Pushing on Personality Detection from Verbal Behavior: A Transformer Meets Text Contours of Psycholinguistic Features

Elma Kerz
RWTH-Aachen University
elma.kerz@ifaar.rwth-aachen.de

Yu Qiao
RWTH-Aachen University
yu.qiao@rwth-aachen.de

Sourabh Zanwar
RWTH-Aachen University
sourabh.zanwar@rwth-aachen.de

Daniel Wiechmann
University of Amsterdam
d.wiechmann@uva.nl

Abstract

Research at the intersection of personality psychology, computer science, and linguistics has recently focused increasingly on modeling and predicting personality from language use. We report two major improvements in predicting personality traits from text data: (1) to our knowledge, the most comprehensive set of theory-based psycholinguistic features and (2) hybrid models that integrate a pre-trained Transformer Language Model BERT and Bidirectional Long Short-Term Memory (BLSTM) networks trained on within-text distributions ('text contours') of psycholinguistic features. We experiment with BLSTM models (with and without Attention) and with two techniques for applying pre-trained language representations from the transformer model - 'feature-based' and 'fine-tuning'. We evaluate the performance of the models we built on two benchmark datasets that target the two dominant theoretical models of personality: the Big Five Essay dataset (Pennebaker and King, 1999) and the MBTI Kaggle dataset (Li et al., 2018). Our results are encouraging as our models outperform existing work on the same datasets. More specifically, our models achieve improvement in classification accuracy by 2.9% on the Essay dataset and 8.28% on the Kaggle MBTI dataset. In addition, we perform ablation experiments to quantify the impact of different categories of psycholinguistic features in the respective personality prediction models.

1 Introduction

Personality is broadly defined as the combination of a person's behavior, emotions, motivation, and characteristics of thought patterns (Corr and Matthews, 2020). Our personality has a major impact on our lives, influencing our life choices, well-being, health, and preferences and desires (Ozer and Benet-Martinez, 2006). Specifically, personality has been repeatedly linked to individual (e.g., happiness, physical and mental health), interpersonal (e.g., quality of relationships with peers, family, and romantic partners), and social-institutional outcomes (e.g., career choice, satisfaction and achievement, social engagement, political ideology) (Soto, 2019).

While there are several models of human personality, the predominant and widely accepted model is the Big Five or Five Factor Model (McCrae and John, 1992; McCrae, 2009). In this model, personality traits are divided into five factors: (1) Extraversion (assertive, energetic, outgoing, etc.), (2) Agreeableness (appreciative, generous, compassionate, etc.), (3) Conscientiousness (efficient, organized, responsible, etc.), (4) Neuroticism (anxious, self-pitying, worried, etc.), and (5) Openness (curious, empathetic, imaginative, etc.). These five personality traits are commonly assessed by questionnaires in which a person reflects on his or her typical patterns of thinking and behavior, such as the NEO Five Factor Inventory (Costa and McCrae, 1992), and the Big-Five Inventory (John et al., 1991); (see Matthews et al., 2009, for a comprehensive overview). The Myers–Briggs Type Indicator (MBTI) is another widely administered questionnaire, in particular in applied settings (Meyers et al., 1990). In contrast to the Big Five personality taxonomy, which conceptualizes human personality as latent trait scores, the MBTI model describes personality in terms of 16 types that result from combining binary categories into four dimensions: (a) Extraversion/Introversion (E/I) - preference for how people direct and receive their energy, based on the external or internal world, (b) Sensing/Intuition (S/N) - preference for how people take in information, through the five senses or through interpretation and meanings, (c) Thinking/Feeling (T/F) - preference for how people make decisions, relying on logic or emotion over people and partic-

Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis, pages 182 - 194

ular circumstances, and (d) Judgment/Perception (J/P) - how people deal with the world, by ordering it or remaining open to new information.

Given its central importance in capturing the essential aspects of human life, increasing attention is being paid to the development of models that can leverage behavioral data to automatically predict personality. Data obtained from verbal behavior is one of the key types 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 (see Tausczik and Pennebaker, 2010; Boyd and Pennebaker, 2017, for historical overviews). Early approaches to automatic personality prediction (APP) – also referred to as automatic personality prediction or recognition – from textual data have relied on machine learning models based on psycholinguistic features, whereas more recent approaches to APP typically draw on deep learning techniques that use pre-trained word embeddings (see Vinciarelli and Mohammadi, 2014, for an overview of the former) (see Mehta et al., 2020b, for an overview of deep learning-based APP).

In this paper, we make a valuable contribution to this dynamic area of APP research by presenting two important improvements in predicting personality traits from textual data: (1) to our knowledge, the most comprehensive set of psycholinguistic features and (2) hybrid models that integrate a pre-trained Transformer Language Model BERT and Bidirectional Long Short-Term Memory (BLSTM) networks trained on in-text distributions ('text contours') of psycholinguistic features. Since our goal is to demonstrate the utility of our modeling approach, we conduct our experiments on two widely used benchmark datasets: the Big Five Essay dataset (Pennebaker and King, 1999) and the MBTI-Kaggle dataset (Li et al., 2018), which align with the dominant personality models described above. The remainder of this paper is organized as follows: In Section 2, we briefly review recent related work on these two benchmark datasets. Then, in Section 3, we present the two benchmark datasets and the extraction of psycholinguistic features using automated text analysis based on a sliding window approach. In Section 4, we describe our modeling approach, and in Section 5, we present and discuss the results. Finally, we conclude with possible directions for future work in Section 6.

