A Personality Mining System for German Twitter Posts With Global Vectors Word Embedding

HENNING USSELMANN1, RANGINA AHMAD1, AND DOMINIK SIEMON2

1Department of Business Science, Technische Universität Braunschweig, 38106 Braunschweig, Germany
2Department of Software Engineering, LUT University, 15210 Lahti, Finland

Corresponding author: Dominik Siemon (dominik.siemon@lut.fi)

ABSTRACT People’s personality influences their behaviors, attitudes, beliefs, and feelings. Therefore, many scientific studies already benefit from easy ways of measuring personality. By analyzing the written text of a person, it is possible to derive Big Five personality traits. One approach to this is to apply the unsupervised learning algorithm Global Vectors Word Embedding (or Representation), abbreviated GloVe, to English Twitter posts. The overall objective of our research is to show that this algorithm can also be applied to German Twitter posts. Therefore, we built a framework for training and applying machine learning models for personality predictions. We tested if a working prediction model for English Twitter users can be adapted for German users. This could reduce efforts for collecting training data. We evaluated our models based on a personality survey with a sample of German users. The method of adapting an existing model does not perform as well as expected but helps prepare the framework for higher volumes of data. In the end, the final model is based on the evaluation data, which results in an acceptable performance. Via a web application (https://www.miping.de) anyone can easily retrieve personality scores for any public German Twitter user. Altogether, it is shown that GloVe is suitable to predict personality based on German language. The published framework and source code allow for independent improvements to and easy application of the trained model. Now, scientific studies and other applications, e.g., chatbots, could easily incorporate personality data.

INDEX TERMS Personality mining, GloVe, LIWC, big five, machine learning.

I. INTRODUCTION

A. PROBLEM AND MOTIVATION

Language is one of mankind’s main ways to express itself. A common language helps to communicate with others and to share knowledge. But the words we use also enable deeper insights into people’s minds [1]. A person’s personality has a direct influence on how they speak and what words they use [2], [3]. The other way round, analyzing the everyday language of a person, for example, which specific words they use, tells a lot about that person’s personality [4]. Personality defines the core of our behaviors, attitudes, beliefs and feelings. It is a very stable concept in comparison to, for example, quickly changing emotions [5], [6]. On the other hand defining and measuring personality is naturally a very complex effort. Ref. [7] developed the Five-Factor model (also known as OCEAN model or Big Five model) based on multiple other approaches to defining personality. It is now a widely accepted and validated model even across different cultures [8]. By using five independent, basic dimensions of personality traits, namely Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness to Experience, it describes a model of human personality. All traits are usually examined via personality inventories, with the NEO Personality Inventory-3 being the current standard questionnaire for the Big Five model [9].

Measuring personality helps in the areas of recruiting employees, understanding health and providing better healthcare, as well as optimizing marketing strategies. Companies are able to carefully manage their teams’ diversity in terms of personality in addition to other capabilities and thereby increasing the overall performance [10]. Studies also show that Big Five personality traits have a significant impact on both mental and physical health. This enables new strategies in treating and researching illness [11]–[13]. The influence of character traits reaches even the domain of e-commerce, moderating the effect of customer relationship concepts and receptivity for advertisement [14], [15]. The problem being that the before mentioned personality inventories can take at least 30 minutes for each test person to complete [9]. With
every additional survey item, researchers increase the risk of participants quitting the survey before completion [16]. This might inhibit the incorporation of personality data in the mentioned contexts although it would provide useful insights.

To address the problem of extensive surveys, researchers started to exploit the relationship between language and personality traits. Instead of letting respondents fill in questionnaires, these new models make use of the already existing data in social media networks (e.g., Twitter) and other areas of the internet (e.g., blog posts) [17], [18]. These systems save time when dealing with personality, thereby enabling practical, efficient applications for both research and practice. One approach is to analyze the specific word usage and build statistical models that estimate character traits for each user. Ref. [19] implemented the basis for that with the Linguistic Inquiry and Word Count (LIWC) which quickly gained in importance for linguistic analysis [20], [21]. Based on the proprietary LIWC many studies have shown that it is possible to estimate the authors’ personality traits based on their written texts [18], [22], [23]. Further improvements have been reached with the use of so called word embeddings like Google’s word2vec or Stanford’s GloVe [24], [25]. The advantage of such models, is that the semantic similarity between words is determined unsupervised, whereas LIWC relies on human judges and psychologists to obtain the meaning of words [22], [26]. Ref. [22] suggest that models with word embeddings predict personality even better than models based on LIWC. With Watson Personality Insight, IBM developed a commercial, closed-source software as a service for ready-to-use personality predictions based on GloVe and Twitter posts [22], [27]. A factor of vital importance for all mentioned models is the test person’s native language because the analysis is based on subtle differences in word usage. Foreign speakers cannot be included in such approaches due to their limited vocabulary [28]. Unfortunately, research often focuses on English language, limiting the pool of participants to English native speakers and therefore also limiting further research in different cultural or national domains. IBM’s Personality Insight, for example, currently supports only five languages: Arabic, English, Japanese, Korean, and Spanish [29]. To increase the accessibility to the latest discoveries in personality research, more languages should be considered and researchers should be enabled to easily utilize new models. For increased transparency and reproducibility of scientific results, the whole source code for creating new models should be published in addition to the models themselves. The usage of new models has to be simple, in order to expand the circle of users beyond machine learning experts.

B. DEFINITION OF GOALS

In the course of this paper, multiple challenges will be addressed. The overall goal is to simplify the use of personality mining for the German language. As of now, there is no proof-of-concept in existence that shows personality mining based on GloVe for the German language. GloVe was chosen over other word embeddings, such as Google’s word2vec, because it is open source, the most current and best suited for our approach. Furthermore, existing ready-to-use applications for personality mining like LIWC or the one from IBM are proprietary and closed-source. This might inhibit widespread usage, independent improvements and scientific peer reviews [30]–[32]. Therefore, this paper aims at the following goals:

1) Development of a proof-of-concept model that is able to estimate a user’s personality based on Twitter posts for the German language with the help of the GloVe word embedding.

2) Make this trained model and its code publicly available for independent use and review.

3) Provide a ready-to-use web application for the model to enable interested stakeholders to retrieve instant personality estimations.

The fulfillment of these goals will show that the GloVe word embedding can be used in the context of personality prediction for the German language. It will set an example of how related research could be published to increase accessibility to the latest results. Furthermore, the process could be reused for applying the method to additional languages. Especially the web application will expand the potential audience of users to groups that are not technically adept. For example, a marketing study interested in the effectiveness of ads on social media could easily incorporate users’ personality scores in addition to the primary questionnaire.

C. METHODOLOGY AND STRUCTURE

The structure of this research project is based on the Design Science Research approach described by [33]. The knowledge contribution will be in the improvement quadrant as the problem of estimating a person’s personality is well known and there are already existing solutions, which will be improved in the course of this research [33]. In the Design Science Research terminology the results of the above mentioned goals are also called artifacts [33]. In section II, the Five-Factor model for personality is described in order to have a consistent definition of personality. There will be an explanation of natural language processing as well as a short review of already existing personality mining tools. The actual design and implementation of this paper’s artifacts will be based on machine learning approaches. Therefore, this development process follows the methodology of the Cross Industry Standard Process for Data Mining (CRISP-DM), which will be outlined in a separate section [34], [35]. This is followed by section III with subsections to determine requirements, develop a suitable technical architecture and provide a breakdown of all used input data. The modeling and deployment subsections consider actual training of machine learning algorithms and how the result is published. The training is comprised of three steps (also see figure 1):
from these posts, and train a machine learning model to deduce personality traits based on LIWC categories.

2) Use this trained model to deduce personality traits based on LIWC categories derived from German Twitter posts. This serves as ground truth for the final step.

3) Map text of German Twitter posts into Global Vector model and use these vectors to train a new machine learning model for deducing personality traits.

FIGURE 1. Step visualization.

To show the positive knowledge contribution the artifacts are examined with regards to the previously defined requirements. Furthermore, a small sample of German Twitter users will be asked to fill in a personality questionnaire to evaluate the artifact’s accuracy. In the end, all results will be discussed and a conclusion will be drawn.

II. THEORETICAL BACKGROUND
A. FIVE-FACTOR MODEL FOR PERSONALITY
1) DEFINITION OF PERSONALITY
Personality research is one of many fields in the area of scientific psychology [36]. It observes the stable characteristics that are relevant for an individual person’s behavior. Characteristics include both observable and non-observable properties [37]. In contrast to clinical psychology, personality research focuses mainly on non-pathological personalities, which excludes, for example, diseases like schizophrenia. Since it is difficult to define the boundaries of abnormal personalities there is some overlap of these two disciplines [37].

Throughout the history of personality research, numerous theories have been developed. Ref. [38] lists the following six major perspectives: biological, cognitive, humanistic, learning, psychodynamic and trait. Each perspective has its own concepts and definitions of personality. Some assumptions have been proven wrong or being unlikely, but the major challenge of defining personality is that there is no method of measuring it directly [38], [39]. One common determinator of all perspectives is that personality refers to the aspects of a person that are relatively stable over time [39]. Personality is both a general theory of a person as a whole entity and something that serves to show differences between individuals [40].

Among the listed perspectives on personality, the trait perspective emphasizes the importance of empirical research and objective observations [39]. The trait approach describes a person’s personality using different adjectives based on self-reports and third party observer descriptions. Those descriptions are usually related to specific situations. Through statistical analysis, different persons can be described and compared. The underlying theory states that individuals have certain properties that influence their reactions in specific situations [37], [39]. Traits are stable over time, meaning that the behavior of an individual person varies from situation to situation, but overall there exists a consistency. By observing many situations it is possible to find the differences between two peoples’ behavior, which is an expression of the underlying personality. This consistency distinguishes traits from relatively short lasting emotions. Although behavior might influence traits as well, trait theory assumes that the dominant direction is the trait’s influence on behavior [41].

The Five-Factor model (also called Big Five model) is one of many implementations of the trait perspective. Currently, it is widely used and has been applied in many studies, proving its usefulness in personality research [42], [43]. The Big Five model unified different approaches of describing personality with as few traits as possible. Factor analysis showed that five independent factors are sufficient. More traits, as used in other models, could be expressed by a combination of only those five factors [7], [44]. Ref. [7] are often associated with the model as they significantly contributed to its discovery and their Revised NEO Personality Inventory (NEO-PI-R) is widely used for measuring the factors [42], [43], [45]. The five independent factors are [7]:

- Extraversion
- Agreeableness
- Conscientiousness
- Neuroticism
- Openness (to experience)

Sometimes the five factors are abbreviated with the term OCEAN. It is argued, that even three factors or 6 to 7 factors would be a better personality description model. Moreover, such factor models are sometimes criticized for being too simple and superficial to meet the complexity of human personality [37]. But the Five-Factor model already shows good results in scientific practice and the measured values correlate with real life behavior [6], [46]. Therefore, in this work personality is used as described by the Big Five model. Each factor can be broken down into six facets each, which allows a more subtle analysis if needed [45]. The measurement and scales are explained in section II-A3. The Five-Factor model itself is a descriptive model of measuring personality. It is not a complete theory for explaining personality. With the Five-Factor Theory of Personality [46] developed a comprehensive theory of the underlying concepts that the Five-Factor model measures. Since the focus of this work is measuring person-
ality and not explaining it, this theory will not be described further.

The meaning of each trait in the Big Five model is briefly explained in the following paragraph [7], [43]. People with a high degree of Extraversion are typically more outgoing, talkative and energetic. They tend to have a larger social network. A high degree of Agreeableness might be expressed by modesty, trustworthiness and willingness to compromise. People with lower values show signs of the exact opposite. High Conscientiousness is represented by a sense of duty, discipline and reliability. Individuals with strong Conscientiousness tend to strictly follow rules. Neuroticism is a scale for emotional stability. Nervous, anxious people score high values in Neuroticism. Recovery from stressful events tends to take longer. Openness describes the degree of openness for new experiences. People with high degrees are often described as creative, artistic and imaginative.

Important aspects of the Five-Factor model are its usability across different cultures and ages, allowing cross-cultural studies with the same model. Furthermore, it has been observed that a human’s personality changes over a lifetime, but the overall tendencies in each factor are robust [43].

2) RELATIONSHIP BETWEEN PERSONALITY AND LANGUAGE
One important aspect in the context of the Big Five model is the lexical hypothesis. It states that characteristics, that are important for a group of people, will become part of this group’s language. Furthermore, the most relevant differences are likely to be encoded in a single word [47], [48]. That implies that all relevant words and phrases to describe important personality traits should be incorporated in the daily language already. By systematically analyzing and reducing a language’s vocabulary, five factors can be found, that correspond to the factors of the Big Five model [37].

Ref. [49] found out that language not only contains words to describe personality but people can be differentiated by their daily language use across a variety of situations. Unwillingly, individuals convey their inner thoughts and feelings with every sentence they speak and write regardless of the involved topics [49]. Interestingly, it is sufficient to just analyze the word usage without any context to find correlations with a person’s personality [3], [21], [50]. “There is a subtle, yet important difference between saying, ‘I am not happy’ versus ‘I am sad’” [50, p. 334]. This example shows two statements with a similar meaning, but it suggests that the first person is thinking more on a scale of happiness and the second person more on a scale of sadness. Which is among other factors a result of the influence of personality on language use [50]. The other way round, by measuring and analyzing individuals’ word usage, their personality can be derived [4]. This will be detailed in sections II-B1 and II-D.

3) MEASURING PERSONALITY
There are several methods of measuring personality: self-reports, interviews and third-party observations [51]. Self-reports ask the subject directly, either verbally or in written questionnaires, about their thoughts, feelings or expected behavior. One problem of this approach is the deliberate or unconscious manipulation of answers, because, for example, test subjects want to appear in a favorable light. To circumvent this challenge persons with frequent contact, such as friends or colleagues of the test subject, could be questioned. Although asking multiple persons could provide more empirical data, observations could be biased, because colleagues, for instance, get to know test subjects only in their working environment. Another method is the observation of behavior during experiments. Those measurements are more objective, but often do not provide the true insight of a person’s motivations, leading to a broader scope of interpretation [43]. If sufficient data is already available, the relationship between language and personality can be exploited with computer programs, e.g. via analyzing word usage, providing a relatively new approach to measuring personality [4], [17], [52]. This will be detailed in section II-D. Self-reports with both paper and electronically based questionnaires are still the major form of personality assessment as they are easy-to-use for both the facilitator and the test subject, economically cheap and no additional person except the test subject is needed [51]. Although self-reports are a common research method, it is criticized that they do not measure a comprehensive picture of human personality and rely on the subject’s willingness for cooperation, in order to get valid results. Furthermore, the fact that data is usually collected just at one point in time, instead of over a longer period, is another point of criticism [53].

Since personality cannot be measured directly, indicators have to be defined, which are dependent on the definition of personality itself. For the Five Factor model personality consists of the five OCEAN factors, but these are also not directly measurable [43]. The NEO-PI-R provides a questionnaire for indirectly measuring those factors [45]. It is widely used and relies on many empirical research projects [41]. There exist multiple validated translations of the NEO-PI-R. Most of these are proprietary and fee-based. Therefore, additional scientifically free-to-use inventories are available, like the Big Five Inventory 2 (BFI-2) [54]. Since the NEO-PI-R represents an important benchmark for personality inventories, it is exemplarily described to show the functioning of these tests.

