♠ SPADE: A Big Five-Mturk Dataset of Argumentative Speech Enriched with Socio-Demographics for Personality Detection

Kerz, E.; Qiao, Y.; Zanwar, S.; Wiechmann, D.

Publication date
2022

Document Version
Final published version

Published in
Conference proceedings : Language Resources and Evaluation Conference

Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 6405–6419
Marseille, 20-25 June 2022

♠ SPADE: A Big Five-Mturk Dataset of Argumentative Speech Enriched with Socio-Demographics for Personality Detection

Elma Kerz[1], Yu Qiao[1], Sourabh Zanwar[1], Daniel Wiechmann[2]
[1]RWTH Aachen University, [2]University of Amsterdam
yu.qiao@rwth-aachen.de, sourabh.zanwar@rwth-aachen.de,
elma.kerz@ifaar.rwth-aachen.de, d.wiechmann@uva.nl

Abstract

In recent years, there has been increasing interest in automatic personality detection based on language. Progress in this area is highly contingent upon the availability of datasets and benchmark corpora. However, publicly available datasets for modeling and predicting personality traits are still scarce. While recent efforts to create such datasets from social media (Twitter, Reddit) are to be applauded, they often do not include continuous and contextualized language use. In this paper, we introduce ♠ SPADE, the first dataset with continuous samples of argumentative speech labeled with the Big Five personality traits and enriched with socio-demographic data (age, gender, education level, language background). We provide benchmark models for this dataset to facilitate further research and conduct extensive experiments. Our models leverage 436 (psycho)linguistic features extracted from transcribed speech and speaker-level metainformation with transformers. We conduct feature ablation experiments to investigate which types of features contribute to the prediction of individual personality traits.

Keywords: Automatic personality detection, language-based personality, argumentative speech, Big Five, dataset

1. Introduction

People's personality comprises a set of individual differences, often referred to as psychological traits and/or dimensions, that describe and explain their behavior, emotions, motivation, and thought patterns (Funder, 2001; Wilt and Revelle, 2015). Personality traits are not observable directly, but are reflected in the recurring behavioral patterns of people and can thus be derived from them. These traits are associated with a variety of important life outcomes and decisions (Ozer and Benet-Martinez, 2006). Specifically, they have been repeatedly linked to individual (e.g., well-being), interpersonal (e.g., relationship satisfaction), and social-institutional outcomes (e.g., career choice and career success) (Soto, 2019). In addition, attitudes and social behavior toward a particular person depend significantly on the impression others have of that person (Uleman et al., 2008). In view of its importance in capturing the essential aspects of a person, increasing attention is being paid to the development of models that can leverage data on human behavior to automatically predict personality. Language data - i.e. data obtained from verbal behavior, including written text or audio recordings - is a key type of such data. Even in the early years of psychology, a person's use of language was seen as a distillation of his or her underlying drives, emotions, and thought patterns (Freud, 1915; Rorschach, 1921; Allport, 1942). In the 1960s, the so-called General Inquirer was introduced, which emerged as the first general and systematic approach to psychological language analysis (Stone et al., 1962). The introduction of Linguistic Inquiry and Word Count (LIWC) (Pennebaker and King, 1999; Pennebaker et al., 2007) was a milestone in transparent psychological text analysis, and LIWC-like features continue to figure prominently in automatic personality detection (APD) approaches today. A common approach to APD is to leverage such features by feeding them machine learning classifiers, such as a Sequential Minimum Optimizer, Support Vector Machine or Naïve Bayes classifier (see Agarwal, 2014, for a review of personality detection from text using shallow learning techniques). More recently, approaches to APD have drawn on context-independent or contextualized word embeddings, e.g. GloVe, Word2Vec or BERT, or have combined handcrafted linguistic features with word embeddings (see, e.g., Majumder et al., 2017, Mehta et al., 2020). Although significant progress has been made in this area, publicly available datasets for modeling and predicting personality traits are still scarce. While recent efforts to create such datasets from social media (Twitter, Reddit) are to be applauded, they often do not include continuous and contextualized language use. Here we introduce ♠ SPADE, the first dataset with continuous samples of argumentative speech labeled with the Big 5 personality traits and enriched with socio-demographic data (age, gender, education level, language background). We provide benchmark models for this dataset and perform feature ablation experiments to investigate which types of features contribute to the prediction of individual personality traits. ♠ SPADE is made available for research purposes upon request.

2. Related Work

Research on personality and language has drawn on a variety of sources ranging from stream-of-consciousness essays (Pennebaker and King, 1999) to emails (Oberlander and Gill, 2006), and blogs (Iacobelli et al., 2011). More recently, the attention has shifted towards language use on social media including data obtained from Facebook (e.g. Kosinski et al., 2015), Twitter (e.g. Plank and Hovy, 2015), and Reddit (e.g. Gjurkovic et al., 2020). Apart from the types of texts that these datasets represent, they also differ with respect to (1) the sheer amount of data they comprise (both in terms of number of texts and amount of words per text), (2) the way in which personality labels were obtained (e.g. validated questionnaire versus human labeled versus inferred from accompanying information, e.g. behaviors on social network), and (3) the amount of speaker- or document-level meta-information available (see Wiegmann et al., 2019 for a survey of datasets). We focus here on a concise overview of some of the most widely used publicly available datasets as well as some recent additions.[1] One of the most widely utilized datasets in automatic personality prediction research is the **Essay dataset** (Pennebaker and King, 1999), which comprises 2,467 stream-of-consciousness texts produced by as many individuals students between 1997 and 2004. The Essay dataset is annotated with the binary labels of the Big Five personality traits that were obtained using a standardized self-report questionnaire. Despite its advanced age, the Essay dataset is still one of the most well-established benchmark datasets due to the relatively large amount of continuous language use and the fact that its personality labels are derived using a validated instrument. Based on Myers–Briggs Type Indicator (MBTI), another widely used dataset is the **MBTI kaggle dataset** (Li et al., 2018). It consists of snippets - usually whole sentences - of social media interaction of over 8,600 users of Personality Cafe[2], an online forum community dedicated to all ranges of personality types and people, all of whom have indicated their MBTI type. The total size of this dataset is approximately 11.2 million words with an average combined length of post samples of 1,288 words per individual. Gjurkovic and Snajder (2018) introduced the **MBTI9k dataset**, which was derived from Reddit. In the MBTI9k, personality labels are derived from special user descriptors on Reddit called 'flairs', an icon or text that appears next to a username. Many users used flairs to report their MBTI type and often also information about their age, gender, personality types of their partners, marital status, medical diagnoses etc. The MBTI9k datasets contains over 583 million words from 9,111 Reddit users. Recently, Gjurkovic et al. (2020) extended the MBTI9k dataset and introduced the **Personality ANd Demographics Of Reddit Authors (PANDORA)**. PANDORA comprises over 17M comments written by more than 10k Reddit users, annotated with MBTI and Enneagram personality labels, alongside age, gender, location, and language. The datasets further contains Big 5 labels for approximately 1.6k users that were extracted from textual information in comments replying to posts which mention a specific online tests. An advantage of such large-scale multilabeled datasets such as PANDORA is that they can be used to develop new deep-learning architectures. Once developed, the resulting models can be fine-tuned and used on smaller datasets that contain validated questionnaire-based personality assessments and/or richer meta-information. Datasets constructed from social media typically contain only short snippets of connected text (e.g. comments, status updates), which may not contain enough linguistic signals for personality detection. In an effort to test this possibility, Stajner and Yenikent (2021) introduced the **MBTI-MTurk dataset** based on data collection via human intelligence tasks (HITs) using the Amazon Mechanical Turk (MTurk) crowdsourcing platform. They used two text prompts (topics: favourite type of vacation; favourite hobbies) to elicit samples of written text with an minimal length of 300 characters long. The dataset consists of 96 HITs completed by different participants (MTurk IDs). Personality labels for the participants were obtained by two trained human annotators who were instructed to look for linguistic signals of MBTI traits in the texts of a given participant. Finally, we note that the vast majority of datasets used in personality prediction - including the ones overviewed here - are confined to written language. Datasets of spoken language that also enable the development of models of personality detection from speech are scarce. A notable exception is the **Electronically Activated Recorder (EAR) corpus** (Mehl et al., 2001), which contains both sound extracts and transcripts from 96 psychology students at the University of Texas at Austin. This corpus contains a total of 97,468 words from 15,269 utterances. Personality labels of the dataset were obtained based on the Big Five model, using the 44-item Big Five Inventory (John et al., 1991).

