# Enriching Social Media Personas with Personality Traits: A Deep Learning Approach Using the Big Five Classes

Joni Salminen[1,2](✉), Rohan Gurunandan Rao[3], Soon-gyo Jung[1], Shammur A. Chowdhury[1], and Bernard J. Jansen[1]

[1] Qatar Computing Research Institute,
Hamad Bin Khalifa University, Doha, Qatar
jsalminen@hbku.edu.qa
[2] University of Turku, Turku, Finland
[3] Indian Institute of Technology Madras, Tamil Nadu, India

**Abstract.** To predict personality traits of data-driven personas, we apply an automatic persona generation methodology to generate 15 personas from the social media data of an online news organization. After generating the personas, we aggregate each personas' YouTube comments and predict the "Big Five" personality traits of each persona from the comments pertaining to that persona. For this, we develop a deep learning classifier using three publicly available datasets. Results indicate an average performance increase of 4.84% in F1 scores relative to the baseline. We then analyze how the personas differ by their detected personality traits and discuss how personality traits could be implemented in data-driven persona profiles, as either scores or narratives.

**Keywords:** Personas · Design · Personality detection · Neutral networks

## 1 Introduction

A persona is defined as a fictitious person that describes user or customer segments of a software system, product, or service [1, 2]. Personas are widely used in many professional fields, including e.g. software development and design [3], marketing and advertising [4], health informatics [5], and so on. A persona simplifies user-centric numbers into an easy-to-understand representation - another human being [6]. Through this property, personas aim to facilitate the communication about users within an organization, so that user-centric decisions (e.g., product development, design, marketing) can be made keeping the end user in mind [7].

Persona design, in turn, deals with the design of persona profiles that support persona users' tasks and goals, while not distracting them [8]. To be useful, personas should contain all information decision makers need to better understand the group of users the persona portrays. This is known as the *rounded persona principle* [9]. A typical persona profile includes a name, a picture, and a description detailing attitudes, needs, wants, and behaviors of the persona [10]. Rounded persona profiles can also include personality traits [11, 12], such as the "Big Five" (BF) traits: *extroversion*

H. Degen and L. Reinerman-Jones (Eds.): HCII 2020, LNCS 12217, pp. 101–120, 2020.
https://doi.org/10.1007/978-3-030-50334-5_7

*(EXT)*, *agreeableness (AGR)*, *openness (OPE)*, *conscientiousness (CON)*, and *neuroticism (NEU)* [13]. Together, these traits are seen to reflect one's personality, defined as a fairly stable state of mind of an individual, where "state of mind" refers to how the individual approaches the world and interacts with others [14].

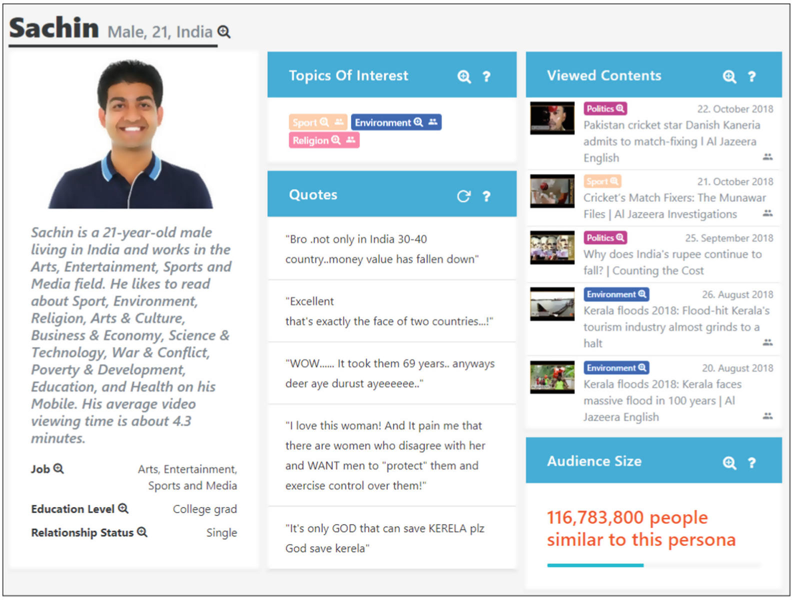
Personality traits can help predict user and customer behavior under different circumstances and use cases, including shopping behavior [15], problem solving patterns [16], task performance [17], voting behavior [18], and so on. Knowing the personality traits can also aid in targeting the users with tailored advertising messages [19]. These linkages between personality traits and behavior imply that including personality traits in persona profiles can be useful, as personas are used for *predicting* user behavior in different situations [1]. Therefore, there are multiple potential benefits to the inclusion of personality traits to personas as completing information.

A related development in computer science is the progress made in automatic personality detection (APD). Overall, the factors that positively contribute to on-going research efforts on APD are (1) the availability of textual data (i.e., social media posts such as tweets, Facebook updates, and YouTube (YT) comments), (2) the personal nature of that data, and (3) the increase of popularity in "sharing" one's thoughts and feelings via social media platforms. Interestingly, these drivers strongly overlap with the drivers of "quantified" or data-driven personas (DDPs) [20], defined as personas that represent social media and online analytics user data in the form of personas [21]. Thus, there is an opportunity to *combine APD and DDPs for more efficient, more complete, and more useful personas that enhance decision makers (e.g., software developers, designers, marketers, etc.) understanding about their users and customers*.

However, in our knowledge, no previous work on DDPs has incorporated personality traits into the personas. Rather, the DDP layouts have focused on containing other information, such as text description, topics of interest, and audience size (see Fig. 1). This is somewhat surprising since APD techniques have rapidly evolved [22–24]. In particular, using neural networks (NNs) to analyze social media texts has yielded positive results for ADP [24, 25]. As put by Carducci et al. [26], social media platforms provide a rich source of user-generated texts that reflect "many aspects of real life, including personality" (p. 127). The fact that users freely share their opinions, moods, and feelings makes these users prime candidates for personality detection, as the personal information they share can be analyzed for personality cues [24].

By definition, DDPs are created using both numerical online analytics data that describes user behavior in an aggregated fashion [27] *and* user-generated social media posts that describe the persona's attitudes [28]. Therefore, APD from a persona's social media content appears as a prominent research gap. In particular, three aspects support this: (1) personality information can support the understanding and use of personas to predict user behavior (i.e., there is both information needs of persona users, (2) textual social media data that can be used for *both* persona creation and APD is widely available, and (c) technology is granting opportunities for merging efficient creation of DDPs with an increasing accuracy of APD from user-generated social media texts.

Therefore, in this research, we combine APD and DDPs to design personas with personality traits that could be automatically generated using numerical and textual social media data. Our research questions are as follows:



Sachin Male, 21, India

Sachin is a 21-year-old male living in India and works in the Arts, Entertainment, Sports and Media field. He likes to read about Sport, Environment, Religion, Arts & Culture, Business & Economy, Science & Technology, War & Conflict, Poverty & Development, Education, and Health on his Mobile. His average video viewing time is about 4.3 minutes.

| | |
|---|---|
| Job | Arts, Entertainment, Sports and Media |
| Education Level | College grad |
| Relationship Status | Single |

Topics Of Interest

Sport · Environment · Religion

Quotes

"Bro .not only in India 30-40 country..money value has fallen down"

"Excellent
that's exactly the face of two countries...!"

"WOW...... It took them 69 years.. anyways deer aye durust ayeeeeee.."

"I love this woman! And It pain me that there are women who disagree with her and WANT men to "protect" them and exercise control over them!"