Consisting of 240 question items it assesses both the five factors and each of its six facets. There exists both a self-report and observer rating form, depending on the needs of the facilitator. One test might take up to 45 minutes. A shorter test not including the facets exists as well [45]. Each question is answered on a five-point Likert scale asking the participants whether they strongly disagree or strongly agree with a given statement, which describes a specific behavior or
attitude [55]. Calculating the mean value of all items assigned to a factor, the raw value for this factor can be derived. The raw value itself does not provide much useful information. Therefore, it has to be translated into a norm value [56]. The norm value allows comparisons with a representative sample population, for example, building rankings on the scale of Extraversion [57]. In the case of personality measures, this is often the population of a country, sometimes divided into age groups [56]. It is noteworthy, that all five factors are normally distributed, just like the intelligence quotient [43]. Norm values are typically T-scores, stanines or percentile ranks. A percentile value states that the given percentage of the norm group scored the same or a lower score [58]. That means, for instance, test subjects who score 90% on the Extraversion scale, have a more extrovert personality than 89% of the people in the norm group. This is useful in multiple contexts, which will be explained in the next section.

4) PRACTICAL APPLICATION
Since personality is defined as characteristics that influence an individual’s behavior, it is of great interest in many aspects of life [37]. Many questionnaires have been developed to detect personality or psychological disorders in clinical psychology [51]. But the influence of psychological factors on the physical health as well has been long known. Personality, as part of the emotional reaction to certain events or stress, influences the body’s biological reaction, such as blood pressure or heart rate. These in turn affect coronary heart diseases, a major cause of death in developed countries [59]. It has been shown that the five factor domains correlate with instances of longevity and physical illness. Further on, their connection with healthy and unhealthy behaviors, which again influence the tendency of illness, has been shown in several studies [12]. A relationship between the risk for major depression during the lifetime and the five factors have been established as well [60]. Thus, personality is relevant in the scope of researching and detecting both physical and mental health problems.

In many companies, personality tests are already part of the personnel selection and hiring process. The trait Conscientiousness shows a correlation with general work performance. All factors show an impact on long term career success and job satisfaction, which can be another factor in recruiting personnel [61]. In today’s working environment, not only an individual’s performance but a whole team’s performance is vital for success. Therefore, team composition is another factor for recruitment. Via personality tests, the overall team composition can be controlled and the individual’s ability to work in teams can be deduced [10], [62].

Not only job performance but also performance in sports is correlated with personality traits, as athletes with successful performance during a season have higher levels of Conscientiousness and lower levels of Neuroticism than less successful athletes [63]. Just as certain personality traits influence healthy behaviors, it is the hypothesis that certain personality traits influence behaviors that lead to greater fitness and higher ambition, which is necessary for professional sports [64]. In practice personality measures can help to optimize team composition and provide targeted psychological support for each athlete during difficult times [65].

In the domain of e-commerce personality measures can help to focus marketing investments on the right customer relations. The individual customer’s personality has a moderating effect on loyalty and receptivity for advertisement [14]. Going forward, personality might influence the way product information is presented, optimized to increase each customer’s willingness to buy the product [15]. Chatbots that take customer’s personality into account can increase overall customer satisfaction and in turn might increase sales for online shops [66]. In the context of the hotel industry, hoteliers could try to attract guests with a personality that assures a higher probability for customer satisfaction and consequently higher loyalty. Their return on marketing could increase with the same input [67].

Summing up, this section shows the great relevance of personality in both research and practical applications with measurable results. Therefore, it is important to increase accessibility and ease of use for personality tests. New measurement methods are based on natural language processing, which is explained in the next chapter.

B. NATURAL LANGUAGE PROCESSING
1) LINGUISTIC INQUIRY AND WORD COUNT
Natural language processing describes techniques and computer systems focused on learning, understanding and producing human language content [68]. Input data involves both written and spoken text of all kinds of patterns and languages. Goals might be automatic translations between foreign languages, text comprehension, and tagging or conducting human-computer conversations as seen in chatbots [68], [69]. Depending on the context and task to be accomplished different approaches and algorithms have emerged, often involving applied statistics and machine learning [68].

One approach for the detection of sentiment and emotions in human language is the LIWC by [19], [68]. It provides an automated method to get insights into “various emotional, cognitive, and structural components present in individuals’ verbal and written speech samples” [20, p. 1]. It leverages the relationship between language and people’s minds as described in section II-A2. Several versions of the closed-source software exist, with the LIWC2015 being the latest revision, which is only commercially available [20]. There exist several, validated translations from the English language, including a German translation [20], [70].

LIWC consists of a dictionary and related processing software. The English dictionary contains about 6,400 word patterns. The patterns ensure that variants of a word, such as singular and plural, are still recognized as that base word. Each word pattern is assigned to one or more of 93 categories. Categories cover both summary variables, such as total word count or count of words with more than six letters, as well as
grammars related categories, such as the use of prepositions or comparisons. Most categories fall under psychological processes, which include emotions, swear words or words with a reference to religion [20]. Applying LIWC to a sample text will identify words in the sample that exist in the dictionary and count the number of words matched for each category. In the end, these key figures are converted to percentages, summarizing the share of words that fall into each category [20], [71]. The dictionary creation and the assignment of words to categories was an iterative, manual process, that involved many human judges to validate each category [20]. Only words considered in this process are included in the dictionary and therefore only these words can contribute to the result in text analysis. This is called a closed vocabulary [72]. Some example areas LIWC has been successfully used in are psychological disorder research, personality mining or lie detection [21]. LIWC is not the only method that tries to capture emotional states in texts. The “Dictionary of Affect” works with the same dictionary method, but provides only two output categories for emotions [68], [73].

2) GLOBAL VECTORS WORD EMBEDDING

A more generic way of converting text input into numerical values are so called word embeddings. Word embeddings are a technique to map a vocabulary of a certain size into a lower-dimensional space. As a result, each word of the input vocabulary is represented by a numeric vector, which is often the required input format for further statistical calculations [74]. There exist multiple algorithms for word embeddings that are optimized for different tasks. Common tasks are [75]:

- Modeling the relatedness between words in such a way humans would assess the similarity between two words
- Analogy tasks, such as x is to y as a to b, where b is answered by the algorithm
- Clustering words or documents into categories

Word embeddings usually work unsupervised, meaning they only need the text itself as input and no human based judgments or labels, which makes extending the vocabulary easier. Hence, this is called an open vocabulary approach [72].

One recent method is GloVe for word representation by [25], which improved the previous popular word2vec by [24]. GloVe performs well for both word analogy tasks as well as word similarity tasks. Its goal is to create a model in such a way that words that often occur together in the input data, should be close together in the lower dimensional output matrix. Words that are syntactically or semantically similar, will be closer to each other [25]. GloVe needs to be trained on a large corpus of texts, for example, an extract of English Wikipedia entries, which contains over one billion words. Of this corpora, the most frequent words, e.g. the top 400 000, are selected to build the vocabulary. For this vocabulary, the algorithm builds a co-occurrence matrix, which counts how often words appear together in a ten word context window. Thereby, words that are often mentioned together can be identified. This matrix is the input for the actual model, which transforms the co-occurrence statistics into a 300 dimensional vector for each word [25].

Figure 2 shows an example of how the concept of sex and gender is encoded in the final vectors for each word. The vector difference between woman and man is similar to the one between queen and king. All other shown word pairs show the same difference between the opposite genders [25], [76]. With this structure the answer to the question “Which word is to Germany, what Paris is to France?” can be calculated by subtracting the vector France from Paris and adding Germany. That results in a vector close to the answer Berlin. This is an analogy task [24]. Since GloVe tries to incorporate as many interpretations and meanings of a word as possible, it implicitly allows to build categories similar to the ones in LIWC [72].

![Visualization of gender concept in GloVe](source: with modifications [76].)

C. CROSS INDUSTRY STANDARD PROCESS FOR DATA MINING

1) OVERVIEW

Data mining\(^1\) is defined as “the process of discovering useful patterns and trends in large data sets” [35, p. 2]. Its purpose is to transform vast amounts of data that is already stored in existing systems into knowledge that leads to informed business decisions or insights [35]. Personality mining is therefore the process of deriving personality related information associated with individuals from data that is originally intended for different purposes, for example, social media data.

Regardless of the contextual domain, there are several steps every data mining project needs to consider. Therefore, standardized processes have been developed to structure data mining projects. [77] described Knowledge Discovery in Databases (KDD), [78] developed the Big Data Analysis Pipeline and [34] created the Cross Industry Standard Process for Data Mining (CRISP-DM) The latter one will be the guideline for the development of a personality mining application.

\(^1\)Data mining and machine learning are used interchangeably in this work.
CRISP-DM is a context neutral methodology that aims to suit all industry areas and is independent on the specific technology used for data mining. It consists of six phases that are traversed iteratively, which can be seen in figure 3 [34].

The process is adaptive, meaning steps backward into a previous phase are allowed and often reasonable, because insights in later phases might lead to changes in earlier stages. Furthermore, the outer circle in the figure represents the knowledge transfer and lessons learned inside an organization across organizations in between data mining projects (as displayed in figure 3) [34], [35]. Each phase will be described in the following sections.

2) BUSINESS UNDERSTANDING
The business understanding phase is the first and initial phase of a data mining project, laying the groundwork for the whole project. Initially, the customer’s objectives need to be identified in order to assess if these objectives can be met by a data mining project in the first place. Part of this is to identify possible constraints that might limit certain approaches and influence the output of the subsequent phases. Following this, more detailed requirements and assumptions should be listed. This initial assessment identifies possible sources of data as well. Only after business related questions have been addressed, specific data mining goals should be phrased. These goals describe the final outcome of the project which enables the customers to achieve their business objectives. If possible, suitable success criteria should be defined. A project plan marking important milestones and highlighting dependencies between tasks concludes the business understanding phase [34], [35].

3) DATA UNDERSTANDING
Following business understanding, the data understanding phase explores all data used for the data mining approach. As a first step, data needs to be collected from sources identified in the previous phase. Data understanding and data preparation are often interlinked tasks, because, for example, tools used for data understanding also need some basic data preparation in order to load the data into the tools [34]. Data sources, methods of how data is collected, and related problems should be recorded for transparency reasons. Next, follow basic data descriptions such as number of datasets, data format, and data fields. It should be evaluated whether requirements can be fulfilled with this data or if different and additional data needs to be collected. If no impediments occurred so far, detailed data exploration will be carried out. Depending on the goals, this can include visualization to detect obvious correlations, simple aggregations such as average, minimum and maximum of important data fields, or other descriptive statistical analysis. Of great interest are certain subgroup breakdowns as well, for example, to find differences between age groups or customer segments. In the end, an important prerequisite for data preparation is a data quality report that examines if there are any missing values if these errors follow any patterns, and if the collected data is correct [34].

4) DATA PREPARATION
Using insights obtained during data understanding, the data preparation phase considers all aspects of transforming collected raw data into suitable variables for the core analysis of the modeling phase [34]. At first, data is selected from the pool of collected data according to the data mining goals and technical constraints, such as data volume. If necessary, missing values are treated to be replaced by some estimate or to be explicitly marked as missing. This depends on the previous analysis of why these values are missing in the first place. Since several models are prone for outliers, it needs to be decided if and how these are treated. Since outliers might also carry valuable information for special cases in the dataset, it could make sense to take these values as they are [79].

An important step is to integrate, unify and transform data from different sources. For instance, for fields containing text, as it is the case with Twitter posts, all characters should be converted to upper or lower case to account for inconsistent user input. For certain use cases it might be necessary to remove special characters and correct spelling mistakes with dictionary approaches [79]. To be able to reproduce these steps a detailed description of all steps, decisions and assumptions should be created [34].

Once the whole dataset is unified and integrated the actual features for the modeling phase can be extracted. This involves selecting fields from the integrated data, but also deriving new attributes as a combination of existing ones. Different strategies can be applied for feature selection in order to select the ones which promise the best performance. One possible way of conversion for text based features is word embeddings as explained in section II-B2. The last steps to create the final dataset for the data mining models are
normalization or standardization. In normalization, such as min-max scaling, all values are shifted and rescaled inside a given range. This is useful if two features should be equally important but have a very different value range. Standardization, such as z-score standardization, requires normally distributed data. For this method, values are centered around the mean with a unit standard deviation. Since values are not capped to upper or lower limits, outliers are preserved. The final dataset serves as input for the modeling phase [79].

5) MODELING

a: MODELING PHASE

Data mining tasks can be roughly divided into two categories of tasks: supervised and unsupervised. In unsupervised methods exists no target variable to be predicted. Contrary, often no correct solution to a given problem exists and the goal is to find hidden patterns in data. The most common representative for unsupervised learning is the process of clustering. Supervised data mining methods have a predefined target variable associated with each dataset. Typical tasks are classification and regression. For classification, each dataset belongs to one or sometimes multiple classes and the goal of an algorithm is to predict this class based on the remaining attributes of that dataset. If the associated classes are continuous, this is called regression [35]. Predicting personality can both be a classification task if individuals are assigned either a high or low Extraversion class, and it can be a regression task if a percentage rank should be predicted for each individual.

The overall data mining task type is usually directly derived from the business goals identified in a previous phase. For each task type exist several algorithms, also called models, with different advantages and disadvantages. Some models are more efficient on high volumes of datasets or in higher dimensional feature spaces, which means many different attributes are available for prediction [79]. Relevant model algorithms are briefly described in section II-C5b. The process of creating a model with given data is called training. The used data is called training data [35]. During training many models have the possibility to adjust certain parameters, which influences the performance of the model. This is called parameter tuning [80].

The modeling phase is the core of a data mining project. First, depending on the overall task one or more algorithms are chosen to be on the shortlist. Ideally, data preparation already considered the specific needs of input formats for the selected models. Next, a test design needs to be created in order to select the best model algorithm of the shortlist and to optimize the model’s parameter if there are any available [34]. For this purpose, data is divided into training, test, and validation data. Training data is used for actually training the model, whereas the test data supports the process of selecting and tuning the models by simulating a prediction on unknown data. Since the correct target variables are known for test data, the actual error can be determined and compared to other models [35]. Once a model is selected, fully trained, and tuned, the validation dataset can be used to estimate the performance of that model. To repeat the process of training and testing multiple times a technique called k-fold cross-validation is used. This also helps when the data volume is not high enough to split the data into multiple test sets. Often 10-fold cross validation is applied, meaning the input data is divided into 10 equally sized but randomly created datasets, where 9 datasets are used for training and the last one for testing. After 10 repetitions 10 models, which have the same parameter but different input data, have been created and tested. If the performance of all models is of similar quality without high variance, it is considered an indicator for a good fit [35]. Additionally, cross-validation helps to avoid overfitting, which is the case if models do not find patterns in the data but instead fully memorize the training set. That might lead to very high performance on that training set, but the model would be useless for prediction on unseen data [80]. The opposite of overfitting is underfitting, which is the case if the model’s predictions are almost independent of input data, for example, just a constant value. Both are unwanted effects in data mining [35]. Cross-validation is used as well for tuning the model’s parameters via an exhaustive grid search, where multiple possible parameter values are systematically tested via repeated training and testing [81].

With the help of repeated training and testing the best model type can be selected and its parameters tuned. The metrics that are used to evaluate the selected model are explained in section II-C6.

b: MODEL TYPES

The following section briefly explains different model types that are used in existing applications for personality mining as presented in II-D. For an easier introduction the linear regression is explained first, other models follow alphabetically. Related models, such as specialized models of a parent model, are explained together with the parent model. The linear regression combines all given attributes in a linear fashion with fixed weights [80]. A simple example is the equation

\[ y = ax + b \]  \hspace{1cm} (1)

which predicts the target variable \( y \) based on one attribute \( x \) [79]. To find the values for parameters \( a \) and \( b \) usually the sum of squares of the difference between the actual and predicted value will be minimized. There is no parameter available for tuning, the best regression line is directly derived from the training data [80]. The ridge regression is a special case of linear regression which introduces a penalty parameter to control a tradeoff between closeness of fit and model complexity. This parameter helps to avoid overfitting and is tuned via cross-validation [80].

