Semi-Supervised Psychometric Scoring of Document Collections

Burak Suyunu
Computer Engineering
Bogazici University
Istanbul, Turkey
burak.suyunu@boun.edu.tr

Gonul Ayci
Computer Engineering
Bogazici University
Istanbul, Turkey
gonul.ayci@boun.edu.tr

Mine Öğretir
Computer Engineering
Bogazici University
Istanbul, Turkey
mine.ogretir@boun.edu.tr

Ali Taylan Cemgil
Computer Engineering
Bogazici University
Istanbul, Turkey
taylan.cemgil@boun.edu.tr

Suzan Uskudarli
Computer Engineering
Bogazici University
Istanbul, Turkey
suzan.uskudarli@boun.edu.tr

Hamza Zeytinoglu
CLARIFIX LTD
London, United Kingdom
hzeytin@gmail.com

Bulent Ozel
Department of Banking and Finance
University of Zurich
Zurich, Switzerland
bulent.ozel@gmail.com

Arman Boyacı
Bogazici University
Istanbul, Turkey
armanboyaci@gmail.com

***Abstract*—We describe a generic computational approach that can be used in developing methods for psychometric profiling. Our approach is based on semi-supervised analysis of document collections using topic modeling. The method depends on a supervisor providing a set of seed documents, grouped by abstract themes, such as Schwartz values or personality traits; and possibly a separate background document corpus. Instead of casting the problem into a standard classification framework, we interpret the group labels as a guide for finding distinguishing features. During training, we train each group of documents associated with a theme separately by using nonnegative matrix factorization to obtain theme specific topic distributions. In the analysis, we decompose a new document using the model learned during training to arrive at the theme scores. We demonstrate our approach on two psychometric profiling theories (Schwartz and Big Five) and evaluate our Schwartz scores with leave-one-out cross-validation method and compare Big Five scores to independent surveys, which are much more costly to carry out.**

***Keywords*—non-negative matrix factorization, semi-supervised learning, Schwartz theory of basic human values, big five personality traits, psychometric profiling, personality recognition**

I. Introduction

The importance of psychometrics in understanding and predicting behaviors of individuals and groups is now well recognized. From workplaces to business relations, from politics to culture and social organizations applications are studied increasingly. As such, an important trend in current data analytics applications is the automated, or semi-automated generation of psychometric profiles of individuals using unstructured textual data. This trend is in stark contrast to traditional approaches based on surveys and questionnaires, which significantly limit the amount of data that can be collected about a particular individual. Data-driven automated approaches also enable the dynamic updating of profiles by learning from streaming data, which is significantly more challenging and costly to achieve through surveys.

As Cambria, Poria, Gelbukh and Thelwall state in their article [1], sentiment analysis in natural language processing (NLP) has many components. They organized NLP problems for human level sentiment analysis in three layers, which are syntactics, semantics, and pragmatics. The first step in pragmatics layer is personality recognition. The characteristics of an individual's emotions, behaviors, cognitions, and thought patterns are defined as personality recognition, and psychometrics are used for identifying personality types.

In this paper, we propose a pragmatic approach for assigning psychometric scores to documents. Such scores can be utilized in the psychometric profiling of individuals via the documents they are associated with (i.e. author, reviewer or propagator) where the aggregated scores of such documents serve as a proxy for an individual's psychometric profile. Here, psychometric profiles are elicited in an implicit fashion instead of explicit approaches that rely on surveys and questionnaires where profiles are created based on self-reported characteristics of individuals.

The proposed document scoring approach is based on the Schwartz Value Theory (Schwartz, 1992, 2006a) that adopts a conception of values that specifies six main features implicit in the writings of many theorists (Allport, 1961; Feather, 1995; Kluckhohn, 1951; Morris, 1956; Rokeach 1973) [2], namely that values: are beliefs linked inextricably to affect; refer to desirable goals that motivate action; transcend specific actions and situations; serve as standards or criteria; are ordered by importance relative to one another; and guide action in conjunction with other values. This theory categorizes ten basic human values (BHVs) in five higher order groups [3]:

- **Openness to change**: *Self-Direction* and *Stimulation*.
- **Self-enhancement**: *Achievement* and *Power*.
- **Hedonism**: *Hedonism* (considered to be shared among *Openness to change* and *Self enhancement*).
- **Conservation**: *Security*, *Conformity*, and *Tradition*.
- **Self-transcendence**: *Benevolence* and *Universalism*.