"It's only GOD that can save KERELA plz God save kerela"

Viewed Contents

Politics 22. October 2018
Pakistan cricket star Danish Kaneria admits to match-fixing l Al Jazeera English

Sport 21. October 2018
Cricket's Match Fixers: The Munawar Files | Al Jazeera Investigations

Politics 25. September 2018
Why does India's rupee continue to fall? | Counting the Cost

Environment 26. August 2018
Kerala floods 2018: Flood-hit Kerala's tourism industry almost grinds to a halt

Environment 20. August 2018
Kerala floods 2018: Kerala faces massive flood in 100 years | Al Jazeera English

Audience Size

116,783,800 people similar to this persona

**Fig. 1.** An example persona, containing text description, demographics, topical interests, quotes, viewed content, and representative audience size - but no personality traits

- How can APD be applied to infer the BF personality traits of DDPs from social media posts of the personas?
- How can the personality traits implemented in the DDP design?

To develop a NN classifier, we collect three publicly available datasets with ground truth on personality traits (see Sect. 3). We then develop the classifier based on inspiration from state-of-the-art models, including the code and data from Majumder et al. [24] that predicts the BF personality traits [13].

To predict personas' personality, we generate 15 personas by collecting social media data of an international online news and media company and applying an automatic DDP creation methodology [21]. After generating these 15 personas, we collect their social media posts (i.e., aggregate each personas' YT comments) and predict the BF personality traits of each persona from the social media posts pertaining to that persona. We then interpret the results by analyzing how the personas differ by their detected personality traits.

We choose the BF framework as it is the most commonly chosen framework in the APD literature [24], making it possible to compare our results against a baseline model. However, our main contribution is not the development of the APD classifier – rather, we apply established, previously working methods in the domain of APD towards the goal of enriching DDPs with information on the persona's personality traits. Our contribution, thus, is demonstrating how DDP and APD methods can be combined for richer persona profiles that more realistically reflect the user population.

The remainder of this work is organized as follows. Section 2 summarizes the state-of-the-art research in APD. Section 3 explains our methodology, i.e., how we incorporate the available techniques in our modeling approach, as well as how we collect the data and evaluate the results. Section 4 presents the results in comparison to a baseline model and predicts the personality traits of 15 personas generated automatically from social media data. We conclude by discussing the limitations of the approach, implications for research and practice, and future research avenues.

## 2 Related Literature

### 2.1 Automatic Personality Detection

The roots of APD from social media texts can be found in two main streams of research: (1) *affective computing and sentiment analysis* [29] and (2) *linguistic styles and psycholinguistic databases* [30, 31]. The joint hypothesis of these streams is that language (e.g., spoken, written) or words reveal one's inner thoughts and, therefore, a person's personality [32]. As stated by Xue et al. [33], the textual information widely shared on social network is "the most direct and reliable way for people to translate their internal thoughts and emotions into a form that others can understand." (p. 4239). Thus, social media texts can reveal aspects of one's personality.

For this, research use many analytical predictors, including linguistic cues, syntactic features, and manually and automatically built semantic lexicons [34–36]. In other words, the intuition is that *linguistic markers* reveal one's personality, so that the use of questionnaires such as the BF Inventory [37] is not needed. Rather, the personality is inferred from unstructured text "in the wild".

For APD, anonymity of social media has two main implications. First, it can mitigate the social desirability bias of expressing one's true feelings. In other words, anonymity can decrease filters imposed by the need for pleasing others or the general opinion [38], possibly revealing truthful information about a person's thinking tendencies. Second, when user IDs are not available, identifying the same users brings about challenges in constructing adequate corpora for personality prediction [39]. Thus, anonymity can both hamper and facilitate APD from text.

Several methods and datasets have been applied for APD. These include, at least, Logistic Regression (LR) [40], K-Nearest Neighbors (KNN) [41], Naïve Bayes (NB) [42], Support Vector Machines (SVM) [43], and, more recently, deep neural networks [33] with architectures such as Convolutional Neural Networks (CNN) [44] and Recurrent Neural Networks (RNN) [45], including Long Short-Term Memory (LSTM), a subtype of RNN [46]. Thus, algorithms and technical framing of APD vary greatly – often, different methods are applied in combination, such as combining CNN and RNN [44]. The features used include, e.g., n-grams, Bag of Words (BOW), Linguistic Inquiry and Word Count (LIWC), and word embeddings [47]. These aim to represent different linguistic aspects that would correlate with the target personality trait.

The prominent datasets are, e.g., the myPersonality (MPD) [48] that contains Facebook status updates, the YT personality dataset [49], and the Essays dataset that contains a stream-of-consciousness essays [30]. These datasets are described in Sect. 3.3. Overall, the state-of-the-art performance is achieved using deep NN architectures trained on multiple datasets and feature representations [24, 33, 44, 46]. This is also the approach we take in this research, as explained in Sect. 3.

### 2.2 Personality Traits in Personas

Personality traits can potentially enhance understanding of *who* the persona is [54], enhancing the empathetic benefits attributed to the persona technique in general [55]. The main motive for inclusion of personality traits is the creation of "holistic persona description" [12] (see also the "rounded persona" concept [9]) that includes multiple types of information: (1) personal details such as demographics and interests, (2) personality traits as captured by psychological models ("psychographics"), (3) intelligence and learning styles, (4) knowledge that describes the persona's expertise and experience in a specific domain, and (5) cognitive processes that describe how the persona processes information. Despite the assumption that personality traits can enhance user understanding, as with many persona benefits [56], the potential of personality traits for persona user experience has not been empirically verified.

In their review of 47 persona templates, Nielsen et al. [10] found that personality and psychographics had been incorporated in persona profiles using manual means, often without using professional psychologists [11, 12, 50–52]. However, automatic inference of personality traits for personas has not previously been accomplished, to our knowledge. For this reason, Salminen et al. [53] consider APD as one of the open opportunities in DDP creation.

In the following section, we explain how we combine APD with DDP creation.

## 3 Methodology

### 3.1 Data-Driven Persona Generation

We generated two DDPs using real data from an actual organization, a large international news and media company. For this, we used the YT viewer data of the said organization. The persona generation follows the methodology developed by An et al. [21, 27] and Jung et al. [57], in which data is collected and processed automatically from online analytics platforms.

For this research, we collected 206 K video views from 13 K videos published between January 1, 2016 and September 30, 2018 on the YT channel of Al Jazeera Media Network (AJ+[1]). For the data collection, we used the YT Analytics Application Programming Interface (API[2]) with the channel owner's permission. The use of an API enables automatic updating of the personas at set intervals [58, 59] The dataset includes all the channel's view counts divided by demographic groups (age group × gender × country), of which there are 1631 with at least one view during the data collection period.

[1] https://www.youtube.com/channel/UCV3Nm3T-XAgVhKH9jT0ViRg.

[2] https://developers.google.com/youtube/analytics/.

The DDP creation methodology executes the following steps [60]:

- **Step 1:** Create an interaction matrix with YT videos as columns, demographic user groups as rows, and view count of each group for each video as elements of the matrix
- **Step 2:** Apply non-negative matrix factorization (NMF) [61] to the interaction matrix to infer $p$ latent video viewing behaviors, where $p$ is the number of personas
- **Step 3:** Choose the representative demographic attributes for each behavior by using weights from the NMF computation
- **Step 4:** Create the personas by enriching the representative demographic groups for each $p$ personas with information, e.g., name, picture, topics of interest, etc.

After obtaining a grouped interaction matrix, we apply NMF for identifying latent video viewing patterns. NMF is particularly intended for reducing the dimensionality of large datasets by discerning latent factors [61]. Figure 2 illustrates the matrix decomposition of NMF; the resulting patterns inferred from the matrix discriminate the user groups based on the variation of their content viewing patterns.