The bayesian ridge regression combines the penalty parameter of the ridge regression with a probabilistic approach leveraging Bayes’ theorem. Domain knowledge of the problem is included by passing a prior probability distribution into the model. If no specific distribution is known,
a Gaussian distribution can be chosen. The combination of a penalty parameter and a probabilistic approach leads to less over-fitting for this model type [82], [83]. Gaussian Processes are based on Bayes’ theorem as well, but the key difference is that they leverage the so-called kernel trick for non-linear regressions [83]. A kernel is a mathematical function that maps the input data into a higher dimensional space. While the input data might not be linearly separable in its original dimensions, provided a suitable kernel is used, the data will be linearly separable in the projected space. This allows the application of linear algorithms for non-linear data [80]. Therefore, the kernel needs to be given as an input into the model. The kernel’s parameter can be automatically optimized during training [83].

Another common algorithm is a decision tree, which consists of decision nodes, branches and leaf nodes. A simple example can be seen in figure 4 showing if the outcome of some contract negotiation is considered good or bad [80]. Starting at the top, different attributes are tested following branches with the corresponding values, eventually arriving at a square leaf node which is the prediction for the given dataset. Since decision trees inherently tend to overfitting, the number of leaf nodes is limited by pruning the tree, which effectively removes some of the decision nodes. The optimal pruning parameters are found via cross-validation [35]. Random forests are a so called ensemble method based on decision trees. Ensemble methods combine multiple models into one, trying to compensate disadvantages of just one model. Random forests train multiple decision trees on different subsets of the input data and average the prediction of all trees as the final output [80].

![Decision Tree Diagram](image)

**FIGURE 4. Decision tree example source: [80].**

Lastly, a Support Vector Machine (SVM) is another linear model that employs the kernel trick. It tries to construct a hyperplane which separates two linearly separable classes with a maximum distance to these classes. For this construction only the nearest data points from the training set are necessary, the so called support vectors. This last property makes a SVM memory efficient. Although this model was originally developed for classification, it has been adapted to handle regression tasks as well [80].

6) EVALUATION AND DEPLOYMENT

As suggested in the previous phase, modeling and evaluation are closely interlinked as the models are already partially evaluated during cross-validation. To assess a model’s performance different metrics can be used. Most of them are based on the comparison between the true value and the model’s predicted value for each dataset. The Mean Absolute Error (MAE) averages the absolute value of all errors in the test set. To account for different scales, the MAE can be converted to a relative absolute error, indicating the proportion of error in relation to the used scale. The Mean Squared Error (MSE) works similar to the MAE but takes the squared instead of the absolute error. It is a common measure in linear regression. The Root Mean Squared Error (RMSE) is the root of the MSE taking the metric’s value back to the original dimensions, making it easier to comprehend. Contrarily, a correlation coefficient is dimensionless on a scale from −1 to 1. A value close to 0 indicates no correlation, whereas 1 and −1 indicate perfect positive or negative correlation between true and predicted values [80]. Another common measure is the coefficient of determination, denoted as $R^2$, which shows the share of variance explained by the used model. It is also possible to define custom cost functions to account for the specific business goal, for example, to weigh false positive credit ratings higher than false negative ratings, as these have a higher business impact [35].

All the mentioned metrics serve to assess a model’s performance during testing and validation. This final validation should be done with a separate dataset that was not used during training and testing. Either this dataset already exists because of the previous split of data or it might be sensible to collect a completely new dataset for validation. Additional checks for over- or underfitting can be applied. During the evaluation phase, not only performance but all defined requirements will be included in the assessment. The model should achieve the data mining goal defined in the business understanding phase and support the related business decisions. If certain response times are necessary in time critical processes, these are evaluated as well. Revealed deficiencies are rectified by jumping to the applicable previous phase [34].

After all requirements are met, the last phase of CRISP-DM, the deployment phase, follows. At first, a plan has to be devised for how to make the models and results usable for the business. Depending on the goals deployment can just consist of a detailed report recommending business decisions and providing insight into the hidden patterns of the analyzed data. In other cases, the trained models need to be incorporated in day to day business, so the plan includes training courses for end users and a setup for monitoring and maintenance, as it is necessary for every software. The
last step before a project concludes should be writing down lessons learned for upcoming other data mining projects [34].

D. EXISTING APPLICATIONS FOR PERSONALITY MINING

1) LIWC BASED APPLICATIONS

Having described how LIWC works generally in section II-B1, this section looks at existing literature and applications that utilize LIWC and data mining to derive people’s personalities. Since people’s personality is reflected in their daily language, this fact can be leveraged to replace or at least support traditional questionnaires. Multiple studies show that the categories of LIWC correlate at least to some degree with the Big Five dimensions [4], [49], [50].

One of the first to use regression models for automated personality prediction based on the 2001 version of LIWC were [4]. They did not just evaluate the correlations between LIWC and Big Five dimensions but also simulated the prediction performance on unseen subjects. The study was based on written essays of 2479 US American students as well as 96 transcribed recordings of daily conversations. To estimate accuracy, the model’s prediction for each dimension is compared to the subject’s actual ratings. Reported Pearson’s correlations are in the range of 0.24 to 0.33 for all five dimensions. As regression models a linear regression, a SVM, and a decision tree for regression were utilized, with the SVM having higher error rates [4]. All mentioned model types are explained in section II-C5.

Ref. [50] did not examine personality prediction with LIWC but investigated the correlation between LIWC categories and personality assessments. As personality inventory the California Adult Q-set by [84] was used with 181 US American students. The LIWC version 2001 categories were calculated on an one hour transcribed interview. They showed that multiple categories correlate both with self- and third party personality ratings. Especially less frequently used categories contribute to good results. Therefore, they should not be excluded when predicting personality, regardless for which inventory [50].

So far, many studies relied on offline data provided by students. Ref. [18] confirmed these previous findings for online communication. 694 English speaking blogs were included. Used personality inventories were both NEO-Five-Factor Inventory (NEO-FFI) for the broad Big Five dimensions and NEO-PI-R for the finer 30 facets. Applying the 2001 version of LIWC confirmed correlations between its categories and personality scores. Interestingly, even single words showed high correlations with some of the dimensions, which indicates that some variance is lost due to LIWC’s categories [18].

It could be shown that personality prediction works also on data from Facebook and Twitter, where status updates typically have a shorter length than in blogs [85], [86]). Ref. [85] used LIWC and other platform specific features, such as the number of followers and mentions, to predict the Big Five personality for Twitter users. With 50 English speaking users, the MAE for the regression was in the range of 11% up to 18% [85].

A good source for personality scores linked to social media profiles was created with the MyPersonality project by [87]. It enabled studies such as [72] and [88]. The MyPersonality project asked Facebook users to complete multiple personality inventories, among others the NEO-PI-R, and asked for permission to use the user’s social media profile for research. Over two million people mostly from the USA and the United Kingdom agreed to share their data. The data was available for researchers until May 2018 [87], [89]. Ref. [72] selected almost 75,000 subjects from this data source to compare LIWC with an open vocabulary approach in terms of predicting personality. As a quality measure for the ridge regression models the authors provided R values, meaning the square-root of the coefficient of determination. For LIWC R values range from 0.21 to 0.29 among the Big Five factors. Their open vocabulary approach performed better with values from 0.31 to 0.42 [72].

Currently, the only publicly available source for predicting personality based on German texts and LIWC is the commercially available Receptiviti API [90], [91]. Supported languages are English, Spanish, Dutch, French, and German and values for the Big Five dimensions as well as the facets are provided [91]. Since Receptiviti is mentioned in the official documentation for LIWC, it can be assumed that it is related to Pennebaker’s research, but no documentation is provided to explain in detail how the service derives a user’s personality based on LIWC [20]. Ref. [23] took this as a reason to examine the black box model of the Receptiviti API. They compared the Receptiviti’s results with known Big Five results from self collected datasets and the MyPersonality project. Reported MAEs ranged from 15% up to 30%. The authors state that at that time, other researchers reached MAEs mostly lower than 12.5%. Additionally, reported Pearson correlation coefficients ranged from 0.249 to 0.412 [23]. Although the reported error rates are not state of the art, they support the fact that in general LIWC results are translatable and work cross lingual, meaning models for one language can be used for another language as well [70], [92].

Summing up, this section described multiple successful approaches of deriving personality from written texts with the help of LIWC. Most authors provided some sort of quality indicator that enables comparison across existing and possible new approaches. It shows that LIWC produces useful results but leaves room for improvement. Additionally, the only approach that allows third parties to explore and use its results is Receptiviti, which provides almost no documentation for their used models. All approaches did not publish their used source code, used input data, or the trained models in some form. This reduces accessibility and reproducibility for existing research.

2) GloVe BASED APPLICATIONS

Improving the open vocabulary approach by [22], [72] compared the performance of LIWC based personality
predictions with GloVe based predictions. Their research is the basis for IBM’s cloud service Personality Insight\(^3\) [22], [27], [93]. This service supports the languages Arabic, English, Japanese, Korean and Spanish [29]. They collected tweets and Big Five personality data from 1323 English writing participants. During model creation performance was optimized in such a way that the prediction is valid for users with few tweets only. Predictions are on a scale from 0 to 1. To translate users’ tweets into numerical values, the GloVe vector for each word is determined and the average over all vectors is calculated. This single vector serves as input data for each user for Gaussian Processes as the used machine learning algorithm [22]. The authors compare this model with a Ridge Regression based on LIWC and a Ridge Regression based on a custom open vocabulary approach as introduced by [22], [72]. For the GloVe based model average Pearson correlations of 0.33 are reported, which is 0.33% better than the second best model for this specific data set. Reported MAEs are in line with previous results being in the range of 10% to 15% for all three models. With GloVe 92% of the users’ words are included, in comparison to only 62% for LIWC [22]. Just like Receptiviti, Personality Insight is closed source and commercially available for predicting Big Five dimensions and their facets. There are no details given about the exact models and parameters used for predicting personality, but a demo application exists for testing the service [94]. IBM reports an average MAE of 0.12% and an average correlation of 0.33 for the English language, which are the same values as [22] report and which is an improvement compared to Receptiviti’s results [93].

The best example for transparency is TwitPersonality by [17]. They provide their used source code, describe model parameters in detail, and at the time of publication the used MyPersonality data set was easily available for other researchers. However, the authors did not publish the trained models in themselves and did not provide a public website or API, as Receptiviti and IBM Personality Insight did, which is helpful for testing and using the model. Model creation is based on 250 English speaking Facebook users from the MyPersonality dataset. Eventually, the evaluation is based on a self collected sample of 24 Twitter users. Used model types are a SVM in comparison with other simpler regression algorithms. The SVM performs best for predicting personality on a scale from 1 to 5 with the MSE ranging from 0.33 to 0.71 [17]. Converting text into vectors works similar to the approach by [22], but instead of just calculating the average of all vectors for a user the maximum and minimum are included as well, resulting in a vector with a dimension of 900 instead of 300. Experimental results confirmed this approach. Instead of GloVe the authors chose FastText by [95] as a word embedding, which results in a word coverage of 95.08%. During the data processing some words are excluded, so word coverage is not directly comparable [17].

Using word embeddings instead of LIWC for automatically predicting personality based on text input improves both accuracy and flexibility. Word embeddings are open source, meaning applications and research can be conducted independently of commercial suppliers. Additionally, word embeddings provide a higher word coverage than LIWC and are available for a higher number of languages [95]. However, currently researchers who want to use these results, for example, in social studies, have to rely on commercial suppliers such as IBM, although their model creation process is not fully transparent. Research conducted in German cannot benefit from these improved results at all, as the language is not supported.

III. DESIGN AND IMPLEMENTATION OF A PERSONALITY MINING TOOL

A. BUSINESS UNDERSTANDING

1) REQUIREMENTS

a: FUNCTIONAL

At the beginning of every project the requirements and goals need to be determined. The overall three goals are defined in section I-B. Those will be broken down into concrete requirements, which will be used to evaluate the final solution (section III-G). The requirements needed to fulfill the overall goals are described in the following sections, beginning with functional and followed by technical requirements.

An overview of all requirements can be found in table 7 in the appendix.

At first, a user should be able to get a personality prediction based on a provided public Twitter user name. This is essential functionality for the web application mentioned in goal 3. The application should be easy to use for end-users of the frontend and setting up the whole system should be reasonably quick. Therefore, detailed documentation should be provided, which supports both goal 2 and 3, because good documentation makes independent reviews and improvements to the software easier [96], [97]. Supplemental to personality predictions, the application should return some kind of word coverage statistics, to enable the user to interpret the reliability of the results.

Data used for training and configuring the system should be made publicly available in addition to the code, as long as it does not violate any terms of use for used services such as Twitter [98]. This supports goal 2. If necessary the web application should have its own terms of use to prevent any misuse [99].

Regarding performance the response time for predicting personalities should be less than a minute. It might be useful to provide some sort of progress bar to reduce the perceived waiting time. This increases user acceptance and assists goal 2 [100].

The time spent on the model training process does not matter, as in the end the final trained model is provided. However, it is important that it is possible for other scientists to retrain models with new input data. That way, they can

\(^3\)Website: https://www.ibm.com/watson/services/personality-insights/ and https://personality-insights-demo.ng.bluemix.net/.
improve the proof-of-concept model, which is an outcome of goal 1 [98].

Finally data retrieval, cleansing and preparation should be automated both for the model training process and the final trained model. This supports the above mentioned requirement of providing just a Twitter user name and retrieving the final personality prediction.

b: TECHNICAL

In addition to functional requirements, some technical requirements can be specified derived from standards for high quality software and the described goals. A necessity for goal 2 is the publication of all written code and documentation under a suitable license that allows independent changes and improvements to the software. Used technologies should be used in their latest, stable version to ensure broad compatibility. Niche technologies should be avoided, if possible, to address a large share of researchers, which increases accessibility. To increase maintainability and flexibility of the software a modular architecture should be chosen. For example, separating frontend and backend is a common approach. This is also supported by having an easy way to configure the software, instead of hard coding procedures and variables. Modularization is also supported by object oriented software development. At last, concerning security, encryption should be used wherever possible and viable. Common measures to protect software systems should be implemented, such as a prevention against bots exploiting web applications [97].

2) SOLUTION OUTLINE

Based on the previously defined overall goals and requirements the following section will outline the technical solution that will be implemented. Since machine learning techniques are involved the programming language Python is used (see also section III-B1). There exist countless additional libraries, so called modules, for Python that extend the functionality. With scikit-learn exists a powerful machine learning library. Another advantage is that Python is open source [101].

Basically there are four outcomes of implementing this personality mining tool. The core outcome is the trained model itself for predicting personality. This model and all steps that are needed to leverage it will be bundled in a new Python module called “MiningPersonalityInGerman” MiPinG. This module can be reused in the third outcome in the shape of a web application, which provides easy access to the model’s capabilities. The last outcome bundles all steps for controlling the machine learning process. It is separated from the module to keep the module’s function as reusable as possible. Basically the fourth outcome just calls functions of the module in the correct order and with specific input data. Section III-B2 explains the technical architecture in detail. As outlined in section I-C three steps are necessary for this specific model training. These steps are implemented in section III-D.

There exists no public database that matches Twitter users with Big Five personality profiles. Therefore, this data has to be obtained otherwise. Since written personality tests take a long time, it is difficult to collect enough data needed for machine learning approaches. For a proof-of-concept model, it might be sufficient to have approximate data. If the solution works as intended, future work can focus on improving the accuracy with better input data, which is a common approach in design science research [33]. With IBM Personality Insight there exists a publicly available application that is able to predict the personality for English speaking users based on their posted tweets [93]. Deriving the LIWC categories from the used tweets in this context results in a dataset with corresponding LIWC values and personality predictions. These can be used as input for the first model training. With this model, it is possible to predict users’ personalities based on the LIWC categories of their written tweets. Since the English and German LIWC dictionaries are inter-operable, the trained model can be leveraged to obtain German personality data based on their tweets [70]. This results in a dataset with German texts and their corresponding personalities. Although there exists a service that is able to directly derive personalities via LIWC for German users (as explained in section II-D1), IBM’s service is used because the scientific documentation is better, the expected performance is higher and the models are based on tweets, which is the same domain MiPinG is intended for. In the end, the German personality data can be used to train a GloVe based model, which then would provide a direct way of predicting German personalities without using IBM Personality Insight. The modeling process is described in section III-D.

