



## Information Systems Research

Publication details, including instructions for authors and subscription information:  
<http://pubsonline.informs.org>

### Getting Personal: A Deep Learning Artifact for Text-Based Measurement of Personality

Kai Yang, Raymond Y. K. Lau, Ahmed Abbasi

To cite this article:

Kai Yang, Raymond Y. K. Lau, Ahmed Abbasi (2023) Getting Personal: A Deep Learning Artifact for Text-Based Measurement of Personality. Information Systems Research 34(1):194-222. <https://doi.org/10.1287/isre.2022.1111>

Full terms and conditions of use: <https://pubsonline.informs.org/Publications/Librarians-Portal/PubsOnLine-Terms-and-Conditions>

This article may be used only for the purposes of research, teaching, and/or private study. Commercial use or systematic downloading (by robots or other automatic processes) is prohibited without explicit Publisher approval, unless otherwise noted. For more information, contact [permissions@informs.org](mailto:permissions@informs.org).

The Publisher does not warrant or guarantee the article's accuracy, completeness, merchantability, fitness for a particular purpose, or non-infringement. Descriptions of, or references to, products or publications, or inclusion of an advertisement in this article, neither constitutes nor implies a guarantee, endorsement, or support of claims made of that product, publication, or service.

Copyright © 2022 The Author(s)

Please scroll down for article—it is on subsequent pages



With 12,500 members from nearly 90 countries, INFORMS is the largest international association of operations research (O.R.) and analytics professionals and students. INFORMS provides unique networking and learning opportunities for individual professionals, and organizations of all types and sizes, to better understand and use O.R. and analytics tools and methods to transform strategic visions and achieve better outcomes.




For more information on INFORMS, its publications, membership, or meetings visit <http://www.informs.org>

# Getting Personal: A Deep Learning Artifact for Text-Based Measurement of Personality

 Kai Yang,<sup>a</sup> Raymond Y. K. Lau,<sup>b</sup> Ahmed Abbasi<sup>c,\*</sup>

<sup>a</sup>Department of Accounting, College of Economics, Shenzhen University, Shenzhen, China 518060; <sup>b</sup>Department of Information Systems, College of Business, City University of Hong Kong, Kowloon, Hong Kong; <sup>c</sup>Human-Centered Analytics Lab, Department of IT, Analytics, and Operations, Mendoza College of Business, University of Notre Dame, Notre Dame, Indiana 46556

\*Corresponding author

Contact: [kayang6-c@my.cityu.edu.hk](mailto:kayang6-c@my.cityu.edu.hk),  <https://orcid.org/0000-0002-6371-7741> (KY); [raylau@cityu.edu.hk](mailto:raylau@cityu.edu.hk),  <https://orcid.org/0000-0002-5751-4550> (RYKL); [aabbasi@nd.edu](mailto:aabbasi@nd.edu),  <https://orcid.org/0000-0001-7698-7794> (AA)

Received: November 16, 2020

Revised: July 7, 2021; November 18, 2021

Accepted: January 7, 2022


Published Online in Articles in Advance:  
March 28, 2022

<https://doi.org/10.1287/isre.2022.1111>

Copyright: © 2022 The Author(s)

**Abstract.** Analysts, managers, and policymakers are interested in predictive analytics capable of offering better foresight. It is generally accepted that in forecasting scenarios involving organizational policies or consumer decision making, personal characteristics, including personality, may be an important predictor of downstream outcomes. The inclusion of personality features in forecasting models has been hindered by the fact that traditional measurement mechanisms are often infeasible. Text-based personality detection has garnered attention because of the public availability of digital textual traces. However, the text machine learning space has bifurcated into two branches: feature-based methods relying on manually crafted human intuition, or deep learning language models that leverage big data and compute, the main commonality being that neither branch generates accurate personality assessments, thereby making personality measures infeasible for downstream forecasting applications. In this study, we propose DeepPerson, a design artifact for text-based personality detection that bridges these two branches by leveraging concepts from relevant psycholinguistic theories in conjunction with advanced deep learning strategies. DeepPerson incorporates novel transfer learning and hierarchical attention network methods that use psychological concepts and data augmentation in conjunction with person-level linguistic information. We evaluate the utility of the proposed artifact using an extensive design evaluation on three personality data sets in comparison with state-of-the-art methods proposed in academia and industry. DeepPerson can improve detection of personality dimensions by 10–20 percentage points relative to the best comparison methods. Using case studies in the finance and health domains, we show that more accurate text-based personality detection can translate into significant improvements in downstream applications such as forecasting future firm performance or predicting pandemic infection rates. Our findings have important implications for research at the intersection of design and data science, and practical implications for managers focused on enabling, producing, or consuming predictive analytics.

**History:** Olivia Sheng, Senior Editor; Huimin Zhao, Associate Editor.

 **Open Access Statement:** This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License. You are free to download this work and share with others, but cannot change in any way or use commercially without permission, and you must attribute this work as “*Information Systems Research*. Copyright © 2022 The Author(s). <https://doi.org/10.1287/isre.2022.1111>, used under a Creative Commons Attribution License: <https://creativecommons.org/licenses/by-nc-nd/4.0/>.”

**Funding:** This work was supported by Oracle for Research (NLP for the Greater Good) and the U.S. National Science Foundation’s Division of Information and Intelligent Systems [Grants BDS-1636933 and IIS-1816504]. The work was also partly supported by the Research Grants Council of the Hong Kong Special Administrative Region [Project: CityU 11507219] and the City University of Hong Kong SRG [Project: 7005780].

**Supplemental Material:** The online appendices are available at <https://doi.org/10.1287/isre.2022.1111>.

**Keywords:** personality text mining • predictive analytics • deep learning • design science • NLP • psychometrics

## 1. Introduction

We live in an era of great socio-economic uncertainty. At the same time, datafication, democratization, consumerization, and the ubiquity of social media have created a seemingly insatiable appetite for real-time analysis,

insights, forecasts, and scrutiny of organizational policies, decisions, and performance. Across time zones, industry sectors, and professions, everyone from financial analysts and epidemiologists to policy makers and think tanks are interested in better insight and foresight.

As part of this global sense-making narrative during turbulent times, the importance of styles and traits has once again come front and center (Crayne and Medeiros 2020, Guest et al. 2020). Personality traits affect life choices, business decisions, suitability for certain jobs, health and well-being, protective behaviors, and numerous other preferences (Goldberg 1990, Majumder et al. 2017, Wang et al. 2019b). This is true for top-level management at publicly traded companies (Hambrick and Mason 1984, Hambrick 2007), political leaders of national and state-level governments (Crayne and Medeiros 2020), everyday online consumers (Adamopoulos et al. 2018), and employees adopting new technologies (Devaraj et al. 2008) or seeking to avoid phishing attacks (Parrish et al. 2009). Simply put, automated personality detection can provide rich predictors that can enhance agility and foresight in an array of downstream predictive analytics applications.

For instance, previous empirical studies have shown that executives' personality traits influence their decision making (Nadkarni and Herrmann 2010, Riaz et al. 2012) and leadership styles (Judge et al. 2002, 2009). These studies underscore the possible relation between leaders' personalities and strategic and tactical organizational decision making, with implications for financial forecasting of firm policies and performance (Peterson et al. 2003). In human resource contexts, personality measures could predict a candidate's suitability for a particular job role and/or teamwork performance (LePine and Van Dyne 2001). In digital marketing and online personalization settings, personality can inform product/music recommendations and effectiveness of word-of-mouth (Celli et al. 2013, Farnadi et al. 2013, Adamopoulos et al. 2018). Personality is a type of psychometric dimension: psychometrics are constructs related to attitudes, traits, and beliefs. In the management, marketing, and information systems (IS) literature, the Big Five personality traits (Goldberg 1990) have been used to examine the impact of personality on various outcomes (Devaraj et al. 2008). Like other psychometric dimensions, one obstacle to larger-scale empirical analysis or predictive modeling using personality is that traditional measurement methods, namely, surveys or manual coding of text, are often invasive and infeasible at scale (Peterson et al. 2003, Hambrick 2007, Ahmad et al. 2020b, Crayne and Medeiros 2020).

Given the difficulties in obtaining traditional psychometric data (Hambrick 2007), natural language processing (NLP) methods may represent an alternative mechanism for measuring personality through user-generated content (Ahmad et al. 2020b). However, the text machine learning (ML) space has bifurcated into two branches: feature-based machine learning relying largely on manually crafted human intuition (Pratama and Sarno 2015, Tadesse et al. 2018) or deep learning language models relying heavily on big data and compute (Majumder et al. 2017, Yu and Markov 2017). The

main commonality between the two being that neither branch generates accurate personality assessments, thereby making such measures infeasible for downstream analytics and policy applications. Accordingly, the research objective of this study is to develop a design artifact for text-based personality detection that bridges the schism by leveraging concepts from relevant psycholinguistic theories in conjunction with advanced deep learning strategies.

Following the design science approach (Hevner et al. 2004, Gregor and Hevner 2013), we use a kernel theory from psycholinguistics to develop a robust middle-ground framework called DeepPerson that couples principled, domain-adapted NLP artifacts (i.e., embeddings, encoders, and attention networks) with state-of-the-art end-to-end deep learning concepts for enhanced predictive power. Design science research questions typically center on the efficacy of design elements within a proposed artifact (Abbasi and Chen 2008) and how the artifact can "increase some measure of operational utility" (Gregor and Hevner 2013, p. 343). Accordingly, our research questions focus on personality detection capabilities and the downstream implications of better text-based personality measurement.

**Research Question 1.** *Relative to existing NLP methods, how effectively can DeepPerson detect personality dimensions from user-generated text?*

**Research Question 2.** *Can enhanced personality measurement significantly improve downstream forecasting outcomes?*

To answer these questions, we performed two sets of evaluation. In the first, we examined the personality detection capabilities of DeepPerson and comparison methods. Results reveal that our framework allows markedly more accurate detection of personality factors from text relative to existing methods developed in academia and industry, including 10%–30% improvements over IBM Personality Insights (Liu et al. 2016), Google BERT (Devlin et al. 2019), and Facebook's RoBERTa (Liu et al. 2019). More importantly, our second evaluation involving two case studies shows that this enhanced performance translates into personality variables that can significantly improve forecasting capabilities in finance and health contexts.

The main contributions of our work are three-fold. First, we propose a novel framework for measuring personality from text. Second, as part of our framework, we design novel transfer learning and hierarchical attention network methods. The proposed self-taught personality detection fine-tuning (SPDFiT) method can overcome the labeled data bottleneck encountered in most psychometric NLP problems by generating numerous pseudo-labeled training examples to enhance end-to-end model training. The word-layer-person hierarchical attention

network (wlpHAN) uses word and concept layer embeddings coupled with person-level embeddings to capture key personality cues appearing in text. Third, using a two-part evaluation, we show that more accurate NLP-based personality detection can translate into significant improvements in downstream predictive analytics applications such as forecasting future firm performance or predicting pandemic infection rates. Most notably, as we demonstrate in our evaluation, this is not the case for state-of-the-art methods which are generally incapable of producing meaningful text-based personality measures. Our work has important implications for IS research—we believe NLP at the intersection of design and data science represents a critical opportunity to develop novel, impactful artifacts that amalgamate socio-technical concepts (Abbasi et al. 2016). Furthermore, our work has practical implications for managers focused on enabling, producing, or consuming analytics in a broad array of contexts where the inclusion of personality information for key decision or policymakers may facilitate enhanced insight and foresight.

The remainder of the article is organized as follows. In the ensuing section, we discuss prior work on personality, describe state-of-the-art NLP methods for personality detection, and introduce key research gaps. In Section 3, we introduce our proposed framework, using a design science approach. Section 4 presents evaluation results for our framework relative to existing NLP methods. Section 5 uses an empirical case study to demonstrate the downstream value proposition of enhanced personality measurement, afforded by our proposed design artifact, for two important forecasting problems in the finance and health domains. The implications of our work, and concluding remarks, appear in Section 6.

## 2. Related Work

### 2.1. Importance of Measuring Personality

Prior IS research has studied the importance of personality. It has been shown to influence technology adoption (Devaraj et al. 2008) and impact online word-of-mouth (Adamopoulos et al. 2018). Personality traits can also impact susceptibility to phishing attacks (Parrish et al. 2009) and influence how users react to online recommendations (Celli et al. 2013). Majumder et al. (2017) define personality as the combination of personal behavior, motivation, and thought patterns. In the field of psychology, the Big Five personality traits (often called the five-factor model) have been widely used to characterize individuals' personalities with respect to five dimensions (Goldberg 1990):

1. Extroversion (EXT): attention-seeking, sociable, playful versus introversion (e.g., shy)
2. Neuroticism (NEU): helplessness, depressive, anxious versus emotional stability (e.g., calm)

3. Agreeableness (AGR): friendly, cooperative versus disagreeableness (e.g., suspicious)

4. Conscientiousness (CON): self-disciplined versus unconscientiousness (e.g., rash, careless)

5. Openness (OPN): creative, imaginative, insightful versus conservatism (e.g., unimaginative).

Unlike human emotions, individuals' personalities have been found to be relatively stable over time (Cobb-Clark and Schurer 2012), generally unaffected by adverse events. In studies focused on senior executives, personality traits have been found to influence decision-making style (Nadkarni and Herrmann 2010, Riaz et al. 2012). For instance, Riaz et al. (2012) suggested that extroversion was positively associated with a spontaneous decision-making style, whereas openness was related to intuitive decision making. The relation between agreeableness or conscientiousness and decision-making style has also been examined (Nygren and White 2005). Other studies have explored the relationship between personality and rational decision making (Hough and Ogilvie 2005). As one example, extroversion has been associated with effective leadership (Judge et al. 2002) and transformational leadership (Judge et al. 2009). Research has also linked the Big Five personality dimensions to downstream implications; Peterson et al. (2003) conducted one of the first studies that examined the relationship between chief executive officers' (CEOs') personality traits and firm performance using a small sample of personality information elicited from 17 executives.

It is worth noting that research examining causal relations related to personalities and outcomes have, in certain circumstances, encountered questions related to reverse causality (Hambrick 2007). For instance, certain types of personalities might be more conducive to being appointed or elected into leadership roles or more indicative of the strategic directions that a particular organization wished to take (Hambrick 2007). Although these concerns are well founded in causal modeling contexts, they do not lessen the potential value proposition of measuring personalities or of incorporating such measures in predictive contexts. Prior IS research has carefully delineated between prediction and explanation (Shmueli and Koppius 2011). As our evaluation results presented in Section 5 and Online Appendix C reveal, personality dimensions are significant and powerful predictors of future outcomes with performance/policy implications. For prediction contexts, this simply means that the underlying mechanisms contributing to their viability as key predictors of future downstream outcomes might encompass personal, organizational, contextual, or environmental factors.

The bigger limitation for use of personality dimensions in prediction contexts has been the paucity of available psychometric data (Ahmad et al. 2020b). Traditional survey and manual annotation techniques are



time-consuming and not well suited for large-scale prediction (Hambrick 2007, Crayne and Medeiros 2020). However, with the growth of online user generated content, there is a wealth of social media, online reviews, and public health 3.0 content (DeSalvo et al. 2017). In the context of personality and leadership, social *executives* (Wang et al. 2021) are increasingly communicating with key stakeholders through social media (Heavey et al. 2020). NLP methods applied to such social media text represents a viable approach for measuring personality dimensions (Back et al. 2010, Tadesse et al. 2018). This research avenue is also consistent with the perspective espoused by prior IS design science work related to business analytics, which has called for design artifacts related to text and social media (Chen et al. 2012, Abbasi et al. 2018). In the following section, we discuss the limitations of current automated NLP efforts related to personality mining from text.

## 2.2. Automated NLP-Based Personality Detection

Automated NLP research focusing on text categorization problems can be broadly grouped into two areas: manual feature engineering approaches and deep learning methods that leverage big data and/or extensive compute. Although prior work on automated text-based personality detection has focused more on feature-based techniques, as we discuss later, both categories of methods offer complementary advantages.

Researchers have examined various linguistic features for detecting individuals' personality traits. These features were generally coupled with ML classifiers such as multinomial naive Bayes (MNB), k-nearest neighbors (KNN), support vector machines (SVM), and gradient boosted trees (Pratama and Sarno 2015, Tadesse et al. 2018). For instance, Gill and Oberlander (2003) observed that individuals with the openness trait tend to use words related to insight, whereas those with the neuroticism tendency are more likely to use concrete and common words when composing messages. The neuroticism trait has also been associated with usage of words with negative appraisal and affect (Mairesse et al. 2007). Mehl et al. (2006) found that men with the conscientiousness trait tended to use more filler words, whereas the same did not hold true for females. The syntactic patterns of messages have also been found to contain important personality cues (Mairesse et al. 2007). Automated feature-based detection methods have attempted to leverage these manually inferred insights, and related lexicons, as feature-based inputs for ML classifiers. For example, the linguistic inquiry and word count (LIWC) and Research Council psycholinguistics database (MRC) lexicons have been used in prior work geared toward automated ML-based scoring of social media text (Tausczik and Pennebaker 2010, Farnadi et al. 2013, Vinciarelli and Mo 2014,

Adamopoulos et al. 2018). In addition to lexicons, bag-of-words and part-of-speech tag n-grams have also been used to detect personality traits (Wright and Chin 2014, Pratama and Sarno 2015). Tadesse et al. (2018) used structured programming for linguistic cue extraction (SPLICE), encompassing sentiment, readability, and self-evaluation features, to detect individuals' personalities. The predictive power of such linguistic features could be bootstrapped by resampling methods such as like synthetic minority oversampling (SMOTE) (Wang et al. 2019b). Guan et al. (2020) proposed a Personality2Vec model in which they ran random walks over user content similarity graphs defined using cosine similarity applied to LIWC category vectors of users' text.

Recently, deep learning-based methods have been used to detect individuals' personality traits based on their social media posts (Agastya et al. 2019, Ahmad et al. 2020a, Leonardi et al. 2020). In particular, it was found that deep Convolutional Neural Networks (CNNs) outperformed classical machine learning classifiers in personality detection (Majumder et al. 2017, Yu and Markov 2017, Sun et al. 2018). The main advantages of deep CNNs are that they can use word embeddings to capture richer contextual information appearing in documents, thereby allowing the models to generate rich abstract representations of documents. For personality detection, these capabilities have been further enhanced by combining CNNs with attention networks. For instance, Xue et al. (2018) exploited word-level attention by aggregating the embeddings of words surrounding a target word, whereas Lynn et al. (2020) applied word- and message-level attention. A limitation of the use of learned word embeddings coupled with generic attention-based CNNs, Graph Convolutional Networks (GCNs), and Long Short Term Memory networks (LSTMs) in the personality detection space has been their inability to capture linguistic cues manifesting at different granularities including person-level characteristics, psychological concepts, and syntactic and word-level patterns.