An example of automatically generated persona is provided in Fig. 1. For further technical reference, we refer the reader to the articles by An et al. [21, 27], as these report the technical details and validation of the method. This research focuses on adding personality traits to these automatically generated DDPs. We utilize the persona's quotes to predict the personality traits. The quotes are comments retrieved from the most viewed content of the persona.

### 3.2 Retrieving Social Media Comments for Each Persona

The process for retrieving the social media comments for the personas is as follows:

- **Step 1:** Generate 15 personas using the dataset described previously
- **Step 2:** Take the top 10 YT videos that each persona has viewed the most
- **Step 3:** Take all comments from these videos and save them in a data structure

Moreover, we ensure that these videos do not overlap, i.e., one persona's top videos do not contain any videos from another persona's top videos. The purpose of this is to ensure that the social media comments between the personas vary adequately to detect any linguistic patterns. The comments of each persona are then used to generate a score for each BF personality trait using the NN described in the following.

### 3.3 Data Collection for Model Development

We combine three publicly available[3] datasets: essays dataset [30], YouTube personality dataset [49], and MPD [48]. These datasets were selected for two reasons. First, they are commonly used in APD research (see, e.g., [22, 24, 34]. Second, they all use

[3] The MPD dataset was previously available on the Web (http://mypersonality.org), but at the time of writing it has been withdrawn. The YT dataset is available upon request (https://www.idiap.ch/dataset/youtube-personality), and the essays dataset can be readily downloaded (https://github.com/SenticNet/personality-detection/blob/master/essays.csv).

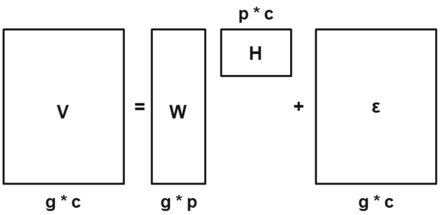


**Fig. 2.** Matrix decomposition carried out using NMF [21]. Matrix *V* is decomposed into *W* and *H*. *g* denotes demographic groups in the dataset, *c* denotes content (e.g., videos), and *p* is the number of latent interaction patterns that are used to create the personas. For this research, we set *p* = 15 to generate enough personas for comparison of their personality traits.

the same predicted classes, i.e., the BF personality traits. Because there are several personality trait taxonomies, it is important for model development that the textual data is associated with the same classes across the datasets used for training the model. The BF framework originates from the 1960s [62] and is continuously garnering research interest from psychologists and researchers in other fields to date [63–65].

In these datasets (see Table 1 for description), the BF traits form the five predicted classes. The definitions of the classes are, according to Agarwal [47] (p. 2):

- **OPE:** Artistic, curious, imaginative, curious, intelligent, and imaginative. Open individuals tend to be artistic and have sophisticated taste. They appreciate diverse views, ideas, and experiences.
- **CON:** Efficient, organized, responsible, organized, and persevering. Conscientious individuals tend to be reliable and focused on achieving, working hard, and planning for the future.
- **EXT:** Energetic, active, assertive, outgoing, amicable, assertive. These individuals are friendly and energetic, drawing inspiration from social situations.
- **AGR:** Compassionate, cooperative, cooperative, helpful, nurturing. Individuals that score high in agreeableness are peacekeepers. They are generally optimistic and trusting of others.
- **NEU:** Anxious, tense, self-pitying, anxious, insecure, sensitive. Neurotics are moody, tense, and easily tipped into experiencing negative emotions.

### 3.4 Strategy for Model Development

Each BF trait is modeled separately as a multiple binary classification task. We follow the approach taken by Majumder et al. [24] because of three reasons: (a) their method for personality detection is well documented in the related research article [24], (b) their code is publicly available in a GitHub repository[4] along with the training data, and (c) their model achieves good performance relative to other models in the field.

Thus, the GitHub code of Majumder et al. [24] served as a starting point for developing the classifier. Their results also form the baseline for benchmark

[4] https://github.com/senticnet/personality-detection.

**Table 1.** Description of the datasets used for model development

| Name | Source | Description | Size (words) |
|---|---|---|---|
| Essays | Students | Students wrote stream-of-consciousness texts. The Big Five ratings were obtained using a questionnaire | 2,467 essays (1,609,042) |
| myPersonality | Facebook | The dataset contains Facebook status posts as raw text, author information, and gold standard labels. The Big Five ratings were obtained using a questionnaire [47] | 9,880 status updates (143,639) |
| YouTube personality | YouTube | The Big Five ratings were obtained from crowd raters predicting personality traits of vloggers "based on what they say in their YouTube videos." [49] (p. 1) | 404 vlog transcripts (240,580) |
| | | | *Total* = 1,993,261 words |

comparison. Due to practical reasons, our approach differs from Majumder et al. [24] in two main aspects. First, we were not able to replicate the process for obtaining the Mairesse features [36] used by Majumder et al. [24]. These features represent the linguistic cues of personality in text. Thus, we had to model the data without the Mairesse features, representing the text using word embeddings instead (see Sect. 3.5).

Second, we use two modeling choices from Sun et al. [66]: (a) an LSTM architecture and (b) multiple datasets. The reason of the former is that LSTM shows good performance in APD [66], while the intuition of the latter is that more data enables the model to train on more linguistic cues [22]. In addition, we use a CNN component, as combined processing steps in NN architectures have shown to increase performance [47]. The model architecture is described in Sect. 3.7.

## 3.5 Data Cleaning and Preprocessing

We followed the text cleaning and processing steps of Majumder et al. [24]. This included sentence splitting as well as data cleaning and unification, such as reduction to lower case. We then tokenized all the text. The words were then converted to fixed length vectors as in Majumder et al. [24], using the GoogleNews Word2Vec model, and using a fixed vector length of 300 dimensions. All sequences were set to a fixed maximum length, which was varied between 2,000 and 12,000 to identify the optimal length to model the data. Sequences longer than the maximum length were truncated, while those shorter than the maximum length were padded. We did not use any hand-picked features like word-count or Mairesse features; instead, we passed the entire data to the NN for feature learning.

### 3.6 Data Partitioning and Model Training

Having performed the preprocessing, we used 10-fold cross validation [67] for checking the accuracy of our model. We split the data during the training phase into ten folds (i.e., parts). Each time, a fold was isolated, and a model was trained on the remaining nine folds. Then, the validation accuracy was tested on the $10^{\text{th}}$ fold. This was repeated ten times, keeping different folds aside for validation. We also trained different models for comparison of their predictive performance – this included comparing individually trained models with a combined model (i.e., five models that predict different traits individually vs. one that predicts all five traits simultaneously).

### 3.7 Neural Network Architecture

We developed a NN with two major sub-architectures: a single dimensional CNN since there is a spatial structure in the input text, and an LSTM network since there is also a temporal correlation between the words in the input text.

After training the model, the NN can identify the relationship between different words in the text and predict the personality trait to which it belongs. This is done by taking an arbitrary length input text, removing fillers, stop words, and foreign characters, tokenizing it, truncating or padding it to the maximum sequence length, then converting it into a matrix of shape (word embedding length = 300) by using the Word2Vec model [68]. The network transforms this matrix by passing it through a Convolution (32 filters of size 3 × 3) and Max Pooling (2 × 2 filters) layers to identify structural features in the text, then through a Spatial Dropout with a rate of 0.2. Then, the matrix is passed through a Bi-directional LSTM layer with 64 units to identify temporal features, and finally through the Dropout (rate of 0.5), Batch Normalization and Dense layers to predict the personalities. A Dense layer is where all neurons in the input layer are connected to all neurons in the output layer with the BF traits.