In section III-C all used input data is described. It is explained how the data is retrieved, in what format it is provided and what measures are taken to prepare it for model training. Additionally, the evaluation data is described which will be used to objectively rate the model’s accuracy. Section III-F describes the process of persisting the model itself and how the web application is implemented. The actual evaluation is carried out in sections III-E and III-G.

B. TECHNICAL ARCHITECTURE

1) USED TECHNOLOGIES

During the development of the personality mining application several common technologies are used. According to the requirements in section III-A1 those technologies that are up-to-date, wide spread and ideally easy to use should be chosen.

The main programming language is Python. According to the TIOBE index Python was the third popular programming language in May 2020 [102]. Although measuring popularity for programming languages depends on the context those are used in and the methods used to measure the popularity, they provide an indicator for the distribution of each language [103]. Python is developed and published under open source license, therefore allowing both scientific and commercial usage. It is an interpreted language, saving separate compiling steps during the development. One of Python’s huge advantages is its platform independence and an
extensive number of third party libraries, which are open-source as well and can easily be imported into an existing Python project [101].

One of those libraries is Scikit-learn\(^4\) which contains many features supporting machine learning projects [104], [105]. Scikit-learn is used in most steps along the CRISP-DM process, which will be explained in detail in the upcoming sections. One of Scikit-learn’s useful features are so-called pipelines. They describe a processing graph the data takes during the modeling steps. For example, a pipeline can be used to calculate all necessary features at first and afterwards apply some kind of feature scaling. The pipeline ensures the correct order of all calculation steps. It is also possible to nest one pipeline into another, creating complex but easy to maintain logic. Pipelines also help to avoid mixing training and testing data, for example, during k-fold cross-validations. Scikit-learn offers ready-to-use algorithms for many common classification and regression models, such as Linear Regression, Support-Vector-Machines or Decision Trees, as well as unsupervised methods like k-Means clustering [106]. Although it provides methods for neural network models, it is not suitable for large scale deep learning. This is a trade-off for its simplicity [107]. Since the goal is to provide a proof-of-concept, the models provided by Scikit-learn should be sufficient. Further improvements of the resulting model could also consider deep neural networks.

For the personality mining application data is collected from multiple platforms. This is accomplished via so-called Application Programming Interface (API). An API separates the call and implementation of functions. Input and output format are specified, so that users of a software function are able to neglect the specific operations behind the interface. Offering a certain set of API methods limits the ability of users to deviate from standard procedure, which often is a desirable outcome. APIs are used to enclose protocol implementations as well, in order to make software development easier [108].

Another important library for the MiPinG application is one to communicate with Twitter’s API. There exist several official and community libraries for Python [109]. In this context Tweepy\(^5\) is selected, as it offers all needed features in one library and is easy to use [110]. The only information Tweepy needs for initialization are the API keys generated on Twitter’s developer platform granting access to the API itself [111]. Thus, focus can be set on the actual data acquisition process. The main access point for Twitter data are public APIs. The company divided its APIs into free to use and different fee-based interfaces. The main difference between those are the amount of data that can be retrieved in a certain time window, some exclusive search features, and the ability to search tweets as far back as 2006, when the platform was established [112]. The free interface offers searches up to 7 days ago [113]. To make the data collection process as reproducible as possible only the free to use APIs will be used. As described in section III-C1 the 7 days limitation does not affect the collection process as the search API endpoint is not used during this process.

The major part of the APIs are based on Representational State Transfer (REST). In addition to that there are streaming APIs that utilize the streaming option of the Hypertext Transfer Protocol (HTTP) [112]. REST is a HTTP based interface design style optimized for distributed client-server communication. REST is not a fixed standard, so each implementation has its own technical specifics. One main property of a RESTful API is its stateless design. That means each request from a client to the server has to contain all information the server needs to fulfill this request. Statelessness simplifies the API, as the server does not need to manage client specific data. Moreover, every request could be processed by a different server of a cluster, as all information needed is included in that request. This supports scalability and performance [114].

Using REST each resource gets a Unique Resource Identifier (URI) that references to one specific resource at any given time, meaning it does not change. Resources can be any media, files, or other data objects stored on the web server. Since HTTP is a basic protocol spoken by the web server, Via standard HTTP verbs like POST, GET and DELETE a tweet can be posted, retrieved and deleted using its URI [115]. Figure 5 shows an example tweet by Barack Obama, former president of the USA.

![Tweet example barack obama source: [116]](https://twitter.com/BarackObama/status/121242416716247040)

The URI of this tweet is [https://twitter.com/BarackObama/status/121242416716247040]. So calling this URI with a GET request, always results in that specific tweet until it is deleted. After deletion that URI will not be reused. This URI can be used across all of Twitter’s API endpoints, so users can not only GET that tweet, but also POST a like, retweet, or reply. Other endpoints of the REST API enable, for example, searches with keywords or time spans [117].

One trade-off of RESTful APIs is the additional overhead in network traffic because every request has to contain all necessary information. This is especially true if large volumes of data are transferred via repeated requests [114]. Therefore,

\[4\] Project website: https://scikit-learn.org/

\[5\] Project website: http://www.tweepy.org/

---

\(^6\) An actual API call would be an authorized GET request to the API endpoint (https://api.twitter.com/1.1/statuses/show.json?id=1212422416716247040), but to keep the example simple the non-API URI is used.
Twitter has separate streaming APIs enabling real-time acquisition of tweets.

Across all of Twitter’s APIs the main response format is **JavaScript Object Notation (JSON)** [118]. Although it originates from the programming language JavaScript, it established itself as a text-based data exchange format in the context of web applications. In many APIs the response is sent via HTTP with the response body as JSON format. It consists of key-value pairs that are readable by both humans and machines. Each value can again be another JSON document creating a complex hierarchy with an arbitrary number of levels [119]. Using the same tweet as shown above an example of the JSON format is shown in figure 6 as it would be returned via an API call.

Each JSON document starts and ends with curly brackets ({}). For better readability the key of each key-value pair is highlighted in light blue text. The first key is `created_at` followed by the timestamp the tweet was posted. Line 5 holds the actual tweet text. Line 19 shows the key `user` which again is a JSON document, indicated by the curly brackets. When working with Twitter’s APIs all tweet and user objects are represented in JSON format, therefore all relevant fields for personality mining have to be extracted as a first step.

One important fact to consider are the rate limits enforced by Twitter. Ignoring these can lead to a blacklisted application, which will block all further communication with the API. Usually these limits are tracked over a 15 minute window. For example, an application can make 900 requests per time window to get tweets by their IDs. With each request up to 100 IDs may be retrieved, effectively limiting applications to get a maximum of 90,000 tweets via this API endpoint in 15 minutes [120]. Streaming endpoints only have rate limits for failed connections, but do not limit the number of tweets to retrieve. The disadvantage is that only current activities are streamed and no historic content can be accessed via streaming [121].

The used library Tweepy helps to automatically enforce rate limits for each API endpoint [122].

**YAML Ain’t Markup Language (YAML)** is another data serialization language very similar to JSON with focus on human readability. In fact JSON is a subset of YAML, meaning every JSON file is also a valid YAML file. Figure 7 shows an example of a YAML file with comments marked with the number sign (#). Since YAML is supported by many modern programming languages, it is often used as a configuration file [123]. Configuration files enable users to control the flow of an application without having to change any actual source code [124]. For sensitive data, like private API keys, it is common practice to avoid configuration files and use environment variables instead [125].

![FIGURE 6. Tweet JSON example.](image6.png)

![FIGURE 7. YAML file example.](image7.png)

In the context of deploying the MiPinG application the software **nginx** is used as a web server [126]. Besides the Apache HTTP server** nginx** was the most used web server in a survey in June 2020 [127]. Since nginx has no Python capabilities, **Gunicorn** is used as a Web Server Gateway Interface (WSGI) server handling the communication with the actual Python program [128]. WSGI is a Python standard that specifies the way of communication between web server and Python web applications [124]. In the Python application itself the library **Flask** is used to implement the other part of WSGI. Via this a custom REST API can be implemented [129]. Details of this structure are explained in section III-B2. The full list of used Python libraries and its versions are listed in the appendix in section V-A.

2) APPLICATION STRUCTURE

As described in the previous section the web application consists of several parts. In the end, this web application serves as a proof-of-concept of how the trained model for personality prediction can be used in practice. Frontend and backend are separated to easily change the layout of how the personality data is presented or processed further. Moreover, the web server provides an additional layer of security e.g. against malicious requests. Figure 8 shows the technical context diagram of the MiPinG web application and the interaction of the different technologies.

Nginx (shown in orange) is always the first communication partner for potential users. It delivers static HTML and JavaScript files. Requests to the actual Flask REST API implemented in Python (purple box) are proxied by nginx to Gunicorn (backend box). Gunicorn automatically translates this request for Flask. The Flask code is the first part of the actual Python MiPinG module. It handles the request and leverages the other parts of the module to, for example, collect tweets for a given user name and derive the personality afterwards. The response will be sent in JSON format. Via JavaScript the client’s browser will present the result.
appropriately. An example of how to use the application is described in section III-F. The structure of the Python module is described in the next section III-B3.

3) MODULE STRUCTURE
The Python module is the actual MiPinG application. Figure 9 shows a simplified directory structure of the module. Folder names are presented in bold letters. The LICENSE file contains the text for Apache License Version 2.0, a common license allowing flexible use and extension of the software both scientifically and commercially [130]. The README.md file is a common way to provide documentation and explanation for the present software. Docs contains additional, detailed documentation for using and changing the MiPinG application. All used and necessary libraries, e.g., Tweepy, are listed in requirements.txt. This is good practice for Python projects [124].

In order to get the fully trained model for personality prediction, the one-off training process has to be executed. This is done with the code file main.py, which loads the configuration from config.yml. Since this process involves several steps, additional classes are defined in helper. For each step of CRISP-DM a class is created. Wherever possible classes and functions in the miping folder are used to increase modularization. Data necessary to reproduce training steps is saved in data.

The miping directory contains all reusable components. The trained models of the one-off process and the code to apply them are saved in application and trainedModels. Interfaces consists of several classes necessary to communicate with different APIs, such as Twitter, Google Maps, IBM, and the LIWC software. Different data structures used throughout the training process are defined under models. Most important is the profile model, which contains both tweet and personality data for a user. The data collection process is described in section III-C. Under directory webapp the Flask classes can be found that implement the REST API mentioned in III-B2.

End users solely interested in getting personality predictions need code from the miping folder only. The code in main.py is relevant for those who want to analyze or improve the trained models.

C. DATA UNDERSTANDING AND PREPARATION
1) TWITTER DATA
a: USER SELECTION
Twitter is a microblogging platform founded in 2006, where users share short messages (so called tweets) of 140 characters (since 2017 up to 280 characters) with the world [131]. In 2019 it had 330 million monthly active users [132]. The biggest social network Facebook had around 2.38 billion monthly active users at the same time [133]. Twitter established itself as a major source of real-time information both on a local and global level. Users share their personal thoughts, interact with each other and discuss current events [134]. Twitter’s diversity in topics, high usage numbers and the fact that many tweets are publicly accessible make it a good source for social studies [135].

At first, suitable Twitter users need to be identified and selected. Afterwards their tweets need to be collected. The goal is to select roughly 1000 of each German speaking and English speaking users. This sample size is bigger or about equal in comparison to existing research [17], [22]. Initially, the selection aims at 1100 users to be able to strictly sort unsuitable users out.

There are multiple requirements regarding the users:
1) User accounts need to be public.
2) Users need to write tweets on their own, especially not via a social media team.

10The code is available at Github: https://github.com/iUssel/MiningPersonalityInGerman.
3) Users should have enough tweets in their timeline to have a broad basis for prediction.

4) Select only native speakers.

Only for users who meet these requirements the personality can be deducted from tweets. Twitter offers the option to restrict account access to be private allowing only followers to read their timeline. Those users will be excluded in the selection process. It might be that users with a public profile generally have a different personality than private users, but this is negligible since the resulting model is never applied to private users. Furthermore, only native speakers understand the small differences in meaning of words, which are essential in this context. The used words in their tweets need to stem from themselves, which would not be the case if a social media team is involved. To ensure enough tweets in a user’s timeline only users with at least 400 tweets are included. This number also includes retweets, so it is just an indicator of how many original tweets a user wrote. The exact number will be checked later. Additionally, only users with a maximum of 10,000 followers are included during the user selection. This arbitrary limit should help to avoid very popular accounts, which tend to have social media teams involved. This also excludes many company accounts, like news outlets or manufacturers, which often have more than a million followers. For a higher selection quality, the automatically selected users are manually verified as a last step.

For model training both English and German native speakers are required. The process to identify these is based on two steps. First, only tweets from a specific country are collected. Twitter offers a streaming API, where tweets can be filtered by GPS coordinates [136]. This selects only tweets from users which allow Twitter to use their location while posting and which coordinates fall into a specified rectangle. A few thousand tweets are collected with this streaming filters for each country. The above minimum requirements and limits for the total number of tweets and followers apply. Retweets are excluded from this process. Since this process collects only users who share their GPS coordinates (GPS users), there is a risk of selection bias. It might be that exactly those users, who share their location, have certain similar personality traits, e.g. higher Extraversion values, which would distort the training data for model training.

In order to get a more random selection of users, some of their followers are included. Twitter’s rate limit for retrieving a user’s followers are relatively strict (15 calls per 15 minutes returning 5000 followers per request), therefore followers from 30 random GPS users are retrieved. For each user only the most recent 5000 followers are fetched. It might be the case that some of these followers already exist among the GPS users, so duplicates are removed. Tweet numbers and follower numbers are also checked. To obtain the final list of 1100 suitable users, 200 GPS users and 900 of the followers will be mixed. The location of the selected followers has yet to be verified, since they have not been retrieved via the GPS selection process. Twitter profiles have a user filled field called location which is used to verify their location. Some users provide nothing, fantasy or not meaningful places, but many others provide a string like “LA, California” or something similar. The Google Maps API is used to parse and validate this string. Having the location of the selected users verified, their used language needs to be verified. Tweeting from the US does not mean that all tweets are written in English language. For each tweet Twitter delivers a language field describing the language used in this tweet. This field is filled via machine learning processes. If multiple languages are used in one tweet or the language cannot be determined, the result is undefined [137], [138].

This is the case if a user posts only emojis, an image or short phrases that are common in multiple languages. Hence, a user’s timeline consists of tweets in the native language, in other languages and tweets with undefined languages. It is not probable that many users exist with 100 % tweets in their native language. Therefore, users are identified as native speakers if 80 % of their tweets are written in the actual language expected for that country (e.g. English for the USA) and the share of foreign language tweets is not bigger than 5 %. That leaves 15 % for undefined tweets as described above. For each user up to 250 tweets are retrieved. Due to API limitations this includes retweets which will be filtered out. Only users are selected whose timeline matches above criteria for native speakers. Although it could be the case that real native speakers are excluded with these criteria, it is more important to exclude non native speakers to have high quality training data. As a last selection step, all automatically selected users are manually checked to exclude those that are obviously not a personal account. Figure 10 shows an example of an excluded account because it is managed by an institution, association, or political party. This reduces the German user number from 1088 to 785 and the number of American users is reduced from 1093 to 998.