## 2.3. Related NLP Methods: Language Models, Transfer Learning, and Attention

In essence, deep learning has shifted the NLP model-building paradigm from manually weighting low-level linguistic features to automated learning of semantic and syntactic representations. Pretrained, general-purpose language models that attempt to learn broad linguistic patterns and relations applicable to an array of text categorization tasks epitomize this shift. These models leverage the classic concept of transfer learning—improving classification performance for a target task in a target domain by acquiring prior classification knowledge from one or more source tasks in corresponding source domains (Pan

and Yang 2009, Torrey and Shavlik 2010). Deep learning has taken transfer learning to a new level, allowing larger models (millions of parameters) trained on larger source data (millions of general-purpose documents). Examples include universal language models such as ULMFiT (Howard and Ruder 2018), deep contextualized representations such as ELMo (Peters et al. 2018), and powerful transformers capable of learning longer sequential patterns, such as BERT (Devlin et al. 2019). ULMFiT uses inductive transfer learning to fine-tune the learning rates at different layers of a deep recurrent neural network (RNN) for enhanced NLP classification (Howard and Ruder 2018). ELMo uses different levels of abstraction knowledge captured at various layers of a deep Bi-LSTM to boost performance (Peters et al. 2018). Similarly, BERT (Devlin et al. 2019) transfers prior knowledge (based on source data) to the bottom layers of a deep transformer network and then allows the top layers to be fine-tuned using a small number of labelled training examples from the target domain and task. Recently, Leonardi et al. (2020) performed text-based personality detection using the BERT transformer embeddings as input for a basic multilayer neural network. A more common domain-adaptation strategy has been to further pretrain BERT models on task-specific corpora (unsupervised) before fine-tuning on the supervised training data (because the original model was trained on Wikipedia and BookCorpus). For instance, BioBERT further pretrained the BERT-Base model on billions of tokens from PubMed articles (Lee et al. 2020), whereas SciBERT did the same on more than a million computer science and biomedical papers from Semantic Scholar (Beltagy et al. 2019). FinBERT is further pretrained on corporate filings, financial analyst reports, and earnings conference call transcripts (Huang et al. 2020). In our evaluation section, we also include a BERT model further pretrained on data more closely aligned with personality detection (we call this benchmark method PersonaBERT).

Apart from pretraining language models, another transfer learning approach is to fine-tune deep learning models using data augmentation methods (Lee 2013, Laine and Aila 2016, Xie et al. 2020). Examples include unsupervised data augmentation (UDA) (Xie et al. 2020) and Self-Ensembling (Laine and Aila 2016). These methods use consistency regularization to avoid disruption from the data augmentation process. A limitation of pseudo-labeling methods in general has been the quality of data generated, which often produces noisy signals that offset the predictive power gains (Lee 2013). This issue can certainly come into play on social media and user-generated text, where data quality is often lower.

A related ML advancement of interest to personality detection has been attention mechanisms. As noted, some

prior personality detection methods have used basic one-dimensional attention, such as AttRCNN (Xue et al. 2018), which uses exploited word-level attention by aggregating the embeddings of words surrounding a target word. The aspect-oriented sentiment analysis literature has also used one-dimensional aspect attention for words within a phrase surrounding opinion source/target keywords, including aspect-aware functions (Zhou et al. 2019) such as dot-product, concat, and general attention. Recognizing that for many tasks, text patterns manifest at the message versus word levels, the state-of-the-art has been hierarchical attention networks (HAN) and self-attention based extensions such as hierarchical convolutional attention networks (HCAN) (Gao et al. 2018). The Msg-Attn (Lynn et al. 2020) approach uses word- and message-level attention for personality detection. However, personality is a person-centric trait manifesting collectively in terms of the psychological concepts conveyed (Goldberg 1990, Cobb-Clark and Schurer 2012). Existing attention mechanisms ignore key person-level information and the organic *concept* construct, instead focusing on the more arbitrary “message” unit of information.

#### 2.4. Limitations of Current Personality Detection and General NLP Methods

The performance of existing ML-based automated personality detection methods has been inadequate. Gjurković et al. (2021) observed that feature-based text classification methods’ predictions often had correlation rates of under 0.2 with gold-standard Big Five traits. Accuracies for industry-leading personality detectors such as IBM personality insights have been observed to be equally low (Jayaratne and Jayatilleke 2020). Similarly, a recent survey found that deep learning-based methods attained mean accuracies of 58%–63% when detecting Big Five traits from text (Mehta et al. 2020). They acknowledge this poor performance as a bottleneck for downstream use and utility of automated detection methods (Mehta et al. 2020, p. 2333–2334), noting “If an individual’s personality could be predicted with a little more reliability, there is scope for integrating personality detection in almost all agents dealing with human-machine interactions such as voice assistants, robots, cars, etc.”

We believe the issue is one of *representational richness*—effective personality detection necessitates machine learning with enhanced expressive power. There is a need to include rich psychological concepts, methods to capture patterns at different granularities, and techniques for overcoming limitations in available psychological/directional training data for individuals. To illustrate this limitation in the state-of-the-art, Table 1 summarizes existing methods covered in Sections 2.2 and 2.3 in terms of four important dimensions: the type of method, the language representations, use of attention mechanisms, and transfer learning. In some respects, existing methods are

limited by the *Goldilocks principle*; each type of method generally does well on one of these dimensions, resulting in a smorgasbord of opportunities and limitations. Feature-based methods use rich, domain-specific lexicons but are limited in the extensiveness of patterns learned because of reliance on feature-based ML classifiers. Deep learning personality detectors use more robust sequential, spatial, and convolutional representational learning, even incorporating basic attention, but lack inclusion of rich psychological concepts, multilevel attention, person-centric patterns, or transfer learning. Language models use powerful self-attention but do not consider patterns at different granularities and are designed for standard word tokens. Relevant hierarchical/aspect attention use general word embeddings, do not go beyond word-sentence-message level attention, and have typically not been used in conjunction with transfer learning. Similarly, relevant transfer learning methods have their limitations, namely learning from noisy data such as user-generated social media (Lee 2013). However, integrating psycholinguistic concepts, state-of-the-art deep learning artifacts for multigranularity patterns/attention, and personality-

appropriate transfer learning is nontrivial. As one example, even IBM moved away from LIWC in recent years toward GloVe word embeddings (Jayaratne and Jayatilke 2020, p. 115347), noting “Earlier versions of the service used the LIWC psycholinguistic dictionary with its machine-learning model. However, the open-vocabulary approach outperforms the LIWC-based model.”

The IS discipline has a rich history of design research using concepts from language, communication, and psychology (Janson and Woo 1996; Lyytinen 1985), including ML work geared toward NLP artifacts (Abbasi and Chen 2008, Abbasi et al. 2018, Li et al. 2020). There is no question that recent advancements in deep learning, namely language models driven by transformers (Devlin et al. 2019), have disrupted NLP design research. In essence, the domain-adapted feature engineering paradigm that was pervasive for many years in text categorization studies, where researchers developed and applied carefully constructed knowledge bases and lexical thesauri, has seemingly been rendered extinct by models capable of using millions, even billions, of parameters tuned on massive text corpora

**Table 1.** Strengths and Limitations of Prior Personality Detection and Related NLP Methods

Category	Example papers/methods	Linguistic representations	Attention mechanisms	Transfer learning
Feature-based personality detection	IBM (Liu et al. 2016) KNN (Farnadi et al. 2013) SVM (Wright and Chin 2014) XGBoost (Tadesse et al. 2018) SMOTETomek (Wang et al. 2019b) Personality2Vec (Guan et al. 2020)	LIWC category feature vectors, GloVe word embeddings, or learned n-grams.	No use of attention. Feature or random walk patterns are learned.	No use of transfer learning. All patterns are learned on task-specific training data.
Deep learning for personality detection	CNN-1 (Majumder et al. 2017) CNN-2 (Yu and Markov 2017) GRU (Yu and Markov 2017) LSTM+CNN (Sun et al. 2018) AttRCNN (Xue et al. 2018) Msg-Attn (Lynn et al. 2020) GCN (Wang et al. 2020)	One-hot representations of words, word2vec applied to training data, or pretrained GloVe word embeddings.	GRUs and LSTMs use gates for retention. AttRCNN inputs word embeddings into GRUs with attention layers. Msg-Attn uses word and message-level attention.	No use of transfer learning. All patterns are learned on task-specific training data.
Language models	BERT (Devlin et al. 2019) Domain-adapted BERT (Beltagy et al. 2019, Huang et al. 2020, Lee et al. 2020) BERT+NN (Leonardi et al. 2020)	Contextualized word embeddings learned via transformer encoders.	BERT models use bidirectional self-attention with multiheaded attention	BERT uses 3.3 million tokens from BooksCorpus and Wikipedia. Domain-adapted BERTs such as SciBERT, BioBERT, and FinBERT are pretrained on task-specific corpora.
Hierarchical and aspect attention	HAN (Yang et al. 2016) HCAN (Gao et al. 2018) SATT-LSTM (Jing 2019) Aspect Attention (Zhou et al. 2019)	Word and sentence embeddings learned from text.	Either word and sentence, word and message, aspect, or self-attention.	No use of transfer learning. All patterns are learned on task-specific training data.
Transfer learning	Self-Ensembling (Laine and Aila 2016) UDA (Xie et al. 2020)	Word or contextualized embeddings.	Not explored.	Transfer learning methods that can generate pseudo-labels.



(Brown et al. 2020). However, we believe this demise has been grossly exaggerated. From a design science perspective, if we define the effectiveness of an artifact based on its level of operational utility (Gregor and Hevner 2013), neither existing feature and deep learning personality detectors or general-purpose language models are well suited for text-based personality detection. As we later demonstrate, existing NLP methods in both branches fail to produce personality measures that can improve downstream prediction outcomes. In fact, we evaluate and markedly outperform every bolded method presented in Table 1. NLP artifacts are inherently socio-technical, and opportunities for human-centered machine learning persist (Abbasi et al. 2016). There is a need to couple the power of state-of-the-art machine learning NLP methods with principled, theory-driven domain adaptation. This is precisely the research gap we aim to address with our proposed framework.

### 3. Deep Learning Framework for Personality Detection

Many prior design science studies have used kernel theories to guide the design of novel artifacts (Li et al. 2020). According to Walls et al. (1992), kernel theories are derived from the natural and social sciences and are used to govern meta-requirements. Arazy et al. (2010) stated that theories from those domains are rarely used as-is because their scope and granularity are often inadequate for a specific design problem. As noted, a fundamental problem with the state-of-the-art for NLP-based personality detection is a lack of representational richness. Existing manual feature engineering approaches lack the breadth of patterns needed to effectively capture personality traces from text, whereas the deep learning-based language models are better suited for learning general NLP patterns but lack contextualization. By focusing on the meta-functions of language, systemic functional linguistic theory (SFLT) provides a theoretical lens for how to think about representational richness in language (Halliday and Hasan 2004). SFLT, which has been used in prior IS design work (Abbasi and Chen 2008), argues that language encompasses three core meta-functions (Halliday and Hasan 2004): ideational, interpersonal, and textual. The ideational meta-function stems from the notion that language provides a mechanism for describing “human experience,” including experiential and logical ideas and concepts (Halliday and Hasan 2004, p. 29). The interpersonal meta-function relates to “enacting our personal and social relationships”; it is both interactive and personal. The textual meta-function focuses on “the construction of text” as “an enabling or facilitating function” (Halliday and Hasan 2004, p. 30).

Table 2 shows how we use SFLT as a kernel theory to guide the design of DeepPerson, our middle-ground framework that combines problem domain adapted design with advanced machine learning techniques. Our main design intuition is that enhancing text-based personality detection necessitates effective representation of the ideational, interpersonal, and textual meta-functions of language as they relate to personality trait traces appearing in natural language. The middle-ground domain adaptation happens as a result of incorporating psychological encoders, a proposed word-layer-person hierarchical attention network (wlpHAN) that includes word, a broader text layer for syntax/semantics/concepts, and person-level information, and our novel transfer learning method for learning robust personality traces.

Building on the design guidelines in Table 2, Figure 1 shows an overview of the proposed DeepPerson framework, which includes three main components: CNN-LSTM, wlpHAN, and transfer learning via SPDFiT. The CNN-LSTM network consists of a CNN-based character encoder and two multilayer Bi-LSTM networks. The first Bi-LSTM takes the character encoder and word CNN embeddings as input. This component is intended to capture language use related to the logical ideational (Word CNN) and textual (character encoder) meta-functions (Mairesse et al. 2007, Kim et al. 2016). The second Bi-LSTM incorporates the psychological concept encoder to capture personality traces related to the experiential ideational facet (Pennebaker and King 1999).

wlpHAN uses word and layer-level attention to capture personality cues appearing at various linguistic granularities for better representation of the ideational meta-function of language. Moreover, because personality traits are speaker-level constructs, wlpHAN also uses a person-level embedding for measuring an individual’s cues across documents to better capture person-specific facets of the interpersonal meta-function of language (from an SFLT perspective).

Finally, because rich psychometric dimensions such as personality traits entail careful examination of context, semantics, lexicogrammar, and expression (Halliday and Hasan 2004), limited training data can pose as a bottleneck (Chen et al. 2018). Accordingly, we propose self-taught personality detection fine-tuning (SPDFiT), a novel inductive transfer learning method that uses a domain adapted pseudo-labeling data augmentation technique to expand available training data by using massive unlabeled domain-specific data to fine-tune the wlpHAN component. In other words, SPDFiT enables the transfer of domain-specific knowledge from similar source problem domains to enhance the target task of personality detection.



**Table 2.** Design Guidelines for DeepPerson Framework

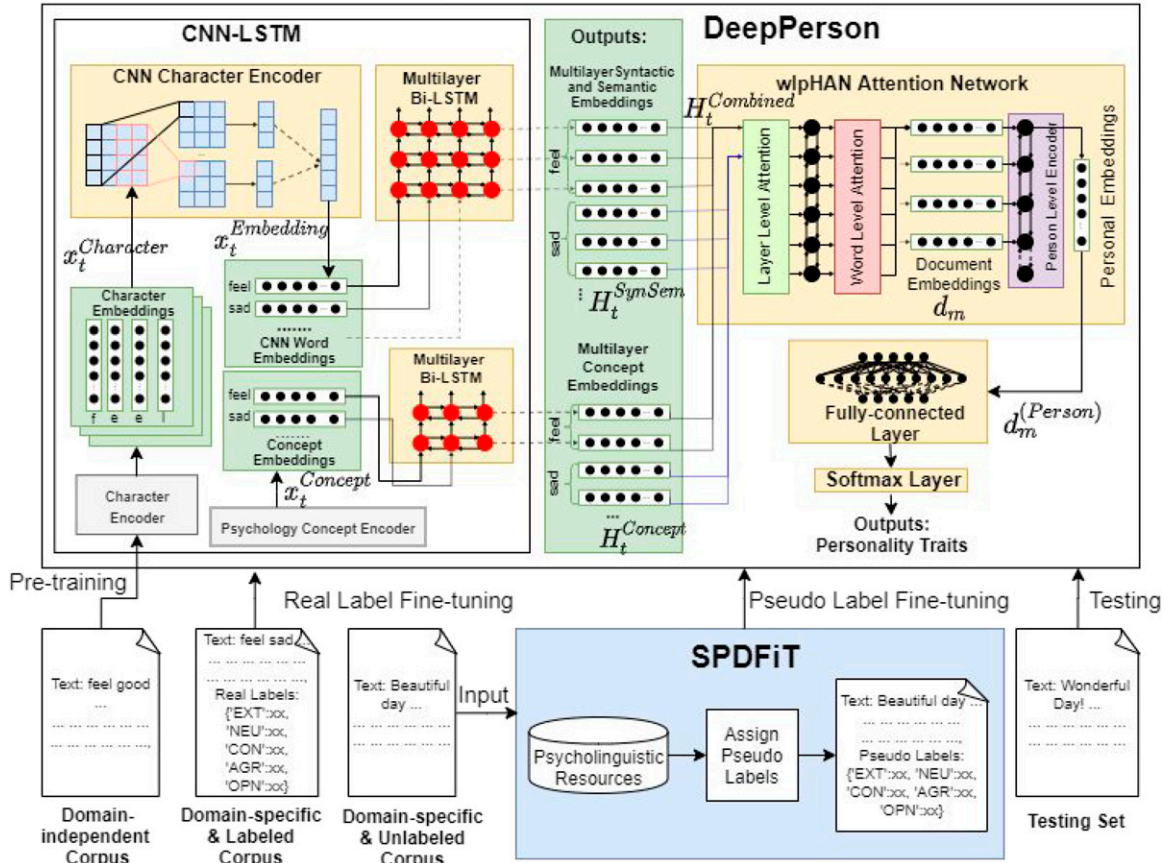
SFLT-based design guidelines	DeepPerson middle-ground framework component	Research gaps explored
Effectively representing the ideational meta-function of language entails consideration for experiential and logical concepts conveyed in text.	<i>Psychological Concept Encoder</i> : The encoder leverages well-established psychometric dictionaries, lexical thesauri, and carefully crafted self-evaluation features in conjunction with task/data specific learning through the “embeddings from language models” idea.	By combining manually crafted psychometric resources capable of capturing experiential ideas related to psychological processes (e.g., affective, cognitive, perceptual, and personal), with generic language models for logical concepts, DeepPerson can better represent the ideational meta-function.
The ideational meta-function manifests at different levels, including word, phrase, clause, sentence, and across sentences.	<i>Word-Layer-Person Hierarchical Attention Network (wlpHAN)</i> : The network uses multiple attention levels to capture personality cues appearing at various linguistic granularities including concept and syntax patterns.	Hierarchical attention has received limited focus in personality detection. Furthermore, prior hierarchical attention work has focused on word or sentence-level attention.
The interpersonal meta-function states that capturing person-specific characteristics entails accounting for speaker cues.	<i>Personal Embeddings</i> : The aforementioned hierarchical attention network also employs a person-level embedding for measuring an individual’s cues across documents.	Incorporating user level characteristics across documents is important for personality detection but has received limited attention in prior studies.
The textual meta-function requires consideration of character and morpheme level patterns.	<i>CNN Character Encoder</i> : The encoder captures spatial patterns at the character level to account for symbols and informal language commonly used in online social media.	Character CNNs have been used in prior NLP studies (Ahmad et al. 2020b, including ones appearing in IS (Li et al. 2020)). Nevertheless, character encoders are important to capture syntactic and morphological patterns related to the textual meta-function.
The three language meta-functions are instantiated through user-generated text via context, semantics, and expression. Due to the richness of, and variance in language usage, limitations on available labeled data can impede the ability to derive robust linguistic patterns.	<i>Self-taught Personality Detection Fine-tuning (SPDFiT)</i> : This inductive transfer learning method uses a novel domain adapted pseudo-labeling data augmentation technique with an entropy-based quality metric to expand the available psychometric NLP training data in a high-fidelity manner.	State-of-the-art inductive transfer learning methods do not include any domain-adapted labeling techniques, and consequently, underperform on text-based personality detection tasks. Existing data augmentation methods lack appropriate quality control resulting in noisily generated data.