Max Pooling, Spatial Dropout, Dropout and Batch Normalization layers are added to prevent overfitting and control the number of parameters in the network. Note that a bidirectional LSTM can incorporate hidden states for both past and future information, and in cases where all the text is pre-specified, like for us, it enables better predictive performance. In turn, Batch Normalization transforms and normalizes the output of the network layer to stabilize and accelerate the training process.

### 3.8 Model Optimization

To optimize the model, we use binary cross-entropy loss, since the output for each personality trait is either 0 or 1. Cross-entropy loss measures the performance of a model that outputs probabilities between 0 and 1. The loss increases when the prediction diverges from the actual label, so the goal of the network is to learn weights that minimize the loss. We use the Adam optimizer [69] based on the stochastic gradient descent algorithm with adaptive learning rate. Adam uses two parameters in conjunction with the first and second moments of the gradient to increase speed of learning.

Similar to Majumder et al. [24], we use a learning rate of 0.001, clipping norm of 0.25 (this is the maximum absolute gradient value to prevent large spikes or updates of the function), and two beta parameters: $\text{Beta}_1$ of 0.70 (a parameter that controls the first moment of the gradient in the Adam optimizer) and $\text{Beta}_2$ of 0.99 (a parameter that controls the second moment of the gradient in the Adam optimizer).

These parameters lead the model to convergence with early stopping, meaning that the validation accuracy stopped improving. Epoch means a full model update run over the training data [70]. We took the results at the 10th epoch, since they did not improve from the 10th to the 15th epoch. Cross-validation was done with 80-10-10 split (train, development, test) and early stopping was done using the development dataset.

# 4 Results

## 4.1 Technical Performance

We use the F1 macro score for evaluating our model, as this metric considers both the precision (i.e., percentage of correct positive results) and recall (i.e., percentage of samples that are correctly identified as positive). The F1 score is the harmonic mean of precision and recall and reaches the best value of 1 with perfect precision and recall, and the worst at 0. This score was also available for the baseline are comparing with, hence allowing us to use it to compare fairly.

Results (see Table 2) show that our model trained combining the three datasets (i.e., Model 2) provides better scores than the baseline model for three personality traits: EXT (an increase of 26.1% in F1 score), OPE (18.1% increase), and AGR (24.1% increase). In contrast, our model loses to the baseline model in prediction of two traits: CON (a slight decrease of −1.8%) and NEU (a large decrease of −42.3%). The results indicate that our classifier achieves a good performance relative to the baseline (an increase of 4.84% in F1 scores on average). Yet, the mixed performance indicates the difficulty of correctly predicting all personality traits. The fact that Model 2 (trained with data from more sources) outperforms Model 1 indicates that using more sources increases the signal for the NN and, therefore, improves the results.

We also predicted each dataset separately to examine how well Model 2 performs by data source (see Table 3). In this, we ensure no data leakage takes place (i.e., separating training and prediction instances) by doing cross validation splits and training a model on the rest, then predicting the isolated fold.

The results in Table 3 show that EXT and OPE receive the highest scores on the YouTube dataset, suggesting that the network find the clearest signal for these traits in that social network. In turn, CON, AGR, and NEU are most easily detected in the essays dataset. Based on the results, personality traits are the hardest to infer from the Facebook status updates. Overall, the results support the application of the model on our DDPs, as their comments originate from YT and the model performs relatively best for YT.

**Table 2.** F1 scores for each BF trait using the essays dataset. Highest values bolded.

| | EXT | OPE | CON | AGR | NEU |
|---|---|---|---|---|---|
| Baseline [9] | 0.525 | 0.553 | **0.553** | 0.486 | **0.575** |
| Model 1* | 0.541 | 0.529 | 0.538 | 0.553 | 0.484 |
| Model 2** | **0.662** | **0.653** | 0.543 | **0.603** | 0.332 |

*trained with essays, **trained with essays + MPD + YouTube

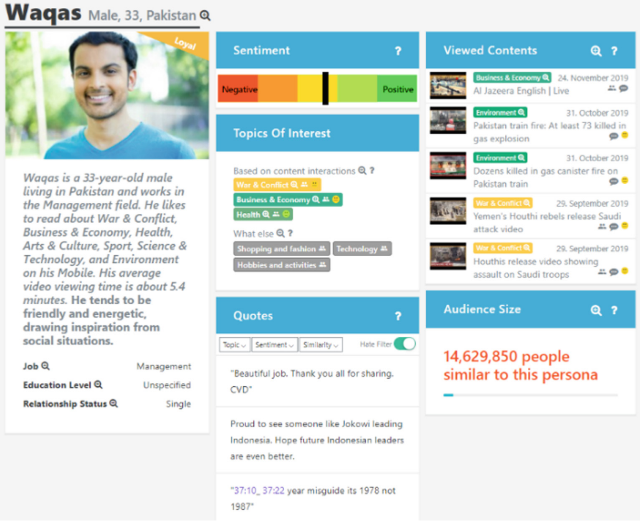
**Table 3.** The F1 scores of Model 2 (trained with essays + MPD + YouTube data) on each dataset. Cross-validation with 10 folds and 80-10-10 split was applied. Highest values bolded.

| | EXT | OPE | CON | AGR | NEU |
|---|---|---|---|---|---|
| Essays | 0.662 | 0.653 | **0.543** | **0.603** | 0.332 |
| MPD | 0.681 | 0.576 | 0.406 | 0.559 | 0.357 |
| YT | **0.719** | **0.686** | 0.485 | 0.444 | **0.403** |

### 4.2 Personality Traits of Personas

As the comments from the same user are scarce (the dataset has typically only one comment per User ID), we predict the personas' personality traits from the aggregated comments of the users corresponding to a given persona. Thus, we group the collected comments (see Sect. 3.2) by persona and predict the personality traits of each grouped collection using Model 2 (e.g., "Persona 15" contains the combined comment texts of Persona 15). This yields a score for each personality trait of each persona, which can be used for enriching persona profiles with personality traits (see Fig. 3).

The results (Fig. 4) show that five personas (out of 15) score lower than average on EXT (P1-2, P4, P7-8), while six score higher (P5, P6, P12-15). Scoring lower on both EXT and NEU seems to be associated, as four personas out of the five that score low on EXT also score lower than average on NEU (P1-2, P4, P8). The personas that score higher than average for EXT and NEU tend to score lower than average on OPE and CON (e.g., P1-2). However, there are deviations from this, such as P7 that scores close to average on every trait. Four personas (P6, P9, P10, and P12) tend to score low on AGR. The observed variation indicates that the APD method applied produces variation in the detected personality traits among the personas, implying that the personas do differ by personality, at least to some degree. Future analyses are needed to understand comprehensively where these differences originate from (e.g., analyzing why the APD method gives these scores on the dataset).



Waqas is a 33-year-old male living in Pakistan and works in the Management field. He likes to read about War & Conflict, Business & Economy, Health, Arts & Culture, Sport, Science & Technology, and Environment on his Mobile. His average video viewing time is about 5.4 minutes. **He tends to be friendly and energetic, drawing inspiration from social situations.**

| | |
|---|---|
| **Job** | Management |
| **Education Level** | Unspecified |
| **Relationship Status** | Single |

"Beautiful job. Thank you all for sharing. CVD"

Proud to see someone like Jokowi leading Indonesia. Hope future Indonesian leaders are even better.