[1] As the MyPersonality dataset Kosinski et al., 2015 has become unavailable to the research community, we do not include it here.

[2] https://www.personalitycafe.com/

3. ♠ SPADE Dataset

3.1. Dataset construction

Our goal in constructing ♠ SPADE was to compile a dataset of longer samples of monological, argumentative speech that maximize the potential of finding linguistic correlates of human evaluations of speech as well as signals of personality traits. The dataset consists of 20 hours of speech from 220 individuals (aged 18-78 years) collected through the Amazon Mechanical Turk (AMT) crowdsourcing platform. The speech samples were elicited through prompts relating to three debating topics: (A) Climate change is the greatest threat facing humanity today, (B) People should be legally required to get vaccinated and (C) The development of artificial intelligence will help humanity. Each participant was asked to talk about a given topic for about three to five minutes. The data collection was done in two rounds. In round 1 (110 participants) participants were asked to record one speech sample on a topic of their choice. In round 2 (110 participants), each participant was asked to record two speech samples on two different topics. Each round comprised two batches: an initial batch of 15 Human Intelligence Tasks (HITs) was run to test the data collection procedure before running the remaining 95 HITs. AMT participants were filtered based on location to include only participants from the US and UK and from Germany. Participants were filtered based on their MTurk IDs so that a participant could only submit the HIT once. We included data from participants after verifying that all the information asked for was present and that the recorded language was comprehensible. After filtering the submissions for various missing files and data, we retained the samples from 214 MTurk participants, representing 333 speech samples covering three topics. All of the data collection was done in compliance with ethical considerations for NLP crowdsourcing (see Shmueli et al., 2021). Informed consent was obtained from all participants.[3]. Following considerations regarding fair payment, we conducted a survey with 100 crowdworkers, where we asked for the amount of compensation (in US dollars) they considered adequate for the task at hand. Based on the results of the survey, the HIT was set up with a compensation of \$5, corresponding to an average pay rate of \$15/hour. All speech samples were manually transcribed by two trained transcribers. We used a rather simple set of transcription guidelines that state that all words spoken have to be transcribed, even if repeated or if fillers (e.g., uh, um). Transcribers could further mark words where they were unsure with a special symbol and also longer stretches of unintelligible speech. The inter-transcriber disagreements, measured by word error rates (WER), was estimated on the basis of a subset of speeches (N=35) that were transcribed by both transcribers. Inter-transcriber disagreements was 8.2%, which is in line with the results obtained in previous studies involving comparable samples of native (WER = 5%) and non-native (WER = 15-20%) spontaneous speech (Zechner et al., 2009). The resulting dataset comprises 848,827 words of transcribed speech. The mean length of the speech transcripts was 2493 words (SD = 752.1 words), ranging between a minimum of 921 words and maximum length of 6123 words. The language samples of the present dataset are thus substantially larger than those of standardly used personality datasets: For example, the texts of the widely-used Essays dataset (Pennebaker and King, 1999) comprise 652 words on average; the combined user-posts of the MBTI Kaggle dataset (Li et al., 2018) comprise 1264 words on average. Table 1 shows the distribution of the demographic variables collected from the participants (age, gender, education, language status).

Age	Gender	Education	Language Status
Min. :18.00	Male: 129	BA : 138	Mono: 122
1st Qu.:25.00	Female: 88	HS : 40	Biling: 57
Mean :32.27	Diverse: 3	MA : 40	L2 Eng: 38
3rd Qu.:35.00		PHD: 1	NA: 3
Max. :78.00		NA: 1	

Table 1: Distribution of demographic variables (Mono = monolingual English speakers (N = 122), Biling = participants who report English as their first language who speak one or more additional languages (N = 57), L2 Eng = participants that use English as a foreign language (N=38))

3.2. Personality labels and additional speaker-level variables

Participants were also asked to fill in three questionnaires: (1) the Language Experience and Proficiency Questionnaire (Kaushanskaya et al., 2020), which was used to assess language profiles of the participants: Participants rated from 0 (none at all) to 10 (very much) their daily exposure to English in different contexts/activities (family, friends, reading, formal education, self-education, watching TV, and listening to the radio/music) as well as their proficiency with regard to their skills in reading, listening, speaking and writing English. The further completed (2) a reading habits questionnaire, which was used to assess the average time spent reading English per week (in hours) across different media both at the present time and in the past. Finally, (3) personality labels were obtained via the Big Five Inventory (BFI) questionnaire (John et al., 1991). The BFI is an efficient and frequently used instrument to assess the big five personality traits (Extraversion, Neuroticism, Conscientiousness, Openness to Experience, and Agreeableness). It consists of 44 self-rating statements, such as 'I see myself as someone who generates a lot of enthusiasm' or 'I see myself as someone who remains calm in tense situations'. Participants rated each statement on a 5-point Likert-scale ranging from 1 (disagree strongly) to 5 (strongly agree). Participants' scale scores for each of the five dimensions were expressed as person-centered z-scores that were adjusted for differences in acquiescent response styles ('yea-saying' vs. 'nay-saying') (for details, see John et al., 2008).

The distributions of the participants self-reported information their proficiency and experience with English and their reading habits are provided in Tables 5 and 6 in the appendix. We observed some asymmetries between subgroups with respect to their exposure to English and self-rated English proficiency: For example,

[3]Before starting the task, participants accepted the following disclaimer: "We do not collect personally identifiable information such as name, address, contact information, or other data that can be associated with a specific individual. The data we collect is encrypted and stored anonymously on a secure server and can only be identified by a system-generated, unique random code. We will use the collected data for research purposes only. Please continue if you agree to these conditions."

regarding the sources of current exposure to English the monolingual group reported highest scores for family (M=9.02), friends (M=8.79) and reading (M=8.9), whereas the 'English as a foreign language' group reports highest scores in reading (M=7.79) and watching TV (M=7.63). With regard to self-rated proficiency the participants in the monolingual and bilingual groups report higher scores on average on all four skills (speaking, listening, reading, writing) (all M>9.28) that their L2 English peers, for whom speaking represents the lowest rated English skill (M=7.66). The distributions of the personality scores are presented in Table 2 and visualized in Figure 1. Table 2 shows that the participants in our dataset are on average more open and agreeable and less conscientious, extraverted and neurotic, when compared to a neutral score of zero. Figure 3 in the appendix presents a visualization of the frequencies of all trait combinations that occur at least twice in the dataset, when participants were classified as having any of the five personality traits based on median splits. The participants shows high frequencies of particular trait combinations, e.g. +open, +agreeable (N=8), whereas others, e.g. +extraverted, +neurotic are unattested (N=0). Extraversion tends to be attested together with conscientiousness (N=9), while neuroticism tends co-occur with agreeableness (N=6). These tendencies for typical and atypical trait combinations support that personality prediction tasks are more adequately modeled as a multi-label classification task.