Before delving into the detailed formulations and intuition behind CNN-LSTM, wlpHAN, and SPDFiT, we present an example to illustrate the enhanced representational richness afforded by these key components of DeepPerson. The wlpHAN component is able to weight syntactic and semantic elements input by the CNN-LSTM at different layers of the attention network, as shown in Figure 2. The illustration depicts the highly weighted elements for detecting the “extroversion” (EXT) and “unconscientiousness” (UNCON) personality dimensions, from two tweets, respectively, for the former U.S. president. An individual with the “extroversion” personality trait tends to be attention-seeking, sociable, and playful, whereas the “unconscientiousness” personality trait is often associated with being reckless and impulsive (Goldberg 1990). By using wlpHAN (e.g., word and layer-level attention coupled with the personal embeddings) in conjunction with the CNN-LSTM, our proposed framework can correctly detect these (and other) personality trait “digital traces” manifesting in documents based on word use, syntax/semantic (synsem) use, and psychological concepts (e.g.,

self-focus, positive emotion, affect, and social process). Although SPDFiT is not explicitly depicted in the example, it has a moderating effect on the accuracy and quality of patterns derived. We later empirically demonstrate the predictive power of each component via aggregate level ablation analysis and instance-level error analysis, including how the concept and syntactic-semantic embeddings learned contribute to the overall effectiveness of DeepPerson.

### 3.1. CNN-LSTM Network for Detecting Hidden Personality Traits

We use a Bi-LSTM network known as “embeddings from language models” (ELMO), which has been successfully applied to NLP tasks (Peters et al. 2018). Each term  $t$  of a sentence is first fed into the CNN-based character encoder to produce the corresponding encoding  $x_t^{Embedding}$ . The encoded term sequences are then input into the first multilayer Bi-LSTM network that captures implicit syntactic patterns embedded in documents. Each Bi-LSTM cell produces two hidden

**Figure 1.** (Color online) Overall Architecture of the Proposed Deep Learning Model

outputs, namely  $\overleftarrow{h}_{t,l}^{SynSem}$  and  $\overrightarrow{h}_{t,l}^{SynSem}$ . In particular,  $\overleftarrow{h}_{t,l}^{SynSem}$  represents the hidden output of term  $t$  at the  $l$  th layer, and  $\overrightarrow{h}_{t,l}^{SynSem}$  represents the hidden output  $t$  for the opposite direction. Hence, the aggregated output of the multilayer Bi-LSTM network is as follows.

$$H_t^{SynSem} = \{x_t^{Encoded}, \overleftarrow{h}_{t,l}^{SynSem}, \overrightarrow{h}_{t,l}^{SynSem} \mid l = 1, \dots, L^{SynSem}\} \\ = \{h_{t,l}^{SynSem} \mid l = 0, \dots, L^{SynSem}\}, \quad (1)$$

where  $h_{t,0}^{SynSem} = x_t^{Embedding}$  is held when  $l = 0$  is true, and  $h_{t,l}^{SynSem}$  represents the combination of  $\overleftarrow{h}_{t,l}^{SynSem}$  and  $\overrightarrow{h}_{t,l}^{SynSem}$  of each hidden layer. The size of the output vector of the Bi-LSTM network is 1,024.

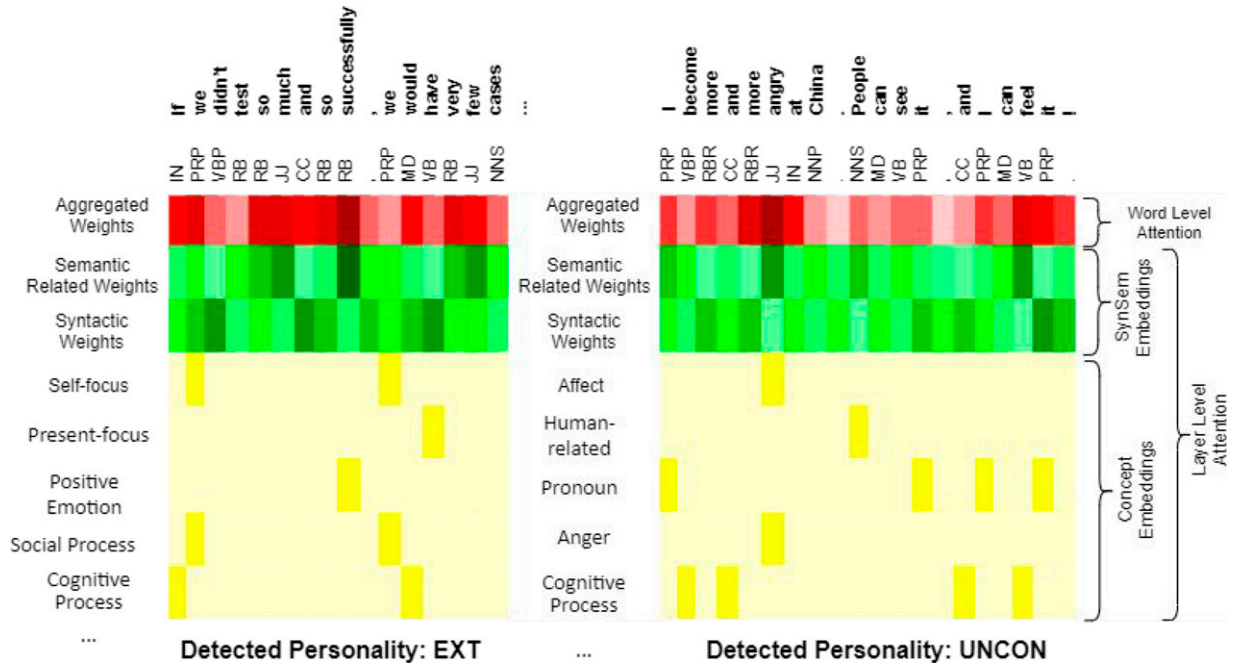
As noted, psychological concepts are an important aspect of the experiential aspect of the ideational meta-function in the context of personality detection (Pennebaker and King 1999). Accordingly, we propose a psychological concept embedding to enhance representational richness for personality detection. The psychological concepts pertaining to each term are

identified using existing psycholinguistic resources (e.g., LIWC, MRC, and SPLICE). This mapping from word/tokens to psychological concepts is a critical mechanism for enabling domain-adapted learning that leverages human knowledge and expertise in conjunction with robust algorithms. As shown in Figure 1, a concept embedding is produced via the psychological concept encoder powered by existing psycholinguistic resources. Let  $x_t^{Concept}$  denote the concept embedding of a term  $t$ . The second multilayer Bi-LSTM network is designed to capture the sequential relationships among concepts expressed in a document, with the output denoted as

$$H_t^{Concept} = \{x_t^{concept}, \overleftarrow{h}_{t,l}^{Concept}, \overrightarrow{h}_{t,l}^{Concept} \mid l = 1, \dots, L^{Concept}\} \\ = \{h_{t,l}^{Concept} \mid l = 0, \dots, L^{Concept}\}, \quad (2)$$

where  $h_{t,0}^{Concept} = x_t^{Concept}$  is held when  $l = 0$  is true;  $h_{t,l}^{Concept}$  represents the combination of  $\overleftarrow{h}_{t,l}^{Concept}$  and  $\overrightarrow{h}_{t,l}^{Concept}$  of each hidden layer, and  $L^{Concept}$  is the number of layers of the Bi-LSTM network. The output dimension of  $H_t^{Concept}$  is the same as that of  $H_t^{SynSem}$ . Finally, the two Bi-LSTM networks are aggregated:

**Figure 2.** (Color online) Visualization of Weighted Elements at Various Layers of the Attention Network



$$\begin{aligned}
 H_t^{Combined} &= \{h_{t,l} | l = 0, \dots, L^{Combined}\}, \text{ where } L^{Combined} \\
 &= L^{SynSem} + L^{Concept}.
 \end{aligned} \tag{3}$$

### 3.2. Word-Layer-Person Hierarchical Attention Network (wlp-HAN)

Although the CNN-LSTM network can generate rich syntactic and semantic representations, previous work in social psychology has shown that individuals' psychological states are related to their personalities (Pennebaker and King 1999), and traces of these can appear at different granularities within text. Attention mechanisms can help capture personality cues appearing at various linguistic levels for better representation of such psychological state information related to the ideational meta-function of language, which can manifest at the word, phrase, clause, sentence, and cross-sentence levels. However, existing attention networks mainly deal with word-based or sentence-based attention (Yang et al. 2016, Gao et al. 2018, Jing 2019). Accordingly, our proposed wlpHAN uses attention at the word and layer levels, as well as a personal embedding to capture speaker level linguistic cues associated with personality traits (which are part of the interpersonal meta-function from an SFLT perspective). As we later demonstrate empirically, the inclusion of layer and person-level attention enhances personality detection capabilities.

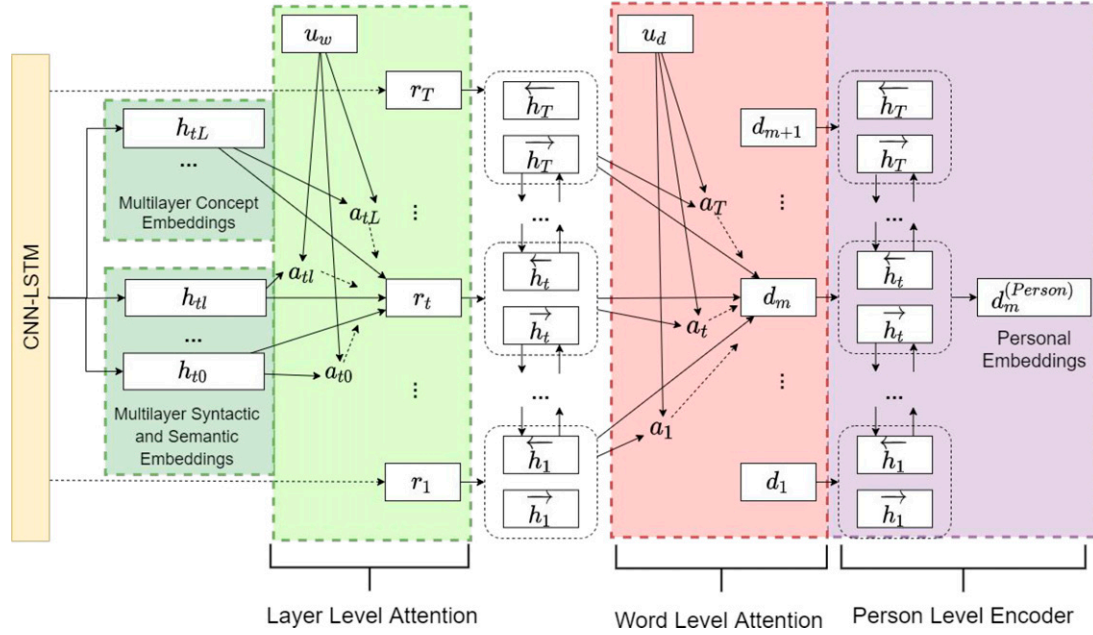
The architectural design of the proposed attention network is outlined in Figure 3. The output from each

layer of the multilayer Bi-LSTMs in the CNN-LSTM network is input to the wlpHAN, which infers appropriate weights for various psycholinguistic elements appearing in different granularities within documents. Let  $T$  denote the set of terms of a document  $m$ . For each term  $t \in T$ , an annotation set  $H_t$  is generated by each Bi-LSTM network according to Equation (3), including both multilayer concept embeddings and multilayer syntactic and semantic embeddings. Let  $h_{t,l}$  be the hidden output corresponding to term  $t$  input into the  $l$ th layer of the attention network. Similar to the approach proposed by Yang et al. (2016), our attention network assigns a higher weight to a layer if  $h_{t,l}$  is similar to the context vector  $u_w$  measured by the inner product of these vectors, whereas  $u_w$  is randomly initialized. A Sigmoid function is then applied to normalize the weights inferred by the attention network. Let  $\alpha_{t,l}$  represent the derived attention score for term  $t$  at the  $l$ th layer of the attention network. The annotation  $r_t$  of term  $t$  is the weighted sum of all hidden annotations of the set  $H_t$ .

$$r_t = \sum_l \alpha_{t,l} h_{t,l}, \text{ Given } \alpha_{t,l} = \frac{\exp(h_{t,l}^\top u_w)}{\sum_{l'} \exp(h_{t,l'}^\top u_w)} \tag{4}$$

Given the term annotation  $r_t$ , a single Bi-LSTM layer is invoked to incorporate the contextual information of a document into the word-level representation. For each term  $t$ , the corresponding hidden output generated by the Bi-LSTM layer of the attention network is denoted  $h_t$ , it is defined as the concatenation of the



**Figure 3.** (Color online) Word-Layer-Person Hierarchical Attention Network (wlpHAN)

hidden output  $\overleftarrow{h}_t$ , and the hidden output of the opposite direction  $\overrightarrow{h}_t$  of this layer.

$$h_t = \{\overleftarrow{h}_t, \overrightarrow{h}_t\} = \{\overleftarrow{LSTM}(r_t), \overrightarrow{LSTM}(r_t)\} \quad (5)$$

The attention mechanism applied to the word-level is similar to that applied to the layer level. The word-level input  $h_t$  is first fed into a fully connected layer to derive the partial document representation  $d_t$  for each term  $t$ . Then, a context vector  $u_d$  is constructed, and its similarity with  $h_t$  is measured in terms of the inner product of the corresponding vectors. A Sigmoid function is then applied to normalize the weights inferred by the word-level attention mechanism. Let  $\alpha_t$  denote the overall attention score for term  $t$ . The final document representation  $d_m$  is derived by summing the weight of each term-based partial document representation  $d_t$ .

$$d_m = \sum_t \alpha_t d_t, \text{ given } d_t = \tanh(W_d h_t + b_d) \text{ and } \alpha_t = \frac{\exp(d_t^\top u_d)}{\sum_{t'} \exp(d_{t'}^\top u_d)} \quad (6)$$

To account for person-level contextual factors, the document representation  $d_m$  is passed into a single-layer Bi-LSTM network that acts as a person-level encoder (Feng et al. 2019). The associated personal embeddings are especially important because social media posts are often short and devoid of sufficient broader text cues related to personality traits. Our person-level context-

aware representation is as follows:

$$d_m^{(Person)} = \{\overleftarrow{LSTM}(d_m), \overrightarrow{LSTM}(d_m)\}. \quad (7)$$

Finally, this representation  $d_m^{(Person)}$  is fed into a Soft-max layer to generate a probability distribution against the Big-Five personality categories  $C = \{EXT, NEU, AGR, CON, OPN\}$ . Let  $D$  denote the set of documents composed by an individual. The probability that a document  $m \in D$  is composed by the individual with a personality trait  $c$  is inferred according to Equation (8). Moreover, the individual's personality score  $\overrightarrow{pScore}$  with respect to the Big-Five personality categories is estimated according to Equation (9). To train the proposed hierarchical attention-based deep learning model, we adopt the common cross entropy loss function (Majumder et al. 2017). Further model details appear in Online Appendix A.

$$p(c|m, \theta) = \frac{\exp(W_{mc} d_m^{(Person)} + b_m)}{\sum_{c \in C} \exp(W_{mc} d_m^{(Person)} + b_m)} \quad (8)$$

$$\forall c \in C: pScore[c] = \frac{\sum_{m \in D} p(c|m, \theta)}{|D|} \quad (9)$$

### 3.3. Self-Taught Personality Detection Fine-Tuning

Effectively training supervised deep learning models usually entails use of a large number of labeled training examples (Chen et al. 2018). Although the first two components of DeepPerson are designed to provide powerful personality detection capabilities, the



paucity of available labeled data for psychometric NLP tasks such as personality detection can be a major impediment (Hambrick 2007, Ahmad et al. 2020b). From an SFLT perspective (Halliday and Hasan 2004), learning is difficult if there is not enough contextual, semantic, expression, and lexicogrammar content to sequence over (i.e., for the CNN-LSTM) and pay attention to (e.g., for the wlpHAN). State-of-the-art NLP language models such as ULMFiT (Howard and Ruder 2018), ELMo (Peters et al. 2018), and BERT (Devlin et al. 2019) bolster the amount of data on which sequence and attention weights can be learned by using inductive transfer learning to pre-train deep neural networks. Although these methods work well for a breadth of NLP problems, their propensity to adapt to a specific domain or task (e.g., psychometric NLP) is constrained by the availability of labeled training examples necessary to fine-tune the models. To alleviate this problem, we design a novel inductive transfer learning method named self-taught personality detection fine-tuning (SPDFiT) for generating pseudo-labeled training examples to enhance the fine-tuning of the first two components of DeepPerson.

The basic intuition behind SPDFiT is as follows. First, it uses existing psycholinguistic resources to derive a good representation  $\vec{t}$  for each unlabeled document  $d_m^{(u)} \in D^{(u)}$ , where  $D^{(u)}$  is an unlabeled domain-specific corpus. Second, it estimates the prior probability  $p(\vec{t} | c)$  based on a small number of labeled training examples  $d_n^{(l)} \in D^{(l)}$ . Third, the posterior probability  $p(c | \vec{t})$  (i.e., a pseudo-label) is derived using Bayes theorem. Fourth, a novel entropy-based measure  $s_m \in [0, 1]$  is applied to assess the quality of each pseudo-labeled training example. Finally, pseudo-labeled examples are selected for model fine-tuning with selection probabilities proportional to their quality measure  $s_m$ . This measure is also used to dynamically adjust the learning rate of the stochastic gradient descent (SGD) process to ensure that the model can incorporate the quantity (of data) and quality (of labeling) tradeoff as part of its learning.