"37:10_ 37:22 year misguide its 1978 not 1987"

Business & Economy 24. November 2019
Al Jazeera English | Live

Environment 31. October 2019
Pakistan train fire: At least 73 killed in gas explosion

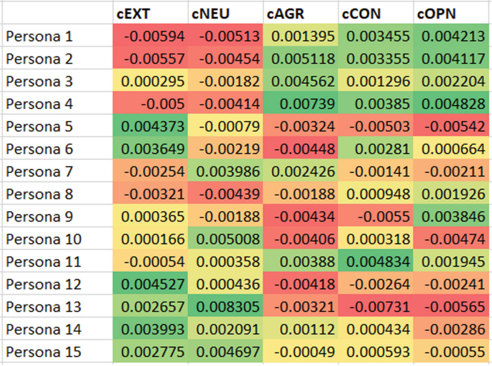
Environment 31. October 2019
Dozens killed in gas canister fire on Pakistan train

War & Conflict 29. September 2019
Yemen's Houthi rebels release Saudi attack video

War & Conflict 29. September 2019
Houthis release video showing assault on Saudi troops

14,629,850 people similar to this persona

**Fig. 3.** An example of incorporating textual description of personality in data-driven personas. The bolded text "**He tends to be friendly and energetic, drawing inspiration from social situations.**" corresponds to the general description of high extroversion [47], reflected in the personas' comments based on automatic personality detection. The generic personality trait descriptions can be automatically inserted based on the personality scores obtained.



| | cEXT | cNEU | cAGR | cCON | cOPN |
|---|---|---|---|---|---|
| Persona 1 | -0.00594 | -0.00513 | 0.001395 | 0.003455 | 0.004213 |
| Persona 2 | -0.00557 | -0.00454 | 0.005118 | 0.003355 | 0.004117 |
| Persona 3 | 0.000295 | -0.00182 | 0.004562 | 0.001296 | 0.002204 |
| Persona 4 | -0.005 | -0.00414 | 0.00739 | 0.00385 | 0.004828 |
| Persona 5 | 0.004373 | -0.00079 | -0.00324 | -0.00503 | -0.00542 |
| Persona 6 | 0.003649 | -0.00219 | -0.00448 | 0.00281 | 0.000664 |
| Persona 7 | -0.00254 | 0.003986 | 0.002426 | -0.00141 | -0.00211 |
| Persona 8 | -0.00321 | -0.00439 | -0.00188 | 0.000948 | 0.001926 |
| Persona 9 | 0.000365 | -0.00188 | -0.00434 | -0.0055 | 0.003846 |
| Persona 10 | 0.000166 | 0.005008 | -0.00406 | 0.000318 | -0.00474 |
| Persona 11 | -0.00054 | 0.000358 | 0.00388 | 0.004834 | 0.001945 |
| Persona 12 | 0.004527 | 0.000436 | -0.00418 | -0.00264 | -0.00241 |
| Persona 13 | 0.002657 | 0.008305 | -0.00321 | -0.00731 | -0.00565 |
| Persona 14 | 0.003993 | 0.002091 | 0.00112 | 0.000434 | -0.00286 |
| Persona 15 | 0.002775 | 0.004697 | -0.00049 | 0.000593 | -0.00055 |

**Fig. 4.** Personality scores (probability of a persona's aggregated comments reflecting a BF trait) of personas based on their aggregated social media comments. The cells show absolute differences from the mean score of the personality trait. Color coding indicates the size of the difference, with positive values in green and negative in red. (Color figure online)

# 5 Discussion

## 5.1 Contribution to Persona Research

Theoretically, the variability in the personality traits among the personas provides an interesting outlook of the "collective personality" of groups interested in the same online content. Perhaps this collective personality could be termed as *persona personality*, i.e., a grouped understanding of personality traits of a user segment. Traditionally, personality traits have been associated with *individuals*, not groups, in social psychology. Thus, the fact that persona can thus portray collective patterns of personality among different user segments is an interesting notion in the cross-section of HCI and social psychology. This notion could be empirically investigated by, e.g., analyzing the relationship between the persona's topical interests and personality. Perhaps certain groups are more drawn to some online content, and the topic of the content thus becomes a proxy measure for users' personality.

Conceptually, there are two challenges pertaining to the amalgamation of APD and DDPs. The first challenge is that a persona, by definition, consists of *several* individuals that are portrayed as *one* persona. However, individuals *within* that group may vary by personality traits. Thus, is it possible to construct "average" personality traits for a persona? How meaningful would this construction be?

Our exploratory results (Fig. 4) indicate that the APD methods can produce variability among the groups. For this method to work, it is required that the social media posts made by the users reveal observable trends toward a certain personality trait. It is also possible that, with other datasets, "averaging" the posts of many individuals would cancel out the individual personality differences.

Regarding the meaningfulness, this can be addressed by presenting the personality traits in "plain language", as demonstrated in Fig. 3.

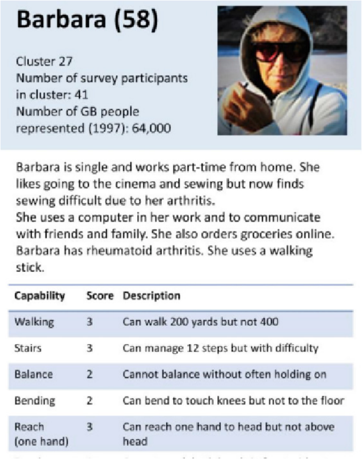
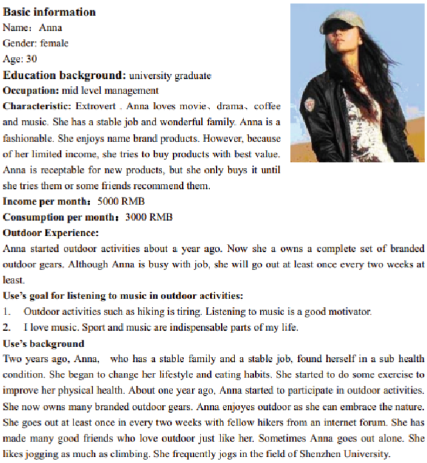
The second challenge is the meaningfulness of the persona personality prediction altogether. In other words, does the endeavor have practical value? Theoretically, psychological information in persona profiles can enhance the understanding of persona users about the persona, and provide utility in various design/development tasks, as well as advertising purposes [19]. Thus, in terms of the possibility of inferring the personality traits, the *opportunity* of predicting a persona's personality is highly prominent. Yet, it is unclear if the personality traits are indeed needed or wanted by end users of personas. Rather, empirical results of these ideas are missing. This implies that user studies on how persona users engage with DDPs enriched with personality traits are direly needed.

Personality traits are not only potentially impactful information for persona profiles (as argued above) but analyzing the persona's behavior by personality traits opens a multitude of related research avenues. For example, can we find substantial personality differences between the personas? How do the personality traits of a persona correlate with the persona's online content consumption patterns? Addressing these questions would shed light on how personas are engaging (i.e., watching and commenting) with different online content and if this behavior differs by the personas' personality traits – in other words, enhancing user understanding.

## 5.2 Design Implications

There are at least two approaches that could be implemented for showing the personality traits in DDPs: (1) quantitatively inspired and (2) qualitatively inspired. In the former, the personality traits are shown as "scores" (see inspiration from previous research in Table 4A), while in the latter they are written in a form of a narrative to describe the persona (Table 4B).