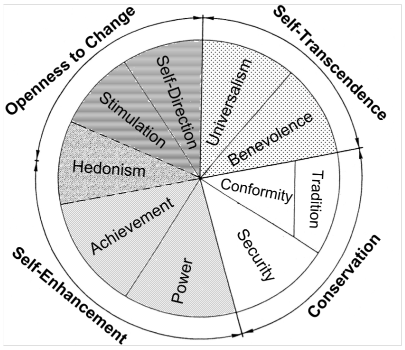


Fig. 1. The relationships among the BHVs and higher order groups in Schwartz's Theory of Basic Human Values.

Our aim is to estimate psychometric scores of documents to enable a soft categorization of individuals from a potentially large and unstructured collection such as online news articles, speeches, and blogs.

The driver of our studies in BHV centers around bridging the gaps between values and their relevant behaviors as a model for enhancing cooperation in social networks [4]. The long-term motivation is to address real-world issues where a) similarity and complementarity can be identified to enhance cooperation, b) individuals can receive feedback about their values base and how to communicate with others that come from distinct bases, and c) trends on how the values of individuals and groups of individuals (i.e. working in an organization such as media, government, commercial company) change and are communicated can be identified as a function of time.

This paper presents three main tasks of estimating psychometric scores: (i) the creation of a corpus from Wikipedia articles; (ii) a semi-supervised nonnegative matrix factorization model to construct a term-topic matrix; and (iii) the computation of psychometric scores for documents.

The rest of this paper is organized as follows: Section II provides an overview of related work, Section III describes the creation of the the corpus, Section IV introduces a semi-supervised nonnegative matrix factorization model to learn psychometric theme specific topic distributions, Section V defines our scoring metrics, Section VI presents an evaluation of the proposed approach, and Section VII presents future directions and concluding remarks.

II. Related Work

The correlation between word use and personality traits has been studied for decades [5], [6]. While previous studies were performed under laboratory conditions, with the extensive use of social media and other online platforms, recent studies are performed under larger and more natural, real-world conditions [7], [8].

Correlations between word use and personality traits allow the prediction of individual characteristics using machine learning methods. The ground truth for these tasks are obtained through questionnaires, such as *Portrait Values Questionnaire* (PVQ) [9] or *The Revised NEO Personality Inventory* (NEO-PI-R) [10]. PVQ measures BHVs according to Schwartz's theory of Basic Human Values. NEO-PI-R measures the *Big Five personality traits*, namely *Extraversion*, *Agreeableness*, *Conscientiousness*, *Openness*, and *Neuroticism*. Big Five Personality traits, generally referred to as the Five Factor Model (FFM), is a well-studied approach to representing human personality traits.

Early work approached the profiling task as a classification problem [11], [12]. Argamon, Dhawle, Koppel, and Pennebaker used Support Vector Machines on four different sets of lexical features [11] to measure the *Extraversion* and *Neuroticism* traits of FFM. Oberlander and Nowson used binary and multiple classifications on different sets of n-gram features to predict four of the FFM traits [12]. In 2013, an author profiling task was introduced in the *Workshop on Computational Personality Recognition*, where the *Essays and MyPersonality* datasets [13] were released. The best performers used tri-grams as features and ensembled trained support vector machine classifiers [13]. In 2014, a labeled dataset of Youtube video transcripts was released to predict traits [14]. The best performers, Alam and Riccardi, extracted psycholinguistic, emotional and part-of-speech features and used audio-visual features. In 2015, in the PAN workshop of CLEF, a challenge was defined to predict the age, gender, and personality traits (according to FFM) of the author of a set of tweets [15]. The corpus used in this task consisted of tweets of Twitter users. These users also took online FFM personality tests, which served as ground truths. Most task participants approached this challenge as a regression problem.

Recently, Majumder, Poria, Gelbukh, and Cambria used convolutional neural networks (CNN) to predict personality traits for a labeled corpus of essays, where a binary classifier network was trained for each trait [16]. For the vectorization of the text, Xue et al. used a recurrent CNN with attention mechanism [17]. For the deep semantic feature extraction, they trained a hierarchical deep neural network and a variant of the inception structure. To predict FFM traits, they combined the extracted deep features with statistical linguistic features in gradient boosting regression [18]. Carducci, Rizzo, Monti, Palumbo, and Morisio trained word vector representations as embeddings after tokenization and used a supervised learning classifier [19].

Most of the author profiling tasks for the psychological traits are based on FFM with a focus on extracting semantic, lexical and psycholinguistic features regarding language use via supervised learning methods.