At a high level, SPDFiT works with the CNN-LSTM and wlpHAN components within the DeepPerson framework as follows: (1) a large unlabeled data set from a similar source NLP domain (e.g., the 1B Word benchmark collection; Chelba et al. 2013) is used to pre-train the CNN-LSTM network; (2) SPDFiT is used to generate pseudo-labeled examples from a large unlabeled social media corpus (i.e., the Go et al. (2009) Sentiment140 corpus) for initial fine-tuning of the whole model; and (3) we apply a small number of labeled training examples from the training set to further fine-tune the model. Although state-of-the-art inductive transfer methods such as ULMFiT (Howard

and Ruder 2018), ELMo (Peters et al. 2018), and BERT (Devlin et al. 2019) include steps (1) and (3) for model pretraining and fine-tuning, these methods do not use pseudo-labeling (step 2). Conceptually, this is a critical domain-adaptation bridge between powerful (generic) universal language modeling and task-specific contextualization using seed manually labeled data rich in human insight. As we later demonstrate empirically, this step allows SPDFiT to markedly outperform state-of-the-art models developed in industry and academia.

The detailed formulations are as follows. The CNN-LSTM network is first pretrained on the 1B Word benchmark collection (Chelba et al. 2013). CNN-LSTM generates two term-based probability distributions: the forward distribution  $p(w_t | w_1, w_2, \dots, w_{t-1})$  and the backward distribution  $p(w_t | w_{t+1}, \dots, w_{|T|})$ , where  $w_t$  is a term weight. For each document, we jointly maximize the likelihood of the forward and the backward probability distributions as follows:

$$\Theta_{new} = \arg \max \left( \sum_{t=1}^{|T|} (\log p(w_t | w_1, w_2, \dots, w_{t-1}; \Theta_{old}) + p(w_t | w_{t+1}, \dots, w_{|T|}; \Theta_{old})) \right). \quad (10)$$

The computational details of the SPDFiT method are shown in Algorithm 1. It first uses existing psycholinguistic resources (e.g., LIWC, MRC, and SPLICE) to extract discriminative features (e.g., psychological features) from a large unlabeled social media data set (i.e., line 3 of Algorithm 1). Meanwhile, the model parameters of the Gaussian distribution are approximated through Gibbs sampling (i.e., line 6 of Algorithm 1). Then, the proposed algorithm computes the priori probability  $p(\vec{t} | c)$  according to the estimated Gaussian distribution (i.e., line 7 of Algorithm 1). For the unlabeled social media data set, the proposed algorithm infers the probability distribution of personality categories according to the Bayes theorem (i.e., line 11 of Algorithm 1). In particular, each unlabeled training example is assigned the personality category with the highest probability (i.e., pseudo-labeling) in line 12 of Algorithm 1. SPDFiT uses Bayesian learning because it is a solid decision theoretic framework that offers an intuitive and principled way of combining prior evidence (e.g., psycholinguistic patterns) to infer the most probable outcomes (pseudo-labels) (Haussler et al. 1994) and has been used effectively in prior deep learning contexts involving limited labeled data (Gal et al. 2017). As we demonstrate empirically in our ensuing evaluation, it outperforms other learning approaches such as logistic regression-based pseudo-labeling.

**Algorithm 1** (Self-Taught Personality Detection Fine-Tuning: SPDFIT)

**Input:** A labeled training set  $D^{(l)}$  with  $N$  documents and  $L$  features, a large unlabeled training set  $D^{(u)}$  with  $M$  documents and  $L$  features, a set of psycholinguistic resources *Lexicon*, a set of personality categories  $C$ , the learning rate  $r$  for SGD

**Output:** DeepPerson with initially fine-tuned parameters  $\theta$

1. Let  $\vec{t} = \langle t_1, t_2, \dots, t_L \rangle$ , where  $t_i$  is the  $i$  feature of the feature vector  $\vec{t}$
2. **FOR** each labeled document  $d_n^{(l)} \in D^{(l)}$  **DO**
3. Extract features of  $d_n^{(l)}$  using psycholinguistic resources:  $\vec{t} = \text{extract}(d_n^{(l)}, \text{Lexicon})$ , where  $\vec{t} \in \mathbb{R}^L$
4. **END FOR**
5. **FOR** each personality category  $c \in C$  **DO**
6. Estimate parameters  $(\mu_c, \Sigma_c)$  of the Gaussian distribution  $\mathcal{N}(\mu_c, \Sigma_c)$  by Gibbs Sampling
7. Compute the prior probability  $p(\vec{t} | c) \sim \mathcal{N}(\mu_c, \Sigma_c)$ , where  $\mu_c \in \mathbb{R}$ ,  $\Sigma_c \in \mathbb{R}^{L \times L}$
8. **END FOR**
9. **FOR** each unlabeled document  $d_m^{(u)} \in D^{(u)}$  **DO**
10. Extract features of  $d_m^{(u)}$  using psycholinguistic resources:  $\vec{t} = \text{extract}(d_m^{(u)}, \text{Lexicon})$ , where  $\vec{t} \in \mathbb{R}^L$
11. Compute posterior probabilities:  $\forall c \in C$ :  

$$p(c | \vec{t}) = \frac{p(\vec{t} | c, \mu_c, \Sigma_c) p(c)}{\sum_{c' \in C} p(\vec{t} | c', \mu_{c'}, \Sigma_{c'}) p(c')}$$
12. Set pseudo label  $l_m = \text{argmax}_c p(c | \vec{t})$
13. Compute Entropy-based quality score:  

$$s_m = \frac{H(\varphi_{\max}) - H(\varphi_m)}{H(\varphi_{\max})}$$
14. Stochastic selection of pseudo-labeled training instance  $(d_m^{(u)}, l_m)$  based on  $s_m$
15. **IF**  $d_m^{(u)}$  is selected **THEN**
16. Predict personality label  $p(c | \theta, d_m^{(u)})$  by invoking the DeepPerson framework
17. Compute the gradient:  $g = \nabla_{\theta} \mathcal{L}(p(c | \theta, d_m^{(u)}), l_m)$
18. Update parameter:  $\theta = \theta - r * s_m * g$
19. **END IF**
20. **END FOR**

Quality is always an important consideration with semisupervised and unsupervised approaches such as pseudo-labeling (Lee 2013). Based on the maximum likelihood assumption, pseudo-labeled training examples with relatively large probabilities with respect to a certain class are more likely to be assigned the correct class labels. Accordingly, we use an information theoretic metric ( $s_m \in [0, 1]$ ) to estimate the quality of pseudo-labeled training examples (i.e., line 13 of Algorithm 1). In information theory, “entropy” denoted

$H(S) = -\sum_{i=1}^{|S|} p_i \log_2 p_i$  has been widely used to measure the uncertainty of a system  $S$ , where a probability distribution  $\varphi$  is often used to characterize various states  $i$  of the system  $S$ . Given the class distributions of pseudo-labeled training examples (i.e.,  $\varphi_m$ ), the instances with relatively low entropy (i.e., low uncertainty or high quality) are more likely to be selected for fine-tuning the proposed deep learning model. Let  $\varphi_{\max}$  denote the most uncertain pseudo-labeling (i.e., an even probability distribution) of any unlabeled examples and  $\varphi_m$  denote the probability distribution of pseudo-labeling for an arbitrary unlabeled example  $m$ . The proposed information theoretic metric for estimating the certainty (quality) of pseudo-labeled training examples is defined as follows:

$s_m = \frac{H(\varphi_{\max}) - H(\varphi_m)}{H(\varphi_{\max})}$ . Furthermore, this quality metric is also used to control the learning rate of the SGD process during model fine-tuning (i.e., lines 17 and 18 of Algorithm 1). Hence, the pseudo-labeled training examples with relatively high certainty scores will trigger higher learning rates in the SGD process and thereby exert greater influence during model fine-tuning.

## 4. Design Evaluation

Following the design science approach, we evaluate the operational utility of our proposed artifact in two ways (Gregor and Hevner 2013). First, we use a design evaluation to show that the DeepPerson framework, grounded in SFLT, outperforms existing feature and deep learning methods for text-based detection of personality dimensions. As part of this evaluation, we also show that this performance lift is attributable to the effectiveness of its key components, namely, wlpHAN and SPDFIT. Our second evaluation uses empirical case studies to demonstrate the downstream implications of these performance deltas. We show that text personality variables developed using DeepPerson can significantly improve forecasting in financial and health contexts where executive decision making can shape outcomes. Our design evaluation is discussed in the remainder of this section (Section 4), whereas one of the case studies appears in Section 5.

### 4.1. Data Sets and Evaluation Procedures

To evaluate the design of DeepPerson, we used three well-known benchmark collections, namely PAN-DORA (Gjurković et al. 2021), myPersonality (Celli et al. 2013), and the Essays data set (Mairesse et al. 2007). PANDORA is a large-scale collection of 3,000,566 Reddit comments from 1,568 users and their corresponding personality traits elicited using surveys involving the same Big-Five constructs (Goldberg 1990). The myPersonality data set contains 10,000 status updates contributed by 250 Facebook users (Celli et al. 2013) and their accompanying Big-Five personality survey results. In contrast, the Essays corpus contains 2,479 essays that

capture a total of 1.9 million words composed by 2,479 psychology students (Mairesse et al. 2007). Similarly, students’ personality traits were elicited by using questionnaires that incorporated the Big-Five constructs. Table 3 depicts basic descriptive statistics for each of the data sets.

In our main evaluation, we compared DeepPerson against feature-based and deep learning methods used in prior personality detection studies, as well as state of the art universal language models (all previously discussed in Table 1). Feature-based methods included KNN coupled with LIWC categories (Farnadi et al. 2013), SVM using word n-grams (Wright and Chin 2014), gradient boosted trees (Tadesse et al. 2018), and the synthetic minority over-sampling and Tomek Link (SMOTETomek) personality detector (Wang et al. 2019b). As noted in our discussion of related work, such LIWC and n-gram-based features input into classical machine learning methods have been used extensively for personality detection (Iacobelli et al. 2011). Our deep learning-based benchmark personality detectors included CNN-1 (Majumder et al. 2017), CNN-2 (Yu and Markov 2017), gated recurrent unit (GRU) network (Yu and Markov 2017), AttRCNN (Xue et al. 2018), LSTM+CNN (Sun et al. 2018), and the graph convolutional networks GCN (Wang et al. 2020). We also included IBM Personality Insights (Liu et al. 2016), Personality2Vec (Guan et al. 2020), and the well-known BERT neural language model developed at Google (Devlin et al. 2019), which has outperformed other methods for many NLP tasks. BERT-Base was simply fine-tuned on our training data sets (no further pretraining). Conversely, PersonaBERT further pretrained the BERT-Base model from checkpoints using the same Sentiment140 and 1BWord corpora used by DeepPerson, before fine-tuning on our training data sets. BERT+NN used the BERT-Base transformer embeddings as input for a multilayer neural network (Leonardi et al. 2020).

Consistent with previous studies (Alam et al. 2013, Farnadi et al. 2013, Majumder et al. 2017, Yu and Markov 2017, Wang et al. 2019b), the personality label of a post/document was considered to be a binarized (median split) representation of the survey-based gold-standard personality label of the user who contributed the post/document; hence, personality detection was considered a binary classification problem. The class label  $c \in \{0, 1\}$  was assumed for each of the Big Five dimensions, and in each run, a personality detector classified whether a document contained that particular personality dimension. Following the common evaluation process for machine learning models involving user-centric data (Prechelt 1998, Ahmad et al. 2020b), our data set was divided into a training set (50% of users), a validation set (25% of users), and a test set (25% of users). Training was performed on

**Table 3.** Basic Descriptive Statistics of the Three Adopted Data Sets

	PANDORA (Reddit) (1,568 users, 3,000,068 posts)			myPersonality (Facebook) (250 users, 9,917 updates)			Essays (2,479 users, 2,479 documents)		
	Mean	Standard deviation	Min-max	Mean	Standard deviation	Min-max	Mean	Standard deviation	Min-max
No. of posts per user	1,917.5	4,242.7	1–52,406	39.7	43.6	1–223	1.0	0	1–1
EXT	0.37	0.30	0–1	3.29	0.86	1.33–5.00	0.52	0.50	0–1
NEU	0.50	0.32	0–1	2.63	0.78	1.25–4.75	0.50	0.50	0–1
AGR	0.42	0.31	0–1	3.60	0.67	1.65–5.00	0.53	0.50	0–1
CON	0.40	0.30	0–1	3.52	0.74	1.45–5.00	0.51	0.50	0–1
OPN	0.63	0.28	0–1	4.07	0.58	2.25–5.00	0.52	0.50	0–1
No. of words per post	0.39	70.2	1–5,306	14.74	12.76	1–113	663.1	267.5	34–3,836
No. of nouns per post	0.65	12.1	0–334	2.81	2.68	0–37	80.66	34.5	5–294
No. of verbs per post	0.24	4.2	0–64	0.81	1.24	0–12	41.10	19.1	1–178
No. of adjectives per post	0.32	6.2	0–168	1.02	1.35	0–36	37.78	16.8	2–165
No. of adverbs per post	0.34	5.6	0–78	0.99	1.39	0–15	63.66	29.2	3–290
No. of concepts per post	0.17	10.1	0–52	11.62	6.80	0–39	45.74	3.4	22–51



all documents associated with users in the training set, parameter tuning occurred on the validation users' documents, models were evaluated on the test users' documents. To make the evaluation more robust, a repeated random subsampling validation process was invoked where the training-validation-testing user splits were randomly shuffled 10 times. For design evaluation, standard document classification metrics such as precision, recall,  $F$  score, accuracy, area under the curve (AUC) were macro-averaged across the Big-Five personality categories (Alam et al. 2013). We also report performance on each of the five dimensions separately. Moreover, we adopted a non-parametric statistical test, namely the Wilcoxon signed-rank test (Wilcoxon 1992) to evaluate the statistical significance of the different performance scores achieved by various models. DeepPerson was implemented on the ELMo architecture in Pytorch. Consistent with prior studies, a grid search was used to tune parameters on the validation set. A mini-batch size of 500 and dropout rate of 0.5 were used.

#### 4.2. Comparing DeepPerson to Benchmark NLP Methods

In this section, we describe the overall design evaluation results for DeepPerson relative to the aforementioned feature-based, deep learning, and language modeling methods. We present results for the PANDORA and myPersonality data sets related to personality traces appearing in social media posts (Tables 4 and 5). The results on the essay data can be found in Online Appendix B. The first two columns in Tables 4 and 5 depict the category of method and specific method name. The next five columns show  $F$  scores for individual Big-Five dimensions, whereas the last six columns display macro-averaged  $f$  score, precision, recall, accuracy, AUC, and percentage improvement in AUC.

The results appearing in Tables 4 and 5 reveal that DeepPerson significantly outperforms all comparison methods in terms of AUC, macro  $F$  score, precision, recall, and accuracy. These performance deltas are consistent across individual personality dimensions. DeepPerson outperforms the best comparison methods, namely AttRCNN (Xue et al. 2018), CNN-1 (Majumder et al. 2017), and PersonaBERT, by 5–15 percentage points across all measures. Using IBM Personality Insights (i.e., the weakest comparison method) as a reference point for percentage lift in AUC, DeepPerson is +25% to +33% higher on the two data sets. This is nearly 13% to 20% relative percentage points higher than the best comparison methods, respectively. The Wilcoxon signed-rank tests reveal that DeepPerson's gains are significant, for instance, compared with CNN-1 ( $W = 0, p < 0.01$  for EXT, NEU, CON, AGR, and OPN).

Although not depicted here, the results on the Essay data are comparable; DeepPerson significantly outperforms all comparison methods (see Online Appendix B). Finally, because our ultimate goal for downstream tasks is to try to approximate a user's personality dimensions (averaged over all document-level scores), we also report results for user-level approximation on PANDORA and myPersonality in Online Appendix B (Tables B3 and B4). DeepPerson attains Pearson's correlation values that are at least 10–18 points higher than the best comparison method, and mean square error (MSE) values that are also at least 10% lower. The results seem to support the efficacy of middle-ground frameworks that harness rich domain knowledge and context-relevant NLP theory in conjunction with powerful state of the art machine learning approaches. In the ensuing section, we use ablation analysis to show that the performance of DeepPerson is attributable to its key components that support the SFLT-based design guidelines: CNN-LSTM, wlpHAN, and SPDFiT.

#### 4.3. Ablation Analysis of SPDFiT

Two key components of DeepPerson are the wlpHAN attention network and the pseudo-labeling SPDFiT transfer learning method. To evaluate their additive impact on DeepPerson, we ran experiments where wlpHAN was removed and SPDFiT was replaced with other baseline methods. The results on the PANDORA data are presented in Table 6; the myPersonality results can be found in Online Appendix B (Table B1). DeepPerson devoid of wlpHAN appears as the first setting: CNN-LSTM (SPDFiT). The absence of wlpHAN does reduce AUC by about five percentage points (relative to the first row in Table 5), underscoring the importance of wlpHAN. The second and third settings depict DeepPerson with wlpHAN and SPDFiT removed. In these settings, the CNN-LSTMs were pretrained using the 1B Word benchmark collection (Chelba et al. 2013) before fine-tuning with the PANDORA training data, and in the case of row 2 (i.e., 1BWord+Sentiment140), further pretrained with the Sentiment140 corpus (Go et al. 2009). More details of the experiments are reported in Online Appendix I. We also report the basic descriptive statistics of the 1B Word and Sentiment140 corpora in Table I1 (Online Appendix I).