**Table 4.** Examples of implementing persona information

<table>
<tr><th><b>A: Implementing Persona Information Quantitatively</b> [71]</th><th><b>B: Implementing Persona Information as a Narrative</b> [72]</th></tr>
<tr><td>

<b>Barbara (58)</b><br>
Cluster 27<br>
Number of survey participants in cluster: 41<br>
Number of GB people represented (1997): 64,000<br>
Barbara is single and works part-time from home. She likes going to the cinema and sewing but now finds sewing difficult due to her arthritis.<br>
She uses a computer in her work and to communicate with friends and family. She also orders groceries online. Barbara has rheumatoid arthritis. She uses a walking stick.
<table>
<tr><th>Capability</th><th>Score</th><th>Description</th></tr>
<tr><td>Walking</td><td>3</td><td>Can walk 200 yards but not 400</td></tr>
<tr><td>Stairs</td><td>3</td><td>Can manage 12 steps but with difficulty</td></tr>
<tr><td>Balance</td><td>2</td><td>Cannot balance without often holding on</td></tr>
<tr><td>Bending</td><td>2</td><td>Can bend to touch knees but not to the floor</td></tr>
<tr><td>Reach (one hand)</td><td>3</td><td>Can reach one hand to head but not above head</td></tr>
</table>
</td><td>

<b>Basic information</b><br>
Name: Anna<br>
Gender: female<br>
Age: 30<br>
<b>Education background:</b> university graduate<br>
<b>Occupation:</b> mid level management<br>
<b>Characteristic:</b> Extrovert . Anna loves movie、drama、coffee and music. She has a stable job and wonderful family. Anna is a fashionable. She enjoys name brand products. However, because of her limited income, she tries to buy products with best value. Anna is receptable for new products, but she only buys it until she tries them or some friends recommend them.<br>
<b>Income per month:</b> 5000 RMB<br>
<b>Consumption per month:</b> 3000 RMB<br>
<b>Outdoor Experience:</b><br>
Anna started outdoor activities about a year ago. Now she a owns a complete set of branded outdoor gears. Although Anna is busy with job, she will go out at least once every two weeks at least.<br>
<b>Use's goal for listening to music in outdoor activities:</b><br>
1. Outdoor activities such as hiking is tiring. Listening to music is a good motivator.<br>
2. I love music. Sport and music are indispensable parts of my life.<br>
<b>Use's background</b><br>
Two years ago, Anna, who has a stable family and a stable job, found herself in a sub health condition. She began to change her lifestyle and eating habits. She started to do some exercise to improve her physical health. About one year ago, Anna started to participate in outdoor activities. She now owns many branded outdoor gears. Anna enjoyes outdoor as she can embrace the nature. She goes out at least once in every two weeks with fellow hikers from an internet forum. She has made many good friends who love outdoor just like her. Sometimes Anna goes out alone. She likes jogging as much as climbing. She frequently jogs in the field of Shenzhen University.
</td></tr>
</table>

These ideas follow the division between different persona types, with the quantitative approach [71] resulting in a chart-like presentation of the details, with "scores" directly representing the quantitatively inferred information. In contrast, the qualitative approach [72] results in a persona layout enriched with more narrative-like, in-depth descriptions. The first option is supported by the fact that the scores of the personality traits are readily available following the application of the neural network.

The second option, in turn, is supported by the previous research that tends to infuse personality and psychographic information into a narrative format [11, 12, 50–52]. Writing the personality traits open might also be better for empathetic understanding of the persona – consider "Mary is an extrovert, enjoying discussions with new acquaintances" vs. "Extroversion$_{\text{MARY}}$ = 0.67". However, the disadvantage of the narratives is that their creation represents an additional step that seemingly requires manual effort. This would take us further from the overarching goal of fully automatic persona creation [73]. Yet, there are implementations of dynamic text templates [74] that could possibly be used for combining the narrative format and automation.

Previous research analyzing the role of text vs. numbers in persona profiles shows that this decision does not critically influence user perceptions of personas [75]. Nonetheless, more research is needed for testing which method of showing personality traits in personas, or combinations thereof, would provide the best option for optimal persona user experience. Moreover, the general question of "how does showing personality traits influence persona users' perceptions and/or actions?" requires an answer, to actually discover the impact of personality traits in personas. Thus, there are several open research questions that require empirical user studies.

### 5.3 Future Work on Improving Persona Personality Detection

The model we have used is quite simple relative to the state-of-the-art approaches in deep learning (see, e.g., [33, 44, 66]). We have assumed a relatively simple model because such a model is easily trainable without designing features and is computationally lightweight. Our results indicate that given sufficient data, a NN can be trained to predict personalities without hand-designed features being provided. At the same time, if hand-designed features are considered for a model like ours, it may be possible to increase accuracy or train a smaller model with less data and achieve similar results. Overall, while for the purpose of this research (i.e., demonstrating the APD for DDPs), the NN's performance is seen as adequate, results should be revisited as new algorithms and feature presentations for APD become available.

### 5.4 Ethical Considerations

APG preserves the privacy of individual users when generating the personas [28], because the information is collected as aggregated user statistics. For example, we can see information such as women aged 25–34, from New York, have *in aggregate* viewed Video X in total of Y times. This information, even though used for persona generation, does not violate the privacy of individual users, as it contains no personally identifiable information. The comments do contain User ID, but this ID tends to be in the form of a pseudonym, and the comments are further anonymized by removing the User IDs from the generated personas [21].

Regarding the ability of DDPs to represent minority groups (so-called "fringe personas" [71]), previous research has shown that the method applied here accurately replicates demographic characteristics of the data [76]. Thus, the choice of dataset dictates the characteristic of the output personas. A selection of an underrepresented subset of data, for example, would yield personas only from underprivileged subjects in the data. In this study, we generated personas using the whole dataset, as we were interested in average or typical users rather than minority subsets. However, future studies could use DDP techniques to generate "minority personas".

The use of personas enhanced with psychological traits comes with added responsibility. There may be a risk of manipulation. However, the story is not black-and-white, as there is a strong argument that people are not as easily gullible based on their psychologic profiles as is commonly presumed in the popular press [77]. As with most tools and applications of HCI, personas can be used "for good" and "for bad" – the final responsibility falls for the person wielding the tool.

## 6 Conclusion

Enriching personas with personality traits can enhance decision makers' understanding about users. Thus far, personality traits in personas have been based on manual data analysis. To provide a more efficient solution, we demonstrate how user-generated social media texts can be used to automatically assign Big Five personality traits to data-driven personas using a neural network classifier. The classifier trained on pre-existing datasets and achieved a good technical performance. In addition, the results show variation among personas in the detected personality traits.

**Acknowledgments.** We thank Dr. Lene Nielsen for discussions and inspiration on how to potentially display the automatically inferred personality traits in data-driven personas. We thank Al Jazeera Media Network for sharing the data that made this research possible.

## References

1. Cooper, A.: The Inmates Are Running the Asylum: Why High Tech Products Drive Us Crazy and How to Restore the Sanity. Sams - Pearson Education, Indianapolis (1999)
2. Pruitt, J., Grudin, J.: Personas: practice and theory. In: Proceedings of the 2003 Conference on Designing for User Experiences, pp. 1–15. ACM, New York (2003). https://doi.org/10.1145/997078.997089
3. Nielsen, L.: Personas - User Focused Design. Springer, London (2013)
4. Salminen, J., Jansen, B.J., An, J., Kwak, H., Jung, S.: Are personas done? Evaluating their usefulness in the age of digital analytics. Persona Stud. **4**, 47–65 (2018). https://doi.org/10.21153/psj2018vol4no2art737
5. LeRouge, C., Ma, J., Sneha, S., Tolle, K.: User profiles and personas in the design and development of consumer health technologies. Int. J. Med. Inform. **82**, e251–e268 (2013). https://doi.org/10.1016/j.ijmedinf.2011.03.006
6. Pruitt, J., Adlin, T.: The Persona Lifecycle: Keeping People in Mind Throughout Product Design. Morgan Kaufmann, Boston (2006)
7. Nielsen, L., Storgaard Hansen, K.: Personas is applicable: a study on the use of personas in Denmark. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 1665–1674. ACM (2014)
8. Salminen, J., Jung, S., An, J., Kwak, H., Nielsen, L., Jansen, B.J.: Confusion and information triggered by photos in persona profiles. Int. J. Hum.-Comput. Stud. **129**, 1–14 (2019). https://doi.org/10.1016/j.ijhcs.2019.03.005
9. Nielsen, L.: Personas - User Focused Design. Springer, New York (2019). https://doi.org/10.1007/978-1-4471-4084-9
10. Nielsen, L., Hansen, K.S., Stage, J., Billestrup, J.: A template for design personas: analysis of 47 persona descriptions from Danish industries and organizations. Int. J. Sociotechnol. Knowl. Dev. **7**, 45–61 (2015). https://doi.org/10.4018/ijskd.2015010104