2 Related work

Majumder et al. (2017) used a convolutional neural network (CNN) feature extractor in which sentences were fed to convolution filters to obtain n-gram feature vectors. Each individual text of the Big Five Essay dataset was represented by aggregating the vectors of its sentences and the obtained vectors were concatenated with psycholinguistic (Mairesse) features (Mairesse et al., 2007). For classification, they fed the resulting document vector to a fully connected neural network with one hidden layer. Using this method, they were able to achieve an average classification accuracy of 58% for the Big Five personality traits on the Essays dataset. Kazameini et al. (2020) were the first to use a Transformer-Based Language model to extract contextualized word embeddings. Specifically, they built a Bagged-SVM classifier fed with contextualized embeddings extracted from BERT, a pre-trained language model with a Bidirectional Encoder from Transformers (Devlin et al., 2018). Their model outperformed the CNN-based model proposed by the Majumder et al. (2017) model by 1.04%. Amirhosseini and Kazemian (2020) used a Gradient Boosting Model (GBM) based on Term Frequency–Inverse-Document-Frequency features (TF-IDF) to predict personality dimensions in the Kaggle MBTI dataset. Their modeling approach achieved an average classification accuracy across all dimensions of 76.1%. Using both the Big Five Essay dataset and the Myers-Briggs' type indicator Kaggle Dataset, Mehta et al. (2020a) proposed the integration of deep learning models and psycholinguistic features with language model embeddings for APP. They extracted a total of 123 psycholinguistic features, including the Mairesse features set (Mairesse et al., 2007), SenticNet (Cambria et al., 2010), NRC-Emotion Lexicon (Mohammad and Turney, 2013), and NRC-VAD Lexicon (Mohammad, 2018). Language model features were extracted using BERT. Their experiments compared the performance of BERT-base and BERT-large in synergy with SVM or Multi-layer Perceptron (MLP) classifiers. BERT-base + MLP yielded an average score of 60.6 on the Essay dataset, while BERTlarge + MLP yielded an average score of 77.1 on the Kaggle dataset. The approach taken in Mehta et al. (2020a) outperformed the previously best-performing model by Amirhosseini and Kazemian (2020) by 1%. Zooming on classification accuracy for specific personality traits, the

models in Mehta et al. (2020a) achieved the highest performance on two of the Big Five personality traits in the Essays dataset (openness, accuracy = 64.6%, and conscientiousness, accuracy = 59.2%) and on three of the four MBTI dimensions in the Kaggle MBTI dataset (Intuitive/Sensing (N/S), accuracy = 86.6%, Thinking/Feeling (T/F), accuracy = 76.1% and Perception/Judging (P/J), accuracy = 67.2%). The highest performance on the Introversion/Extraversion (I/E) MBTI dimension (79%) was obtained by the 'GBM + TFIDF' model reported in Amirhosseini and Kazemian (2020). The highest performance on the three remaining Big Five dimensions was achieved recently by Ramezani et al. (2021), which used an ensemble modeling approach (stacking) to combine linguistic and ontology-based features with deep learning-based methods based on a hierarchical attention network as a meta-model. Although the overall performance of SOTA on the Essay dataset was not superior - mainly due to relatively poor performance on the Openness trait (accuracy = 56.3%), this work has demonstrated the utility of model stacking as an effective way to boost the prediction of personality traits. For a performance overview of the models reviewed here for different data sets and personality dimensions, see Table 1 in Section 4.

3 Method

3.1 Datasets

We conducted our experiments with two widely used personality benchmark datasets: (1) The Essays Dataset (Pennebaker and King, 1999) and (2) Kaggle MBTI Dataset (Li et al., 2018). (1) Essays: This stream-of-consciousness dataset consists of 2468 essays written by students and annotated with the binary labels of the Big Five personality traits, which were obtained through a standardized self-report questionnaire. The average text length is 672 words and the total size of the dataset is approximately 1.6 million words. One of the reasons why Essays is an established benchmark dataset is the relatively large amount of continuous language use and the fact that the personality traits were obtained using a validated instrument. (2) Kaggle MBTI: This dataset was collected through the PersonalityCafe forum[1] and thus provides a diverse sample of people interacting in an informal online social environment. It consists of samples of social media interactions from 8675 users, all of whom indicated their MBTI type. The average text length is 1,288 words. The total size of the entire dataset is approximately 11.2 million words.

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

3.2 Measurement of text contours of psycholinguistic features

The texts from both datasets (the Big Five Essay dataset and the MBTI Kaggle dataset) were automatically analyzed using an automated text analysis (ATA) system that employs a sliding window technique to compute sentence-level measurements. These measurements capture the within-text distributions of scores for a given psycholinguistic feature, referred to here as 'text contours' (for recent applications of the ATA system in the context of text classification, see (Kerz et al., 2020; Qiao et al., 2021a,b). We extracted a set of 437 theory-based psycholinguistic features that can be binned into four groups: (1) features of morpho-syntactic complexity (N=19), (2) features of lexical richness, diversity and sophistication (N=77), (3) readability features (N=14), and (4) lexicon features designed to detect sentiment, emotion and/or affect (N=326). Tokenization, sentence splitting, part-of-speech tagging, lemmatization and syntactic PCFG parsing were performed using Stanford CoreNLP (Manning et al., 2014). The group of **morpho-syntactic complexity features** includes (i) surface features related to the length of production units, such as the average length of clauses and sentences, (ii) features of the type and frequency of embeddings, such as number of dependent clauses per T-Unit or verb phrases per sentence and (iii) the frequency of particular structure types, such as the number of complex nominals per clause. This group also includes (iv) information-theoretic features of morphological and syntactic complexity based on the Deflate algorithm (Deutsch, 1996). The group of **lexical richness, diversity and sophistication features** includes six different subtypes: (i) lexical density features, such as the ratio of the number of lexical (as opposed to grammatical) words to the total number of words in a text, (ii) lexical variation, i.e. the range of vocabulary as manifested in language use, captured by text-size corrected type-token ratio, (iii) lexical sophistication, i.e. the proportion of relatively unusual or advanced words in a text, such as the number of words from the New General Service List (Browne et al., 2013), (iv) psycholinguistic norms of words, such as the