In settings 4 and 5, SPDFiT was replaced with other state-of-the-art transfer learning methods: UDA (Xie et al. 2020) and Self-Ensembling (Laine and Aila 2016). We implemented UDA and Self-Ensembling using an open-source back-translation tool for data augmentation (Edunov et al. 2018). UDA used a loss function based on Kullback-Leibler (KL) divergence, whereas Self-Ensembling used mean square error as the loss function. Because UDA and Self-Ensembling are not



**Table 4.** Evaluation of DeepPerson and Comparison Methods on PANDORA (Reddit)

Paradigm	Method	EXT	NEU	CON	AGR	OPN	Av. F	Av. P	Av. R	Acc	AUC	Imp.
Transfer learning	DeepPerson	64.9	64.3	63.8	66.5	66.1	65.1	67.8	62.7	69.9	75.0	+33.7%
	Represent. learning											
	CNN-1	58.6	58.7	57.8	59.9	60.5	59.1	60.6	57.7	65.2	64.0	+14.1%
	CNN-2	57.7	57.9	57.1	59.6	58.4	58.1	59.5	56.8	64.4	62.7	+11.8%
	AttRCNN	59.2	59.0	57.0	61.9	60.5	59.5	61.0	58.2	65.6	64.6	+15.2%
	Msg-Attn	56.2	56.8	55.6	58.3	57.3	56.9	57.9	55.9	63.4	60.5	+7.8%
	GRU	56.3	54.1	53.7	58.3	55.1	55.5	56.2	54.9	61.8	57.8	+3.0%
	LSTM+CNN	57.0	57.2	56.5	59.1	58.0	57.5	58.8	56.3	64.0	61.2	+9.1%
	GCN	56.7	56.3	56.2	58.4	56.8	56.9	58.0	55.8	63.4	60.5	+7.8%
Language model	PersonaBERT	58.2	58.3	57.5	60.1	59.5	58.7	60.2	57.4	65.0	63.4	+13.0%
	BERT-Base	55.2	56.7	56.7	58.8	57.9	57.1	58.3	55.9	63.7	61.5	+9.6%
	BERT+NN	55.5	56.9	57.0	58.3	58.2	57.2	58.5	56.0	63.8	60.6	+8.0%
	RoBERTa	58.4	58.0	57.9	59.7	60.1	58.8	60.3	57.5	65.0	63.2	+12.7%
Feature-based	IBM	55.2	53.0	52.9	49.7	47.6	53.4	52.8	54.1	57.5	56.1	—
	KNN	56.3	55.6	53.9	57.6	57.3	56.1	57.1	55.2	63.0	58.6	+4.5%
	SVM	56.2	55.7	54.9	56.6	51.9	55.1	56.0	54.2	61.7	56.9	+1.4%
	XGBoost	56.2	56.7	54.2	57.6	56.2	56.2	57.2	55.3	62.9	58.9	+5.0%
	Personality2Vec	58.3	58.2	58.0	60.2	58.4	58.6	60.0	57.3	64.8	62.9	+12.1%
	SMOTETomek	57.4	56.8	55.7	57.4	53.3	56.1	57.2	55.1	62.5	59.4	+5.9%

Notes. Av. F, Av. P, Av. R, Acc, and AUC refer to macro-averaged *F* score, precision, recall, accuracy, and area under the ROC curve w.r.t five personality categories. Imp. refers to percentage improvement in terms of AUC. All numbers are shown in % format. CNN-1 (Majumder et al. 2017), CNN-2 (Yu and Markov 2017), GRU (Yu and Markov 2017), BERT-Base (Devlin et al. 2019), BERT+NN (Leonardi et al. 2020), RoBERTa (Liu et al. 2019), KNN (Farnadi et al. 2013), SVM (Wright and Chin 2014), XGBoost (Tadesse et al. 2018), AttRCNN (Xue et al. 2018), Msg-Attn (Lynn et al. 2020), GCN (Wang et al. 2020), Personality2Vec (Guan et al. 2020), SMOTETomek (Wang et al. 2019b), and LSTM+CNN (Sun et al. 2018).

specifically designed for personality detection tasks, to have a fair comparison, they used the same exact psychological lexicons as SPDFit (i.e., the LIWC, MRC, and SPLICE). Settings 6–8 depict alternative pseudo-labeling methods that use logistic regression (Lee’s 2013), Lasso regression (Hastie et al. 2009), or Ridge regression for pseudo-labeling. Unlike SPDFit, these pseudo-labeling methods are not equipped with a quality assessment metric to filter out low-quality labels. We also included three BERT (Devlin et al.

2019) settings, the aforementioned BERT-Base and PersonaBERT, plus an intermediate setting only further pretrained on 1BWord (but not Sentiment140) before being fine-tuned on PANDORA training data (settings 9–11). In setting 12, we replaced CNN-LSTM with just a Bi-LSTM. Finally, setting 13 used a Doc2vec (Le and Mikolov 2014)—like BERT-Base, this setting too signified the impact of no domain-specific pretraining. The BERT, Doc2Vec, and Bi-LSTM settings did not use character-level embeddings.

**Table 5.** Evaluation of DeepPerson and Comparison Methods on myPersonality (Facebook)

Paradigm	Method	EXT	NEU	CON	AGR	OPN	Av. F	Av. P	Av. R	Acc	AUC	Imp.
Transfer learning	DeepPerson	67.4	66.8	66.6	66.3	67.7	67.0	69.5	64.8	70.3	70.7	+25.1%
	Represent. learning											
	CNN-1	58.2	57.0	58.0	54.3	59.5	57.4	58.8	56.1	62.2	62.7	+11.0%
	CNN-2	57.0	55.7	56.0	53.3	58.1	56.0	57.4	54.8	61.0	60.2	+6.5%
	AttRCNN	58.7	58.5	57.7	60.3	59.8	59.0	60.5	57.6	64.2	63.3	+12.0%
	Msg-Attn	56.0	56.9	56.1	53.5	57.2	55.9	56.8	55.2	51.1	61.2	+8.3%
	GRU	55.1	50.3	53.6	51.2	57.2	53.5	54.2	52.8	59.1	59.0	+4.4%
	LSTM+CNN	56.0	55.9	55.9	53.3	60.3	56.3	56.6	56.0	54.3	62.5	+10.6%
	GCN	57.1	56.1	54.7	54.9	60.0	56.5	56.4	56.6	58.8	61.2	+8.3%
Language model	PersonaBERT	59.5	53.2	57.0	55.1	60.6	57.1	58.3	56.0	62.1	61.2	+8.3%
	BERT-Base	57.2	53.9	56.0	53.2	60.6	56.2	58.1	54.4	61.6	60.6	+7.3%
	BERT+NN	57.6	53.9	56.1	53.1	60.5	56.2	59.0	54.1	61.0	60.1	+6.4%
	RoBERTa	58.7	55.2	56.7	56.3	59.9	57.4	59.4	55.5	62.5	62.7	+11.0%
Feature-based	IBM	56.2	51.0	52.2	45.5	42.3	49.4	50.1	48.8	53.0	56.5	—
	KNN	56.5	58.3	54.1	54.2	58.3	56.3	57.8	55.0	54.4	62.0	+9.7%
	SVM	54.5	54.8	51.5	52.2	60.6	54.7	54.7	54.8	51.6	61.0	+8.0%
	XGBoost	57.3	57.0	54.3	55.1	56.2	56.0	57.2	55.0	55.3	60.9	+7.8%
	Personality2Vec	57.8	57.0	59.0	55.0	58.4	57.5	58.0	56.9	57.8	61.7	+9.2%
	SMOTETomek	54.7	55.2	53.4	52.3	60.1	55.8	56.2	54.2	47.5	61.0	+8.0%

Notes. Av. F, Av. P, Av. R, Acc, and AUC refer to macro-averaged *F* score, precision, recall, accuracy, and area under the ROC curve w.r.t five personality categories. Imp. refers to percentage improvement in terms of AUC.

The improvement column in Table 6 shows that DeepPerson devoid of wlpHAN improves AUC by 20% ( $F$  score by +11.7%) compared with the Doc2Vec (pretrained) approach and is at least +8% better than all ablation settings in terms of relative percentage improvement. The exclusion of SPDFiT after wlpHAN has already been removed (settings 2–8) degrades performance by five to seven points in terms of AUC (relative improvement of at least +8%). This includes alternative pseudo-labeling methods such as CNN-LSTM(Logistic), CNN-LSTM(Lasso), and CNN-LSTM(Ridge) and state-of-the-art transfer learning methods like UDA and Self-Ensembling. Although not depicted, with SPDFiT and wlpHAN, this relative delta is about +28%. SPDFiT (setting 1) also outperforms all BERT models (settings 9–11), including when further pretrained on the same domain-specific corpora (and fine-tuned on personality training data), by at least 11% in terms of relative percentage improvement. Finally, CNN-LSTM (setting 2) outperforms the use of Bi-LSTM (setting 12), suggesting that even without wlpHAN and SPDFiT, the CNN-LSTM setting still works well. Collectively, the results of this first ablation analysis underscore the importance of all three key components of DeepPerson, and SPDFiT in particular. Wilcoxon signed-rank tests show that these deltas are significant (all  $p < 0.01$ ).

An important consideration for transfer learning approaches is the amount of unlabeled data needed to garner enhanced predictive power. We performed additional analysis to examine the impact of the proportion of pseudo-labeled data on the performance of SPDFiT. We varied the percentage of unlabeled training examples from 10% to 100% (i.e., 100% denotes the full unlabeled data set), in increments of 10%. To isolate the impact of just using unlabeled data, for all methods evaluated, no fine-tuning was performed on labeled training data. Hence, unlabeled data were used to train the models, which were then evaluated

on the PANDORA and myPersonality test data across the various folds. For each increment, DeepPerson and comparison methods were trained for 20 epochs. The top two charts in Figure 4 depict plots of the classification performance when using SPDFiT versus comparison transfer-learning alternatives. The results reveal that SPDFiT is able to garner fairly good results when using as little as 50% of the full unlabeled training set; moreover, it outperforms all comparison methods in terms of overall  $F$  score when using 40% or more of the unlabeled data on PANDORA or 30% or more of the data on myPersonality. The bottom two charts depict the performance of SPDFiT on the five individual personality dimensions. Although not shown here, SPDFiT outperformed all comparison methods on all five dimensions when using just 50% of the unlabeled data. Given the wide range over which SPDFiT works well, we believe the results further underscore the robustness of the SPDFiT component of DeepPerson.

#### 4.4. Ablation Analysis of wlpHAN

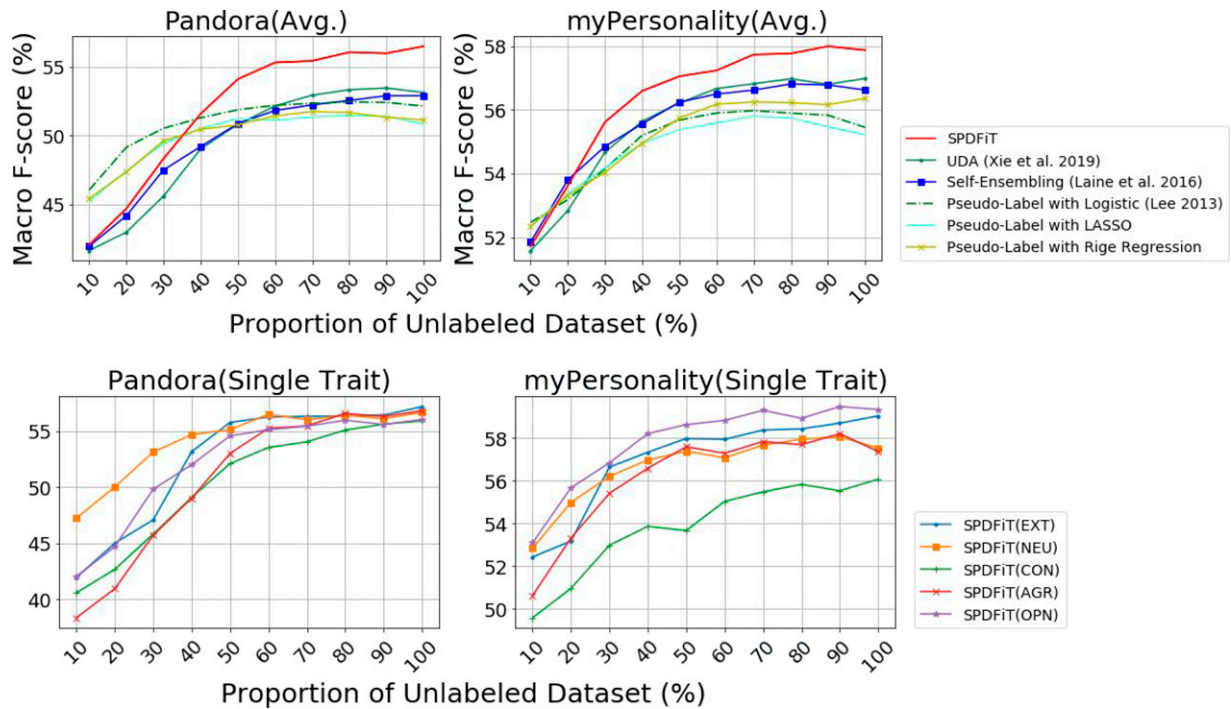
For the second ablation analysis, we examined the effectiveness of the word-, layer-, and person-based components of wlpHAN (depicted in Figure 3). For all settings, DeepPerson was invoked without the SPDFiT module to better isolate the performance impact of wlpHAN. In particular, we compared the detection performance of CNN-LSTM with full wlpHAN (setting 1) against a word-based attention only (i.e., no layer or person-level attention, setting 4), one with synsem+word (no concept embedding in the layer level attention, setting 3), and one with synsem+concept+word but no person-level encoder (setting 2). As noted in our related work section, incorporating psychological concepts into our deep learning model might be construed as being somewhat analogous to aspects-level sentiment classification (Cheng et al. 2017, Wang et al. 2019a,

**Table 6.** Comparative Evaluation of SPDFiT and Its Variants (Without wlpHAN)

Method	EXT	NEU	CON	AGR	OPN	Av. F	Av. P	Av. R	Acc	AUC	Imp.
1. CNN-LSTM (SPDFiT)	62.1	61.8	61.4	63.7	62.7	62.3	64.5	60.4	67.7	70.0	20.1%
2. CNN-LSTM (1BWord+Sentiment140)	58.1	58.7	57.6	61.0	59.5	59.0	60.4	57.6	65.1	63.6	9.1%
3. CNN-LSTM (1BWord)	57.7	57.9	57.0	59.7	58.9	58.2	59.6	57.0	64.6	62.3	6.9%
4. CNN-LSTM (UDA)	59.2	59.2	58.5	61.2	61.0	59.8	61.5	58.3	65.8	65.1	11.7%
5. CNN-LSTM (Self-Ensembling)	59.4	59.1	58.2	61.0	60.8	59.7	61.3	58.2	65.8	65.2	11.8%
6. CNN-LSTM (Logistic)	58.9	59.1	58.1	61.6	60.4	59.6	61.2	58.1	65.6	64.9	11.3%
7. CNN-LSTM (LASSO)	58.8	59.0	58.0	61.5	60.1	59.5	61.0	58.0	65.5	64.7	11.0%
8. CNN-LSTM (Ridge)	58.8	59.1	58.1	61.4	60.3	59.5	61.1	58.1	65.6	64.7	11.0%
9. PersonaBERT	58.2	58.3	57.5	60.1	59.5	58.7	60.2	57.4	65.0	63.4	8.7%
10. BERT (1BWord)	56.7	56.7	55.8	61.0	58.9	57.8	59.0	56.8	64.2	61.3	5.1%
11. BERT (Base)	55.2	56.7	56.7	58.8	57.9	57.1	58.3	55.9	63.7	60.2	3.3%
12. Bi-LSTM (1BWord)	56.0	55.9	56.6	57.3	57.1	56.6	57.7	55.5	63.2	59.1	1.4%
13. Doc2Vec (Pretrained)	54.6	55.0	55.9	56.6	57.2	55.8	56.8	54.9	62.5	58.3	—

Notes. Av. F, Av. P, Av. R, Acc, and AUC refer to macro-averaged  $F$  score, precision, recall, accuracy, and area under the ROC curve w.r.t five personality categories. Imp. refers to percentage improvement in terms of AUC. All numbers are shown in % format.

**Figure 4.** (Color online) Impact of Proportion of Unlabeled Data on Performance for SPDFiT



Galassi et al. 2020, Li et al. 2020). Accordingly, in settings 5–7, in place of wlpHAN, we substituted three aspect attention methods based on the notion of aspect-aware functions (Zhou et al. 2019): dot-product attention (DPA), concat attention (CA), and general attention (GA). In settings 8–12, we swapped out wlpHAN for other state-of-the-art attention networks such as HAN (Yang et al. 2016), SATT-LSTM (Jing 2019), HCAN (Gao et al. 2018), AttRCNN (Xue et al. 2018), and message-level attention (Msg-Attn) (Lynn et al. 2020).

As shown in Table 7 (settings 2–4), the syntax/semantic layer, concept, and person level encoders each contribute about two percentage points to wlpHAN’s overall AUC. wlpHAN also outperforms other state-of-the-art attention networks depicted in settings 8–12 such as HAN, SATT-LSTM, AttRCNN, Msg-Attn, and HCAN by three to six percentage points. Furthermore, when replacing wlpHAN with aspect-level attention networks (i.e., settings 5–7 in Table 7), performance degrades by five to six percentage points. The relative percentage improvements for wlpHAN compared with all existing attention models is 5% to 11%, with all differences significant ( $p < 0.01$ ). This performance improvement can be attributed to wlpHAN’s capability to incorporate syntax, psychologic concepts, and person-level contextual information into the personality detection process; these are all elements shown to be important for personality detection and are well aligned

with our SFLT-based design guidelines (Gill and Oberlander 2003, Mairesse et al. 2007).

#### 4.5. Error Analysis of DeepPerson vs. Benchmark Methods

As noted in Figure 2 and related discussion, and shown empirically with ablation results presented in Tables 6 and 7, the psychological concepts and patterns derived using CNN-LSTM coupled with wlpHAN (with performance boosted by SPDFiT) are critical to the performance of DeepPerson relative to the state-of-the-art. To delve deeper into these results, we conducted a series of pair-wise comparisons of instance-level error rates for DeepPerson versus CNN-1, CNN-2, PersonaBERT, and AttRCNN. In each comparison, we identified the 25% of instances on PANDORA with the widest prediction error margins between DeepPerson and each comparison method (i.e., the cases where DeepPerson was most accurate *relative* to the comparison method in terms of MSE or mean absolute error (MAE)). For these instances, we then used the following additive ablation settings to identify how various components of DeepPerson contributed to these deltas: CNN-LSTM (word), CNN-LSTM (SynSem+Word), CNN-LSTM (SynSem+ Concept+Word), CNN-LSTM (+wlpHAN), and CNN-LSTM (+SPDFiT), which is the full DeepPerson. Furthermore, this analysis was performed within each of the Big Five traits (i.e., for

**Table 7.** Comparative Evaluation of wlpHAN

Method	EXT	NEU	CON	AGR	OPN	Av. F	Av. P	Av. R	Acc	AUC	Imp.
1. CNN-LSTM (wlpHAN)	62.4	61.5	61.6	64.5	64.0	62.8	64.9	60.9	68.2	70.3	+20.0%
2. CNN-LSTM (SynSem+Concept+Word)	61.1	60.6	60.0	62.7	62.6	61.4	63.3	59.7	67.1	68.0	+16.0%
3. CNN-LSTM (SynSem+Word)	60.3	59.7	59.4	61.5	61.6	60.5	62.1	59.0	66.3	66.2	+13.0%
4. CNN-LSTM (Word)	58.3	58.6	57.8	60.6	60.0	59.0	60.4	57.8	65.1	64.0	+9.2%
5. Aspect-Attention (DPA)	57.7	57.1	62.3	58.4	62.6	59.6	61.0	58.3	65.5	64.7	+10.4%
6. Aspect-Attention (CA)	56.8	57.5	60.3	58.7	61.6	59.0	60.3	57.7	65.1	63.9	+9.0%
7. Aspect-Attention (GA)	57.2	58.6	61.7	58.7	61.5	59.5	61.1	58.1	65.4	64.8	+10.6%
8. HAN	56.2	57.5	55.7	58.3	57.3	57.0	58.1	56.0	63.5	60.7	+3.6%
9. SATT-LSTM	55.2	55.9	54.8	57.5	56.3	55.9	56.8	55.1	62.7	58.6	—
10. HCAN	56.2	57.1	56.1	58.8	57.9	57.2	58.4	56.2	63.7	61.1	+4.3%
11. AttRCNN	59.2	59.0	57.0	61.9	60.5	59.5	61.0	58.2	65.6	64.6	+10.2%
12. Msg-Attn	56.2	56.8	55.6	58.3	57.3	56.9	57.9	55.9	63.4	60.5	+3.2%

Notes. Av. F, Av. P, Av. R, Acc, and AUC refer to macro-averaged  $F$  score, precision, recall, accuracy, and area under the ROC curve w.r.t five personality categories. Imp. refers to percentage improvement in terms of AUC. All numbers are shown in % format.

all five dependent variables) to allow better understanding of how learned patterns/components improve identification of different personality traits. The results for MSE appear in Figure 5. The  $y$  axis shows relative improvements compared with the previous component.