Our work aims to predict BHV of documents to be further utilized in profiling users. For this purpose, we developed a novel model called Semi-Supervised Nonnegative Matrix Factorization (SS-NMF) via latent topic modeling with NMF. We approached the problem as a set of multi-label classification tasks. Previous studies focus on supervised approaches with specific corpora. Our perspective is to train a semi-supervised generalized model via content-specific training corpus. Our model allows a wide range of texts to be processed in order

TABLE I
DISTRIBUTION OF DOCUMENTS FOR EACH BHV IN THE TRAINING DATASET

BHV	**Universalism**	**Benevolence**	**Conformity**	**Tradition**	**Security**
Seed	10	5	4	3	5
All	136	45	32	17	28
BHV	**Power**	**Achievement**	**Hedonism**	**Stimulation**	**Self-direction**
Seed	3	4	4	2	5
All	21	44	77	7	27

to predict the BHV of the document collections and authors. The aim is to learn the relationships among the concepts in the texts and their corresponding BHV concepts.

III. CORPUS

In this section, we explain the data collection and preprocessing techniques.

A. Data Collection

To obtain the training data for our system, Wikipedia articles were crawled for the ten BHVs. For each value, a few key seed articles were qualitatively selected in order to construct a value-specific corpus of Wikipedia articles. The Short Schwartz Values Survey (SSVS) which is a validated tool based on the original SVS (Schwartz Values Survey) and PVQ [9] was used to select the seed documents. A custom crawler for this purpose was developed that exploits the structural characteristics of Wikipedia articles. It traverses the URLs within the *See also*, Relevant topics and *References* sections of documents. The distance (depth) is determined for each newly encountered article from its corresponding seed document. This distance is used to indicate an article's semantic relevance to a core value. In this work, for our training set, only articles with distance equals to one are considered (434 documents). Table I shows the number of documents for each BHV according to the seed and crawled documents.

B. Preprocessing of Data

The crawled documents are converted to plain text, then the keywords are extracted and the term frequency (tf) values are computed. Each document is tokenized, after which punctuations, numbers, and stopwords are removed[1]. Words with apostrophes, such as *isn't*, are normalized to *is not*. The words whose character size are greater than 2 are retained. Finally, words are lemmatized in order to obtain the base form of words.

The data is represented as bag-of-words with their tf values. An n-gram model was used combining unigrams, bigrams, and trigrams. To prevent overpopulating the vocabulary, only the most frequent 50,000 n-grams in the vocabulary are considered.

A standardized text processing module for processing documents is made available[2].

IV. MODEL

Our approach is based on the Nonnegative Matrix Factorization (NMF) [20] that decomposes a given nonnegative X matrix into two non-negative W and H factors. More formally, given a $V \times T$ nonnegative matrix $X = \{x_{\nu,\tau}\}$ where $\nu = 1 : V, i = 1 : I$ and $\tau = 1 : T$, our aim is to find nonnegative factors W and H such that

$$x_{\nu,\tau} \approx [WH]_{\nu,\tau} = \textstyle\sum_i w_{\nu,i} h_{i,\tau}$$

The approximate decomposition is obtained by solving the minimization problem:

$$(W,H)^* = \arg\min_{W,H} D(X \parallel WH), \text{ subject to } W,H \geqslant 0 \quad (1)$$

In Equation 1, the function D is a suitably chosen error function. In this work, we choose the information (Kullback-Leibler) divergence, which is defined as:

$$D(X \parallel \Lambda) = -\sum_{\nu,\tau}(x_{\nu,\tau} \log \frac{\lambda_{\nu,\tau}}{x_{\nu,\tau}} - \lambda_{\nu,\tau} + x_{\nu,\tau}) \qquad (2)$$

Fixed point iteration is a popular method for this minimization. In this paper, we will refer to the $V \times I$ matrix W as the *template matrix*, and $I \times T$ matrix H the *excitation matrix*.

A. NMF for Topic Modeling

The goal of topic modelling is to explore the hidden thematic structure of documents. There are popular topic modelling techniques such as Principal Component Analysis (PCA), Vector Quantization (VQ), Latent Dirichlet Allocation (LDA) and NMF. NMF is preferable since it provides more coherent topics [21], [22].

In topic modelling, the *template* and *excitation* matrices hold the latent semantic relationships between documents and terms by means of topics. In our model, the template and the excitation matrices are document-topic and topic-term matrices, respectively. That is, the νth row of the *template matrix* is the latent representation for document ν and τth column of the *excitation matrix* shows the latent relationships of term τ with regard to I topics.

B. Semi-Supervised NMF (SS-NMF) model for Psychometric Scoring of Document Collections

In the SS-NMF model, a collection of documents of various themes is trained to associate themes with new documents. A *template* and *excitation* matrix is created for each theme in the corpus. The matrix, X, is the document-term representation of the training corpus related to all trained themes. Each element in X ($x_{\nu,\tau}$) corresponds to the tf value of the τth term in the νth document. In the case of BHV, the themes correspond to basic human values.