average age of acquisition of the word (Kuperman et al., 2012) and two recently introduced types of features: (v) word prevalence features that capture the number of people who know the word (Brysbaert et al., 2019; Johns et al., 2020) and (vi) register-based n-gram frequency features that take into account both frequency rank and the number of word n-grams ($n \in [1, 5]$). The latter were derived from the five register subcomponents of the Contemporary Corpus of American English (COCA, 560 million words, Davies, 2008): spoken, magazine, fiction, news and academic language (see Kerz et al., 2020, for details see e.g.). The group of **readability features** combines a word familiarity variable defined by a prespecified vocabulary resource to estimate semantic difficulty along with a syntactic variable, such as average sentence length. Examples of these measures include the Fry index (Fry, 1968) or the SMOG (McLaughlin, 1969). The group of **lexicon-based sentiment/emotion/affect features (SentiEmo)** was derived from a total of ten lexicons that have been successfully used in personality detection, emotion recognition and sentiment analysis research: (1) The Affective Norms for English Words (ANEW) (Bradley and Lang, 1999), (2) ANEW-Emo lexicons (Stevenson et al., 2007), (3) DepecheMood++ (Araque et al., 2019), (4) The Geneva Affect Label Coder (GALC) (Scherer, 2005), (5) The General Inquirer (Stone et al., 1966), (6) The LIWC dictionary (Pennebaker et al., 2001), (7) The NRC Word-Emotion Association Lexicon (Mohammad and Turney, 2013), (8) The NRC Valence, Arousal, and Dominance (NRC-VAD) lexicon (Mohammad, 2018), (9) SenticNet (Cambria et al., 2010), and (10) the Sentiment140 lexicon (Mohammad et al., 2013). The feature value for each subcategory in a given lexicon is the mean value of all rated/scored words in a given sentence. The informational gain of 'text contours' compared to text-averages is illustrated in Figure 1. The Figure shows the distribution of z-standardized values of three selected features for a randomly selected text from the Essay dataset. The red line represents the average feature value of the text. As can be seen from the graphs, all feature values fluctuate within the text, with high values for one feature often offset by lower values for another. The contour-based classifiers, discussed in more detail in Section 3, can take advantage of this high-resolution assessment of psycholinguistic features.

4 Modeling approach

Our models are constructed from three components: (a) a 'contour encoder' that converts a sequence of psycholinguistic features into a hidden representation vector, (b) a pre-trained transformer-based language model, BERT, that converts a sequence of tokens into a hidden representation vector, and (c) a classifier that outputs the probability of a personality feature given the hidden representation of the sample. We conduct experiments with three types of personality prediction models: (1) contour encoder + classifier, (2) hybrid models that combine the contour encoder with a transformer-based language model + classifier, and (3) a stacking model that combines ten repetitions of the best performing model. As for the contour encoder, we experiment with BLSTM and BLSTM with attention models. Attention-based models have been successfully used in a variety of tasks, including machine translation (Bahdanau et al., 2014), speech recognition (Huang and Narayanan, 2016) and relation classification (Zhou et al., 2016). In the context of personality classification, learning a scoring function gives sentence weighting to the attention mechanism and allows a model to pay more attention to the most influential sentences in a text for a personality trait. As for the hybrid models, we experiment with different strategies for applying the pre-trained language model - 'feature-based' and 'fine-tuning': In the feature-based approach, we freeze model weights during training and use the pre-trained contextualized word embeddings from BERT. In the 'fine-tuning' approach, we unfreeze all 12 layers and fine-tune towards the personality detection task (see Devlin et al., 2018).

All models are implemented using PyTorch (Paszke et al., 2019). Unless specifically stated otherwise, we use binary cross entropy as our loss function, 'AdamW' as optimizer, a fixed learning rate of 8×10^{-4} and $dropout = 0.1, l2 = 1 \times 10^{-4}$ as the regularization. The optimal network structures and values of hyperparameters were found by grid-search. The performance of the models is evaluated by 10-fold cross-validation (ten repetitions) to counter variability due to initialization of the weights. We report the results of the best performing models in comparison to the performance of the APP systems presented in Section 2 Table 1.

4.1 Components

Contour Encoder: The contour encoder, $\text{Encoder}_{PSYLING}(X)$, transforms a sequence of

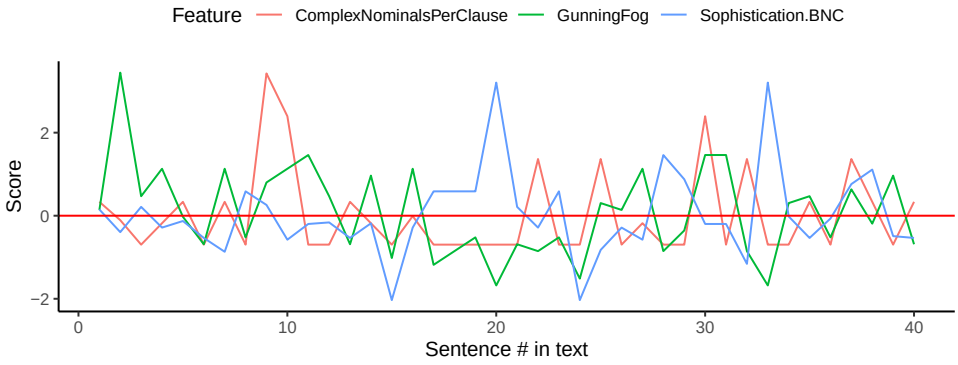

Figure 1: Text contours for three selected features of first 40 sentences of a randomly selected text from the Essays dataset (ID: 2004 499).

psycholinguistic features $X = (x_1, x_2, \ldots, x_n)$ to a hidden psycholinguistic representation vector $P_{PSYLING}$ of a given text. Here, x_i is a 436 dimensional vector for the ith sentence obtained from the APA system described in Section 3.2. In this paper, two architectures of contour encoder are applied: BLSTM and BLSTM with attention (ATTN). The BLSTM contour encoder is a L-layer BLSTM with number of hidden states of d_h. The hidden representation from this model is a $d_o = 2d_h$ dimensional vector, which is a concatenation of the last hidden states of the last layer in forward ($\overrightarrow{h_n}$) and backward direction ($\overleftarrow{h_1}$). Specifically, $X \mapsto \text{Encoder}_{BLSTM}(X) = P$:

$$[\overrightarrow{H}, \overleftarrow{H}] = BLSTM(X)$$
$$P = [\overrightarrow{h_n}^T | \overleftarrow{h_1}^T]^T$$

where $[\cdot|\cdot]$ is concatenation operator, $\overrightarrow{H} = (\overrightarrow{h}_1, \overrightarrow{h}_2, \ldots, \overrightarrow{h}_n)$ and $\overleftarrow{H} = (\overleftarrow{h}_1, \overleftarrow{h}_2, \ldots, \overleftarrow{h}_n)$ are BLSTM model's last layer hidden states in the forward and backward direction.