Summarizing, for all the next steps only tweets are selected which are written in the native language. Via this selection process the data basis for all upcoming steps is determined. Only users with a minimum number of tweets and
a maximum number of followers are considered. Native speakers are identified by the given locations and a high percentage of tweets in the target language.

b: TWEET TEXT PREPARATION
For all tweets the following attributes are saved: unique id, creation timestamp, corresponding user id, flag indicating if it is a retweet, actual text body, language and flag indicating if it is a response to another tweet. Both user list and tweet list are exported as CSV files, to skip the selection process in consecutive program runs. To prepare and clean the Twitter data, all tweets are combined in a single string for each user. All characters are converted to lower case. URLs are removed with regular expressions, as they do not add any value during text analysis [17]. Since hashtags (”#”) are an essential feature of Twitter, often carrying meaning for a post, they are not removed. Instead the hashtag symbol is padded with spaces so that “#hashtag” becomes “ # hashtag”. That enables LIWC and word embeddings to recognize the word correctly and at the same time the hashtag symbol is included in the machine learning process. It could be that excessive usage of hashtags indicates a certain personality trait. Other punctuation characters are padded with spaces as well. Whereas mentions (indicated with “@username”) are removed from texts, as the username is arbitrary and therefore has no deeper meaning in this text analysis. In the end, consecutive spaces are removed leaving only one space between words. At this point, all user profiles are filtered out that contain less than 600 words. This is a threshold under which IBM Personality Insight returns a warning that not enough words are provided [93]. This affects only a few users.

Table 1 shows information about how many words and tweets are included. As the number of tweets was artificially limited, the maximum for both countries is 250 tweets. The average number of tweets is almost identical with a similar standard deviation. There are just a few outliers with fewer tweets, as can be seen by the first and third quartile. The word counts for both countries are similar, although Germany has a higher mean of 5466 words. This might be due to cultural or language differences and not necessary because of different personalities. Standard deviation is relatively high, but the minimum number of words is well above the limit of 600 words for both countries.

2) LIWC DATA
The LIWC software comes as a standalone package. There exists a LIWC API via a website but access to it was limited at the time of data preparation and had to be requested via an application form, which was not granted. Therefore, the standard LIWC software is used [90], [140]). This is the only processing step that cannot be integrated into the overall architecture. LIWC comes with several built-in dictionaries: three versions for the English language and the latest 2015 version for the German language, which was developed by [70]. For further analysis the 2015 version is used for both languages. To get all LIWC output variables, the previously exported CSV files are used as input for the LIWC program. Each user’s text, the single field containing all posts of a user described in the previous section, is saved in one column of the file and each row represents one user. In total there is one file for all German users and one file for all American users. According to the recommendations the files are UTF-8 encoded. Special German characters like “ä”, also written “ae”, are included in both spelling variants, as well as common abbreviations [70]. Automatic spell checking and correction will not be applied to the input text, as LIWC is robust to natural occurring spelling mistakes [141]. That means there are no further text preparations before the LIWC categories are calculated. The German LIWC dictionary has some additional categories in comparison with the English version. These regard the differentiation of personal pronouns between formal and informal use (“Sie” versus “du”) [70]. Since the English dictionary is the common basis for modeling personality, these categories are ignored. The results are again saved as CSV files, in order to import them into the Python program. All category values are added to the user’s individual profile object. This intermediate result is again exported as CSV files, to be able to skip the LIWC category generation process.

3) IBM PERSONALITY INSIGHT DATA
IBM’s Personality Insight API is a REST API similar to Twitter’s API. There are no ready to use modules available to control the communication, so the HTTP requests and responses have to be submitted and analyzed directly. Basically, the API takes UTF-8 encoded text, some meta and authentication data as input and returns the user’s personality profile as a JSON response. The service offers 1000 free requests per month, which is sufficient for the English users. German users are not relevant for this step, as IBM does not support their language [27]. The API returns the total word count of each analyzed text, the processed language and Big Five personality traits as percents. Included in each of the traits are additional six facets which are more detailed. An example of the response can be seen in figure 11. In the end, each response contains a warnings field, indicating, for example, if not enough words have been provided as input. Since only users are selected whose tweets make up more than 600 words, this warning is never encountered [29]. The JSON response is parsed and the percentile information of each trait saved in the user’s profile together with the basic information, the input text and the LIWC categories. The collection of all profiles is again exported as CSV files.

Figure 12 shows the boxplot of the Big Five dimensions for the American users as retrieved via the IBM API. For all dimensions we can see a maximum value of 1 and a minimum of 0. The upper bound of the boxes indicate the 75% percentile and the lower bound the 25% percentile. Additionally, the median is indicated by the horizontal line. Across all dimensions boxes are roughly equally sized and except for dimension Openness the median is close to 50%. This is an expected distribution since Big Five personalities...
TABLE 1. Number of Tweets and words in German und U.S. Twitter data.

|          | Minimum | Q1    | Q3    | Maximum | Mean  | Standard Deviation |
|----------|---------|-------|-------|---------|-------|-------------------|
| # Words  | USA     | 691   | 2517.75 | 4202.23 | 13673 | 4436.1            | 2039.3 |
|          | Germany | 755   | 4162.0 | 6567.0 | 16927 | 5466.3            | 2136.9 |
| # Tweets | USA     | 53    | 250    | 250    | 247.3 | 16.6              |
|          | Germany | 31    | 250    | 250    | 247.0 | 18.6              |

FIGURE 11. Part of IBM JSON response.

are normally distributed and in a random sample this distribution should be seen as well. The slightly higher tendency for the dimension Openness, could indicate a selection bias in the user selection as explained in section III-C1. On the other hand, it is possible that all public Twitter users generally have a higher Openness value, therefore the risk of a small selection bias in this input data is accepted for the time being.

4) GloVe DATA
As explained in section II-B2, the GloVe word embedding needs to be created from suitable input data before it can be applied to a specific problem. Since huge amounts of data need to be processed for this creation, time can be saved by leveraging already existing vectors. In [76], the creators of GloVe, provide pre-trained word vector files derived from the English Wikipedia, news articles or Twitter posts, but no vectors for the German language. Ref. [142] take the German Wikipedia content to create trained GloVe vectors. Their used code is open-source and apart from preprocessing the Wikipedia texts the model training is identical to the one by [76], [143].[11]

The word embedding is saved as a plain text file containing about 1.3 million entries each including a word and its associated 300 dimensional vector. When deploying the final model to a web server the GloVe data will be part of that deployment. Main memory is a significant cost factor for servers [144]. Therefore, the text file is inserted into a simple database because via this handling the data will require less main memory. Now, converting a given German word into its vector representation can be achieved by a simple query to the database.

5) SURVEY EVALUATION DATA
Since the personality data for the German users is indirectly deduced via the English users’ personalities, it is important to collect evaluation data reflecting a real world sample. A survey based on the publicly available BFI-2 by [54] is conducted among German Twitter users. The BFI-2 consists of 60 questions with a 5 point Likert-scale, which are listed in the appendix V-C. Besides the Big Five domains the inventory captures 15 facets. For evaluation the facets are ignored. Ref. [54] validated their instrument in a study with a sample size of 1224. Additionally, norm values for all Big Five dimensions are supplied, enabling a conversion from raw values to T-scores and percentile ranks. These norm values are further separated by sex and age [54].

In the end, 15 users participate in the survey for collecting evaluation data. This emphasizes the difficulty of collecting enough data as input for machine learning projects. Also [17] use only 24 users in their evaluation. All users except one are male. About half of the participants are aged below 30 years. Details can be found in section V-D in the appendix. Using the individuals’ answers to the survey and the norm values, a percentile rank for each domain can be derived. Figure 13 shows the boxplot of the Big Five domains for these 15 German users.

11 Website: https://gitlab.com/deepset-ai/open-source/glove-embeddings-de.
Since the sample size is relatively small, it is expected to show deviations from a normal distribution. The dimensions Openness and Agreeableness appear to have a similar range as the English users’ personalities. It is noteworthy that the dimension of Neuroticism shows considerable lower values, but the small dot in the upper right corner indicates one extreme outlier with very high Neuroticism. This personality data is used in section III-E to evaluate the real world performance of the created models. The small sample size and distorted distribution has to be kept in mind for later discussion.

6) FEATURE EXTRACTION

In the previous sections all relevant input data for modeling has been described. To apply machine learning algorithms the final features have to be extracted. For the first step in modeling a model based on LIWC data will be trained. The LIWC data already has a numeric format for all 93 categories. In order to ensure equal weighting across all 93 categories, a z-score scaling is applied to each field. This feature extraction is used in the second step as well, in order to derive the German users’ personalities.

For the third modeling step the Twitter posts need to be converted into numerical values via the GloVe word embedding. For each user exists one field containing all prepared texts from the Twitter posts. This long string is split into tokens separated by space. Each word and each punctuation or other special character forms a separate token. Next, each token will be looked up in the previously created GloVe database. Since not all tokens will result in a match, the coverage of matched tokens is calculated in this step as well. Once all tokens are replaced by their corresponding 300 dimensional vector, the final feature vector for each user is calculated. This is achieved by calculating the average, the maximum and the minimum values for each dimension separately and combining the result in a 900 dimensional vector as described by [17]. No scaling is applied to the final features, as information is encoded in the vector differences of GloVe and scaling would deteriorate this encoding. The average word coverage for all 785 German users is 71.1%.

D. MODELING PERSONALITY

1) REVERSE ENGINEERING OF IBM PERSONALITY INSIGHT WITH LIWC

For this first modeling step, the features generated by LIWC for the 998 American users are used to create a model that predicts the Big Five personality dimensions. The target variables are defined by the Big Five dimensions returned by IBM Personality Insight. The goal is to find the best model type and its optimal tuned parameters for good performance across all dimensions. Following model types are considered for model selection, as they are mentioned in the existing personality mining literature (see section II-D) or are common basic models in data mining (see section II-C5b):

- Bayesian Ridge
- Decision Tree
- Gaussian Process
- Random Forest
- Ridge Regression
- Support Vector Machine

Although it would be possible for some model types to incorporate all five dimensions into a single model, in favor of reduced complexity and higher flexibility there will be a separate trained model for each dimension. As part of the model selection an exhaustive grid search is performed. The initial parameters for each model are listed in the appendix in table 9. After each iteration the parameters will be manually adjusted, in order to get closer to a local performance maximum for each model type.

The following steps are repeated for all five dimensions. For each model type the exhaustive grid search uses the RMSE as a metric to identify the best parameter options during a 10-fold cross-validation. This results in the currently best tuned model for each type. Via another cross-validation the mean RMSE and the doubled standard deviation of the RMSE are determined for each best tuned model. Both scores help to compare the models among each other. Overfitting can be identified by a relatively high standard deviation, as this indicates the model is strongly fitted to each training set. From those best tuned models for each type the best model for the current Big Five dimension is selected. To compare this model’s performance with the existing approaches from the literature the MAE, MSE, RMSE and $R^2$ are calculated. Additionally, the average RMSE over all five dimensions is calculated for each model type for easier comparison.

After four iterations, the optimal parameters for each model type have been identified and each model’s performance has converged. The final parameters are listed in the appendix in table 10. In table 2 the average scores of all model types are listed. For each iteration the best model type is the Support Vector Machine (bold highlighted) with
a RMSE of around 0.154. All models are already close to their final score after iteration 1, indicating that the first grid search already yielded parameters close to their optimum. A small decrease in performance can be observed for some models because parameters have been adjusted for a smaller chance of overfitting. This is especially true for Decision Trees and Random Forests. Their first iteration has been conducted without any pruning, in contrast to the last iteration which produced similar results with some degree of pruning. Detailed results separated by each dimension are listed in the appendix in tables 11 to 15.

As a result the Support Vector Machine is selected as the best model for reverse engineering IBM Personality Insight. The optimal parameters are listed in table 3 and apply for all five dimensions. It is noteworthy that the number of iterations for finding the optimal hyperplane is limited to 100. This significantly reduces training time from several hours to a few minutes without having a measurable impact on performance.

Figure 14 shows MAE, MSE, RMSE and $R^2$ for the final Support Vector Machine as calculated during cross-validation. The colored bars indicate the average scores separated by Big Five dimension (abbreviated with OCEAN) and the doubled standard deviation is shown by the black bars. Except for the dimension Neuroticism this deviation is relatively small. Ref. [23] explained that state of the art models reach MAEs around 12.5%, which the present Support Vector Machine meets. $R^2$ lies well above 0.5 across all dimensions indicating a good fit of the model.

As a last step for reverse engineering the Support Vector Machine is trained with the whole dataset. The export of the five resulting models is explained in section III-F. Using the fully trained model and the whole training dataset Pearson correlation coefficients can be calculated. All p-values are 0 and the correlation coefficient itself ranges from 0.89 up to 0.93. This is well above the values reported by [23] with 0.249 up to 0.412. Reasons for this are discussed in section IV. Summing up, this section showed the creation of a model that predicts personality based on LIWC input data. Its preliminary evaluation indicates good performance on a level similar to models in existing literature.

2) DERIVING PERSONALITY OF GERMAN USERS AS GROUND TRUTH

After creating a LIWC based model for personality prediction, this model is applied to the selected 785 German users as the second step of modeling. LIWC categories have been generated with the latest German dictionary as pointed out in section III-C2. As before, the same preparation in the form of z-score standardization is applied. In the end, the five trained SVM models predict all Big Five dimensions for each user and this information is again exported as a CSV file to be able to skip this prediction step later on. Figure 15 shows the boxplot for the derived German personal- ities. Apparently, the box sizes are smaller in comparison to the American users’ distribution, indicating a tendency of the model to rather predict a mean value. Nevertheless, the German distribution still covers a high share of the valid range of values between 0 and 1. There exist a few outliers, that are even negative or just above the maximum value of 1.0. According to the definition of Big Five percentiles, these values are invalid. Although it would be possible to apply a lower and upper limit for these predictions in order to avoid invalid values, the data is not changed because the variance should be maintained. Otherwise, another bias is being risked for training the Glove model if extreme values are capped. The invalid values have to be kept in mind for deploying the final model because for the final prediction only valid values should be returned.

3) TRAINING GloVe BASED MODEL FOR GERMAN LANGUAGE

The third step of modeling is based on the Glove features described in section III-C6 for the German users. Target variables are the personality predictions from the previous section. In general, the same procedure as for the LIWC model creation applies and the same model types are considered in model selection. The initial parameters for each model are listed in the appendix in table 16. After six iterations optimal parameters have been found. Those final parameters are listed in the appendix in table 17. As table 4 shows, iteration one already provided scores close to the optimum. For each iteration the best model type is a Random Forest (bold highlighted) with a RMSE of around 0.1024. In iteration four the Random Forest is excluded to get more detailed results for the second best model, a Decision Tree. Additionally, a Random Forest without pruning resulted in a RMSE of 0.0831 for dimension Openness. Since training time for just one dimension in this scenarios is 1.5 days, a small decline in performance is accepted for a training time of 5 minutes for the other iterations. Detailed results separated by each dimension are listed in the appendix in tables 18 to 22.

Resulting from grid search and cross-validation the Random Forest is selected as the best model to predict personality

|TABLE 2. LIWC average scores of all models.|
|---|
|**Model Type** | **Average Score (RMSE)** in Iteration 1 | **2** | **3** | **4** |
|Bayesian Ridge | 0.1655 | 0.1655 | 0.1655 | 0.1654 |
|Decision Tree | 0.2091 | 0.2091 | 0.2019 | 0.2026 |
|Gaussian Process | 0.182 | 0.182 | 0.182 | 0.182 |
|Random Forest | 0.1673 | 0.1678 | 0.1714 | 0.1714 |
|Ridge Regression | 0.1657 | 0.1657 | 0.1634 | 0.1661 |
|Support Vector Machine | 0.1535 | 0.1535 | 0.1534 | 0.1539 |

|TABLE 3. LIWC support vector machine final parameters.|
|---|
|**Parameter Name** | **Parameter Values** |
|kernel type | rbf |
|regularization parameter (C) | 10 |
|margin of tolerance (epsilon) | 0.1 |
|maximum iterations | 100 |
based on GloVe features. The optimal parameters are listed in table 5 and apply for all five dimensions.