Dimension	Mean	SD	Min	Max
Openness	0.107	0.343	-1.239	1.268
Conscientiousness	-0.104	0.461	-2.053	1.968
Extraversion	-0.090	0.437	-1.953	1.051
Agreeableness	0.161	0.577	-3.045	1.721
Neuroticism	-0.049	0.477	-2.177	1.363

Table 2: Distribution of BFI scores in the dataset. BFI scores shown are person-centered standard (or Z) scores

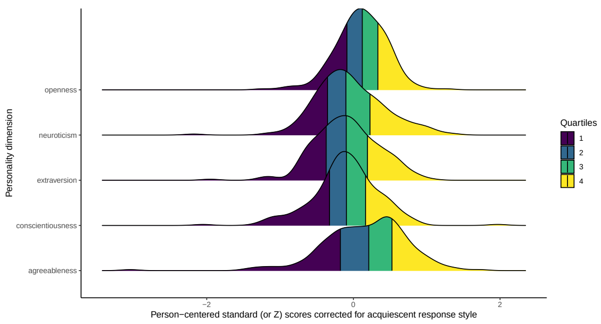

Figure 1: Distribution of BFI scores in the dataset. BFI scores shown are person-centered standard (or Z) scores.

4. Extraction of linguistic features

The speech transcripts were automatically analyzed using CoCoGen (Ströbel et al., 2016), a computational tool that implements a sliding window technique to calculate sentence-level measurements that capture the within-text distributions of scores for a given language feature (for current applications of the tool in the context of text classification, see Kerz et al., 2020; Kerz et al., 2021). We extract a total of 436 features that fall into nine categories: (1) measures of syntactic complexity (N=16), (2) measures of lexical richness (N=15), (3) information theoretic measures (N =3), (4) register-based n-gram frequency measures (N=25), (5) readability measures (N=14), (6) psycholinguistic measures (N=37), (7) LIWC-style (Linguistic Inquiry and Word Count) features (N=61), (8) sentiment related features (N=209) and (9) emotion related features (N=56). Tokenization, sentence splitting, part-of-speech tagging, lemmatization and syntactic PCFG parsing were performed using Stanford CoreNLP (Manning et al., 2014). The syntactic complexity measures comprise (i) surface measures that concern the length of production units, such as the mean length of words, clauses and sentences, (ii) measures of the type and incidence of embeddings, such as dependent clauses per T-Unit or verb phrases per sentence or (iii) the frequency of particular types of particular structures, such as the number of complex nominal per clause. These features are implemented based on descriptions in (Lu, 2010) using the Tregex tree pattern matching tool (Levy and Andrew, 2006) with syntactic parse trees for extracting specific patterns. Lexical richness measures fall into three distinct sub-types: (i) lexical density, i.e. the ratio of the number of lexical (as opposed to grammatical) words to the total number of words in a text, (iii) lexical variation, i.e. the range of vocabulary as displayed in language use, captured by text-size corrected type-token ratio and (iii) lexical sophistication, i.e. the proportion of relatively unusual or advanced words in the learner's text, such as the number of New General Service List (Browne and others, 2013). The operationalizations of these measures follow those described in Lu (2012) and Ströbel (2014). The information theoretic measures have been used as 'holistic' measures of linguistic complexity. These measures use the Deflate algorithm (Deutsch, 1996) to compress a text and obtain complexity scores by relating the size of the compressed file to the size of the original file (for the operationalization and implementation of these measures see (Ströbel, 2014)). The register-based n-gram frequency measures are derived from the five register sub-components of the Contemporary Corpus of American English (COCA, (Davies, 2008)): spoken, magazine, fiction, news and academic language. The total of 25 measures results from the combination of (a) a 'reference list' containing the top 100k most frequent n-grams and their frequencies from one of five registers of the COCA corpus and (b) the size of the n-gram ($n \in [1, 5]$) (see Kerz et al., 2020 for details). The readability measures combine a word familiarity variable defined by prespecified vocabulary resource to estimate semantic difficulty together with a syntactic variable, such as average sentence length. Examples of these measures are the Fry index (Fry, 1968) or the SMOG (McLaughlin, 1969). The psycholinguistic measures capture cognitive aspects of reading not directly addressed by the surface vocabulary and syntax features of traditional formulas. These measures include a word's average age-of-acquisition (Kuperman et al., 2012) or prevalence, which refers to the number of people knowing the word (Brysbaert et al., 2019; Johns et al., 2020). The LIWC feature set (Pennebaker et al., 2001) is one of the most common closed-vocabulary methods for personality detection from text. These features concern frequency counts of words that are associated with 60 psychologically relevant relevant subgroups like 'function words' (e.g., articles, conjunctions, pronouns), 'affective processes' (e.g., happy, nervous, cried) and 'social processes' (e.g., mate, talk, friend). The 209 features from the 'sentiment' category were derived from a total of five lexicons that have been successfully employed in sentiment analysis research (ANEW (Bradley and Lang, 1999), General Inquirer (Stone et al., 1966), NRC-VAD (Mohammad, 2018), SenticNet (Cambria et al., 2010), Sentiment140 (Mohammad et al., 2013). Finally, the 48 features from the emotion feature group were obtained from three dictionaries – EmoLex (Mohammad and Turney, 2013) and GALC (Scherer, 2005), and DepecheMood++ (Araque et al., 2019) – that have been successfully employed in emotion and personality recognition research.

All features are computed at the level of individual sentences. The resulting sequences of sentence-level scores capture the progression of a feature score from the beginning of a text to its end and are referred to here as 'text contours'. In addition to these high-resolution measurements of text features, we also computed text-average scores by aggregating all sentence-level scores to the mean value for each text. The informational gain of 'text contours' relative to text-averages is illustrated in Figure 2. The top panel in Figure 2 shows the distribution of scores of three selected features for a randomly selected text (transcript) from the dataset: One syntactic complexity feature (Complex Nominals per Clause), one psycholinguistic lexical feature (Word Prevalence), and one readability feature (the Gunning Fox Index). For the purposes of this illustration, these features were z-standardized. The black line represents the mean feature value of the text. As is evident in the graphs, all features score fluctuate within the text, with high values on one feature often being compensated for by lower values on another. A contour-based classifier can capitalize on this higher-resolution assessment of language features. The graph in the lower panel of Figure 2 illustrates that the individual features of a given lexicon typically yield rather uninformative mean scores, as most sentences do not contain any of the relevant terms. The graph shows the feature score

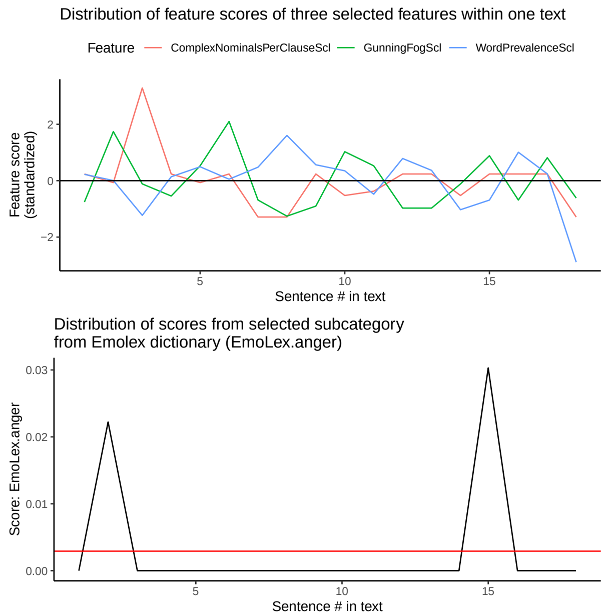

Figure 2: Text contours for selected features of a single text. Top panel: Distribution of three z-standardized language feature scores from three different feature groups (red:syntactic, green: psycholinguistic, blue: readability). Bottom panel: Text contour of an individual feature of a closed-vocabulary feature (anger words from EmoLex dictionary).

for terms from the EmoLex lexicon that are indicative of anger, which are zero in 16 out of the 18 sentences in the example text. So, while a means-based classifier would evaluate the text on the basis of its text-average score – given by the red line in Figure 2 – the contour-based approach can identify and capitalize on information bearing-signals in the shape of local peaks.