The ATTN contour encoder model was constructed as follows: Given a input sequence X, a sequence of weights will be computed with the help of a BLSTM model. Then the hidden representation of a given text can be obtained by computing the weighted sum of (a) concatenated hidden vectors from the last layer of the BLSTM model in forward and backward direction (b) feature vectors in X. We also experimented with (c) computing weights for each individual dimension of x_i and then taking weighted sum of X by applying this weights. Our experiments shows, that the approach (c) works best for both dataset. So in this paper, we define $X \mapsto \text{Encoder}_{ATTN}$(X)=P:

$$H = BLSTM(X)$$
$$M = \text{Tanh}(W_{att}H + b_{att})$$
$$\boldsymbol{\alpha} = \text{Softmax}(M)$$
$$V = \sum_{i=1}^{n} \boldsymbol{\alpha_i} \odot \boldsymbol{x_i}$$
$$P = \text{Tanh}(W_{pool}V + b_{pool})$$

where $W_{att} \in \mathbb{R}^{436 \times d_o}, b_{att} \in \mathbb{R}^{436}$. H and d_o is defined as in BLSTM encoder description. Softmax is defined as: $\alpha_{ij} = \frac{e^{m_{ij}}}{\sum_{k=1}^{n} e^{m_{kj}}}$

BERT Language Model: We use a pre-trained BERT transformer model, 'bert-base-uncased', from Huggingface's transformers library (Wolf et al., 2019). The model consists of 12 transformer layers with a hidden size of 768 and 12 attention heads. Texts are tokenized using BERT's BPE tokenizer. We use as input to BERT language model the initial 512 tokens $T = (t_1, t_2, \ldots, t_m)$ of a given text, i.e. up to 510 word tokens plus the [cls] token at the beginning and the [sep] token at the end of a given text). Assuming the output of the l layer of BERT is $H^{(l)} = (h_1^{(l)}, h_2^{(l)}, \ldots, h_n^{(l)})$, then a hidden vector is computed by either (a) the output for the [cls]-token, i.e. i.e., $V = h_1^{(l)}$ or by (b) averaging the output at the position of the actual tokenized words, i.e., $V = \frac{1}{m-2} \sum_{i=i}^{m-2} h_i^{(l)}$. Experiments with both approaches for $l \in [1, 12]$ revealed that that (a) the latter approach consistently works better than the former and (b) that $l = 11$ works best for the Essays dataset, wheras $l = 12$ works best for the MBTI dataset. So we define $X \mapsto \text{Encoder}_{BERT}(T) = P$

$$H^{(l)} = \text{BERT}(T)$$
$$V = \frac{1}{m-2} \sum_{i=i}^{m-2} h_i^{(l)}$$
$$P = \text{Tanh}(W_{pool} V + b_{pool})$$

Classifier: We use a multi-layer feed-forward neural network as our classifier component. The input to the classifier has a dimension corresponding to the underlying encoder's output dimension. We use PReLU as the activation function. Batch normalization was applied between layers of the classifier. All hidden layers share a same hidden size.

4.2 Models

We first construct models based solely on psycholinguistic features. These models (1) serve as interpretable baselines for the hybrid prediction models and (2) allow us to determine feature importance of individual features groups in predicting personality traits. To fully utilize the information provided by the contour-based measurement of text features, the models rely on BLSTM or BLSTM with attention architecture, i.e. at position of $\text{Encoder}_{PSYLING}$, Encoder_{BLSTM} or Encoder_{ATTN} is applied.

$$P = \text{Encoder}_{PSYLING}(X)$$
$$y = \text{Classifier}(P)$$

Encoder_{BLSTM} has 3 layers with 256 hidden states. We applied a learning rate of 0.001 during training of this model. The BLSTM in Encoder_{ATTN} has 3 layers with 512 hidden states. The classifier has 3 layers with hidden size of 512.

Our hybrid architecture combines text contours of psycholinguistic features with Transformer-based language models using a late-fusion method by concatenating the hidden representations from the psycholinguistic contour encoder and BERT, specifically

$$P_{PSYLING} = \text{Encoder}_{PSYLING}(X)$$
$$P_{BERT} = \text{Encoder}_{BERT}(T)$$
$$P = [P_{PSYLING}^T | P_{BERT}^T]^T$$
$$y = \text{Classifier}(P)$$

At the position of $\text{Encoder}_{PSYLING}$, Encoder_{BLSTM} can be used, which has 3 layers with hidden states of 256, or Encoder_{ATTN}, of which BLSTM also has 3 layers with hidden states of 256 with $dropout = 0.2$. During training, parameters of BERT has a fixed learning rate of 2×10^{-5} while learning rate of 8×10^{-5} is applied to other parameters. The classifier has 3 layers with hidden size of 512.

The final model used in our experiments employed a stacking approach to ensemble our best performing models (Wolpert, 1992), which has been shown to effectively increase the accuracy of the ensembled individual models. Specifically, we employed model stacking to combine BERT+ATTN-PSYLING (FT) model instances for both dataset.

The training procedure consists of two stages: In stage one, we take the model prediction on the dev-fold of each model trained on the train-fold of a k-fold CV. These predictions are then concatenated and constitute the one dimension out of 10 of the input data in a subsequent stage (stage 2). We did the same for all 10 iterations. The final predictions of the model are derived from another logistic regression model trained on the concatenated prediction vectors from stage 1 (10-fold CV).

4.3 Feature importance

To assess the relative importance of the feature groups, we employed Submodular Pick Lime (SP-LIME; Ribeiro et al. (2016)). SPLIME is a method to construct a global explanation of a model by aggregating the weights of linear models, that locally approximate the original model. To this end, we first constructed local explanations using LIME. Analogous to super-pixels for images, we categorized our features into four groups – lexical richness, morphosyntactic complexity, readability, sentiment/emotion (see section 3.2). We used binary vectors $z \in \{0, 1\}^d$ to denote the absence and presence of feature groups in the perturbed data samples, where d is the number of feature groups. Here, 'absent' means that all values of the features in the feature group are set to 0, and 'present' means that their values are retained. For simplicity, a linear regression model was chosen as the local explanatory model. An exponential kernel function with Hamming distance and kernel width $\sigma = 0.75\sqrt{d}$ was used to assign different weights to each perturbed data sample. After constructing their local explanation for each data sample in the original dataset, the matrix $W \in \mathbb{R}^{n \times d}$ was obtained, where n is the number of data samples in the original dataset and W_{ij} is the jth coefficient of the fitted linear regression model to explain data sample x_i. The global