Figure 16 shows again the error scores for the final Random Forest model. The MAE and MSE are consistently low across all dimensions with reasonable deviation. Ref. [22] reported an average MAE across all dimensions of just above 10%, meaning the Random Forest appears to perform even better. Since the German target variables are based on IBM Personality Insight, which utilizes the model of [22], the newly created model cannot be better than its origin. Possible reasons for this constellation are discussed in section IV. The Support Vector Machine trained by [17] reaches values of RMSE between 0.115 and 0.168,\textsuperscript{12} which the present Random Forest model outperforms slightly. $R^2$ shows an exception for dimension Neuroticism. This dimension showed anomalies

\begin{table}[h]
\centering
\caption{GloVe average scores of all models.}
\begin{tabular}{|c|c|c|c|c|c|c|}
\hline
Model Type & Average Score (RMSE) in Iteration \\
\hline
Bayesian Ridge & 0.1254 & 0.1254 & 0.1254 & 0.1254 & 0.1254 & 0.1254 \\
Decision Tree & 0.1257 & 0.1257 & 0.1257 & 0.1257 & 0.1257 & 0.1257 \\
Gaussian Process & 0.1274 & 0.1274 & 0.1274 & 0.1274 & 0.1274 & 0.1274 \\
Random Forest & 0.1024 & 0.0993 & 0.0994 & N/A & 0.1006 & 0.1024 \\
Ridge Regression & 0.1257 & 0.1248 & 0.1247 & 0.1247 & 0.1248 & 0.1248 \\
Support Vector Machine & 0.1277 & 0.1256 & 0.1259 & 0.1259 & 0.1256 & 0.1262 \\
\hline
\end{tabular}
\end{table}

\begin{table}[h]
\centering
\caption{GloVe random forest final parameters.}
\begin{tabular}{|c|c|}
\hline
Parameter Name & Parameter Values \\
\hline
number of trees & 50 \\
maximum depth & 5 \\
minimum number of samples for split & 2 \\
minimum number of samples at leaf & 10 \\
maximum number of leaf nodes & no limit \\
maximum number of used features & 450 \\
\hline
\end{tabular}
\end{table}
already during LIWC training. Otherwise $R^2$ are well above 0.5.

As preparation for deployment and in order to calculate Pearson correlation coefficients, the whole dataset is used for training the Random Forest with optimal parameters. All p-values are smaller than 0.001 and the correlation coefficient ranges from 0.86 to 0.92. This is again well above the range reported by [22] (0.25 to 0.42), which will be discussed in section IV. In summary, this section showed the creation of a GloVe based model for personality prediction, whose preliminary evaluation indicates good performance.

E. PERFORMANCE EVALUATION

1) VALIDATION AGAINST EVALUATION DATA

The evaluation phase is split into two parts. First, the real world performance of the model is estimated by validating it against the separately collected data described in section III-C5. These results will show potential for optimization which will be realized in section III-E2. Secondly, since some requirements identified during business understanding are referencing the deployment phase, the second evaluation part for a holistic assessment takes place in section III-G.

For evaluation the previously created model is used to predict personality scores for all 15 evaluation users. These scores are then compared to the scores obtained via the personality questionnaire. Feature extraction follows the same process as explained during data preparation. The latest tweets are retrieved for each user, cleaning transformations are applied and finally GloVe features are extracted. Figure 17 shows the error scores for comparing predictions against true values. Across all metrics and all dimensions the scores are worse than expected via cross-validation. MAE values are still comparable to the ones of the Receptiviti API described in section II-D1. But still, with RMSE values above the threshold of 20%, the ability to distinct between users of different personalities diminishes. Remarkably, $R^2$ lies in the negative range, indicating a bad fit of the model. However, it has to be kept in mind that the $R^2$ metric is not reliable for small sample sizes [145]. The same holds true for the Pearson correlation coefficient [146]. It yields values between $-0.33$ and $+0.1$, but p-values are mostly above 0.7, which indicates no significant correlation between prediction and true value.

This real world evaluation shows that the trained model might not be suitable for deployment. For a successful proof of concept, the source for this error needs to be identified. Therefore, the evaluation data is used to derive the users’ personality based on the LIWC model trained in section III-D1. This model generated the original German ground truth for training the GloVe model, so now the output of the LIWC model is treated as the true value for the evaluation users and compared to the same GloVe predictions. One user is excluded because the tweets did not contain at least 600 words. Figure 18 shows the error scores achieved via this comparison. Now, MAE and RMSE values are on the same level as during cross-validation. For most dimensions $R^2$ improved as well, except for Openness. Pearson correlations range from 0.39 up to 0.73, mostly significant on a 5% level, except for Neuroticism, which has a p-value of 0.17.

In total, this comparison shows that the general model creation, selection and training process is valid, since it yielded in a model that is able to reliably predict its ground truth data. The evaluation also showed that the German ground truth personality data, which was generated via the reverse engineered LIWC based model, does not match the evaluation data retrieved via personality questionnaires. Reasons for this
are reviewed in the general discussion in section IV. In order to create a better model, the evaluation data is used for training a new model, which is described in the next section.

2) MODELING AND EVALUATION WITH SURVEY DATA
Since the previous evaluation showed that the general model creation process is valid but the input data is not suitable, a new GloVe based model is created and evaluated using the data collected via survey as described in section III-C5.

Obviously, 14 users are not sufficient for a sophisticated model, but it demonstrates a path to obtain a better model. In order to evaluate real world performance of the new model 4 users are separated as a validation data set. These are not considered in model selection and training. Data preparation and feature extraction follows the same process as before. Target variables are defined by the personality scores obtained via personality questionnaires. Cross-validation is reduced to a 3-fold cross-validation to adjust for the smaller

FIGURE 17. First GloVe model error scores for evaluation data.

FIGURE 18. GloVe model error scores for evaluation data via LIWC.
| Model Type        | Average Score (RMSE) |
|-------------------|----------------------|
| Bayesian Ridge    | 0.2660               |
| Decision Tree     | 0.2574               |
| Gaussian Process  | 0.2642               |
| Random Forest     | 0.2609               |
| Ridge Regression  | 0.2599               |
| Support Vector Machine | 0.2645         |

For evaluation, the Ridge Regression model is trained with all 10 users. This model is used to predict the personality for the validation data set consisting of the remaining 4 users. Even after the evaluation has been conducted those 4 users must not be included in the model, as this would change the whole model and render the evaluation invalid. Figure 19 shows the error scores for this final assessment. For completeness $R^2$ are shown, although the sample size is too small for being informative. MAE values range from 0.08 up to 0.22, but they are consistently lower than the first GloVe based model. The same is true for RMSE. Previously, error scores over 0.30 have been reached, now the maximum is 0.25. Openness, Agreeableness and Neuroticism lie even significantly below 0.20, making predictions useful for real-life applications. Again, despite the small sample sizes Pearson correlation coefficients are supplied. They range from $-0.98$ up to 0.95 with $p$-values from 0.02 to 0.91.

All in all, the evaluation of this survey data based model indicates a better performance than the previous model. Obviously, this performance is currently limited by the very small sample size used for training. Nevertheless, this model is used for deployment to finalize the proof of concept of a personality mining system.

F. DEPLOYMENT OF MODEL

The following section describes all necessary steps for deploying the trained model. As a first step, the model needs to be exported because otherwise it would have to be trained again by potential users. With Open Neural Network Exchange (ONNX) exists an exchange format for machine learning projects [147]–[149]. Since there exists one model for each dimension, there will be five exported files. As explained in section III-C4, the GloVe vectors are supplied as a SQLite database file. SQLite is an open-source, free to use format that helps increasing performance and reducing memory requirements in this context [150].

To show the ease of use of the created personality mining system and to make the results more accessible, a website is created according to the architecture outlined in section III-B2. The website is available at www.miping.de. A potential user accesses the website via encrypted HTTPS. Initially, information about the Big Five model for personality and its dimensions is provided. The core functionality is to get a personality score for an arbitrary German Twitter user. Figure 20 depicts an example for the German government spokesperson Steffen Seibert, whose Twitter account name is RegSprecher. Before the process is started, the user has to agree to the disclaimer and the privacy notes, which is important to limit the legal risks for the operator of the website. Legal notice and privacy notes have been created via online generators [151], [152].

Once the user agreed to the terms, Google’s reCaptcha is loaded. This technology helps to recognize automated and malicious requests via bots [153]. Figure 21 shows the process of reCaptcha. To start the process, the user has to confirm that they are no robot. Depending on different indicators, such as saved cookies, mouse movement and the user’s IP address, reCaptcha either automatically jumps to step (c) or the user has to solve a task as shown in step (b). The reCaptcha step can also be deactivated, for example if the service is only running locally.

After reCaptcha is successfully processed, the user can request a personality score by clicking the “Send” button. This sends a reCaptcha token and the entered Twitter user name to the Nginx webserver. The request is proxied to the locally running Gunicorn backend server, which triggers the actual Python module to process the request. Initially, the reCaptcha token is verified and the format of the passed input data is checked. In this case, Twitter user names are at maximum 15 characters long [154]. Longer requests are denied, as they could be malicious requests. With the given user name, the same data preparation tasks as during training are performed. Therefore, the Twitter API is used to request the latest 200 tweets for that user, apply data cleansing and combine all tweets into a single string. This is used to derive the GloVe features, which serves as input for the trained models. The prediction for all Big Five dimensions along with the word coverage statistics is returned to the client’s browser via JSON, where it is presented on the user interface. Usually, the result is ready after less than 15 seconds during which a rotating loading icon is shown. An example of result presentation can be seen in figure 22.

The necessary configuration and Python backend files are included in the overall code repository, making it easy for third parties to host a similar website as presented above. The complete source code including the trained models is published at www.github.com/iUssel/MiningPersonalityInGerman. Because Twitter restricts sharing complete datasets retrieved via its API, only user ids are published [155]. Nevertheless, it is still possible to recreate the training process with this information. Due to its file size of 5.3 gigabyte, the GloVe database file is hosted only on the MiPinG.
FIGURE 19. Survey based GloVe model error scores.

Get Your Twitter Personality

Username

@ regspreecher

By pressing the "Agree" button you agree to the following terms: This website is offered as is by a private person. You use it at your own risk. There is no guarantee for correct results. Provide only user names of which you have consent to use it on this site. To derive the personality, public Tweets are temporarily saved on the server and immediately deleted afterwards. Please read the Privacy page. reCaptcha is loaded once you press the button.

FIGURE 20. MiPmG website - entering Twitter name.

FIGURE 21. MiPmG Website - process of reCaptcha.

Additionally, a Python package is published at www.pypi.org/project/miping/. This includes only the necessary resources for applying the created models and excludes code that would be necessary for training or improving the models. A Python package increases availability for Python developers, as access is managed via a package manager, which handles installation and updates for the user.

With all the steps described in this section, the whole personality mining system is made available to the public. An example of simple usage has been provided in the form of a website. Via disclosure of the whole source code independent peer reviews are simplified and improvements can be conducted by other researchers or users.

G. REQUIREMENTS EVALUATION

Usually, the last phase of CRISP-DM is the deployment phase. Since some of the defined requirements identified during business understanding affect the deployment itself, the whole personality mining system and its deployment are evaluated against these requirements in this section. The complete list of requirements is shown in the appendix in table 7. Starting with the functional requirements, the user should be easily able to get a personality prediction for a Twitter user name without registration on the website. Although, the
accuracy of the predictions has potential for improvement, the basic process is working. Section III-F shows suitable accuracy of the predictions has potential for improvement, H. Usselmann results. The evaluation of the LIWC based model in language independent model does not provide the expected deployed to the web application. The evaluation did not meet the requirements, the test subject data test subjects, who filled in BFI-2 questionnaires. As this Performance was measured against a small sample of German obtain German personality data as input for a GloVe model. dent personality prediction model. This model was used to and IBM Personality Insight to create a language indepen- train models as ONNX format, allowing independent use training data. None of the existing solutions in literature provided this feature. With Python a popular, modern language has been chosen. The web application runs on the latest Linux Ubuntu version 20.04.1 and supports both Firefox (version 81.0) and Chrome (version 85.0) as browsers in their latest versions using HTTPS. This fulfills the require- front end and backend are separated and connected via a REST API. Configuration is isolated in YAML files, allowing fine control over the process flow. Lastly, misuse is prevented by the integration of Google’s reCaptcha and by the fact that no personal data is permanently saved on the server. In conclusion, both functional and technical requirements are fulfilled for this proof of concept. The next section looks at limitations and discusses overall results.

IV. DISCUSSION AND ASSESSMENT OF RESULTS
In the introduction three main goals are defined. These include the creation of a proof-of-concept model for personality prediction based on GloVe, the publication of its source code, and a ready-to-use web application for utilizing the trained model. Since no German personality data is publicly available, a methodology was designed to leverage LIWC and IBM Personality Insight to create a language independent personality prediction model. This model was used to obtain German personality data as input for a GloVe model. Performance was measured against a small sample of German test subjects, who filled in BFI-2 questionnaires. As this evaluation did not meet the requirements, the test subject data was used to create a new, more reliable model. This model is deployed to the web application.

It appears that the original methodology for creating a language independent model does not provide the expected results. The evaluation of the LIWC based model in section III-D1 showed remarkably good results and outstanding Pearson correlation coefficients. One possible explana- tion for this above average performance, might be that the input data is only the output of IBM’s machine learning models. These models already smoothed the real life personality data, making it easier for another algorithm to make predictions on this data. That could also be the reason for the outstanding performance of the first created GloVe model in section III-D3, whose input is ultimately based on the output of two machine learning models, namely IBM Personality Insight and the LIWC based model. This chain of trained models might explain the high performance during model training and lower performance during the final evaluation as small errors could add up along this chain but are only detected when comparing the first input with the last output. Last but not least, combining the English and German LIWC dictionary introduces another source of uncertainty along the chain.

The final evaluation against survey based data has limited explanatory power due to its small sample size. It indicates that the first GloVe model does not provide state of the art per- formance. On the other hand, it has to be noted that the ques- tionnaire measures the German BFI-2 which shows strong correlations with the Big Five factors of the widespread NEO-PI-R, but still the factors in both questionnaires are not identical. Since IBM Personality Insight is based on a different questionnaire than BFI-2, the evaluation measures the exact target variable from the input data. Nevertheless, it should provide an accurate enough estimate to assess the quality of the models. Additionally, measuring the models' performance by using RMSE as the main selection criteria might have influenced the final model, for example, predictions could have a tendency towards the mean value of the training data.

The first iteration for creating a GloVe model helped cre- ating an automated process for data preparation, modeling, and evaluation. Part of this includes optimization for higher volumes of data, e.g. when querying the Twitter API. The second and final iteration results in a proof-of-concept model with acceptable error scores, for example with the RMSE being at maximum 0.25%. Applying the German GloVe to Twitter posts results in a word coverage of about 71% in comparison to just above 90% for the English language. Ref. [17] removed common stop words, punctuations, and hashtags for their approach. Additionally, they applied automatic spelling correction. In order to not manipulate the input data accidentally, these preparations have not been applied for the German approach, which could explain the difference in word coverage.