5. Experimental setup

In this section, we describe six benchmark models for the introduced dataset: (1) a fine-tuned Bidirectional Encoder Representations from Transformers (BERT) model, (2) and (3) two convolutional neural network classifiers trained on text-average ('means-based') and sentence-level ('contour-based') language features respectively, (4) a hybrid model integrating BERT predictions with the language features, (5) a hybrid model combining language features and sociodemographic features and (6) a full model integrating language features and sociodemographic features with BERT predictions. In each model, each speaker in the training data is considered as a data point. The input of the model consists of all the text sequences of each speaker. The output is the class of the corresponding speaker, 0 for "low" on a given personality dimension and 1 for "high" on that dimension, where binary classes are derived from continuous BFI scores using median splits. Following recent work in personality prediction research (e.g. Başaran and Ejimogu, 2021, Ramezani et al., 2020), all models were trained in a multi-label classification setting (Tsoumakas and Katakis, 2007), as personality labels cannot be assumed to be statistically independent (see section 3.2). We evaluated all models using 20 times repeated 10-fold crossvalidation to counter variability due to weight initialization. We report performance metrics averaged over all runs. All models are implemented using PyTorch (Paszke et al., 2019). Unless specifically stated otherwise, we use 'BCELoss' as our loss function, 'AdamW' as optimizer, one cycle learning rate scheduler (OCLR)(Smith and Topin, 2017) and $dropout = 0.3$, $L2 = 1 \times 10^{-4}$ as the regularization. The optimal network structures and values of hyperparameters are found by grid-search.

5.1. Fine-tuned BERT Model (BERT-BLSTM)

Since their inception, transformer-based pretrained language models such as BERT (Devlin et al., 2018) have achieved state-of-the-art performance in various classification tasks. Here we used the Huggingface Transformers library (Wolf et al., 2020) for fine-tuning a pretrained 'bert-base-uncased' model. The model consists of 12 Transformer layers with hidden size 768 and 12 attention heads. We run experiments with (1) a linear fully-connected layer for classification as well as with (2) an intermediate bidirectional LSTM layer with 256 hidden units (Al-Omari et al., 2020) (BERT-BLSTM). The following hyperparameters are used for fine-tuning: a fixed learning rate of 2×10^{-5} is applied and $L2$ regularization of 1×10^{-6}. All models were trained for 8 epochs, with batch size of 4 and maximum sequence length of 512. No dropout is used. We focus here on the results of the best-performing transformer-based model, namely BERT-BLSTM.

5.2. BLSTM Contour-based language features (BLSTM-CBLF)

To utilize the information carried by the contour-based measurement of language features (cf. Figure 2), we build a bidirectional LSTM neural network (BLSTM). Specifically, a 2-layer BLSTM with a hidden state dimension of 32. The input to the model is a sequence $X = (x_1, x_2, \ldots, x_n)$, where x_i, the output of CoCoGen for the ith sentence of a document, is a 436 dimensional vector and n is the sequence length. To predict the labels of a sequence, we concatenate the last hidden states of the last layer in forward ($\overrightarrow{h_n}$) and backward directions ($\overleftarrow{h_n}$). The result vector of concatenation h_n is then transformed through a fully connected layer, whose activation function is Parametric Rectifier Linear Unit (PReLU). The output of this network is a vector of the five binarized personality traits. More precisely:

$$\begin{aligned}[\overrightarrow{h_n}, \overleftarrow{h_n}] &= \text{BLSTM}(X)\\ f &= \text{PReLU}(W_f h_n + b_f)\\ y &= \sigma(W_o f + b_o)\end{aligned}$$

where $[\cdot|\cdot]$ is concatenation operator and σ is sigmoid function. $\overrightarrow{h_n}, \overleftarrow{h_n}$ are 64 dimensional vectors and their concatenation $h_n = [\overrightarrow{h_n}^T | \overleftarrow{h_n}^T]^T$ is a 128 dimensional vector. $W_f \in \mathbb{R}^{128\times64}$ and $W_o \in \mathbb{R}^{64\times5}$. Bias terms b_f and b_o are of dimension 64 and 5 respectively. The min and max OCLR learning rates are 1×10^{-5} and 1×10^{-3}

5.3. BLSTM Means-based language features (BLSTM-MBLF)

To evaluate the utility of adopting the contour-based approach, we also build a means-based model for purposes of comparison. This model is a 4-layer feed forward neural network:

$$\begin{aligned}f_i &= \text{Tanh}(W_i f_{i-1} + b_i), i = 1, 2\\ f_3 &= \text{PReLU}(W_3 f_2 + b_3)\\ y &= \sigma(W_4 f3 + b_4)\end{aligned}$$

where $W_1 \in \mathbb{R}^{436\times64}, W_2 \in \mathbb{R}^{64\times64}, W_3 \in \mathbb{R}^{64\times32}$ and $W_4 \in \mathbb{R}^{32\times5}$. The bias terms $b1, b2, b3, b4$ are vectors of dimension $64, 64, 32, 5$ respectively. Given a sequence of contour-based language features $X = (x_1, x_2, \ldots, x_n)$, $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is computed and used as feature of our mean based model, i.e. $f_0 = \bar{X}$. While training, we use the same max OCLR learning rate as 5.2 but with a min learning rate of 6×10^{-5}

5.4. BLSTM-CBLF+BERT

Following Lee et al. (2021), we assemble our hybrid model by (1) obtaining soft labels $g \in \mathbb{R}^5$ (probabilities that a text belongs to the corresponding trait class) from the fine-tuned BERT model by applying sigmoid layer on top of its output logits and then (2) interweaving g with $h_n \in \mathbb{R}^{64}$ described in Section 5.2 by concatenating g with h_n after a linear layer with activation funtion PReLU (3) The concatenated vector is fed into a 2-layer feedforward classifier. Specifically:

$$\begin{aligned}f_1 &= \text{PReLU}(W_1 h_n + b_1)\\ f_3 &= \text{PReLU}(W_3 f_2 + b_3)\\ y &= \sigma(W_o f_3 + b_o)\end{aligned}$$

where $f_2 = [f_1^T | g^T]^T$, $W_1 \in \mathbb{R}^{64\times32}, W_3 \in \mathbb{R}^{37\times32}, W_o \in \mathbb{R}^{32\times5}$ and $b_1, b_3 \in \mathbb{R}^{32}, b_o \in \mathbb{R}^5$. The best minimum and maximum OCLR learning rates are found to be 3×10^{-5} and 1×10^{-3}

5.5. BLSTM-CBLF + speaker background (BLSTM-CBLF+BG)

This model has the same structure as BLSTM-CBLF + BERT model described in Section 5.4. However instead of g, we concatenate the speaker background vector $b \in \mathbb{R}^{54}$ with f_1. Correspondingly, $W_1 \in \mathbb{R}^{64\times32}, W_3 \in \mathbb{R}^{86\times32}, W_1 \in \mathbb{R}^{32\times5}$ and $b_1, b_3 \in \mathbb{R}^{32}, b_o \in \mathbb{R}^5$. Same OCLR parameter setups are applied as 5.4.