	Essays						MBTI Kaggle				
	O	C	E	A	N	Avg	I/E	N/S	T/F	P/J	Avg
Majumder et al. (2018)	61.1	56.7	58.1	56.7	57.3	58	-	-	-	-	-
Kazameini et al (2020)	62.1	57.8	59.3	56.5	59.4	59	-	-	-	-	-
Amirhosseini & Kazemian (2020)	-	-	-	-	-	-	79	86	74.2	65.4	76.1
Mehta et al (2020):											
Psycholinguistic + MLP	60.4	57.3	56.9	57	59.8	58.3	77.6	86.3	72	61.9	74.5
BERT-base + MLP	64.6	59.2	60	58.8	60.5	60.6	78.3	86.4	74.4	64.4	75.9
All features (base) + MLP	61.1	57.4	57.9	58.6	60.5	59.1	78.4	86.6	75.9	64.4	76.3
BERT-large + MLP	63.4	58.9	59.2	58.3	58.9	59.7	78.8	86.3	76.1	67.2	77.1
Ramezani et al. (2021)	56.30	59.18	**64.25**	60.31	61.14	60.24	-	-	-	-	-
Psycholinguistic models (ours)											
BLSTM-PSYLING	61.69	59.22	58.12	56.87	57.52	58.68	77.29	86.31	72.91	61.01	74.38
ATTN-PSYLING	63.15	59.79	59.18	58.29	59.79	60.04	77.29	86.19	73.97	63.69	75.29
Hybrid models (ours)											
BERT+BLSTM-PSYLING (FB)	64.25	60.80	60.92	59.26	60.48	61.14	78.39	86.58	74.42	64.17	75.89
BERT+ATTN-PSYLING (FB)	64.78	61.13	60.44	59.30	60.68	61.27	78.82	86.78	76.62	65.78	77.00
BERT+BLSTM-PSYLING (FT)	65.55	60.72	60.72	60.52	**62.14**	61.93	85.78	90.86	83.79	79.79	85.06
BERT+ATTN-PSYLING (FT)	66.23	60.60	61.61	**61.05**	61.65	62.28	**86.25**	90.96	84.66	79.65	85.38
BERT+PSYLING Ensemble	**71.95**	**61.38**	63.01	60.16	60.98	**63.50**	85.47	**92.27**	**85.70**	**82.58**	**86.51**

Table 1: Performance comparison (classification accuracy) of our models (bottom) with previous state-of-the-art-models (top). Best performance indicated in bold.

importance score of the SP-LIME for feature j can then be derived by: $I_j = \sqrt{\sum_{i=1}^{n} |W_{ij}|}$

5 Results and Discussion

An overview of the results of our models in comparison to those reported in the previous studies reviewed above is presented in Table 1. As Table 1 shows, we achieve state-of-the-art (SOTA) results on both benchmark personality datasets: On the Big Five Essay dataset, our best-performing model achieves a classification accuracy of 63.5%, which corresponds to an increase of 2.9% over the previous SOTA. On the MBTI Kaggle dataset, our best model improved the classification accuracy of SOTA by 8.28%. On both datasets the highest classification accuracy was achieved by the ensemble model, which combined ten iterations of a hybrid model integrating a fine-tuned BERT model with an attention-based BLSTM model trained on text contours (see BERT+PSYLING Ensemble in Table 1). Our models achieve the highest performance on four of the Big Five - all except Extraversion - and on all four MBTI dimensions, with the largest increase in performance for the Big Five on the Openness dimension (+7.35%) and for the MBTI on the T/F dimension (+9.6%). Comparing the accuracy for each personality trait from Table 1 for the hybrid models trained with the "feature-based" strategy (denoted by "FB") with the corresponding value for the models trained with the "fine-tuning" strategy (denoted by "FT"), we find that the accuracy of all traits improved when each pre-trained model was fine-tuned on the data set. Comparing the accuracy for each personality trait for the models trained with an attention mechanism (denoted by 'ATTN') to the corresponding value for the models trained without this mechanism (denoted by 'BLSTM'), we find that accuracy on all dimensions except the MBTI N/S improved when an attention mechanism was used. Our results also show that approaches grounded in interpretable features can achieve competitive performance with Transformer-based approaches: Our best-performing model trained solely on psycholinguistic features, the attention-based BLSTM model (ATT-PSYLING), achieved an average classification accuracy of 60.04%, approaching the previous SOTA model, BERT-base + MLP Mehta et al. (2020a), by only 0.54%. This is a promising finding given the need for more interpretable personality prediction models that can provide valuable insights into key psycholinguistic features to drive personality prediction and advance personality psychology research. See e.g. Rudin (2019) for more general calls for using white-box models to solve practical problems, particularly in the context of

O		C		E		A		N	
Group	I	Group	I	Group	I	Group	I	Group	I
SentiEmo	18.49	SentiEmo	21.36	SentiEmo	16.39	SentiEmo	9.28	SentiEmo	16.62
lexical	12.90	lexical	14.48	lexical	10.93	lexical	7.52	lexical	10.23
readability	9.57	readability	9.57	morph.syn	9.17	morph.syn	6.23	morph.syn	8.11
morph.syn	7.08	morph.syn	8.91	readability	7.51	readability	4.21	readability	7.06

I/E		N/S		T/F		P/J	
Group	I	Group	I	Group	I	Group	I
SentiEmo	33.73	SentiEmo	21.32	SentiEmo	45.06	SentiEmo	24.97
lexical	29.94	lexical	14.25	lexical	24.64	readability	17.21
morph.syn	20.65	readability	12.55	morph.syn	20.31	morph.syn	16.02
readability	18.33	morph.syn	10.40	readability	18.76	lexical	14.48

Table 2: Results of the feature ablation for Big Five Essays datset (top) and Kaggle MBTI dataset (bottom): Feature importance (Model: ATTN-PSYLING) macro-averaged across 100 model instances. (10 $\times$ 10-fold CV).

critical industries such as healthcare, criminal justice, and news. This is due to the fact that human experts in a given application domain require both accurate and understandable models (Loyola-Gonzalez, 2019).