V. CONCLUSION
The overall goal for this work was to improve accessibility to results in personality mining research and to transfer those into the German language domain. Based on the existing research for GloVe based personality prediction, mainly [22] and [17], a proof-of-concept model has been created to
predict Big Five personality scores based on German Twitter posts. Four artifacts have been created:

1) Trained model for prediction
2) Published source code for creating and evaluating this model
3) Separate Python module for applying the model
4) Web application as demonstration object

With these artifacts, researchers are able to independently review the produced results and the process for model training. They can alter this process to fit their needs or improve overall prediction results, e.g. with self-collected input data. Researchers, who are only secondary interested in the actual personality predictions, can either utilize the existing website or the separate Python module to integrate these predictions in their primary research. Instead of needing multiple minutes for filling in a personality questionnaire, test subjects would only need a few seconds to provide their Twitter user name. This would shorten overall survey times and might in turn increase participation rates. Additionally, scientists would not have to rely on proprietary, closed-source applications, especially when privacy and control of data might be important.

Going forward, the proof-of-concept model could be enhanced by collecting and using more training data as input. Furthermore, it could be expanded to predict all thirty facets of the Big Five model by [45]. It would be possible as well to take other social networks or communication forms, e.g. instant messaging, into account or combining the forecasting power of these sources [32]. With the improvement of automatic machine translations, for example, via cross-lingual word embeddings, cross language personality predictions could be investigated, similar to the approach of using LIWC for two languages [158].

All in all, this work showed that a GloVe based personality mining system is feasible for the German language and demonstrates how related research might be published for good accessibility.

APPENDIX

A. LIST OF DEPENDENCIES

Full list of dependencies for MiPinG:

- certifi 2020.6.20
- cffi 1.14.1
- chardet 3.0.4
- click 7.1.2
- cryptography 3.0
- fire 0.3.1
- Flask 1.1.2
- googlemaps 4.4.2
- gunicorn 20.0.4
- idna 2.10
- itsdangerous 1.1.0
- Jinja2 2.11.2
- joblib 0.16.0
- keras2onnx 1.7.0
- MarkupSafe 1.1.1
- numpy 1.19.1
- oauthlib 3.1.0
- onnx 1.7.0
- onnxconverter-common 1.7.0
- onnxmltools 1.7.0
- onnxruntime 1.4.0
- pandas 1.1.0
- protobuf 3.12.4
- pycparser 2.20
- pyOpenSSL 19.1.0
- PySocks 1.7.1
- python-dateutil 2.8.1
- python-dotenv 0.14.0
- pytz 2020.1
- PyYAML 5.3.1
- requests 2.24.0
- requests-oauthlib 1.3.0
- scikit-learn 0.23.2
- scipy 1.5.2
- six 1.15.0
- skl2onnx 1.7.0
- supervisor 4.2.0
- termcolor 1.1.0
- threadpoolctl 2.1.0
- tweepy 3.9.0
- typing-extensions 3.7.4.2
- urllib3 1.25.10
- Werkzeug 1.0.1

B. REQUIREMENTS

Table 7 lists the requirements for a personality mining system identified during the business understanding phase.

C. BFI-2 LIST OF QUESTIONS

List of questions for BFI-2 [54]:

1) Ich gehe aus mir heraus, bin gesellig.
2) Ich bin einfühlsam, warmherzig.
3) Ich bin eher unordentlich.
4) Ich bleibe auch in stressigen Situationen gelassen.
5) Ich bin nicht sonderlich kunstinteressiert.
6) Ich bin durchsetzungsfähig, energisch.
7) Ich begegne anderen mit Respekt.
8) Ich bin bequem, neige zu Faulheit.
9) Ich bleibe auch bei Rückschlägen zuversichtlich.
10) Ich bin vielseitig interessiert.
11) Ich schäume selten vor Begeisterung über.
12) Ich bin eher unordentlich.
13) Ich schätze saubere, ordnungsgemäße Keramik.
14) Ich bin systematisch, halte meine Sachen in Ordnung.
15) Ich bin erforderlich, mir fallen raffinierte Lösungen ein.
16) Ich bin eher ruhig.
17) Ich habe mit anderen wenig Mitgefühl.
18) Ich bin systematisch, halte meine Sachen in Ordnung.
19) Ich reagiere leicht angespannt.
**TABLE 7. Complete list of requirements.**

| ID | Title | Description | Functional / Technical |
|----|-------|-------------|------------------------|
| 1  | Personality prediction for Twitter user name | Should provide functionality on website to get a personality prediction for a Twitter user name. | Functional |
| 2  | Website should be easy to use | To increase ease of use no login or registration should be required, results should be both textual and visual. | Functional |
| 3  | Documentation | Provide both inline comments and a separate documentation to make comprehension easier for third parties. | Functional |
| 4  | Should provide word coverage statistic | Additionally to a personality score, the word coverage of GloVe should be supplied to make it easier for users to estimate the validity of the result. | Functional |
| 5  | Documentation of used data | For transparency and reproducibility make used data available, wherever possible. | Functional |
| 6  | Lawful website operation | Terms of use are mandatory for German websites, a privacy note might be appropriate. | Functional |
| 7  | Prediction time should be less than 1 minute for end-users | Training time is not as important as prediction time for end-users. Loading animations might reduce perceived waiting time. | Functional |
| 8  | Simplify retraining of model | Using the training process again with different input (e.g. for improvements) should be possible and easy to do. | Functional |
| 9  | Automated data preparation | Data retrieval and preparation should be included into the code and automated to be able to rerun this with the same data or new data multiple times. | Functional |
| 10 | Code is available as open source with a suitable license | The main goal is to enable peer review, independent improvements and ease of use. Code should be publicly available with a suitable license. Trained models should be available in a suitable format. | Technical |
| 11 | Up-to-date technologies | Used components should be up-to-date, e.g. website should work with latest browser version, programming language should be popular. | Technical |
| 12 | Modular software architecture | Make components reusable, modular software architecture e.g. separated frontend and backend. | Technical |
| 13 | Configuration should be flexible | To adjust for different use cases there should be configuration files to change settings and API keys fast. | Technical |
| 14 | Object oriented software development | Should support modular architecture and embedding in other systems. | Technical |
| 15 | Use encryption | For website communication use HTTPS for safe communication. | Technical |
| 16 | Prevent misuse of website | For cost efficiency the website is not designed for mass data prediction, therefore this should be inhibited. Bots should be identified. | Technical |

**TABLE 8. Evaluation survey results.**

| User Number | Openness | Conscientiousness | Extraversion | Agreeableness | Neuroticism | Sex | Age |
|-------------|----------|-------------------|--------------|---------------|-------------|-----|-----|
| 1           | 0.48     | 0.03              | 0.23         | 0.33          | 0.31        | Male | <30 Years |
| 2           | 0.87     | 0.45              | 0.18         | 0.76          | 0.11        | Male | <30 Years |
| 3           | 0.87     | 0.67              | 0.92         | 0.71          | 0.07        | Male | 31-40 Years |
| 4           | 0.28     | 0.29              | 0.37         | 0.45          | 0.93        | Male | >50 Years |
| 5           | 0.77     | 0.45              | 0.59         | 0.65          | 0.04        | Male | 31-40 Years |
| 6           | 0.68     | 0.51              | 0.58         | 0.52          | 0.27        | Male | 41-50 Years |
| 7           | 0.91     | 0.15              | 0.74         | 0.13          | 0.25        | Female | 31-40 Years |
| 8           | 0.92     | 0.1               | 0.9          | 0.45          | 0.17        | Male | 31-40 Years |
| 9           | 0.2      | 0.17              | 0.18         | 0.71          | 0.07        | Male | 41-50 Years |
| 10          | 0.43     | 0.08              | 0.37         | 0.18          | 0.31        | Male | >50 Years |
| 11          | 0.28     | 0.05              | 0.12         | 0.27          | 0.24        | Male | <30 Years |
| 12          | 0.73     | 0.62              | 0.38         | 0.39          | 0.36        | Male | <30 Years |
| 13          | 0.92     | 0.25              | 0.87         | 0.45          | 0.11        | Male | <30 Years |
| 14          | 0.43     | 0.72              | 0.23         | 0.59          | 0.2         | Male | <30 Years |
| 15          | 0.64     | 0.2               | 0.81         | 0.52          | 0.09        | Male | <30 Years |

20) Ich kann mich für Kunst, Musik und Literatur begeistern.
21) Ich neige dazu, die Führung zu übernehmen.
22) Ich habe oft Streit mit anderen.
23) Ich neige dazu, Aufgaben vor mir herzuschieben.
24) Ich bin selbstsicher, mit mir zufrieden.
25) Ich leide philosophische Diskussionen.
26) Ich bin weniger aktiv und unternehmungslustig als andere.
27) Ich bin nachsichtig, vergebe anderen leicht.
28) Ich bin manchmal ziemlich nachlässig.
29) Ich bin ausgeglichen, nicht leicht aus der Ruhe zu bringen.
30) Ich bin nicht besonders einfallsreich.
31) Ich bin eher schüchtern.
32) Ich bin hilfsbereit und selbstlos.
33) Ich mag es sauber und aufgeräumt.
34) Ich mache mir oft Sorgen.
35) Ich weiß Kunst und Schönheit zu schätzen.
36) Mir fällt es schwer, andere zu beeinflussen.
TABLE 9. LIWC gridsearch initial parameters.

| Model Type             | Parameter Name                        | Parameter Values                        |
|------------------------|---------------------------------------|-----------------------------------------|
| Bayesian Ridge         | iterations                            | 100, 300, 500                           |
| Decision Tree          | maximum depth                         | 5, 10, 20, 50, 100, no limit            |
|                        | minimum number of samples for split   | 2, 5, 10, 20                            |
|                        | minimum number of samples at leaf     | 2, 5, 10                                |
| Gaussian Process       | kernel                                | 1.0 * RBF(1.0) (default)                |
|                        | restarts for kernel parameters        | 0, 2, 10                                |
| Random Forest          | number of trees                       | 10, 50, 100, 200                        |
|                        | maximum depth                         | 2, 5, 10, no limit                      |
|                        | minimum number of samples for split   | 2, 5, 10                                |
|                        | minimum number of samples at leaf     | 1, 2, 5, 10                             |
|                        | maximum number of leaf nodes          | 100, 500, no limit                      |
|                        | maximum number of used features       | no limit                                |
| Ridge Regression       | regularization strength               | 0.00001, 0.001, 0.01, 0.1, 1.0, 10.0    |
| Support Vector Machine | kernel type                           | rbf, sigmoid, linear                    |
|                        | regularization parameter (C)          | 0.001, 0.1, 1.0                         |
|                        | margin of tolerance (epsilon)         | 0.001, 0.1, 10                          |
|                        | maximum iterations                    | 10                                      |

TABLE 10. LIWC gridsearch final parameters.

| Model Type             | Parameter Name                        | Parameter Values                        |
|------------------------|---------------------------------------|-----------------------------------------|
| Bayesian Ridge         | iterations                            | 10                                      |
| Decision Tree          | maximum depth                         | 5, 2                                   |
|                        | minimum number of samples for split   | 2, 50                                  |
|                        | minimum number of samples at leaf     | 50                                     |
| Gaussian Process       | kernel                                | 1.0 * RBF(1.0) (default)                |
|                        | restarts for kernel parameters        | 0                                      |
| Random Forest          | number of trees                       | 100                                    |
|                        | maximum depth                         | 5                                      |
|                        | minimum number of samples for split   | 5, 10                                  |
|                        | minimum number of samples at leaf     | 10                                     |
|                        | maximum number of leaf nodes          | no limit                                |
|                        | maximum number of used features       | no limit                                |
| Ridge Regression       | regularization strength               | 10.0                                    |
| Support Vector Machine | kernel type                           | rbf                                    |
|                        | regularization parameter (C)          | 10                                     |
|                        | margin of tolerance (epsilon)         | 0.1                                    |
|                        | maximum iterations                    | 100                                    |

TABLE 11. LIWC model scores for Openness.

| Openness   | Average Score (RMSE) in Iteration |
|------------|-----------------------------------|
| Model Type | 1 | 2 | 3 | 4 |
| BayesianRidge | 0.1338 | 0.1338 | 0.1338 | 0.1338 |
| DecisionTree | 0.1806 | 0.1806 | 0.1803 | 0.1837 |
| GaussianProcessRegressor | 0.6919 | 0.6919 | 0.6919 | 0.6919 |
| RandomForest | 0.1385 | 0.1386 | 0.1427 | 0.1427 |
| RidgeRegression | 0.1336 | 0.1336 | 0.1325 | 0.136 |
| SupportVectorMachine | 0.1316 | 0.1316 | 0.1316 | 0.1324 |

37) Ich bin manchmal unhöflich und schroff.
38) Ich bin effizient, erledige Dinge schnell.
39) Ich fühle mich oft bedrückt, freudlos.

TABLE 12. LIWC model scores for Conscientiousness.

| Conscientiousness | Average Score (RMSE) in Iteration |
|-------------------|-----------------------------------|
| Model Type        | 1 | 2 | 3 | 4 |
| BayesianRidge     | 0.1649 | 0.1649 | 0.1649 | 0.1648 |
| DecisionTree      | 0.2149 | 0.2149 | 0.203 | 0.203 |
| GaussianProcessRegressor | 0.5884 | 0.5884 | 0.5884 | 0.5884 |
| RandomForest      | 0.17 | 0.1734 | 0.1734 |
| RidgeRegression   | 0.164 | 0.164 | 0.164 | 0.164 |
| SupportVectorMachine | 0.1496 | 0.1496 | 0.1496 | 0.1496 |

TABLE 13. LIWC model scores for Extraversion.

| Extraversion | Average Score (RMSE) in Iteration |
|--------------|-----------------------------------|
| Model Type   | 1 | 2 | 3 | 4 |
| BayesianRidge | 0.1745 | 0.1745 | 0.1745 | 0.1742 |
| DecisionTree | 0.2188 | 0.2188 | 0.2086 | 0.2086 |
| GaussianProcessRegressor | 0.523 | 0.523 | 0.523 | 0.523 |
| RandomForest | 0.1778 | 0.1828 | 0.1828 |
| RidgeRegression | 0.1704 | 0.1704 | 0.1704 | 0.1724 |
| SupportVectorMachine | 0.1589 | 0.1589 | 0.1586 | 0.1592 |

TABLE 14. LIWC model scores for Agreeableness.

| Agreeableness | Average Score (RMSE) in Iteration |
|---------------|-----------------------------------|
| Model Type    | 1 | 2 | 3 | 4 |
| BayesianRidge | 0.1665 | 0.1665 | 0.1665 | 0.1665 |
| DecisionTree | 0.1955 | 0.1955 | 0.1938 | 0.1938 |
| GaussianProcessRegressor | 0.5629 | 0.5629 | 0.5629 | 0.5629 |
| RandomForest | 0.1618 | 0.1659 | 0.1659 |
| RidgeRegression | 0.1674 | 0.1674 | 0.1674 |
| SupportVectorMachine | 0.1516 | 0.1516 | 0.1516 | 0.1516 |

TABLE 15. LIWC model scores for Neuroticism.

| Neuroticism | Average Score (RMSE) in Iteration |
|-------------|-----------------------------------|
| Model Type  | 1 | 2 | 3 | 4 |
| BayesianRidge | 0.1878 | 0.1878 | 0.1878 | 0.1878 |
| DecisionTree | 0.2356 | 0.2356 | 0.2238 | 0.2238 |
| GaussianProcessRegressor | 0.5182 | 0.5182 | 0.5182 | 0.5182 |
| RandomForest | 0.1883 | 0.1891 | 0.1923 | 0.1923 |
| RidgeRegression | 0.1905 | 0.1905 | 0.1853 | 0.1905 |
| SupportVectorMachine | 0.1757 | 0.1757 | 0.1757 | 0.1763 |