5.6. BLSTM-CBLF + speaker background + BERT (BLSTM-CBLF+BG+BERT)

In this model with structure described in Section 5.4 is extended with the prediction of the BERT model described in Section 5.1. In order to interweave both speaker background and BERT soft predictions f_2 is instead computed as: $f_2 = [f_1^T | g^T | b^T]^T$ Correspondingly, $W_1 \in \mathbb{R}^{64 \times 32}, W_3 \in \mathbb{R}^{91 \times 32}, W_1 \in \mathbb{R}^{32 \times 5}$ and $b_1, b_3 \in \mathbb{R}^{32}, b_o \in \mathbb{R}^5$. Same OCLR parameter setups are applied as 5.4.

5.7. Feature ablation

We performed feature ablation studies to assess the informativeness of a feature group in the prediction of each of the five personality traits by quantifying the change in predictive power when comparing the performance of a classifier trained with the all feature groups versus the performance without a particular feature group. Specifically, we employed Submodular Pick Lime (SP-LIME) (Ribeiro et al., 2016), a method to construct a global explanation of a model by aggregating the weights of the linear models. We first construct local explanations using LIME with a linear local explanatory model, exponential kernel function with Hamming distance and a kernel width of $\sigma = 0.75\sqrt{d}$, where d is the number of feature groups. The global importance score of the SP-LIME for a given feature group j can then be derived by: $I_j = \sqrt{\sum_{i=1}^{n} |W_{ij}|}$, where W_{ij} is the jth coefficient of the fitted linear regression model to explain a data sample x_i.

6. Results

The models were evaluated using accuracy, precision, recall and F1 scores as the performance metrics. For reasons of space, we focus here on the discussion of classification accuracy[4] presented in Table 6. An overview of the results of all performance metrics is provided in Table 8 in the appendix. All benchmark models showed a consistent improvement over a majority class baseline model with an average increase in classification accuracy ranging between +6.25%, for the BERT model, to +12.7%, for the hybrid model that integrates the contour-based BLSTM with the speaker-level background variables and BERT predictions. The use of text-contours improved the average classification accuracy of the language-based classification over a means-based approach by +2.16% and improved average precision by +3.06%. At the level of individual personality traits, the contour-based model outperformed the means-based model in four of the five traits (extraversion, conscientiousness, neuroticism and openness), reaching a maximum increase in performance of +4.18%. For the only trait where it did not outperform the means-based model (agreeableness), the difference in classification accuracy was only 0.61%. These results clearly demonstrate the benefits of utilizing the within-text distributions of (psycho-)linguistic features for personality detection. The performance of the fine-tuned BERT model was relatively poor: While the best-performing hybrid model (BLSTM-CBLF+BG+BERT) included BERT predictions, the improvement over a model without the BERT predictions (BLSTM-CBLF+BG) was only marginal (+0.04%). In isolation, the fine-tuned BERT model lagged behind both BLSTM models based on (psycho-)linguistic features, with the contour-based model reaching as much as +3.99% higher average classification accuracy than the BERT model. This result contrasts with previous findings on other datasets (Mehta et al., 2020), which found that language modeling features (BERT word embeddings) consistently beat conventional psycholinguistic features. Of course, the overall success of a model based on engineered features hinges on the predictive value of its features, as well as on the resolution at which these features are measured. This work has shown that a model based on transparent language features is not only more transparent than a neural language model but can also outperform such a model in terms of prediction accuracy. The comparatively weak performance of the BERT model may be related to the fact that spoken language is characterized by many phenomena, such as hesitations, repetition, incomplete utterances, that do not appear within the types of written data (Wikipedia and the Book Corpus) that the BERT model was trained on. Future research may investigate these issues in more detail. The integration of socio-demographic background variables available in ♠ SPADE led to an improvement in average classification accuracy of +1.74% over the next-best model that did not include this information but did include BERT predictions. This result demonstrates the usefulness of incorporating rich socio-demographic background into future datasets for personality prediction.

The results of the feature ablation experiments are presented in Table 6. We find that overall the predictions of the model are most strongly driven by the feature groups 'sentiment', 'LIWC' and 'psycholinguistic'. However, with the exception of the information theoretic feature group, all features groups exhibited relatively high I-values (all I>2.5 relative to maxima of I=5 to 5.1 across personality traits). Pairwise correlations (Spearman rank-order correlation coefficients) of feature importance scores across personality traits were observed to be largest for the pair conscientiousness-openness ($\rho = 0.93$) and lowest for the pair extraversion-neuroticm ($\rho = 0.81$) (see also Figure 6 in the appendix). In follow up analyses, we sought to identify characteristic language features associated with each personality trait by computing the difference between the z-standardized mean scores of high- and low-scoring individuals on a given trait. This analysis revealed some interesting patterns: While

[4] Micro averages were used to evaluate the 100 model instances; macro averages were used to evaluate the model across traits.

Model	E	A	C	N	O	Avg.
Majority class baseline	49.93	49.85	49.93	49.88	49.69	49.86
BERT-BLSTM	54.21	59.45	57.84	50.55	58.5	56.11
BLSTM-MBLF	58.75	62.37	53.11	54.04	61.44	57.94
BLSTM-CBLF	61.06	61.76	57.07	58.22	62.37	60.10
BLSTM-CBLF+BERT	**61.20**	63.51	58.09	58.32	63.01	60.82
BLSTM-CBLF+BG	59.41	65.45	**60.69**	60.13	**66.89**	62.52
BLSTM-CBLF+BG+BERT	59.31	**66.14**	59.95	**61.20**	66.22	**62.56**

Table 3: Evaluation results of the six benchmark models. Numbers represent classification accuracy (%) micro-averaged across 20 times 10-fold cv. In the 'Avg.' column, the macro-averaged classification accuracy (%) across 5 traits are presented.

E		A		C		N		O	
Group	I	Group	I	Group	I	Group	I	Group	I
Sentiment	5.10	Sentiment	5.07	Sentiment	4.89	Sentiment	4.82	Sentiment	5.00
LIWC	3.47	LIWC	3.68	LIWC	3.49	LIWC	3.45	LIWC	3.62
Psycholing	3.36	Psycholing	3.32	Psycholing	3.09	Emotion	3.05	Psycholing	3.19
Readability	3.12	Syntactic	3.16	Ngram	3.07	Ngram	2.95	Ngram	3.10
Emotion	3.04	Emotion	3.14	Emotion	3.04	Psycholing	2.94	Emotion	3.10
Ngram	2.99	Ngram	3.03	Syntactic	2.99	Syntactic	2.90	Syntactic	3.01
Syntactic	2.85	Readability	2.81	Readability	2.96	Readability	2.71	Readability	2.80
Lexical	2.64	Lexical	2.66	Lexical	2.65	Lexical	2.52	Lexical	2.64
InfTheo	1.33	InfTheo	1.48	InfTheo	1.36	InfTheo	1.35	InfTheo	1.40

Table 4: Results of the feature ablation experiment: Feature importance (Model: BLSTM-CBLF) macro-averaged across 200 model instances. (20 × 10-fold CV)

space limitations prevent a more detailed discussion, we observed for example that individuals scoring high on the extraversion scale showed higher proportions of words relating to power and positive emotion - but also to anger - as well as greater lexical diversity. Individuals scoring high on the neuroticism scale showed higher proportions of words related to anxiety and disappointment, as well as words associated with evaluation and conformity, but also larger amounts of n-grams from the register of academic language. Highly conscientious individuals showed high proportions of affiliation words (ally, friend) and high proportions of words referring to men and social roles associated with men. A visualization of some of the top-20 most characteristic individual features per personality trait and the top-2 most characteristic features per feature group by trait is presented in Figure 5 in the appendix. These results partially align with past findings from personality psychology based on language analysis: For example, the finding that extraversion is related to using more positive emotion words (e.g., great, happy, amazing) has been repeatedly obversed across many types of data (for overviews, see, e.g., Park et al., 2015; Boyd and Schwartz, 2021). However, most of the previous research has so far relied on closed-class, word counting approaches – predominantly LIWC. The results of the present work go beyond these findings by showing that individual personality traits are also characterized, for example, by different usage patterns of multi-word combinations (n-grams) as well as by more abstract measures of lexical variety.