Looking at the bar charts, we can see that just using CNN-LSTMs with the word representation underperforms AttRCNN on all five dimensions (even on these instances where overall lifts are highest for DeepPerson). Similarly, lifts versus CNN-1, CNN-2, and PersonaBERT are also modest on these instances where DeepPerson as a whole is most dominant. Interestingly, adding synsem and concept patterns, the personal embeddings in wlpHAN, and SPDFiT all cause large incremental improvements. It is worth noting that the synsem and concept embeddings complement each other. Although both have sizable lifts for all five traits, the former is most effective on the conscientiousness and extraversion traits and the latter on agreeableness and openness. The results also show that the personal embedding lift is most pronounced compared with PersonaBERT, and we see the SPDFiT moderating “boost” across all five traits, in all four comparisons. By comparing results on instances most likely driving relative deltas for DeepPerson against four of the best benchmarks, on all five traits, the results underscore how DeepPerson uses representational richness via its three main components to better infer personality digital traces and reduce error rates.

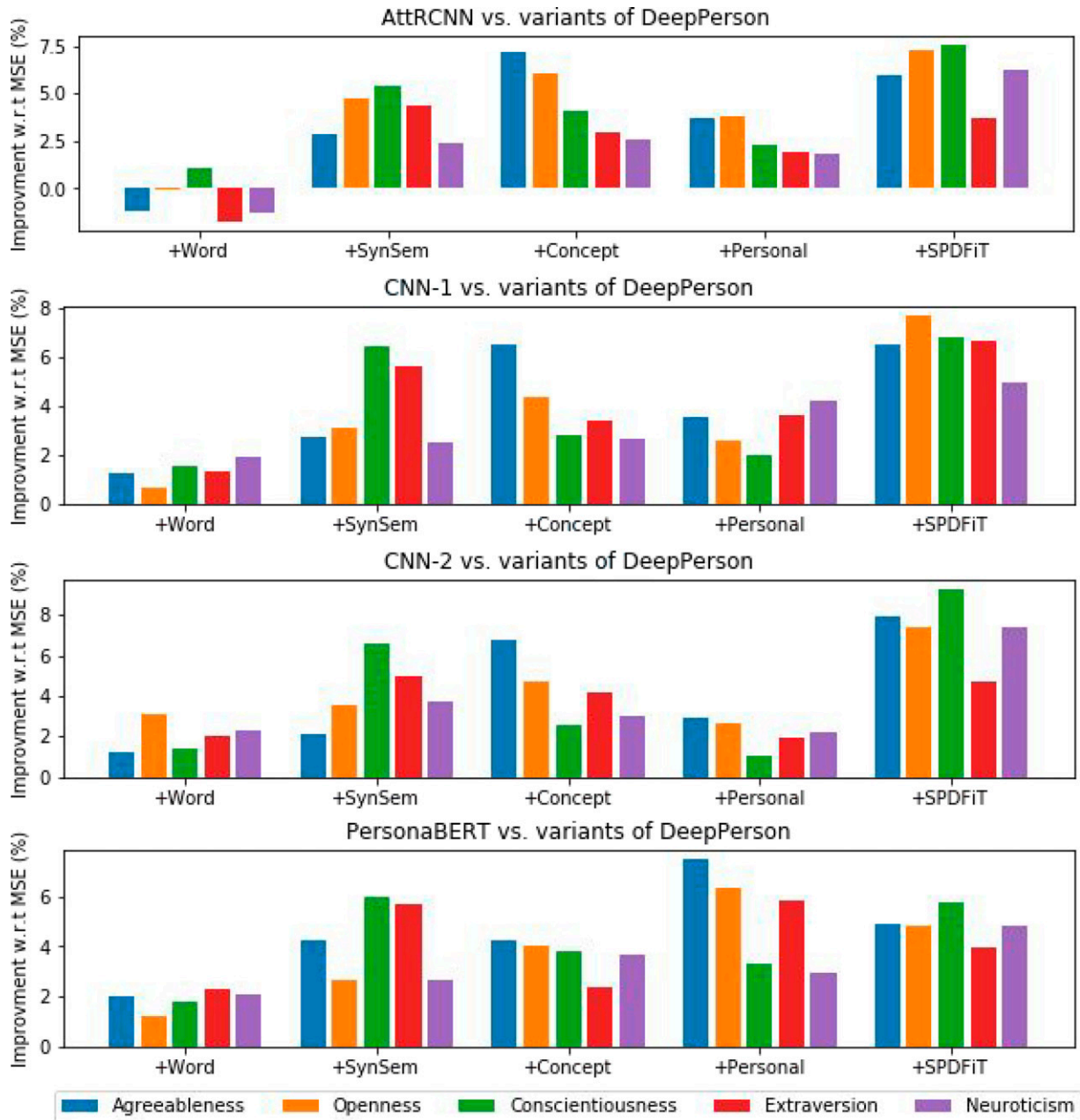
The error analysis in Figure 5 shows the importance of synsem and concept embeddings for improving detection of all five personality traits. To illustrate the types of syntactic/semantic (synsem) and concept patterns learned by DeepPerson, previously highlighted in Figure 2, we performed two additional analyses. In the first, we identified user-trait tuples for which

DeepPerson yielded accurate personality dimension scores (averaged across all their documents) and AttRCNN had high error rates. We then extracted key concept patterns for these users by identifying wlpHAN tokens with high attention scores in the multilayer concept embeddings. The results for three example users with high respective EXT, NEU, and CON appear in Table 8. The concept pattern tags correspond to categories in LIWC. Interestingly, many of the key concept patterns learned are consistent with those observed manually in prior text-based personality analysis. For instance, extroverts (EXT) tend to make positive references to friends and social processes, individuals with neuroticism (NEU) often describe their feelings and exhibit a wider range of emotions including anger and anxiety, and those that are conscientious (CON) make references to responsibilities and time/work related concepts (Mairesse et al. 2007).

Table 9 shows some of the most prevalent synsem patterns for these same traits. For the synsem patterns, we added part-of-speech tag annotations *ex post* (using the Penn Treebank), to better illustrate the syntactic elements of the synsem patterns. These patterns complement the concept embedding based ones. For instance, extroverts make greater use of compound conjunctions (CC) and punctuation that allow conveyance of additional information, neuroticism manifests in the form of greater usage of first-person pronouns (PP), and conscientious writers make greater use of adjectives (JJ) for detail. These results illustrate the types of personality cues learned by DeepPerson (and highlighted by wlpHAN), which relate to ideational, textual, and interpersonal metafunctions alluded to in SFLT. Overall, the ablation and error analysis results lend credence to the utility of our CNN-LSTM, wlpHAN, and SPDFiT components and further highlight the overall efficacy of our



**Figure 5.** (Color online) Relative MSE Improvement by Adding Components of DeepPerson



DeepPerson framework. In the ensuing section, we show that these performance deltas can also translate into downstream value in two forecasting case studies.

### 5. Downstream Predictive Application Using Detected Personality Traits

The enhanced NLP-based personality detection afforded by DeepPerson is only valuable if the generated personality dimension variables can lead to improved descriptive insights or better predictive foresight. We test the latter: the ability of DeepPerson-generated Big-Five personality variables to improve forecasting in financial

and health contexts with implications for business analytics and policy, respectively. In this section, we use DeepPerson to compute Big-Five personality scores for senior executives at S&P 1500 firms based on their Twitter posts. We then use these personality variables, along with other features, to forecast future firm financial performance metrics. In a second case appearing in Online Appendix H, we score the personalities of world and state-level leaders (executives) based on their tweets and use this information to enhance epidemiological forecasts related to the global COVID-19 pandemic.

**Table 8.** Examples of Concept Patterns Learned by DeepPerson

Concept patterns	Example concept text
Trait: EXT; Scores: Actual = 0.85, DeepPerson = 0.90, AttRCNN = 0.46	
[posemo posemo]	Life is much better < posemo> with people to share < posemo> it with!
[posemo friend]	It's nice < posemo> having a partner < friend> to wrestle life with.
[posemo social social]	I'm just glad < posemo> you're < social> readily about to lend a helping < social> hand when asked
Trait: NEU; Scores: Actual = 0.88, DeepPerson = 0.91, AttRCNN = 0.51	
[affect anx]	I had put 6 hours into the game and never enjoyed < affect> much of it, 30 FPS was very distracting < anx>.
[present negemo anx]	Now we are < present> using Facebook's terrible < negemo> freebooting tendencies in order to avoid < anx> copyright.
[sad present]	Downright disappointing < sad> that this is < present> how it has to be.
[feel affect]	It's hard < feel> to stay interested < affect> in something when I can't show that I'm making any progress
[anger feel]	F*** < anger> me for having hobbies, right? How about you, Mr. Too Cool < feel> For School?
Trait: CON; Scores: Actual = 0.81, DeepPerson = 0.86, AttRCNN = 0.33	
[present work time present work]	I've < present> read < work> a little more about Model G and I still < time> have < present> to work < work> out the details, but the models and the theory make sense now.
[present present work]	It does < present> take < present> careful study < work> and a fair amount of self-awareness to confirm the results.
[present family time]	Well all I really can do is go < present> to my home town and see < present> friends and family < family>. I don't have time < time> to go on a vacation for myself.

In the remainder of this section, we demonstrate that senior executives' personality traits derived using DeepPerson can significantly improve our ability to predict firms' policy and financial outcomes, relative to existing personality methods and exclusion of personality information entirely. Such forecasts are of interest to many stakeholder groups, including investors (FinTech) and corporate headhunters (workforce analytics). We focus on the personality traits of senior executives who are employed by the constituent firms of the S&P Composite 1500 Index, which encompasses large corporations, midsize firms, and small firms. Consistent with prior IS studies (Shi et al. 2016), we retrieved information about senior executives at S&P-1500 firms from the company pages of CrunchBase. Using definitions (and job titles) for senior executives as explicated in prior studies (Medcof 2007, Masli et al. 2016), we managed to gather information related to senior executives at 425 of the S&P-1500 firms. This included names, Twitter accounts, education levels, and so on, for employees who had c-suite job titles. These senior executives' demographic and compensation information were also retrieved from the Executive Compensation database. Among the identified senior executives, we selected those who were employed between 1990 and 2017 and who possessed Twitter accounts, resulting in 352 executives: 219 chief executive officers (CEOs), 40 chief financial officers (CFOs), 22 CXOs (e.g., Chief Marketing/Information/Technology/Operating Officers), 188 directors, 62 presidents, and 10 chairmen. All tweets composed by the identified executives between 2006 and 2017 were

retrieved. Retweeted content, URLs, and images were excluded. This resulted in an average of 529 tweets per executive (i.e., ~186,000 data points). Following the experimental procedure described in Section 4, DeepPerson was fine-tuned using the training set of the myPersonality (Facebook) corpus before it was invoked to derive the Big-Five personality dimension scores based on executives' Twitter posts. Prior leader personality studies note the benefits of using models trained on larger sets of general social media data, such as the ability to use personality labels from hundreds or thousands of users for training (Hrazdil et al. 2020). Furthermore, prior work does not note differences in personality trait linguistic patterns and cues based on one's personal status or professional standing (Mairesse et al. 2007). Consistent with prior work, we assume that personalities are relatively stable during the aforementioned analysis period (Cobb-Clark and Schurer 2012). Following the methodology adopted by Bertrand and Schoar (2003), we collected annual financial indicators related to firms' policy and financial outcomes for 1990–2017 using the Compustat database. These indicators were investment (INVEST), cash flow (CF), cash holdings (CH), leverage (LEVER), interest coverage (IC), the ratio of selling, general and administrative expenses (SG&A), the ratio of dividends and earnings over incomes (D&E), and return on asset (ROA) (Bertrand and Schoar 2003). The basic descriptive statistics of the dependent variables and predictor variables/features used in our case study are shown in Table 10.

**Table 9.** Examples of SynSem Patterns Learned by DeepPerson

SynSem patterns	Example SynSem text
Trait: EXT	
[CC VBN]	Feeling loved and (CC) appreciated (VBN)
[RB < p>]	Having a great day so far (RB), < p> thanks to santa paula noon meetings.
[CC RB VB]	has a LONG day in the field tomorrow and (CC) then (RB) is (VB) escaping Isla Vista for the weekend
Trait: NEU	
[VB PP RB]	Did I piss off a gypsy? because there's a fly in my room that won't leave (VB) me (PP) alone (RB).
[PP < p>]	My (PP) brain is like cake batter. (< p>) . (< p>) . (< p>) . (< p>)
[RB PP VB]	hungry and got no food but it is cold out so (RB) I (PP) don't (VB) want to go out to get it!
Trait: CON	
[VB DT JJ RB]	Might be taking the humble food fight (VB) a (DT) little (JJ) too (RB) seriously
[VB JJ NN]	Is wearing (VB) red (JJ) lipstick (NN), watching movies, and her mother screech at the family dog.
[WRB JJ DT VBZ]	first day PhD applications, forgot how (WRB) challenging (JJ) this (DT) is (VBZ).

According to Henderson et al. (2006), senior executives usually learn and exert influence rapidly during their initial employment period. Accordingly, we focus on examining if personalities of senior executives may predict firms' policy and financial outcomes during their initial tenure (i.e., short-term impact). To measure firms' outcomes, consistent with prior studies (Dubofsky and Varadarajan 1987, Li and Simerly 1998), we calculate the first two-year average of each chosen financial indicator after a senior executive has joined a firm. More specifically, the average of the logarithm of the annual measures was used to reduce skewness (Chu et al. 2013). Only those firm-year observations were retained where a single senior executive joined the firm in each two-year observation period. This resulted in 519 total firm-executive-biennial observations in our data set. Following Bonsall et al. (2017), we eliminated instances for a given firm or financial DC if any of the DVs or IVs of interest were missing in that first two-year period. The DV counts in Table 10 reflect the final number of instances incorporated.

Given our stated objective of demonstrating the utility of personality dimensions generated using DeepPerson for predicting firms' policy and financial outcomes, it was important to incorporate a robust set of accompanying predictor variables (i.e., features) and forecasting models such that performance lifts due to DeepPerson were atop reasonable baseline models. Consistent with prior work forecasting financial measures, we used two well-known predictive regression methods well suited for inferring nonlinear patterns: random forest regression (RFR) and gradient boosted decision trees (GBDT), both available in the Scikit-learn package (Pedregosa et al. 2011). We formalize our prediction tasks as follows:

$$I_{CEO\ in} = f(EXT, NEU, AGR, CON, OPN, \text{Baseline Features}), \quad (11)$$

where  $I_{CEO\ in}$  is the logarithm of the first two-year average for each chosen financial indicator after a senior executive has joined a firm, and  $f(\cdot)$  is a nonlinear function capturing the relationship between the predictor variables (i.e., personality traits and baseline features) and dependent variables (i.e., financial indicators). For our baseline feature set, in addition to lagged  $(t - 1)$  performance and  $(t - 1)$  policy indicator values as features, we also incorporated relevant lagged financial measures used in prior studies (Bonsall et al. 2017). These included logarithms of total assets, ROA, and cash flow (Barth et al. 2001, Bertrand and Schoar 2003). To capture industry-specific variations, firm standard industrial classification (SIC) codes were included as a feature. Executives' personal characteristics used in prior studies were also incorporated as features, including age, gender, income, education level, and reputation (Bertrand and Schoar 2003, Brick et al. 2006, Weng and Chen 2017). Adapting the methodology proposed by Weng and Chen (2017), reputation was estimated by counting the frequency of appearance of the executive's name in news articles retrieved from Google. To account for baseline semantic information embedded in executives' tweet text, we also included the sentiment of the tweets given by LIWC and their top-10 topics extracted using latent dirichlet allocation (i.e., from the document-topic vector) (Blei et al. 2003). We report the statistic of top-five topics on Table 10. Finally, basic social media-based features such as the number of tweets, followers, and favorites were also included.

