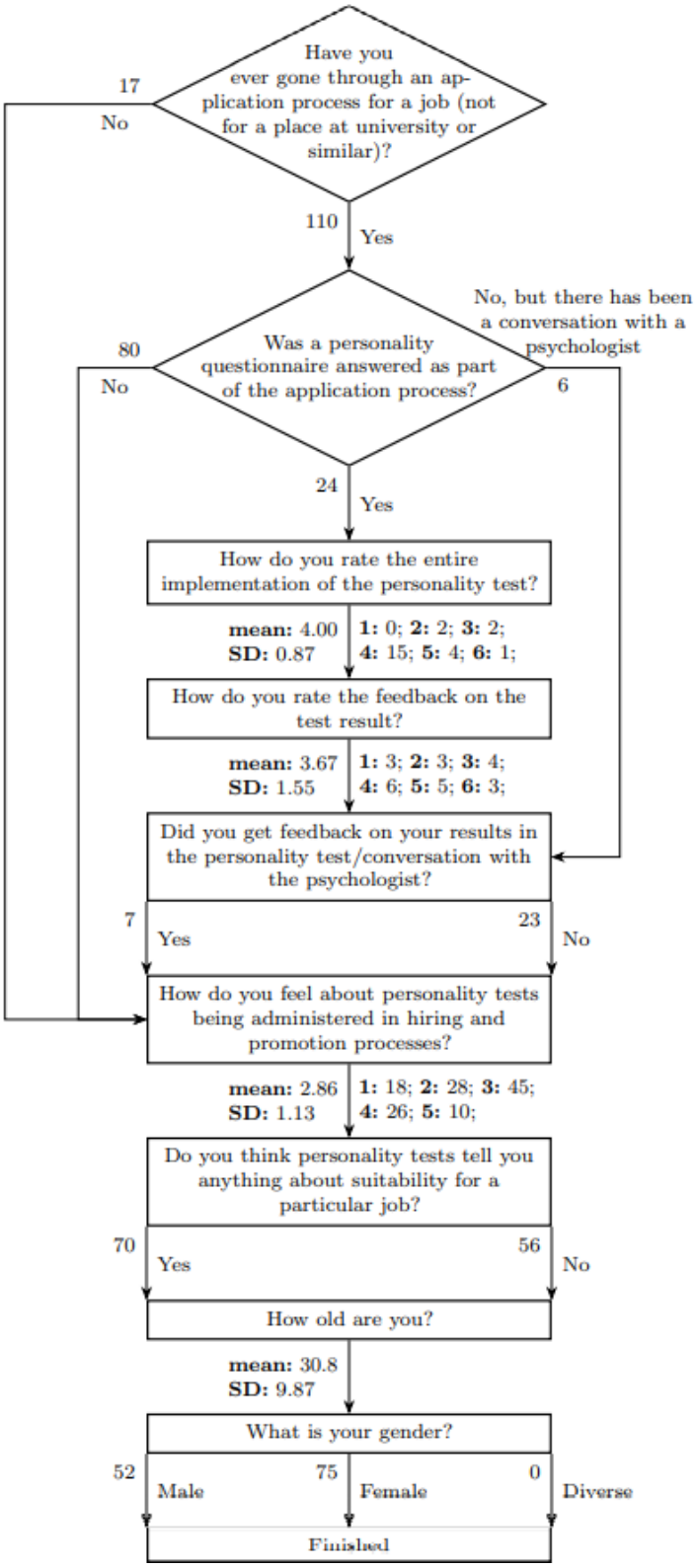

Figure B.1: Online survey structure and results.

Autor:innen

Volker Kempf ist seit 2018 wissenschaftlicher Mitarbeiter am Institut für Mathematik und Computergestützte Simulation der Universität der Bundeswehr München und promoviert dort auf dem Gebiet der numerischen Analysis. Während seiner vorherigen Tätigkeit als Offizier bei der Bundeswehr erlangte er einen Bachelor- und Masterabschluss in der Fachrichtung Mathematical Engineering und absolvierte einen Masterstudiengang Mathematik im Fernstudium. Beginnend im Mai 2021 studierte er an der Wilhelm Büchner Hochschule den berufsbegleitenden MBA-Studiengang Engineering Management, den er im April 2022 abschloss.

Prof. Dr. Helge Nuhn ist seit 2020 Professor für Digital Business Engineering an der Wilhelm Büchner Hochschule in Darmstadt. Er ist Wirtschaftsinformatiker (Dipl., TU Darmstadt), promovierte zum Thema temporärer Organisationsformen an der EBS Universität für Wirtschaft und Recht und hat mehr als zehn Jahre selbständig und in renommierten Unternehmensberatungen gearbeitet (Horváth, PwC, KPMG). Seine praktischen und Forschungsschwerpunkte liegen im Bereich Organisationstheorie, temporäre Organisationsformen und Projektmanagement welche er mit neuesten Erkenntnissen im Bereich der Forschung um Künstliche Intelligenz verknüpft. Er ist Leiter der Fachgruppe Agile Management der GPM Deutsche Gesellschaft für Projektmanagement e.V., Mitglied der GI Gesellschaft für Informatik e.V. und u.a. Dozent an der CBS International Business School.

Ansprechpartner:innen

Prof. Dr. Helge Nuhn
Wilhelm Büchner Hochschule, Hilpertstrasse 31, D-64295 Darmstadt, Germany,
E-Mail: Helge.Nuhn@wb-fernstudium.de

Überblick über die Bände der Schriftenreihe

Band 1 / 2022: **Christoph Sternberg, Ralf Isenmann**
Untersuchung regionaler Besonderheiten im Individualverkehr bei ausgewählten deutschen Smart-City-Projekten

Band 2 / 2022: **Fabian Fries, Manfred Hahn**
Dynamik von Doppelstern-Systemen

Band 3 / 2022: **Stefan Kaden, Ralf Isenmann**
IT based Framework facilitating Technology Roadmapping striving for Sustainability

Band 4 / 2022: **Hannah Seibel, Manfred Hahn**
Von der Raupe zur Drohne – Leichtbau in Anlehnung an die Natur

Band 5 / 2022: **Thomas König, Manfred Hahn**
Statische Festigkeitsberechnung einer 5-Speichen Fahrradfelge aus Faserverbundkunststoff

Band 6 / 2022: **Alrik Selle, Manfred Hahn**
Ertüchtigung der automatisierten Wetterbeobachtung unter extremen Vereisungen

Band 7 / 2022: **Valerie Seitz, Birgit Zimmermann**
Nachhaltiges Energiekonzept für einen Bauernhaushalt im ländlichen Äthiopien

INFORMATIK

INGENIEUR-
WISSENSCHAFTEN

ENERGIE-,
UMWELT- UND
VERFAHRENSTECHNIK

WIRTSCHAFTS-
INGENIEURWESEN
UND TECHNOLOGIE-
MANAGEMENT

WILHELM BÜCHNER HOCHSCHULE
Mobile University of Technology

EINE HOCHSCHULE DER KLETT GRUPPE.

www.wb-fernstudium.de

www.wb-online-campus.de