7. Conclusion

In this paper, we introduce ♠ SPADE, a new data resource for modeling and predicting personality traits from speech behavior. A distinguishing feature of SPADE is the continuous and contextualized nature of speech samples combined with BigFive personality trait information obtained from a standard questionnaire completed by individual speakers. In addition, this dataset is enriched with socio-demographics for each speaker. Our best benchmark model achieved an average classification accuracy of 62.56%, which is in good agreement with SOTA results for personality predictions on the available benchmark datasets (Essay (Pennebaker and King, 1999): 60.6%, MBTI Kaggle (Li et al., 2018): 77.1; see Mehta et al., 2020). In addition, we show that automatic prediction of personality traits benefits from the inclusion of within-text distributions of linguistic features, as evidenced by higher accuracy of prediction models that utilize such text contours. By making the dataset available to the research community, we hope to facilitate research on automatic personality recognition from speech behavior.

8. Bibliographical References

Agarwal, B. (2014). Personality detection from text: A review. *International Journal of Computer System*, 1(1):1–4.

Al-Omari, H., Abdullah, M. A., and Shaikh, S. (2020). Emodet2: Emotion detection in english textual dialogue using bert and bilstm models. In *2020 11th International Conference on Information and Communication Systems (ICICS)*, pages 226–232.

Allport, G. W. (1942). The use of personal documents in psychological science. *Social Science Research Council Bulletin*.

Araque, O., Gatti, L., Staiano, J., and Guerini, M. (2019). Depechemood++: a bilingual emotion lexicon built through simple yet powerful techniques. *IEEE transactions on affective computing*.

Başaran, S. and Ejimogu, O. H. (2021). A neural network approach for predicting personality from facebook data. *SAGE Open*, 11(3):21582440211032156.

Boyd, R. L. and Schwartz, H. A. (2021). Natural language analysis and the psychology of verbal behavior: The past, present, and future states of the field. *Journal of Language and Social Psychology*, 40(1):21–41.

Bradley, M. M. and Lang, P. J. (1999). Affective norms for english words (anew): Instruction manual and affective ratings. Technical report, Technical report C-1, the center for research in psychophysiology ....

Browne, C. et al. (2013). The new general service list: Celebrating 60 years of vocabulary learning. *The Language Teacher*, 37(4):13–16.

Brysbaert, M., Mandera, P., McCormick, S. F., and Keuleers, E. (2019). Word prevalence norms for 62,000 english lemmas. *Behavior research methods*, 51(2):467–479.

Cambria, E., Speer, R., Havasi, C., and Hussain, A. (2010). Senticnet: A publicly available semantic resource for opinion mining. In *2010 AAAI fall symposium series*.

Davies, M. (2008). The Corpus of Contemporary American English (COCA): 560 million words, 1990-present.

Deutsch, P. (1996). Rfc1951: Deflate compressed data format specification version 1.3.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Freud, S. (1915). *The unconscious*. Penguin Classics.

Fry, E. (1968). A readability formula that saves time. *Journal of reading*, 11(7):513–578.

Funder, D. C. (2001). Accuracy in personality judgment: Research and theory concerning an obvious question.

Gjurković, M. and Šnajder, J. (2018). Reddit: A gold mine for personality prediction. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 87–97.

Gjurkovic, M., Karan, M., Vukojevic, I., Bosnjak, M., and Snajder, J. (2020). PANDORA talks: Personality and demographics on reddit. *CoRR*, abs/2004.04460.

Grubinger, T., Zeileis, A., and Pfeiffer, K.-P. (2014). evtree: Evolutionary learning of globally optimal classification and regression trees in r. *Journal of statistical software*, 61(1):1–29.

Iacobelli, F., Gill, A. J., Nowson, S., and Oberlander, J. (2011). Large scale personality classification of bloggers. In *International Conference on Affective Computing and Intelligent Interaction*, pages 568–577. Springer.

John, O. P., Donahue, E. M., and Kentle, R. L. (1991). Big five inventory. *Journal of Personality and Social Psychology*.

John, O. P., Naumann, L. P., and Soto, C. J. (2008). Paradigm shift to the integrative big five trait taxonomy: History, measurement, and conceptual issues. In R. W. Robins O. P. John et al., editors, *Handbook of personality: Theory and research*, page 114–158. The Guilford Press.

Johns, B. T., Dye, M., and Jones, M. N. (2020). Estimating the prevalence and diversity of words in written language. *Quarterly Journal of Experimental Psychology*, 73(6):841–855.

Kaushanskaya, M., Blumenfeld, H. K., and Marian, V. (2020). The language experience and proficiency questionnaire (leap-q): Ten years later. *Bilingualism: Language and Cognition*, 23(5):945–950.

Kerz, E., Qiao, Y., Wiechmann, D., and Ströbel, M. (2020). Becoming linguistically mature: Modeling english and german children's writing development across school grades. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 65–74.

Kerz, E., Qiao, Y., and Wiechmann, D. (2021). Language that captivates the audience: Predicting affective ratings of ted talks in a multi-label classification task. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 13–24.

Kosinski, M., Matz, S. C., Gosling, S. D., Popov, V., and Stillwell, D. (2015). Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American psychologist*, 70(6):543.

Kuperman, V., Stadthagen-Gonzalez, H., and Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 english words. *Behavior research methods*, 44(4):978–990.

Lee, B. W., Jang, Y. S., and Lee, J. H.-J. (2021). Pushing on text readability assessment: A transformer meets handcrafted linguistic features. *arXiv preprint arXiv:2109.12258*.

Levy, R. and Andrew, G. (2006). Tregex and tsurgeon: tools for querying and manipulating tree data structures. In *LREC*, pages 2231–2234. Citeseer.

Li, C., Hancock, M., Bowles, B., Hancock, O., Perg, L., Brown, P., Burrell, A., Frank, G., Stiers, F., Marshall, S., et al. (2018). Feature extraction from social media posts for psychometric typing of participants. In *International Conference on Augmented Cognition*, pages 267–286. Springer.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International journal of corpus linguistics*, 15(4):474–496.

Lu, X. (2012). The relationship of lexical richness to the quality of esl learners' oral narratives. *The Modern Language Journal*, 96(2):190–208.

Majumder, N., Poria, S., Gelbukh, A., and Cambria, E. (2017). Deep learning-based document modeling for personality detection from text. *IEEE Intelligent Systems*, 32(2):74–79.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., and McClosky, D. (2014). The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

McLaughlin, G. H. (1969). Clearing the smog. *Journal of Reading*.