We ran the aforementioned regression models either with or without the DeepPerson personality dimensions as features. The models devoid of personality features included all other variables discussed (i.e., financial, personal, and social media sentiment/topic). We also compared performance using personality dimensions generated with DeepPerson relative to methods benchmarked earlier in our design evaluation: CNN-1, CNN-2, and



**Table 10.** Basic Descriptive Statistics of the Variables Used in This Study

		Count	Mean	Standard deviation	Min	Max
Performance indicators (dependent variables)	<i>D&amp;E</i>	496	0.04	1.94	-42.77	2.33
	<i>ROA</i>	479	0.11	0.09	-0.62	0.45
Policy indicators (dependent variables)	<i>Leverage</i>	519	0.46	0.62	-8.46	3.90
	<i>SG&amp;A</i>	405	0.28	0.41	0.01	7.84
	<i>Cash Holdings (CH)</i>	497	4.84	31.18	0.00	658.78
	<i>Interest Coverage (IC)</i>	458	54.17	385.35	-1,692.22	6,760.74
	<i>Investment</i>	488	0.24	0.15	0.00	1.00
SE particulars (baseline features)	<i>Cash Flow (CF)</i>	486	0.82	3.29	-23.59	49.23
	<i>Has-MBA</i>	352	0.01	0.10	0	1
	<i>Income (K)</i>	352	273.40	172.30	0.00	1,001.92
	<i>Gender</i>	352	0.98	0.13	0	1
	<i>Age</i>	314	63.95	8.88	42.00	87.00
	<i>LOG(Reputation)</i>	352	2.84	2.38	0.00	11.72
	<i>EXT</i>	352	0.36	0.20	0.00	0.80
	<i>NEU</i>	352	0.19	0.12	0.02	0.76
	<i>CON</i>	352	0.52	0.11	0.15	0.84
	<i>AGR</i>	352	0.57	0.20	0.10	0.96
	<i>OPN</i>	352	0.79	0.15	0.40	1.00
	<i>No. of tweets</i>	352	529.40	302.66	11	856
	<i>No. of followers</i>	352	11,380.46	44,364.89	4	494,000
	<i>No. of favorites</i>	352	2,402.13	17,289.02	0	297,000
	<i>Sentiment: Positive</i>	352	0.54	0.25	0.01	1.00
	<i>Sentiment: Negative</i>	352	0.38	0.22	0.00	1.00
	<i>Topic 1</i>	352	0.12	0.11	0.00	0.61
<i>Topic 2</i>	352	0.10	0.11	0.00	0.51	
<i>Topic 3</i>	352	0.08	0.10	0.00	0.66	
<i>Topic 4</i>	352	0.08	0.06	0.00	0.39	
<i>Topic 5</i>	352	0.06	0.08	0.00	0.48	

PersonaBERT. In all experiments, the widely used mean square error (MSE) and MAE metrics were used to measure predictive power. The improvement in performance brought about by inclusion of personality features was once again computed as follows:  $Imp = \frac{MSE_{baseline} - MSE_{experimental}}{MSE_{baseline}} \times 100\%$ . Consistent with our design evaluation, all models were trained on a training split and tested on subsequent instances. Once again, the nonparametric Wilcoxon signed-rank test was used to examine statistical significance.

Tables 11 and 12 show the percentage improvements in MSE and MAE, respectively, and statistical significances when adding DeepPerson-based personality features to the baseline feature set devoid of personality information, as well as the results when using CNN-1, CNN-2, and PersonaBERT Big-Five personality features. The tables report results for GBDT and RFR each run with 20 estimators. In general, the inclusion of the DeepPerson-based personality dimension features improves MSE or MAE by 4% to 15% for each of the eight possible dependent variables (six policy indicators and two performance indicators). The average improvements using DeepPerson are in the 6.1%–14.3% range across the two models and MSE/MAE metrics. Performance gains for all eight dependent variables attributable to inclusion of the five DeepPerson-based personality dimensions were

significant ( $p < 0.05$ ). These results suggest that the personality measures derived using DeepPerson can enhance predictive power in firm policy and performance forecasting contexts. Next, when comparing the results for DeepPerson-based personality dimensions versus those derived using comparison detection methods such as CNN-1, CNN-2, and BERT, there are three important takeaways worth highlighting. First, the RFR and GBDT models using personality features derived via DeepPerson improve MSE and MAE by an average of 4%–14% over the comparison methods. Second, among the three benchmark comparison methods, features generated using BERT and CNN-1 improve average results across the eight firm policy and performance prediction tasks (with average lifts of 2%–8%). However, on average, the use of CNN-2 garners little to no improvement. Although CNN-2 enhances forecasting of performance indicators, it markedly underperforms on policy indicators.

Third, we also comparatively evaluated the classical ARIMA (AutoRegressive Integrated Moving Average) model widely used in predicting financial time series data (Mohamed et al. 2010). Similar to GBDT and RFR, ARIMA parameters were tuned extensively, including the order of the auto-regressive function, the differentiation term, and the order of the moving average. The last row of Tables 11 and 12 shows the MSE and MAE score percentages for ARIMA relative

**Table 11.** Percentage Improvement in Performance (MSE) Across Different Personality Detectors

Models	Policy indicators						Performance indicators		
	CH	CF	INVEST	LEVER	IC	SG&A	D&E	ROA	Average
<b>RFR</b>									
DeepPerson	3.23**	3.15*	8.61*	5.88*	6.47**	10.78**	8.53**	6.73**	6.67
CNN-1	-0.21	-0.92	5.23*	3.80*	2.60	2.36	3.05	4.85	2.60
CNN-2	2.64*	-6.67	3.95	3.04	-6.93	-3.19	5.16*	6.34**	0.54
PersonaBERT	2.53*	1.77	4.74*	4.26**	3.94*	5.68*	3.18	5.35*	3.93
<b>GBDT</b>									
DeepPerson	20.91**	7.49**	11.91**	12.95**	8.65**	31.74**	8.13**	12.52**	14.29
CNN-1	12.57**	1.38	4.4	3.87	0.15	1.84	0.21	2.63	3.38
CNN-2	1.63	-8.61	0.88	-3.36	-9.27	-10.11	6.99*	7.44*	-1.80
BERT	10.28**	5.62**	6.72*	6.93*	4.58	20.34**	5.51	8.97*	8.62
ARIMA	-9.96	-22.75	-9.54	-26.90	-23.20	-7.67	-8.92	-27.48	-17.05

Notes. Each value is a percentage. For each regression model and financial indicator, we estimate the average improvement with or without incorporating senior executives' personality traits into the model. The ARIMA row shows the possible improvement for GBDT relative to the common time-series prediction model (ARIMA). DeepPerson, CNN-1, CNN-2, and BERT refer to predictions using executives' personality traits detected by the respective methods.

\*\* $p < 0.01$ ; \* $p < 0.05$ : Wilcoxon signed-rank test.

to the GBDT-DeepPerson model. ARIMA had significantly lower results across all eight firm policy and performance indicators, with almost 17% worse MSE as a whole (all  $p < 0.05$ ). Although these results were using cross-validation, we also performed a single chronological training-testing split as a robustness check. Those results, in Online Appendix G, are consistent with results appearing here. Collectively, these results further underscore the value of the personality dimensions derived using DeepPerson.

As a robustness check, we repeated the empirical case study using only executives and data from the S&P-500 and garnered similar results. We also examined the impact of specific Big-Five dimensions as features to see which traits are the strongest predictors. We also conducted a sensitivity analysis to evaluate the minimal number of executives' tweets required to produce significant prediction improvement. These results appear in Online Appendices C, D, and E, respectively. In Online Appendix F, we show that these downstream results also hold for DeepPerson ablation settings examined in Sections 4.3 and 4.4. As noted earlier, a second downstream predictive application of DeepPerson in the context of COVID-19 forecasting appears in Online Appendix H. Collectively, our results show that downstream forecasting models using personality dimensions scored by DeepPerson can dramatically enhance their results, whereas this is not the case when using benchmark personality detectors or classic time series forecasting methods. As shown in the user-level results in Online Appendix B, personality scores generated with DeepPerson are better correlated with survey-based personality measurements relative to comparison methods. The imprecision of comparison text-based personality detection methods may lead to incorrect personality traits (i.e., noisy features). It is generally believed that noisy features tend to jeopardize the performance of a

prediction model (John et al. 1994). In other words, the design evaluation deltas reported in the prior section do translate into operational utility in the form of better foresight in an important business analytics context.

## 6. Results Discussion, Limitations, and Concluding Remarks

From a design science perspective, we make three contributions. First, we propose a novel DeepPerson framework that makes personality detection from text possible, practical, and valuable. Second, as part of our framework, we propose two novel machine learning artifacts, namely the SPDFiT transfer learning approach and the word-layer-person attention network. Third, through a robust design evaluation and two case studies, we offer empirical insights on the extent of operational utility afforded by DeepPerson and its key components, including for downstream forecasting tasks in financial and health contexts. Our results also have at least four important implications for IS research and practice.

(1) *Debunking the "Brute Force AI" Fallacy*: In recent years, with the rise of Big Data and cloud computing, it has been suggested that large-scale deep learning models encompassing billions of parameters tuned using millions of documents can address most NLP problems. The idea that such generic language models are "all you need" has been perpetuated by industry research related to powerful artifacts such as BERT and GPT-3 (Devlin et al. 2019, Brown et al. 2020). However, because of the pace of change and lack of thorough benchmarking, the efficacy and utility of such artifacts for a breadth of NLP tasks might be overstated (Zimbra et al. 2018). Our findings suggest that not only are such language models markedly less effective for personality detection than DeepPerson, they are often unable to

**Table 12.** Percentage Improvement in Performance (MAE) Across Different Personality Detectors

Models	Policy indicators						Performance indicators		
	CH	CF	INVEST	LEVER	IC	SG&A	D&E	ROA	Average
<b>RFR</b>									
DeepPerson	4.05**	3.39*	6.63**	5.16*	4.98*	8.15**	7.41**	9.00**	6.10
CNN-1	0.36	-1.83	4.84*	2.60	0.92	2.15*	1.14	3.09	1.66
CNN-2	3.51*	-4.59	3.90*	2.05	-4.27	-1.07	6.43**	8.03*	1.75
BERT	3.91*	1.85	4.44*	3.98*	2.10*	3.54*	3.51	3.89**	3.40
<b>GBDT</b>									
DeepPerson	10.22**	6.40**	8.31**	7.10*	5.27**	18.02**	8.55**	11.78**	9.46
CNN-1	3.58*	1.54	4.93*	0.63	0.60	3.92*	0.19	3.04	2.30
CNN-2	1.36	-7.09	1.80	-1.89	-6.44	-11.75	7.89*	5.75*	-1.30
BERT	2.67*	3.66**	6.43*	2.90*	4.50	12.88**	6.33	6.64*	5.75
ARIMA	-8.98	-16.87	-8.64	-16.85	-11.61	-7.75	-5.67	-20.55	-12.11

Notes. Each value is a percentage. For each regression model and financial indicator, we estimate the average improvement with or without incorporating senior executives' personality traits into the model. The ARIMA row shows the possible improvement for GBDT relative to the common time-series prediction model (ARIMA). DeepPerson, CNN-1, CNN-2, and BERT refer to predictions using executives' personality traits detected by the respective methods.

\*\* $p < 0.01$ ; \* $p < 0.05$ : Wilcoxon signed-rank test.

offer statistical or practical significance for downstream forecasting contexts. This is consistent with recent studies that have warned generic language models are like “stochastic parrots” that might be getting too big by over relying on the sheer number of word tokens used during pretraining (Bender et al. 2021). Case in point, BERT-Base and PersonaBERT relied on 3.3 and 4.1 billion tokens, respectively, whereas DeepPerson only used 800 million. RoBERTa used 10 times as much data as BERT (an estimate 30 billion-plus tokens). As we foreshadowed earlier, we believe the demise of artifacts grounded in principled domain adaption has been overstated.

(2) *Design Science as a Mechanism for Middle-Ground Frameworks*: In contexts where limited labeled data related to the target task is available, brute force learning strategies are less effective. In such cases, representation engineering that adapts machine learning artifacts such as encoders, embeddings, attention mechanisms, and custom transfer learning schemes can present opportunities for effective domain adaptation (Abbasi et al. 2019). By serving as a mechanism for balancing the tradeoffs between data and intuition, socio and technical factors, inductive versus deductive insights, and general versus domain-specific learning, design science represents a robust approach for developing middle-ground frameworks that harness the power of human cumulative tradition in concert with powerful artificial intelligence.

(3) *Importance of Personality for Predicting Policy*: We show that when done correctly, personality dimensions can improve our foresight related to prediction of policy indicators and outcomes. The inclusion of personality measures derived by DeepPerson enhanced forecasts for financial policy indicators by 6 to 14 percentage points on average. Similarly, DeepPerson attained the biggest lifts for health pandemic forecasting relative to alternative epidemiological and data-driven models examined

(see Online Appendix H). Recently, many predictive analytics researchers have noted the challenges related to forecasting complex policy-related outcomes, including noisy input data and the need for a diversity of models (Bertozzi et al. 2020, Hutson 2020). Our results suggest that the traits of leaders tasked with informing policy-related decisions might be another important input for such models. In addition to influencing decisions directly, leaders' traits may often reflect the characteristics of the organizations or populations they lead and represent—for example, advisory boards and employees in firms or the general public and government in states and countries (Hambrick 2007). Whereas the reverse causal relationship between leader personality and outcomes of organizations might be debated in empirical causal inference studies, in prediction contexts (Shmueli and Koppius 2011), our study suggests that the personality of executives might serve as a rich low-dimensional feature representation for forecasting policy-related indicators and outcomes.

(4) *Toward Proactive Personalization*: Accurate automated personality detection has important implications for the broader movement toward “proactive personalization.” In personalized marketing, personality information can enrich predictive models related to various stages of the customer lifecycle including acquisition, retention, and expansion (Gupta et al. 2006, Brown et al. 2015). As cybersecurity moves from reactive to proactive, personality measures could enhance predictive user models in human-in-the-loop frameworks (Parrish et al. 2009, Bravo-Lillo et al. 2010). In human capital management contexts, workforce analytics models already leveraging survey-based personality measures could be made timelier with NLP-based personality scores (Ryan and Herleman 2015). In precision medicine, with the trend toward public health 3.0 (DeSalvo et al. 2017), personality information can help better align preventative



interventions with individual patient characteristics (Friedman 2000). For instance, the conscientiousness trait has been found to be predictive of health and longevity, from childhood to old age (Friedman and Kern 2014). Higher extraversion is linked to greater likelihood of seeking preventative screenings (Aschwanden et al. 2019). Lower conscientiousness and high neuroticism have been associated with greater vaccine hesitancy (Aschwanden et al. 2021, Murphy et al. 2021). Personality could provide a mechanism for measuring heterogeneity in user intent (Ahmad et al. 2022). NLP-based personality detection could inform various such proactive intervention personalization use cases.

Our work is not without its limitations. Bias is an important consideration for NLP models (Lalor et al. 2022). Furthermore, future work on personality across languages and using multimedia input including audio and video would be beneficial. Our design evaluation focused on social media postings, forum messages, and lengthier texts (essays). Other relevant documents might warrant exploration, including speech transcripts and written articles. Nevertheless, we believe this work has important implications for research at the intersection of design and data science that integrates social-technical concepts into novel domain-adapted machine learning artifacts, and for practitioners that enable, produce, or consume predictive analytics where the inclusion of personality information may enhance insight and foresight.

## References

- Abbasi A, Chen H (2008) CyberGate: A design framework and system for text analysis of computer-mediated communication. *Management Inform. Systems Quart.* 32(4):811–837.
- Abbasi A, Kitchens B, Ahmad F (2019) The risks of AutoML and how to avoid them. *Harvard Business Review* (October 24), <https://hbr.org/2019/10/the-risks-of-automl-and-how-to-avoid-them>.
- Abbasi A, Sarker S, Chiang RH (2016) Big Data research in information systems: Toward an inclusive research agenda. *J. Assoc. Inform. Systems* 17(2):3.
- Abbasi A, Zhou Y, Deng S, Zhang P (2018) Text analytics to support sense-making in social media: A language-action perspective. *Management Inform. Systems Quart.* 42(2):427–464.
- Adamopoulos P, Ghose A, Todri V (2018) The impact of user personality traits on word of mouth: Text-mining social media platforms. *Inform. Systems Res.* 29(3):612–640.
- Agastya IMA, Handayani DOD, Mantoro T (2019) A systematic literature review of deep learning algorithms for personality trait recognition. *Proc. 5th Internat. Conf. on Comput. Engrg. and Design* (IEEE, New York), 1–6.
- Ahmad H, Asghar MZ, Khan AS, Habib A (2020a) A systematic literature review of personality trait classification from textual content. *Open Comput. Sci.* 10:175–193.
- Ahmad F, Abbasi A, Kitchens B, Adjeroh DA, Zeng D (2022) Deep learning for adverse event detection from web search. *IEEE Trans. Knowledge Data Engrg.*, Forthcoming.
- Ahmad F, Abbasi A, Li J, Dobolyi DG, Netemeyer R, Clifford G, Chen H (2020b) A deep learning architecture for psychometric natural language processing. *ACM Trans. Inform. Systems* 38(1):Article No. 6.
- Alam F, Stepanov EA, Riccardi G (2013) Personality traits recognition on social network-Facebook. *Proc. 7th Internat. AAAI Conf. on Weblogs and Social Media* (AAAI, California), 1–4.
- Arazy O, Kumar N, Shapira B (2010) A theory-driven design framework for social recommender systems. *J. Assoc. Inform. Systems* 11(9):2.
- Aschwanden D, Gerend MA, Luchetti M, Stephan Y, Sutin AR, Terracciano A (2019) Personality traits and preventive cancer screenings in the Health Retirement Study. *Preventative Medicine* 126:105763.
- Aschwanden D, Strickhouser JE, Sesker AA, Lee JH, Luchetti M, Stephan Y, Sutin AR, et al. (2021) Psychological and behavioural responses to coronavirus disease 2019: The role of personality. *Eur. J. Personality* 35(1):51–66.
- Back MD, Stopfer JM, Vazire S, Gaddis S, Schmukle SC, Egloff B, Gosling SD (2010) Facebook profiles reflect actual personality, not self-idealization. *Psych. Sci.* 21(3):372–374.
- Barth ME, Cram DP, Nelson KK (2001) Accruals and the prediction of future cash flows. *Accounting Rev.* 76(1):27–58.
- Beltagy I, Lo K, Cohan A (2019) SciBERT: A pretrained language model for scientific text. Inui K, Jiang J, Ng V, Wan X, eds. *Proc. Conf. Empirical Methods in Natural Language Processing* (ACL, Pennsylvania), 3615–3620.
- Bender EM, Gebru T, McMillan-Major A, Shmitchell S (2021) On the dangers of stochastic parrots: Can language models be too big? *Proc. ACM Conf. on Fairness, Accountability, and Transparency* (ACM, New York), 610–623.
- Bertozzi AL, Franco E, Mohler G, Short MB, Sledge D (2020) The challenges of modeling and forecasting the spread of COVID-19. *Proc. National Acad. Sci. USA* 117(29):16732–16738.
- Bertrand M, Schoar A (2003) Managing with style: The effect of managers on firm policies. *Quart. J. Econom.* 118(4):1169–1208.
- Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. *J. Machine Learn. Res.* 3:993–1022.
- Bonsall SB, Holzman ER, Miller BP (2017) Managerial ability and credit risk assessment. *Management Sci.* 63(5):1425–1449.
- Bravo-Lillo C, Cranor LF, Downs J, Komanduri S (2010) Bridging the gap in computer security warnings: A mental model approach. *IEEE Security Privacy* 9(2):18–26.
- Brick IE, Palmon O, Wald JK (2006) CEO compensation, director compensation, and firm performance: Evidence of cronyism? *J. Corporate Finance* 12(3):403–423.
- Brown DE, Abbasi A, Lau RY (2015) Predictive analytics: Predictive modeling at the micro level. *IEEE Intelligence Systems* 30(3):6–8.
- Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, et al. (2020) Language models are few-shot learners. *Adv. Neural Inform. Processing Systems* 33:1877–1901.
- Celli F, Pianesi F, Stillwell D, Kosinski M, et al. (2013) Workshop on computational personality recognition (shared task). *Proc. 7th Internat. AAAI Conf. on Weblogs and Social Media* (AAAI, California), 2–5.
- Chelba C, Mikolov T, Schuster M, Ge Q, Brants T, Koehn P (2013) One billion word benchmark for measuring progress in statistical language modeling. Preprint, submitted December 11, <http://arxiv.org/abs/1312.3005>.
- Chen H, Chiang RHL, Storey VC (2012) Business intelligence and analytics: From Big Data to big impact. *Management Inform. Systems Quart.* 36(4):1165–1188.
- Chen D, Wang W, Gao W, Zhou Z (2018) Tri-net for semi-supervised deep learning. *Proc. 27th Internat. Joint Conf. on Artificial Intelligence* (AAAI, Palo Alto, CA), 2014–2020.
- Cheng J, Zhao S, Zhang J, King I, Zhang X, Wang H (2017) Aspect-level sentiment classification with heat (hierarchical attention) network. *Proc. ACM Conf. on Inform. and Knowledge Management* (ACM, New York), 97–106.