40) Es macht mir Spaß, gründlich über komplexes Dinge nachzudenken und sie zu verstehen.
41) Ich bin voller Energie und Tatendrang.
TABLE 16. GloVe gridsearch initial parameters.

| Model Type          | Parameter Name                      | Parameter Values |
|---------------------|-------------------------------------|------------------|
| Bayesian Ridge      | iterations                          | 100, 300, 500    |
| Decision Tree       | maximum depth                       | 5, 10, 20, 50, 100, no limit |
|                     | minimum number of samples for split | 2, 5, 10         |
|                     | minimum number of samples at leaf    | 2, 5, 10         |
| Gaussian Process    | kernel restarts for kernel parameters| 1.0 * RBP(1.0) (default) |
|                     |                                     | 0, 2, 10         |
| Random Forest       | number of trees                      | 100              |
|                     | maximum depth                        | no limit         |
|                     | minimum number of samples for split | 2                |
|                     | minimum number of samples at leaf    | 20               |
|                     | maximum number of leaf nodes         | no limit         |
|                     | maximum number of used features      | no limit         |
| Ridge Regression    | regularization strength              | 0.00001, 0.001, 0.01, 0.1, 1.0, 10.0 |
| Support Vector Machine | kernel type                        | rbf, sigmoid, linear |
|                     | regularization parameter (C)         | 0.001, 0.01, 0.1, 1, 10 |
|                     | margin of tolerance (epsilon)        | 0.001, 0.1, 1, 10 |
|                     | maximum iterations                   | 100              |

TABLE 17. GloVe gridsearch final parameters.

| Model Type          | Parameter Name                      | Parameter Values |
|---------------------|-------------------------------------|------------------|
| Bayesian Ridge      | iterations                          | 25               |
| Decision Tree       | maximum depth                       | 4                |
|                     | minimum number of samples for split | 2                |
|                     | minimum number of samples at leaf    | 50               |
| Gaussian Process    | kernel restarts for kernel parameters| ConstantKernel + Matern + WhiteKernel^14 |
| Random Forest       | number of trees                      | 50               |
|                     | maximum depth                        | 5                |
|                     | minimum number of samples for split | 2                |
|                     | minimum number of samples at leaf    | 20               |
|                     | maximum number of leaf nodes         | no limit         |
|                     | maximum number of used features      | 450              |
| Ridge Regression    | regularization strength              | 20               |
| Support Vector Machine | kernel type                       | linear           |
|                     | regularization parameter (C)         | 0.01             |
|                     | margin of tolerance (epsilon)        | 0.1              |
|                     | maximum iterations                   | 300              |

TABLE 18. GloVe model scores for Openness.

| Openness               | Average Score (RMSE) in Iteration |
|------------------------|-----------------------------------|
| Model                  | 1       | 2       | 3       | 4       | 5       | 6       |
| BayesianRidge          | 0.1007  | 0.1007  | 0.1018  | 0.1007  | 0.1007  | 0.1007  |
| DecisionTree           | 0.1085  | 0.1084  | 0.1074  | 0.1026  | 0.1026  | 0.1026  |
| GaussianProcessRegressor | 0.1023 | 0.0979  | 0.0979  | 0.0979  | 0.0979  | 0.0979  |
| RandomForest           | 0.0862  | 0.0843  | 0.0843  | N/A     | 0.0854  | 0.0868  |
| RidgeRegression        | 0.1006  | 0.1002  | 0.1002  | 0.1002  | 0.1002  | 0.1002  |
| SupportVectorMachine   | 0.101   | 0.1006  | 0.1006  | 0.1006  | 0.1007  | 0.1006  |

TABLE 19. GloVe model scores for Conscientiousness.

| Conscientiousness      | Average Score (RMSE) in Iteration |
|------------------------|-----------------------------------|
| Model                  | 1       | 2       | 3       | 4       | 5       | 6       |
| BayesianRidge          | 0.1269  | 0.1269  | 0.1291  | 0.1269  | 0.1269  | 0.1269  |
| DecisionTree           | 0.1341  | 0.1317  | 0.1317  | 0.1305  | 0.1305  | 0.1305  |
| GaussianProcessRegressor | 0.1342 | 0.1254  | 0.1254  | 0.1254  | 0.1254  | 0.1254  |
| RandomForest           | 0.1086  | 0.1048  | 0.1048  | N/A     | 0.1069  | 0.1082  |
| RidgeRegression        | 0.127   | 0.1258  | 0.1258  | 0.1258  | 0.1258  | 0.1258  |
| SupportVectorMachine   | 0.1274  | 0.1269  | 0.1269  | 0.1269  | 0.127   | 0.1269  |

42) Ich bin anderen gegenüber misstrauisch.
43) Ich bin verlässlich, auf mich kann man zählen.
44) Ich habe meine Gefühle unter Kontrolle, werde selten wütend.
45) Ich bin nicht sonderlich fantasievoll.
46) Ich bin gesprächig.
47) Andere sind mir eher gleichgültig, egal.
48) Ich bin eher der chaotische Typ, mache selten sauber.
49) Ich werde selten nervös und unsicher.
50) Ich finde Gedichte und Theaterstücke langweilig.
51) In einer Gruppe überlasse ich lieber anderen die Entscheidung.
52) Ich bin höflich und zuvorkommend.
53) Ich bleibe an einer Aufgabe dran, bis sie erledigt ist.
54) Ich bin oft deprimiert, niedergeschlagen.
55) Mich interessieren abstrakte Überlegungen wenig.
56) Ich bin begeisterungsfähig und kann andere leicht mitreißen.
57) Ich schenke anderen leicht Vertrauen, glaube an das Gute im Menschen.
58) Manchmal verhalte ich mich verantwortungslos, leichtsinnig.
59) Ich reagiere schnell gereizt oder genervt.
60) Ich bin originell, entwickle neue Ideen.

D. EVALUATION SURVEY RESULTS
Table 8 shows the results of the survey conducted to collect evaluation data. Twitter names have been replaced by an anonymous user number. The subsequent five columns present the Big Five personality scores for each user as a percentage rank. In the last two columns demographic data is shown. Age is given in four different age groups according to [54].

E. LIWC MODEL TRAINING
Table 9 shows all model parameters for the first grid search iteration during LIWC model training.

Table 10 shows all final model parameters as the result of the grid search during LIWC model training. Tables 11 to 15 show the average RMSE score for each model type in all grid search iterations during LIWC model
training. The best model type of each iteration is bold highlighted.

F. GloVe MODEL TRAINING
Table 16 shows all model parameters for the first grid search iteration during GloVe model training.

Table 17 shows all final model parameters as the result of the grid search during GloVe model training.

Tables 18 to 22 show the average RMSE score for each model type in all grid search iterations during GloVe model training. The best model type of each iteration is bold highlighted.

G. GloVe MODEL TRAINING WITH SURVEY DATA
Table 23 shows the RMSE scores of all models across all five dimensions for the GloVe model training with survey data. The best model type of each iteration is bold highlighted.

REFERENCES
[1] H. Kämper, “Sigmund freuds sprachenken ein beitra zur sprachbewussteinsgeschichte,” in Neue Deutsche Sprachgeschichte, D. Chenuhin, K. Jakob, and A. Linke, Eds. Berlin, Germany: DE GRUYTER, 2002, pp. 239–252.
[2] Y. R. Tausczik and J. W. Pennebaker, “The psychological meaning of words: LIWC and computerized text analysis methods,” J. Lang. Social Psychol., vol. 29, no. 1, pp. 24–54, Mar. 2010.
[3] J. W. Pennebaker, M. R. Mehl, and K. G. Niederhoffer, “Psychological aspects of natural language use: Our words, our selves,” Annu. Rev. Psychol., vol. 54, no. 1, pp. 547–577, Feb. 2003.
[4] F. C. Cressey, M. A. Walker, M. R. Mehl, and R. K. Moore, “Using linguistic cues for the automatic recognition of personality in conversation and text,” J. Artif. Intell. Res., vol. 30, pp. 457–500, Nov. 2007.
[5] W. R. Revelle and S. Cherer, “Personality and emotion,” in Oxford Companion to Emotion and the Affective Sciences, D. Sander and K. Scherer, Eds. Oxford, U.K.: Oxford Univ. Press, 2009, pp. 304–305.
[6] C. J. Boyce, A. M. Wood, and N. Powdthavee, “Is personality fixed? Personality changes as much as ‘variable’ economic factors and more strongly predicts changes to life satisfaction,” Social Indicators Res., vol. 111, no. 1, pp. 287–305, 2013.
[7] R. R. McCrae and D. O. John, “An introduction to the five-factor model and its applications.” J. Pers., vol. 60, no. 2, pp. 175–215, 1992.
[8] C. H. Lee, K. Kim, Y. S. Seo, and C. K. Chung, “The relations between personality and language use,” J. Gen. Psychol., vol. 134, no. 4, pp. 405–413, Oct. 2007.
[9] R. R. McCrae and P. T. Costa, Jr., “Brief versions of the NEO-PI-3,” J. Individual Differences, vol. 28, no. 3, pp. 116–128, Jul. 2007.
[10] D. A. Newman and J. S. Lyon, “Recruitment efforts to reduce adverse impact: Targeted recruiting for personality, cognitive ability, and diversity,” J. Appl. Psychol., vol. 94, no. 2, pp. 298–317, 2009.
[11] J. E. Strickhouser, E. Zell, and Z. Križan, “Does personality predict health and well-being? A metasynthesis,” Health Psychol., Off. J. Division Health Psychol., Amer. Psychol. Assoc., vol. 36, no. 8, pp. 797–810, 2017.
[12] E. Ferguson, “Personality is of central concern to understand health: Towards a theoretical model for health psychology,” Health Psychol., Rev., vol. 7, no. sup1, pp. S32–S70, May 2013.
[13] G. N. Marshall, C. B. Wortman, R. R. Vickers, J. W. Kusulas, and L. K. Hervig, “The five-factor model of personality as a framework for personality-health research,” J. Personality Social Psychol., vol. 67, no. 2, pp. 278–286, 1994.
[14] M. T. Adjei and M. N. Clark, “Relationship marketing in a B2C context: The moderating role of personality traits,” J. Retailing Consum. Services, vol. 17, no. 1, pp. 73–79, Jan. 2010.
[15] J. I. Jahng, H. Jain, and K. Ramamurthy, “Personality traits and effectiveness of presentation of product information in e-business systems,” Eur. J. Inf. Syst., vol. 11, no. 3, pp. 181–195, Sep. 2002.

[14] This kernel provided better results than the default kernel [159].
G. Matthews, I. J. Deary, and M. C. Whiteman, *Personality Traits*, Cambridge, U.K.: Cambridge Univ. Press, 2012.

R. R. McCrae, “The five-factor model of personality traits: Consensus and controversy,” in *The Cambridge Handbook of Personality Psychology* (Cambridge handbooks in psychology), G. Matthews and P. J. Corr, Eds. Cambridge, U.K.: Cambridge Univ. Press, 2009, pp. 148–161.

D. Fetenchenauer, Psychologie. München: Verlag Franz Vahlen, 2018.

J. M. Digman and J. Inouye, “Further specification of the five robust factors of personality,” *J. Personality Social Psychol.*, vol. 50, no. 1, pp. 116–123, 1986.

P. T. Costa and R. R. McCrae, “The revised NEO personality inventory (NEO-PI-R),” in *The SAGE Handbook Personality Theory Assessment*, D. H. Saklofske, G. J. Boyle, and G. Matthews, Eds. Los Angeles, CA, USA: SAGE, 2008, pp. 179–198.

R. R. McCrae and J. P. T. Costa, “The five-factor theory of personality,” in *Handbook Personality* O. P. John, R. W. Robins, and L. A. Pervin, Eds. New York, NY, USA: Guilford Press, 2010, pp. 159–181.

O. P. John, A. Angleiter, and F. Ostendorff, “The lexical approach to personality: A historical review of trait taxonomic research,” *Eur. J. Personality*, vol. 2, no. 3, pp. 171–203, Sep. 1988.

G. Saucier and L. R. Goldberg, “The language of personality: Lexical perspectives on the five-factor model,” in *The five-factor model personality* (NEO-FFI), G. Wiggins, Ed. New York, NY, USA: Guilford Press, 1996, pp. 21–50.

J. W. Pennebaker and L. A. King, “Linguistic styles: Language use as an individual difference,” *J. Personality Social Psychol.*, vol. 77, no. 6, pp. 1296–1312, 1999.

L. A. Fast and D. C. Funder, “Personality as manifest in word use: Correlations with self-report, acquaintance report, and behavior,” *J. Personality Social Psychol.*, vol. 94, no. 2, pp. 334–346, 2008.

G. J. Boyle and E. Helmes, “Methods of personality assessment,” in *The Cambridge Handbook of Personality Psychology* (Cambridge handbooks in psychology), G. Matthews and P. J. Corr, Eds. Cambridge, U.K.: Cambridge Univ. Press, 2009, pp. 110–126.

J. W. Pennebaker and A. Graybeal, “Patterns of natural language use: Disclosure, personality, and social integration,” *Current Directions Psychol. Sci.*, vol. 10, no. 3, pp. 90–93, 2001.

Y. S. Ben-Porath, “Self-report inventories,” in *Handbook of Psychology*, I. B. Weiner, Ed. Hoboken, NJ, USA: Wiley, 2013.

D. Danner, B. Rammstedt, M. Bluemke, C. Lechner, S. Berres, T. Knopf, L. A. Fast, and D. C. Funder, “Personality as manifest in word use: Correlations with self-report, acquaintance report, and behavior,” *J. Personality Social Psychol.*, vol. 94, no. 2, pp. 334–346, 2008.
I. Stapleton, “Flake8: Your tool for style guide enforcement,” Tech. Rep., 2016.

O. Adams, A. Makarucha, G. Neubig, S. Bird, and T. Cohn, “Cross-lingual word embeddings for low-resource language modeling,” in Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics, M. Lapata, P. Blunsom, and A. Koller, Eds. Valencia, Spain, 2017, pp. 937–947.

C. Fonnesbeck, “Fitting Gaussian process models in Python,” Tech. Rep., 2017. [Online]. Available: https://blog.dominodatalab.com/fitting-gaussian-process-models-python

HENNING USSELMANN received the B.Sc. degree in business information systems from the Leibniz University of Applied Sciences, Hanover, Germany, in 2016, and the M.Sc. degree in business information systems from the Braunschweig University of Technology, Germany, in 2020. In 2016, he worked abroad in U.K., as a Business Intelligence Developer at Volkswagen Financial Services, for which he has been working continuously since 2013. He specializes in SAP-based data warehouses in combination with HANA and Hadoop technology.

RANGINA AHMAD received the B.Sc. and M.Sc. degrees in business information systems from the Braunschweig University of Technology, Germany, where she is currently pursuing the Ph.D. degree. Since 2018, she has been a Junior Researcher at the Chair of Information Management, Braunschweig University of Technology. Her research focuses on topics, such as human-AI interaction, personality psychology, and e-services, and she has been published in leading information systems conferences, such as Americas Conference on Information Systems and Hawaii International Conference on Information Systems.

DOMINIK SIEMON studied business information systems at the Braunschweig University of Technology, where he received the Dr. rer. pol. degree in business information systems with his dissertation on “IT-supported collaborative creativity.” He worked as a Research Assistant and a Postdoctoral Researcher at the Braunschweig University of Technology and a part-time Professor of business information systems and digital business at IU International University. He is currently an Associate Professor with the Department of Software Engineering, Lappeenranta-Lahti University of Technology (LUT University), Finland. His mainly design-oriented research in the field of information systems addresses human-AI collaboration, conversational agents, innovation management, collaboration technology, and creativity. His work has been published in leading conferences, such as the International Conference on Information Systems, and in journals, such as Education and Information Technologies, AIS Transactions on Human-Computer Interaction, and the Communications of the Association for Information Systems.