In what follows, we present the results of the ablation experiments. Feature group importance was quantified using SP-LIME on the best performing model trained only on text contours of psycholinguistic features, the ATTN-PSYLING model. The results of the feature ablation experiment are presented in Table 2. The table shows that the prediction of personality traits was influenced by all four feature groups (all I $>$ 4.21). Overall, personality traits were best predicted by the sentiment/emotion/affect (SentiEmo) feature group. The lexical richness, diversity and sophistication group consistently ranked second on all traits except the P/J MBTI dimension. This result indicates that in addition to words associated with affective-emotional categories, personality traits are also related to more general aspects of vocabulary. Morphosyntactic complexity and readability play a minor role but still achieve high I-scores compared to the highest scoring group in predicting Extraversion, Neuroticism, and Agreeableness (ratio: I(group$_j$) / I(SentEmo) $>$ 0.45). Finally, zooming in on the specific interactions between psycholinguistic cues and personality traits, we calculated the difference between the average feature scores of text samples with different labels for each personality trait. Visualizations of the most important psycholinguistic features that influence the prediction of personality traits are shown in Figures 4 and in the Appendix. Some interesting patterns emerged: For example, texts produced by extroverts tend to (a) have less complex morphosyntax than those by introverts (as indicated by the lower scores of the information-theoretic complexity measures), (b) contain a greater proportion of positive words, and (c) have a higher proportion of frequently used n-grams from the spoken language, news, and magazine registers. The language use of individuals scoring high on Neuroticism showed (a) a higher proportion of self-referencing words, (b) higher proportions of words related to sadness, anxiety and disappointment, but also (c) a higher proportion of longer n-grams from the fiction register. Highly conscientious individuals showed (a) a higher proportion of words with high prevalence, i.e. words that are known by a larger percentage of the population, (b) more words associated with affiliation (ally, friend) and (c) a higher proportions of frequently used n-grams from the academic register. These results replicate and extend previous findings reported in the literature (for overviews see, e.g., Mairesse et al., 2007; Park et al., 2015; Boyd and Schwartz, 2021).

6 Conclusion

Due to its central importance in capturing the essential aspects of human life, increasing attention is being paid to the modeling and predicting personality traits. In this work, we made valuable contributions to advance the state of the art in automatic prediction of personality traits from verbal behavior. We demonstrated that models trained with a comprehensive set of theory-based psycholinguistic features can compete with a Transformer-based model when their within-text distribution is taken

into account. Moreover, we showed that hybrid models incorporating such features can improve the performance of pre-trained Transformer language models, even when the latter is based on a larger model (BERT-large). We also showed that different techniques for applying pre-trained language representations from the Transformer model have an impact on model performance. Our ablation experiments have yielded interesting insights into the interplay between theory-based psycholinguistic features and personality traits. Here, we decided to focus on the two most widely used benchmark datasets. In our future work, we intend to conduct experiments with more recent, larger personality datasets such as PANDORA (Gjurkovic et al., 2020). Since this dataset also includes metadata (gender, age, and location/region), it would be interesting to see how they contribute to modeling and predicting personality traits from language use.

References

Mohammad Hossein Amirhosseini and Hassan Kazemian. 2020. Machine learning approach to personality type prediction based on the myers–briggs type indicator®. *Multimodal Technologies and Interaction*, 4(1):9.

Oscar Araque, Lorenzo Gatti, Jacopo Staiano, and Marco Guerini. 2019. Depechemood++: a bilingual emotion lexicon built through simple yet powerful techniques. *IEEE transactions on affective computing*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Ryan L Boyd and James W Pennebaker. 2017. Language-based personality: a new approach to personality in a digital world. *Current opinion in behavioral sciences*, 18:63–68.

Ryan L Boyd and H Andrew Schwartz. 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.

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

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

Marc Brysbaert, Paweł Mandera, Samantha F McCormick, and Emmanuel Keuleers. 2019. Word prevalence norms for 62,000 english lemmas. *Behavior research methods*, 51(2):467–479.

Erik Cambria, Robyn Speer, Catherine Havasi, and Amir Hussain. 2010. Senticnet: A publicly available semantic resource for opinion mining. In *2010 AAAI fall symposium series*.

Philip J Corr and Gerald Matthews. 2020. *The Cambridge handbook of personality psychology*. Cambridge University Press.

Paul T Costa and Robert R McCrae. 1992. *Neo personality inventory-revised (NEO PI-R)*. Psychological Assessment Resources Odessa, FL.

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

Peter Deutsch. 1996. Rfc1951: Deflate compressed data format specification version 1.3.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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

Matej Gjurkovic, Mladen Karan, Iva Vukojevic, Mihaela Bosnjak, and Jan Snajder. 2020. PANDORA talks: Personality and demographics on reddit. *CoRR*, abs/2004.04460.

Che-Wei Huang and Shrikanth S Narayanan. 2016. Attention assisted discovery of sub-utterance structure in speech emotion recognition. In *Interspeech*, pages 1387–1391.

Oliver P John, Eileen M Donahue, and Robert L Kentle. 1991. Big five inventory. *Journal of Personality and Social Psychology*.

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

Amirmohammad Kazameini, Samin Fatehi, Yash Mehta, Sauleh Eetemadi, and Erik Cambria. 2020. Personality trait detection using bagged svm over bert word embedding ensembles. *arXiv preprint arXiv:2010.01309*.

Elma Kerz, Yu Qiao, Daniel Wiechmann, and Marcus Ströbel. 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.

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

Charles Li, Monte Hancock, Ben Bowles, Olivia Hancock, Lesley Perg, Payton Brown, Asher Burrell, Gianella Frank, Frankie Stiers, Shana Marshall, et al. 2018. Feature extraction from social media posts for psychometric typing of participants. In *International Conference on Augmented Cognition*, pages 267–286. Springer.

Octavio Loyola-Gonzalez. 2019. Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. *IEEE Access*, 7:154096–154113.

François Mairesse, Marilyn A Walker, Matthias R Mehl, and Roger K Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of artificial intelligence research*, 30:457–500.

Navonil Majumder, Soujanya Poria, Alexander Gelbukh, and Erik Cambria. 2017. Deep learning-based document modeling for personality detection from text. *IEEE Intelligent Systems*, 32(2):74–79.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 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.

G. Matthews, I. Deary, and M. Whiteman. 2009. *Personality Traits*. Cambridge University Press.

Robert R McCrae. 2009. The five-factor model of personality traits: Consensus and controversy. *The Cambridge handbook of personality psychology*, pages 148–161.