11. Anvari, F., Richards, D., Hitchens, M., Babar, M.A.: Effectiveness of persona with personality traits on conceptual design. In: Proceedings of the 37th International Conference on Software Engineering, vol. 2, Piscataway, NJ, USA, pp. 263–272. IEEE Press (2015)
12. Anvari, F., Richards, D., Hitchens, M., Babar, M.A., Tran, H.M.T., Busch, P.: An empirical investigation of the influence of persona with personality traits on conceptual design. J. Syst. Softw. **134**, 324–339 (2017). https://doi.org/10.1016/j.jss.2017.09.020
13. Gosling, S.D., Rentfrow, P.J., Swann, W.B.: A very brief measure of the Big-Five personality domains. J. Res. Pers. **37**, 504–528 (2003)
14. Ardelt, M.: Still stable after all these years? Personality stability theory revisited. Soc. Psychol. Q. 392–405 (2000)
15. Leong, L.-Y., Jaafar, N.I., Sulaiman, A.: Understanding impulse purchase in Facebook commerce: does Big Five matter? Internet Res. **27**, 786–818 (2017)
16. Hoffman, L.R.: Homogeneity of member personality and its effect on group problem-solving. J. Abnorm. Soc. Psychol. **58**, 27 (1959)
17. Barrick, M.R., Mount, M.K.: The Big Five personality dimensions and job performance: a meta-analysis. Personnel Psychol. **44**, 1–26 (1991)
18. Schoen, H., Schumann, S.: Personality traits, partisan attitudes, and voting behavior. Evidence from Germany. Polit. Psychol. **28**, 471–498 (2007)
19. Haugtvedt, C.P., Petty, R.E., Cacioppo, J.T.: Need for cognition and advertising: understanding the role of personality variables in consumer behavior. J. Consum. Psychol. **1**, 239–260 (1992)
20. Salminen, J., Guan, K., Jung, S.-G., Chowdhury, S.A., Jansen, B.J.: A literature review of quantitative persona creation. In: Proceedings of the ACM Conference of Human Factors in Computing Systems (CHI 2020), Honolulu, Hawaii, USA. ACM (2020)
21. An, J., Kwak, H., Salminen, J., Jung, S., Jansen, B.J.: Imaginary people representing real numbers: generating personas from online social media data. ACM Trans. Web (TWEB) **12**, 1–26 (2018)
22. Alam, F., Riccardi, G.: Predicting personality traits using multimodal information. In: Proceedings of the 2014 ACM Multi Media on Workshop on Computational Personality Recognition, pp. 15–18. ACM (2014)
23. Bleidorn, W., Hopwood, C.J.: Using machine learning to advance personality assessment and theory. Pers. Soc. Psychol. Rev. 1088868318772990 (2018)
24. Majumder, N., Poria, S., Gelbukh, A., Cambria, E.: Deep learning-based document modeling for personality detection from text. IEEE Intell. Syst. **32**, 74–79 (2017). https://doi.org/10.1109/MIS.2017.23
25. Kim, J.H., Kim, Y.: Instagram user characteristics and the color of their photos: colorfulness, color diversity, and color harmony. Inf. Process. Manag. **56**, 1494–1505 (2019). https://doi.org/10.1016/j.ipm.2018.10.018
26. Carducci, G., Rizzo, G., Monti, D., Palumbo, E., Morisio, M.: TwitPersonality: computing personality traits from tweets using word embeddings and supervised learning. Information **9**, 127 (2018)
27. An, J., Kwak, H., Jung, S., Salminen, J., Jansen, B.J.: Customer segmentation using online platforms: isolating behavioral and demographic segments for persona creation via aggregated user data. Soc. Netw. Anal. Min. **8** (2018). https://doi.org/10.1007/s13278-018-0531-0
28. Salminen, J., et al.: From 2,772 segments to five personas: summarizing a diverse online audience by generating culturally adapted personas. First Monday **23** (2018). https://doi.org/10.5210/fm.v23i6.8415
29. Cambria, E.: Affective computing and sentiment analysis. IEEE Intell. Syst. **31**, 102–107 (2016)

30. Pennebaker, J.W., King, L.A.: Linguistic styles: language use as an individual difference. J. Pers. Soc. Psychol. **77**, 1296 (1999)
31. Tausczik, Y.R., Pennebaker, J.W.: The psychological meaning of words: LIWC and computerized text analysis methods. J. Lang. Soc. Psychol. **29**, 24–54 (2010). https://doi.org/10.1177/0261927X09351676
32. Tskhay, K.O., Rule, N.O.: Perceptions of personality in text-based media and OSN: a meta-analysis. J. Res. Personal. **49**, 25–30 (2014)
33. Xue, D., et al.: Deep learning-based personality recognition from text posts of online social networks. Appl. Intell. **48**, 4232–4246 (2018). https://doi.org/10.1007/s10489-018-1212-4
34. Howlader, P., Pal, K.K., Cuzzocrea, A., Kumar, S.D.: Predicting Facebook-users' personality based on status and linguistic features via flexible regression analysis techniques. In: Proceedings of the 33rd Annual ACM Symposium on Applied Computing, pp. 339–345. ACM (2018)
35. Luyckx, K., Daelemans, W.: Using syntactic features to predict author personality from text. Proc. Digital Human. **2008**, 146–149 (2008)
36. Mairesse, F., Walker, M.A., Mehl, M.R., Moore, R.K.: Using linguistic cues for the automatic recognition of personality in conversation and text. J. Artif. Intell. Res. **30**, 457–500 (2007)
37. Rammstedt, B., John, O.P.: Measuring personality in one minute or less: a 10-item short version of the Big Five inventory in English and German. J. Res. Pers. **41**, 203–212 (2007)
38. Fang, J., Wen, C., Prybutok, V.: An assessment of equivalence between paper and social media surveys: the role of social desirability and satisficing. Comput. Hum. Behav. **30**, 335–343 (2014)
39. Kozinets, R.V., Dolbec, P.-Y., Earley, A.: Netnographic analysis: understanding culture through social media data. The SAGE Handbook of Qualitative Data Analysis, pp. 262–276 (2014)
40. Plank, B., Hovy, D.: Personality traits on Twitter—or—how to get 1,500 personality tests in a week. In: Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pp. 92–98 (2015)
41. Pratama, B.Y., Sarno, R.: Personality classification based on Twitter text using Naive Bayes, KNN and SVM. In: 2015 International Conference on Data and Software Engineering (ICoDSE), pp. 170–174 (2015). https://doi.org/10.1109/ICODSE.2015.7436992
42. Sewwandi, D., Perera, K., Sandaruwan, S., Lakchani, O., Nugaliyadde, A., Thelijjagoda, S.: Linguistic features based personality recognition using social media data. In: 2017 6th National Conference on Technology and Management (NCTM), pp. 63–68. IEEE (2017)
43. Mitrou, L., Kandias, M., Stavrou, V., Gritzalis, D.: Social media profiling: a Panopticon or Omniopticon tool? In: Proceedings of the 6th Conference of the Surveillance Studies Network, Barcelona, Spain (2014)
44. Darliansyah, A., Naeem, M.A., Mirza, F., Pears, R.: SENTIPEDE: a smart system for sentiment-based personality detection from short texts. J. Univ. Comput. Sci. **25**, 1323–1352 (2019)
45. Tandera, T., Suhartono, D., Wongso, R., Prasetio, Y.L.: Personality prediction system from Facebook users. Procedia Comput. Sci. **116**, 604–611 (2017)
46. Yılmaz, T., Ergil, A., İlgen, B.: Deep learning-based document modeling for personality detection from Turkish texts. In: Arai, K., Bhatia, R., Kapoor, S. (eds.) FTC 2019. AISC, vol. 1069, pp. 729–736. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-32520-6_53
47. Agarwal, B.: Personality detection from text: a review. Int. J. Comput. Syst. **1**, 1–4 (2014)