Mehl, M. R., Pennebaker, J. W., Crow, D. M., Dabbs, J., and Price, J. H. (2001). The electronically activated recorder (ear): A device for sampling naturalistic daily activities and conversations. *Behavior research methods, instruments, & computers*, 33(4):517–523.

Mehta, Y., Fatehi, S., Kazameini, A., Stachl, C., Cambria, E., and Eetemadi, S. (2020). Bottom-up and top-down: Predicting personality with psycholinguistic and language model features. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 1184–1189. IEEE.

Mohammad, S. M. and Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Computational intelligence*, 29(3):436–465.

Mohammad, S., Kiritchenko, S., and Zhu, X. (2013). Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013)*, Atlanta, Georgia, USA, June.

Mohammad, S. (2018). Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184.

Oberlander, J. and Gill, A. J. (2006). Language with character: A stratified corpus comparison of individual differences in e-mail communication. *Discourse processes*, 42(3):239–270.

Ozer, D. J. and Benet-Martinez, V. (2006). Personality and the prediction of consequential outcomes. *Annu. Rev. Psychol.*, 57:401–421.

Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., Ungar, L. H., and Seligman, M. E. (2015). Automatic personality assessment through social media language. *Journal of personality and social psychology*, 108(6):934.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, et al., editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Pennebaker, J. W. and King, L. A. (1999). Linguistic styles: language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296.

Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Pennebaker, J. W., Booth, R. J., and Francis, M. E. (2007). Linguistic inquiry and word count: Liwc [computer software]. *Austin, TX: liwc. net*, 135.

Plank, B. and Hovy, D. (2015). Personality traits on twitter—or—how to get 1,500 personality tests in a week. In *Proceedings of the 6th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 92–98.

Ramezani, M., Feizi-Derakhshi, M.-R., Balafar, M.-A., Asgari-Chenaghlu, M., Feizi-Derakhshi, A.-R., Nikzad-Khasmakhi, N., Ranjbar-Khadivi, M., Jahanbakhsh-Nagadeh, Z., Zafarani-Moattar, E., and Rahkar-Farshi, T. (2020). Automatic personality prediction; an enhanced method using ensemble modeling. *arXiv preprint arXiv:2007.04571*.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Rorschach, H. (1921). *Psychodiagnostik: Methodik und ergebnisse eines warhrnehmungsdiagnostischen Experiments (deutenlassen von zufallsformen)*, volume 2. E. Bircher.

Scherer, K. R. (2005). What are emotions? and how can they be measured? *Social science information*, 44(4):695–729.

Shmueli, B., Fell, J., Ray, S., and Ku, L.-W. (2021). Beyond fair pay: Ethical implications of nlp crowdsourcing. *arXiv preprint arXiv:2104.10097*.

Smith, L. N. and Topin, N. (2017). Super-convergence: Very fast training of residual networks using large learning rates. *CoRR*, abs/1708.07120.

Soto, C. J. (2019). How replicable are links between personality traits and consequential life outcomes? the life outcomes of personality replication project. *Psychological Science*, 30(5):711–727.

Štajner, S. and Yenikent, S. (2021). How to obtain reliable labels for mbti classification from texts? In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1360–1368.

Stone, P. J., Bales, R. F., Namenwirth, J. Z., and Ogilvie, D. M. (1962). The general inquirer: A computer system for content analysis and retrieval based on the sentence as a unit of information. *Behavioral Science*, 7(4):484.

Stone, P. J., Dunphy, D. C., and Smith, M. S. (1966). The general inquirer: A computer approach to content analysis.

Ströbel, M., Kerz, E., Wiechmann, D., and Neumann, S. (2016). Cocogen-complexity contour generator: Automatic assessment of linguistic complexity using a sliding-window technique. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 23–31.

Ströbel, M. (2014). *Tracking complexity of l2 academic texts: A sliding-window approach*. Master thesis. RWTH Aachen University.

Tsoumakas, G. and Katakis, I. (2007). Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13.

Uleman, J. S., Adil Saribay, S., and Gonzalez, C. M. (2008). Spontaneous inferences, implicit impressions, and implicit theories. *Annu. Rev. Psychol.*, 59:329–360.

Wiegmann, M., Stein, B., and Potthast, M. (2019). Celebrity profiling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2611–2618.

Wilt, J. and Revelle, W. (2015). Affect, behaviour, cognition and desire in the big five: An analysis of item content and structure. *European journal of personality*, 29(4):478–497.

Wolf, T., Chaumond, J., Debut, L., Sanh, V., Delangue, C., Moi, A., Cistac, P., Funtowicz, M., Davison, J., Shleifer, S., et al. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Zechner, K., Higgins, D., Xi, X., and Williamson, D. M. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken english. *Speech Communication*, 51(10):883–895.

Appendix

Table 5: Self-reported experience & proficiency of participants.

	Mono		Biling		L2 Eng	
	M	SD	M	SD	M	SD
Age of acquisition	1.21	2.82	1.91	4.64	6.55	3.98
Age became fluent	4.47	3.35	5.32	3.66	14.45	7.83
Age started reading	4.53	1.65	4.91	3.33	8.95	3.97
Age fluent reading	6.8	3.01	7	3.29	12.79	5.56
Current exposure to English (0-10)						
Family	9.02	2.47	7.54	3.68	4.34	3.2
Friends	8.79	2.69	7.7	3.6	5.95	3.17
Reading	8.9	2.51	7.88	3.65	7.79	2.75
Classroom	6.32	4.18	6.26	4.26	6.53	3.63
Workplace	8.25	3.03	7.51	3.59	7.18	3.07
Videogames	7.34	3.66	6.6	3.9	4.84	3.48
Watching TV	8.3	2.82	7.46	3.5	7.63	2.92
Listening to music	8.15	3	7.14	3.53	7.32	2.8
Social media	7.84	3.29	7.16	3.75	7.5	2.86
Self-rated English proficiency (0-10)						
Speaking	9.39	1.65	9.3	1.77	7.66	1.88
Listening	9.52	1.61	9.28	1.73	8.71	1.74
Reading	9.49	1.43	9.51	1.14	9.11	1.66
Writing	9.55	1.36	9.37	1.22	8.16	1.59

Table 6: Self-reported reading habits of participants

	Mono		Biling		L2 Eng	
	M	SD	M	SD	M	SD
Time per week currently spent reading English						
Books	3.43	1.42	3.46	1.54	3.27	1.52
Magazines	1.88	1	2.11	1.22	1.92	0.95
Videogames	2.84	1.39	2.84	1.28	2.44	1.46
Social media	3.54	1.12	3.47	1.07	3.33	1.29
Websites	3.82	0.97	3.57	1.11	4.14	0.93
Time per week spent reading English (past)						
Books	3.04	1.39	2.98	1.29	2.82	1.37
Magazines	1.83	1.11	2.09	1.38	1.79	0.96
Videogames	2.75	1.51	2.86	1.55	2.39	1.57
Social media	3.61	1.37	3.61	1.19	3.42	1.5
Websites	4.03	1.23	3.88	1.3	4.34	1.17

Table 7: Evaluation results of the six benchmark models:

Measure	Model	E	A	C	N	O	Avg.
precision	BLSTM-CBLF	62.79	63.88	57.67	58.63	61.81	60.95
	BLSTM-MBLF	58.74	64.74	52.28	53.12	60.55	57.89
	BERT-FullyConnected	59.62	74.13	60.31	54.32	56.74	61.02
	BERT-BLSTM	47.79	46.01	53.3	29.62	30.2	41.38
	BLSTM-CBLF+BERT	62.32	66.43	57.39	57.89	62.22	61.25
	BLSTM-CBLF+BG	59.48	66.70	60.81	59.66	66.63	62.66
	BLSTM-CBLF+BG+BERT	59.19	67.51	59.68	60.59	65.66	62.53
recall	BLSTM-CBLF	48.02	57.76	46.20	49.67	58.24	51.98
	BLSTM-MBLF	49.67	57.76	48.04	51.79	58.35	53.12
	BERT-FullyConnected	52.14	59.72	49.33	50.04	57.53	53.75
	BERT-BLSTM	58.25	59.31	58.8	49.66	64.44	58.09
	BLSTM-CBLF+BERT	50.16	57.71	55.71	54.40	60.00	55.60
	BLSTM-CBLF+BG	50.66	64.58	55.33	57.23	63.30	58.22
	BLSTM-CBLF+BG+BERT	51.32	64.95	55.98	59.24	63.35	58.97
F1	BLSTM-CBLF	54.42	60.67	51.30	53.78	59.97	56.03
	BLSTM-MBLF	53.83	61.05	50.07	52.45	59.43	55.37
	BERT-FullyConnected	54.64	64.36	52.81	49.38	50.68	54.37
	BERT-BLSTM	49.13	49.65	53.46	32.82	36.76	44.36
	BLSTM-CBLF+BERT	55.59	61.76	56.54	56.09	61.09	58.21
	BLSTM-CBLF+BG	54.72	65.63	57.94	58.42	64.92	60.32
	BLSTM-CBLF+BG+BERT	54.97	66.21	57.77	59.91	64.49	60.67
accuracy	Majority class baseline	49.93	49.85	49.93	49.88	49.69	49.86
	BLSTM-CBLF	61.06	61.76	57.07	58.22	62.37	60.10
	BLSTM-MBLF	58.75	62.37	53.11	54.04	61.44	57.94
	BERT-FullyConnected	52.53	61.42	54.08	51.05	55.84	54.98
	BERT-BLSTM	54.21	59.45	57.84	50.55	58.5	56.11
	BLSTM-CBLF+BERT	**61.20**	63.51	58.09	58.32	63.01	60.82
	BLSTM-CBLF+BG	59.41	65.45	**60.69**	60.13	**66.89**	62.52
	BLSTM-CBLF+BG+BERT	59.31	**66.14**	59.95	**61.20**	66.22	**62.56**

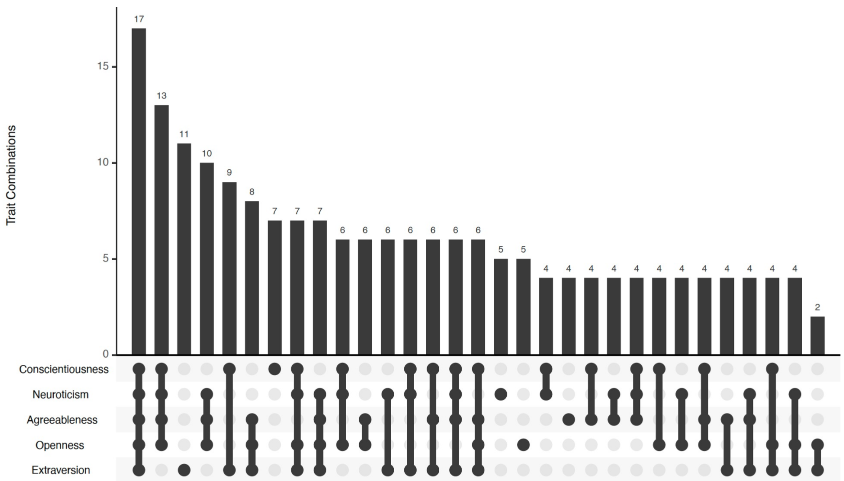

Figure 3: Frequency of personality trait combinations. For each of the five personality dimensions, a trait was considered present when an individual's BFI score was greater than the group median on a given dimension.

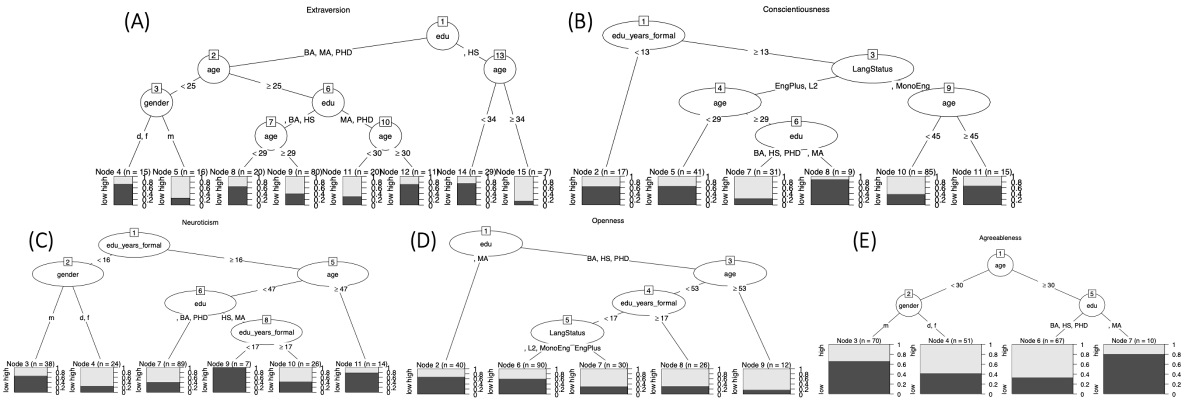

Figure 4: Dependencies of personality on demographic variables: The figure shows the results of a globally optimal classification tree method implemented in the evtree algorithm (Grubinger et al., 2014) used to visualized the interactions among personality and demographic variables. This method searches over the parameter space of trees using an evolutionary algorithms, which can identify patterns hidden by traditional methods that use a greedy heuristic, where split rules are selected in a forward stepwise search for recursively partitioning the data into groups. The analysis revealed several interesting patterns. For example, very high proportions of openness were observed for individuals between 30 and 53 years of age that spoke an additional language next to their L1 English.

Figure 5: Upper panel: Top 20 most characteristic features from each feature group by personality trait. Lower panel: Top 2 most characteristic features from each feature group by personality trait. Plotted scores represent the difference between the z-standardized mean scores of high- and low-scoring individuals on a given personality trait. Positive scores are characteristic of the high-scoring individuals on a given trait (e.g. individuals with high extraversion scores).

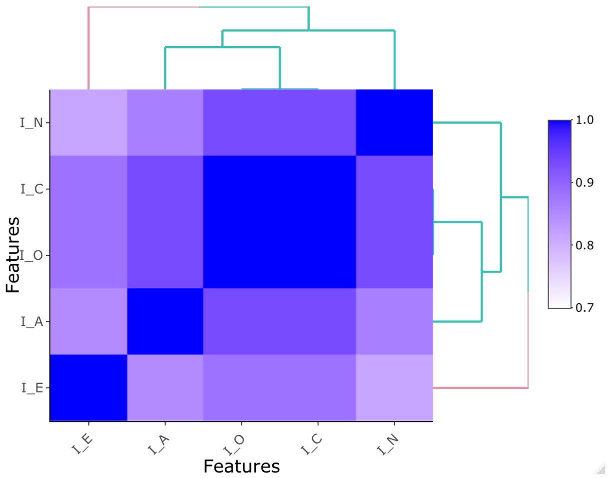

Figure 6: Pairwise correlations (Spearman rank-order correlation coefficients) for feature importance (I) scores across personality traits. Correlations of I scores were observed to be largest for the pair conscientiousness-openness (ρ = 0.93) and lowest for the pair extraversion-neuroticm

.