Robert R McCrae and Oliver P John. 1992. An introduction to the five-factor model and its applications. *Journal of personality*, 60(2):175–215.

G Harry McLaughlin. 1969. Clearing the smog. *Journal of Reading*.

Yash Mehta, Samin Fatehi, Amirmohammad Kazameini, Clemens Stachl, Erik Cambria, and Sauleh Eetemadi. 2020a. 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.

Yash Mehta, Navonil Majumder, Alexander Gelbukh, and Erik Cambria. 2020b. Recent trends in deep learning based personality detection. *Artificial Intelligence Review*, 53(4):2313–2339.

Isabel Briggs Meyers, Mary H McCaulley, and Allen L Hammer. 1990. *Introduction to Type: A Description of the Theory and Applications of the Myers-Briggs Type Indicator*. Consulting Psychologists Press.

Saif Mohammad. 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.

Saif Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 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.

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

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

Gregory Park, H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Michal Kosinski, David J Stillwell, Lyle H Ungar, and Martin EP Seligman. 2015. Automatic personality assessment through social media language. *Journal of personality and social psychology*, 108(6):934.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

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

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

Yu Qiao, Xuefeng Yin, Daniel Wiechmann, and Elma Kerz. 2021a. Alzheimer's disease detection from spontaneous speech through combining linguistic complexity and (dis) fluency features with pretrained language models. *arXiv preprint arXiv:2106.08689*.

Yu Qiao, Sourabh Zanwar, Rishab Bhattacharyya, Daniel Wiechmann, Wei Zhou, Elma Kerz, and Ralf Schlüter. 2021b. Prediction of listener perception of argumentative speech in a crowdsourced data using (psycho-) linguistic and fluency features. *arXiv preprint arXiv:2111.07130*.

Majid Ramezani, Mohammad-Reza Feizi-Derakhshi, Mohammad-Ali Balafar, Meysam Asgari-Chenaghlu, Ali-Reza Feizi-Derakhshi, Narjes Nikzad-Khasmakhi, Mehrdad Ranjbar-Khadivi, Zoleikha Jahanbakhsh-Nagadeh, Elnaz Zafarani-Moattar, and Taymaz Rahkar-Farshi. 2021. Automatic personality prediction; an enhanced method using ensemble modeling. *arXiv preprint arXiv:2007.04571*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 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.

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

Christopher J Soto. 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.

Ryan A Stevenson, Joseph A Mikels, and Thomas W James. 2007. Characterization of the affective norms for english words by discrete emotional categories. *Behavior research methods*, 39(4):1020–1024.

Philip J Stone, Dexter C Dunphy, and Marshall S Smith. 1966. The general inquirer: A computer approach to content analysis.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Alessandro Vinciarelli and Gelareh Mohammadi. 2014. A survey of personality computing. *IEEE Transactions on Affective Computing*, 5(3):273–291.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

David H Wolpert. 1992. Stacked generalization. *Neural networks*, 5(2):241–259.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)*, pages 207–212.