48. Stillwell, D.J., Kosinski, M.: myPersonality project: example of successful utilization of online social networks for large-scale social research. Presented at the International Conference on Mobile Systems (MobiSys) (2012)
49. Biel, J.-I., Gatica-Perez, D., Dines, J., Tsminiaki, V.: Hi YouTube! personality impressions and verbal content in social video. https://infoscience.epfl.ch/record/196978. https://doi.org/10.1145/2522848.2522877. Accessed 07 Jan 2020
50. Jones, M., Marsden, G.: Mobile Interaction Design. Wiley (2006)
51. Negru, S., Buraga, S.: A knowledge-based approach to the user-centered design process. In: Fred, A., Dietz, Jan L.G., Liu, K., Filipe, J. (eds.) IC3K 2012. CCIS, vol. 415, pp. 165–178. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-54105-6_11
52. Pichler, R.: A template for writing great personas (2012)
53. Salminen, J., Jansen, B.J., An, J., Kwak, H., Jung, S.-G.: Automatic persona generation for online content creators: conceptual rationale and a research agenda. Personas - User Focused Design. HIS, pp. 135–160. Springer, London (2019). https://doi.org/10.1007/978-1-4471-7427-1_8
54. Anvari, F., Tran, H.M.T.: Persona ontology for user centred design professionals. In: The ICIME 4th International Conference on Information Management and Evaluation, Ho Chi Minh City, Vietnam, pp. 35–44 (2013)
55. Câmara, M., Signoretti, A., Costa, C., Soares, S.C.: Business Affective Persona (BAP): a methodology to create personas to enhance customer relationship with trust and empathy. Revista Turismo Desenvolvimento, pp. 85–97 (2018)
56. Salminen, J., Jung, S., Chowdhury, S.A., Sengün, S., Jansen, B.J.: Personas and analytics: a comparative user study of efficiency and effectiveness for a user identification task. In: Proceedings of the ACM Conference of Human Factors in Computing Systems (CHI 2020), Honolulu, Hawaii, USA. ACM (2020). https://doi.org/10.1145/3313831.3376770
57. Jung, S., Salminen, J., Kwak, H., An, J., Jansen, B.J.: Automatic Persona Generation (APG): a rationale and demonstration. In: Proceedings of the 2018 Conference on Human Information Interaction & Retrieval, New Brunswick, NJ, USA, pp. 321–324. ACM (2018). https://doi.org/10.1145/3176349.3176893
58. Jung, S., Salminen, J., An, J., Kwak, H., Jansen, B.J.: Automatically conceptualizing social media analytics data via personas. Presented at the International AAAI Conference on Web and Social Media (ICWSM 2018), San Francisco, California, USA, 25 June 2018 (2018)
59. Jung, S., Salminen, J., Jansen, B.J.: Personas Changing over time: analyzing variations of data-driven personas during a two-year period. In: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, New York, NY, USA, pp. LBW2714:1–LBW2714:6. ACM (2019). https://doi.org/10.1145/3290607.3312955
60. Salminen, J., et al.: Generating cultural personas from social data: a perspective of middle eastern users. In: Proceedings of the Fourth International Symposium on Social Networks Analysis, Management and Security (SNAMS-2017), Prague, Czech Republic. IEEE (2017). https://doi.org/10.1109/FiCloudW.2017.97
61. Lee, D.D., Seung, S.H.: Learning the parts of objects by non-negative matrix factorization. Nature **401**, 788–791 (1999)
62. Norman, W.T.: Toward an adequate taxonomy of personality attributes: replicated factor structure in peer nomination personality ratings. J. Abnorm. Soc. Psychol. **66**, 574 (1963)
63. Ashton, M.C., Lee, K.: How well do Big Five measures capture HEXACO scale variance? J. Pers. Assess. **101**, 567–573 (2019)
64. Goldberg, L.R.: The development of markers for the Big-Five factor structure. Psychol. Assess. **4**, 26 (1992)

65. Yin, C., Zhang, X., Liu, L.: Reposting negative information on microblogs: do personality traits matter? Inf. Process. Manag. **57**, 102106 (2020). https://doi.org/10.1016/j.ipm.2019.102106
66. Sun, X., Liu, B., Cao, J., Luo, J., Shen, X.: Who am I? Personality detection based on deep learning for texts. In: 2018 IEEE International Conference on Communications (ICC), pp. 1–6 (2018). https://doi.org/10.1109/ICC.2018.8422105
67. Cawley, G.C., Talbot, N.L.: Efficient leave-one-out cross-validation of kernel fisher discriminant classifiers. Pattern Recogn. **36**, 2585–2592 (2003)
68. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: Proceedings of the 31st International Conference on Machine Learning (ICML 2014), pp. 1188–1196 (2014)
69. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
70. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**, 436 (2015)
71. Goodman-Deane, J., Waller, S., Demin, D., González-de-Heredia, A., Bradley, M., Clarkson, J.P.: Evaluating inclusivity using quantitative personas. Presented at the Design Research Society Conference, 28 June 2018 (2018). https://doi.org/10.21606/drs.2018.400
72. Tu, N., et al.: Combine qualitative and quantitative methods to create persona. In: 2010 3rd International Conference on Information Management, Innovation Management and Industrial Engineering, pp. 597–603 (2010). https://doi.org/10.1109/ICIII.2010.463
73. Salminen, J., Jung, S.G., Jansen, B.J.: The future of data-driven personas: a marriage of online analytics numbers and human attributes. In: ICEIS 2019 - Proceedings of the 21st International Conference on Enterprise Information Systems, Heraklion, Greece, pp. 596–603. SciTePress (2019)
74. Jung, S., An, J., Kwak, H., Ahmad, M., Nielsen, L., Jansen, B.J.: Persona generation from aggregated social media data. In: Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems, Denver, Colorado, USA, pp. 1748–1755. ACM (2017)
75. Salminen, J., Liu, Y.-H., Sengun, S., Santos, J.M., Jung, S.-G., Jansen, B.J.: The effect of numerical and textual information on visual engagement and perceptions of ai-driven persona interfaces. In: Proceedings of the ACM Intelligent User Interfaces (IUI 2020), Cagliary, Italy. ACM (2020)
76. Salminen, J., Jung, S.-G., Jansen, B.J.: Detecting demographic bias in automatically generated personas. In: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, pp. LBW0122:1–LBW0122:6. ACM, New York (2019). https://doi.org/10.1145/3290607.3313034
77. Phillips, M.J.: Ethics and Manipulation in Advertising: Answering a Flawed Indictment. Greenwood Publishing Group (1997)