- Chu CI, Chatterjee B, Brown A (2013) The current status of greenhouse gas reporting by Chinese companies. *Management Audit J.* 28(2):114–139.
- Cobb-Clark D, Schurer S (2012) The stability of big-five personality traits. *Econom. Lett.* 115:11–15.
- Crayne MP, Medeiros KE (2020) Making sense of crisis: Charismatic, ideological, and pragmatic leadership in response to Covid-19. *Amer. Psychologist.* 76(3):462–474.
- DeSalvo KB, Wang YC, Harris A, Auerbach J, Koo D, O'Carroll P (2017) Public health 3.0: A call to action for public health to meet the challenges of the 21st century. *Prevention Chronic Dis.* 14:1–9.
- Devaraj S, Easley RF, Crant JM (2008) Research note—How does personality matter? Relating the five-factor model to technology acceptance and use. *Inform. Systems Res.* 19(1):93–105.
- Devlin J, Chang M, Lee K, Toutanova K (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. *Proc. Conf. of the NAACL-HLT (ACL, Pennsylvania)*, 4171–4186.
- Dubofsky P, Varadarajan PR (1987) Diversification and measures of performance: Additional empirical evidence. *Acad. Management J.* 30(3):597–608.
- Edunov S, Ott M, Auli M, Grangier D (2018) Understanding back-translation at scale. *Proc. Conf. on Empirical Methods in Natural Language Processing (ACL, Pennsylvania)*, 489–500.
- Farnadi G, Zoghbi S, Moens MF, Cock MD (2013) Recognising personality traits using Facebook status updates. *Proc. 7th Internat. AAAI Conf. on Weblogs and Social Media (AAAI, California)*, 14–18.
- Feng S, Wang Y, Liu L, Wang D, Yu G (2019) Attention based hierarchical LSTM network for context-aware microblog sentiment classification. *World Wide Web (Bussum)* 22(1):59–81.
- Friedman HS (2000) Long-term relations of personality and health: Dynamisms, mechanisms, tropisms. *J. Personality* 68(6):1089–1107.
- Friedman HS, Kern ML (2014) Personality, well-being, and health. *Annu. Rev. Psych.* 65:719–742.
- Gal Y, Islam R, Ghahramani Z (2017) Deep Bayesian active learning with image data. *Proc. Internat. Conf. on Machine Learn. (PMLR)*, 1183–1192.
- Galassi A, Lippi M, Torrioni P (2020) Attention in natural language processing. *IEEE Trans. Neural Networks Learn. System* 32(10):1–18.
- Gao S, Ramanathan A, Tourassi G (2018) Hierarchical convolutional attention networks for text classification. *Proc. 3rd Workshop on Representation Learn. for NLP*, 11–23.
- Gill AJ, Oberlander J (2003) Perception of email personality at zero-acquaintance: Extraversion takes care of itself; neuroticism is a worry. *Proc. Cognitive Sci. Soc.*, 456–461.
- Gjurković M, Karan M, Vukojević I, Bošnjak M, Šnajder J (2021) PANDORA talks: Personality and demographics on Reddit. *Proc. 9th Internat. ACL Workshop on Natural Language Processing for Social Media (ACL, Pennsylvania)*, 138–152.
- Go A, Bhayani R, Huang L (2009) Twitter sentiment classification using distant supervision. Accessed December 1, 2009, <https://cs.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision09.pdf>.
- Goldberg LR (1990) An alternative description of personality: The big-five factor structure. *J. Personality Soc. Psych.* 59(6):1216–1229.
- Gregor S, Hevner A (2013) Positioning and presenting design science research for maximum impact. *Management Inform. Systems Quart.* 37(2):337–355.
- Guan Z, Wu B, Wang B, Liu H (2020) Personality2vec: Network representation learning for personality. *Proc. IEEE 5th Internat. Conf. on Data Sci. in Cyberspace (IEEE, New York)*, 30–37.
- Guest JL, Rio CD, Sanchez T (2020) The three steps needed to end the Covid-19 pandemic: Bold public health leadership, rapid innovations, and courageous political will. *JMIR Public Health* 6(2):e19043.
- Gupta S, Hanssens D, Hardie B, Kahn W, Kumar V, Lin N, Ravishanker N, et al. (2006) Modeling customer lifetime value. *J. Service Res.* 9(2):139–155.
- Halliday MAK, Hasan R (2004) *An Introduction to Functional Grammar*, 3rd ed. (Routledge, Oxfordshire, England, UK).
- Hambrick DC (2007) Upper echelons theory: An update. *Acad. Management Rev.* 32(2):334–343.
- Hambrick DC, Mason PA (1984) Upper echelons: The organization as a reflection of its top managers. *Acad. Management Rev.* 9(2):193–206.
- Hastie T, Tibshirani R, Friedman J (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer Science & Business Media, Berlin).
- Haussler D, Kearns M, Schapire RE (1994) Bounds on the sample complexity of Bayesian learning using information theory and the VC dimension. *Machine Learn.* 14:83–113.
- Heavey C, Simsek Z, Kyprianou C, Risius M (2020) How do strategic leaders engage with social media? A theoretical framework for research and practice. *Strategic Management J.* 41(8):1490–1527.
- Henderson AD, Miller D, Hambrick DC (2006) How quickly do CEOs become obsolete? Industry dynamism, CEO tenure, and company performance. *Strategic Management J.* 27(5):447–460.
- Hevner A, March S, Park J, Ram S (2004) Design science in IS research. *Management Inform. Systems Quart.* 28(1):75–105.
- Hough JR, Ogilvie OT (2005) An empirical test of cognitive style and strategic decision outcomes. *J. Management Stud.* 42(2):417–448.
- Howard J, Ruder S (2018) Fine-tuned language models for text classification. Preprint, submitted January 18, <https://arxiv.org/abs/1801.06146>.
- Hrazdil K, Novak J, Rogo R, Wiedman C, Zhang R (2020) Measuring executive personality using machine-learning algorithms: A new approach and validation tests. *J. Bus. Finance Accounting* 47(3–4):519–544.
- Huang A, Wang H, Yang Y (2020) FinBERT—A deep learning approach to extracting textual information. Preprint, submitted July 28, <https://dx.doi.org/10.2139/ssrn.3910214>.
- Hutson M (2020) The mess behind the models: Too many of the COVID-19 models led policymakers astray. Here's how tomorrow's models will get it right. *IEEE Spectrum* 57(10):30–35.
- Iacobelli F, Gill AJ, Nowson S, Oberlander J (2011) Large scale personality classification of bloggers. *Affective Computing and Intelligent Interaction* (Springer, Berlin), 568–577.
- Janson MA, Woo CC (1996) A speech act lexicon: An alternative use of speech act theory in information systems. *Inform. Systems J.* 6(4):301–329.
- Jayarathne M, Jayatilleke B (2020) Predicting personality using answers to open-ended interview questions. *IEEE Access* 8:115345–115355.
- Jing R (2019) A self-attention based LSTM network for text classification. *J. Phys.* 1207(1):1–5.
- John GH, Kohavi R, Pflieger K (1994) Irrelevant features and the subset selection problem. *Proc. 11th Internat. Conf. on Machine Learn.*, 121–129.
- Judge TA, Piccolo RF, Kosalka T (2009) The bright and dark sides of leader traits: A review and theoretical extension of the leader trait paradigm. *Leadership Quart.* 20(6):855–875.
- Judge TA, Bono JE, Ilies R, Gerhardt MW (2002) Personality and leadership: A qualitative and quantitative review. *J. Appl. Psych.* 87(4):765–780.
- Kim Y, Jernite Y, Sontag D, Rush AM (2016) Character-aware neural language models. *Proc. 30th AAAI Conf. on Artificial Intelligence (AAAI, California)*, 2741–2749.

- Laine S, Aila T (2016) Temporal ensembling for semi-supervised learning. Preprint, submitted October 7, <https://arxiv.org/abs/1610.02242>.
- Lalor JP, Yang Y, Smith K, Forsgren N, Abbasi A (2022) Benchmarking intersectional biases in NLP. *Proc. Assoc. Comput. Linguistics* (Association for Computational Linguistics, Pennsylvania).
- Le Q, Mikolov T (2014) Distributed representations of sentences and documents. *Proc. Internat. Conf. on Machine Learn.*, 1188–1196.
- Lee DH (2013) Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks, *Proc. ICML Workshop on Challenges in Representation Learning* 3(2):896.
- Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J (2020) BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36(4):1234–1240.
- Leonardi S, Monti D, Rizzo G, Morisio M (2020) Multilingual transformer-based personality traits estimation. *Information (Basel)* 11(4):179.
- LePine JA, Van Dyne L (2001) Voice and cooperative behavior as contrasting forms of contextual performance: Evidence of differential relationships with big five personality characteristics and cognitive ability. *J. Appl. Psych.* 86(2):326.
- Li M, Simerly RL (1998) The moderating effect of environmental dynamism on the ownership and performance relationship. *Strategic Management J.* 19(2):169–179.
- Li J, Larsen K, Abbasi A (2020) TheoryOn: A design framework and system for unlocking behavioral knowledge through ontology learning. *Management Inform. Systems Quart.* 44(4):1733–1777.
- Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, et al. (2019) RoBERTa: A robustly optimized BERT pretraining approach. Preprint, submitted July 26, <https://arxiv.org/abs/1907.11692>.
- Liu Z, Wang Y, Mahmud J, Akkiraju R, Schoudt J, Xu A, Donovan B (2016) To buy or not to buy? Understanding the role of personality traits in predicting consumer behaviors. Spiro E, Ahn YY, eds. *Social Informatics*, vol. 10047, Lecture Notes in Computer Science (Springer, Cham), 337–346.
- Lynn V, Balasubramanian N, Schwartz HA (2020) Hierarchical modeling for user personality prediction: The role of message-level attention. *Proc. 58th Annual Meeting of the ACL* (ACL, Pennsylvania), 5306–5316.
- Lyytinen KJ (1985) Implications of theories of language for information systems. *MIS Quart.* 9(1):61–74.
- Mairesse F, Walker MA, Mehl MR, Moore RK (2007) Using linguistic cues for the automatic recognition of personality in conversation and text. *J. Artificial Intelligence Res.* 30:457–500.
- Majumder N, Poria S, Gelbukh A, Cambria E (2017) Deep learning-based document modeling for personality detection from text. *IEEE Intelligence Systems* 32(2):74–79.
- Masli A, Richardson VJ, Watson MW, Zmud RW (2016) Senior executives' IT management responsibilities: Serious IT-related deficiencies and CEO/CFO turnover. *Management Inform. Systems Quart.* 40(3):687–708.
- Medcof JW (2007) CTO power. *Res. Tech. Management* 50(4):23–31.
- Mehl MR, Gosling SD, Pennebaker JW (2006) Personality in its natural habitat: manifestations and implicit folk theories of personality in daily life. *J. Personality Soc. Psych.* 90(5):862.
- Mehta Y, Majumder N, Gelbukh A, Cambria E (2020) Recent trends in deep learning based personality detection. *Artificial Intelligence Rev.* 53(4):2313–2339.
- Mohamed N, Ahmad MH, Ismail Z, Suhartono S (2010) Short term load forecasting using double seasonal arima model. *Proc. Regional Conf. on Statist. Sci.* (10):57–73.
- Murphy J, Vallières F, Bentall RP, Shevlin M, McBride O, Hartman TK, et al. (2021) Psychological characteristics associated with COVID-19 vaccine hesitancy and resistance in Ireland and the United Kingdom. *Nature Comm.* 12(1):1–15.
- Nadkarni S, Herrmann POL (2010) CEO personality, strategic flexibility, and firm performance: The case of the Indian business process outsourcing industry. *Acad. Management J.* 53(5):1050–1073.
- Nygren TE, White RJ (2005) Relating decision making styles to predicting self-efficacy and a generalized expectation of success and failure. *Proc. Human Factors and Ergonomics Soc. Meeting*, 432–434.
- Pan SJ, Yang Q (2009) A survey on transfer learning. *IEEE Trans. Knowledge Data Engrg.* 22(10):1345–1359.
- Parrish JL Jr, Bailey JL, Courtney JF (2009) *A Personality Based Model for Determining Susceptibility to Phishing Attacks* (University of Arkansas, Little Rock).
- Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M (2011) Scikit-learn: Machine learning in Python. *J. Machine Learn. Res.* 12:2825–2830.
- Pennebaker JW, King LA (1999) Linguistic styles: Language use as an individual difference. *J. Personality Soc. Psych.* 77(6):1296.
- Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L (2018) Deep contextualized word representations. *Proc. NAACL-HLT* (ACL, Pennsylvania), 2227–2237.
- Peterson RS, Smith DB, Martorana PV, Owens PD (2003) The impact of chief executive officer personality on top management team dynamics: One mechanism by which leadership affects organizational performance. *J. Appl. Psych.* 88(5):795–808.
- Pratama BY, Sarno R (2015) Personality classification based on Twitter text using naive Bayes, KNN and SVM. *Proc. IEEE Internat. Conf. on Data and Software Engrg.* (IEEE, New York), 170–74.
- Prechelt L (1998) Automatic early stopping using cross validation. *Neural Networks* 11(4):761–767.
- Riaz MN, Riaz MA, Batool N (2012) Personality types as predictors of decision making styles. *J. Behav. Sci.* 22(2):99–114.
- Ryan J, Herleman H (2015) A Big Data platform for workforce analytics. Tonidandel S, King EB, Cortina JM, eds. *Big Data at Work: The Data Science Revolution and Organizational Psychology* (Routledge, Oxfordshire, England, UK), 19–42.
- Shi Z, Lee GM, Whinston AB (2016) Toward a better measure of business proximity: Topic modeling for industry intelligence. *Management Inform. Systems. Quart.* 40(4):1035–1056.
- Shmueli G, Koppius O (2011) Predictive analytics in information systems research. *Management Inform. Systems Quart.* 35(3):553–572.
- Sun X, Liu B, Cao J, Luo J, Shen X (2018) Who am I? Personality detection based on deep learning for texts. *Proc. IEEE Internat. Conf. on Comm.* (IEEE, New York), 1–6.
- Tadesse MM, Lin H, Xu B, Yang L (2018) Personality predictions based on user behavior on the Facebook social media platform. *IEEE Access* 6:61959–61969.
- Tausczik YR, Pennebaker JW (2010) The psychological meaning of words: LIWC and computerized text analysis methods. *J. Language Soc. Psych.* 29(1):24–54.
- Torrey L, Shavlik J (2010) Transfer learning. *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques* (IGI Global), 242–264.
- Vinciarelli A, Mo G (2014) Survey of personality computing. *IEEE Trans. Affective Comput.* 5:273–291.
- Walls JG, Widmeyer GR, El Sawy OA (1992) Building an information system design theory for vigilant EIS. *Inform. Systems Res.* 3(1):36–59.
- Wang Q, Lau RYK, Xie H (2021) The impact of social executives on firms' mergers and acquisitions strategies: A difference-in-difference analysis. *J. Bus. Res.* 123:343–354.
- Wang Z, Wu CH, Li QB, Yan B, Zheng KF (2020) Encoding text information with graph convolutional networks for personality recognition. *Appl. Sci. (Switzerland)* 10(12):4081.
- Wang Z, Wu C, Zheng K, Niu X, Wang X (2019b) SMOTETomek-based resampling for personality recognition. *IEEE Access* 7:129678–129689.



- Wang Y, Chen Q, Ahmed M, Li Z, Pan W, Liu H (2019a) Joint inference for aspect-level sentiment analysis by deep neural networks and linguistic hints. *IEEE Trans. Knowledge Data Engrg.* 33(5):1–12.
- Weng PS, Chen WY (2017) Doing good or choosing well? Corporate reputation, CEO reputation, and corporate financial performance. *North Amer. J. Econom. Finance* 39:223–240.
- Wilcoxon F (1992) *Individual Comparisons by Ranking Methods. Breakthroughs in Statistics* (Springer, Berlin).
- Wright WR, Chin DN (2014) Personality profiling from text: Introducing part-of-speech N-grams. *Proc. Internat. Conf. on User Modeling, Adaptation, and Personalization* (Springer, Berlin), 243–253.
- Xie Q, Dai Z, Hovy E, Luong MT, Le QV (2020) Unsupervised data augmentation for consistency training. Ranzato M, Beygelzimer A, Nguyen K, Liang PS, Vaughan JW, Dauphin Y, eds. *Proc. 34th Conf. on Neural Inform. Processing Systems*, vol. 33 (Curran Associates, Red Hook, New York), 6256–6268.
- Xue D, Wu L, Hong Z, Guo S, Gao L, Wu Z, Zhong X, Sun J (2018) Deep learning-based personality recognition from text posts of online social networks. *Appl. Intelligence* 48(11):4232–4246.
- Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E (2016) Hierarchical attention networks for document classification. *Proc. Conf. of the NAACL: Human Language Technologies* (ACL, Pennsylvania), 1480–1489.
- Yu J, Markov K (2017) Deep learning based personality recognition from Facebook status updates. *Proc. 8th IEEE Internat. Conf. on Awareness Sci. and Tech.* (IEEE, New York), 383–387.
- Zhou J, Huang JX, Chen Q, Hu QV, Wang T, He L (2019) Deep learning for aspect-level sentiment classification: Survey, vision, and challenges. *IEEE Access* 7:78454–78483.
- Zimbra D, Abbasi A, Zeng D, Chen H (2018) The state-of-the-art in Twitter sentiment analysis: A review and benchmark evaluation. *ACM Trans. Management Inform. Systems* 9(2):1–29.