A Appendices

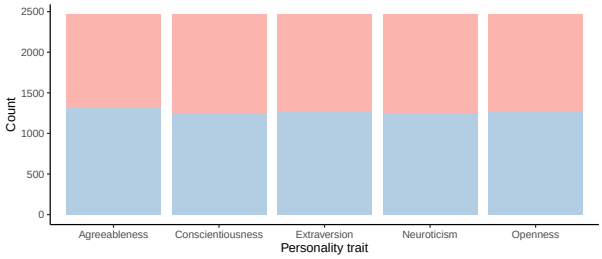

Figure 2: Distribution of labels in the Essay dataset

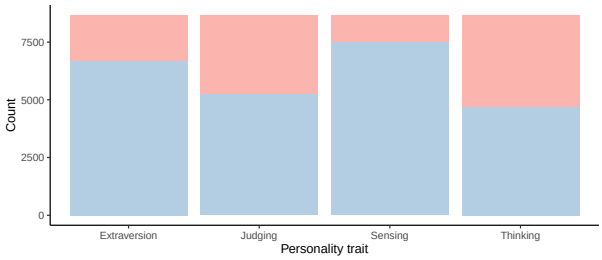

Figure 3: Distribution of labels in the Kaggle MBTI dataset

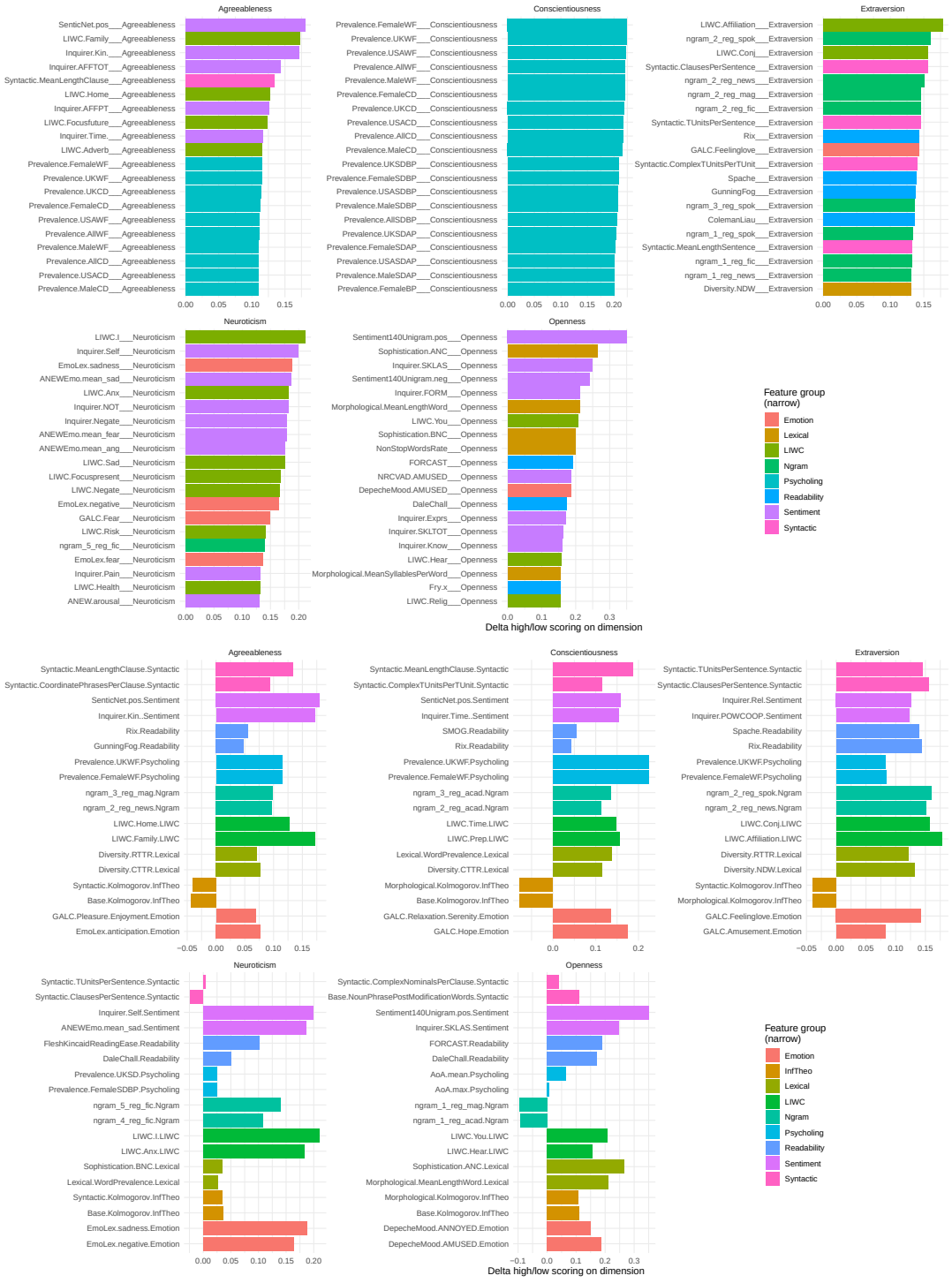

Figure 4: **Essays dataset:** Upper panel: Top 20 most characteristic features from each feature group by personality trait. Lower panel: Top 2 most characteristic features from each feature group by personality trait. Plotted scores represent the difference between the z-standardized mean scores of 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).

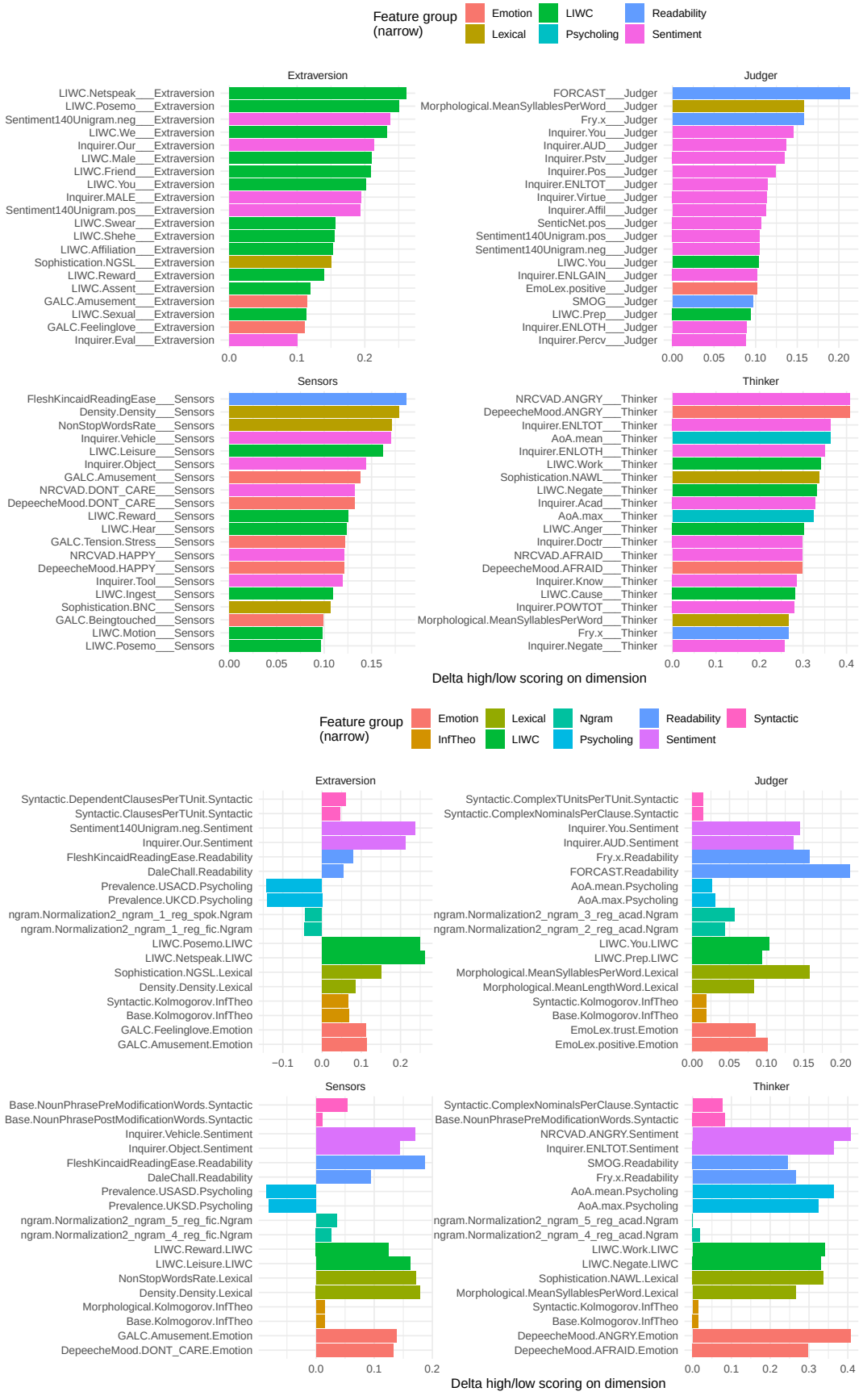

Figure 5: **MBTI Kaggle dataset:** Upper panel: Top 20 most characteristic features from each feature group by personality trait. Lower panel: Top 2 most characteristic features from each feature group by personality trait. Plotted scores represent the difference between the z-standardized mean scores 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).