The following notation is used for indices:

T: for the dictionary,
V: documents related to the trained theme,
$\neg V$: documents related to background,
I: latent topics of the trained theme,
$\neg I$: latent topics of the background.

[1] English stopwords are removed using the Natural Language Toolkit found at https://www.nltk.org/.

[2] https://github.com/bulentozel/omterms

The decomposed matrix, X has two regions: the documents related to the trained theme ($X_{V,T}$) and the documents related to the other themes ($X_{\neg V,T}$).

The excitation matrix, H, has two regions; $H_{I,T}$ and $H_{\neg I,T}$. $H_{I,T}$ relates the trained-theme-related latent topics and the terms in the dictionary. $H_{\neg I,T}$ relates the background latent topics and the terms in the dictionary. The latent topics related the background are associated with all themes and possibly with other topics. Figure 2 shows the *excitation* matrices H_1 and H_2.

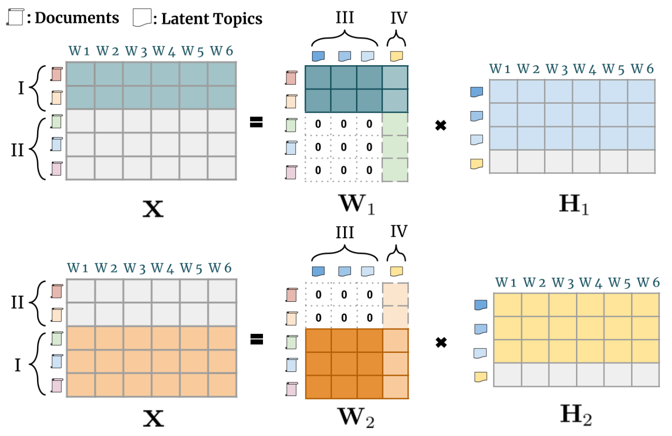


Fig. 2. The illustration of SS-NMF model for two themes. **I**: Documents Related Trained Theme, **II**: Documents Related Other Themes, **III**: Theme Related Latent Topics, **IV**: Other Latent Topics (Background etc.). The regions, $X_{V,T}$ and $X_{\neg V,T}$ are shown in colors for two themes in the corpus representation matrix. The regions, $W_{V,I}$, $W_{V,\neg I}$, $W_{\neg V,\neg I}$ and $W_{\neg V,I}$, of each instance in the template matrix W are denoted with four different colors. The regions $H_{I,T}$ of the excitation matrices, H_1 and H_2 are also denoted in colors.

The template matrix, W, has four regions. $W_{V,I}$, $W_{V,\neg I}$, $W_{\neg V,\neg I}$, $W_{\neg V,I}$. Figure 2 shows the training of SS-NMF model for two different themes. These regions are theme related latent topics for theme related documents, $W_{V,I}$, background latent topics for theme related documents, $W_{V,\neg I}$, theme-related latent topics for the documents from different themes, $W_{\neg V,I}$, and background latent topics for the documents from different themes, $W_{\neg V,\neg I}$.

In W_1 and W_2, the elements with white background correspond to $W_{\neg V,\neg I}$. This region is initialized to zeros so that latent topics of the not-trained-theme-related documents become background topics.

C. Initialization of SS-NMF

NMF has significant advantages in matrix factorization: The factorized matrices maintain sparsity and non-negativity of the X matrix and the interpretability of basis vectors. However, NMF has also its disadvantages. The optimization problem for Equation 1 has been shown to generalize k-means clustering problem which is known to be NP-complete [23]. Multiplicative update rule for this optimization problem only guarantees to find a local minimum, rather than a global minimum, since it is convex in either W or H, but not both. In practice, it is possible to run NMF with different initial setups and choose the one with the best local minimum. However, this reduces the replicability of the solution when even slight changes in parameters may produce different NMF factors. Thus, the initialization of the model is critical to obtain consistent results.

The most simple and preferred initialization method is random initialization where template and excitation matrices are initialized as dense matrices of random numbers between 0 and 1. However, since each random initialization may end up at a different local minimum, we propose a very inexpensive and highly consistent deterministic initialization method called *bCool* to initialize the excitation matrix H (inspired by the works of Langville, Meyer, Albright, Cox and Duling [24]). *bCool* initializes each row of H by averaging the rows of X. The steps to initialize a theme's H are:

1) Group X by themes and choose $p \times N_I$ densest rows of each group (typically the longest rows in sparse matrices), where N_I is the number of theme-related latent topics.
2) Split each group that consists of dense rows into N_I subgroups (each subgroup will have p items) where the cumulative density of each subgroup is approximately equal.
3) Assign the means of each subgroup of the theme to each theme related latent topic rows of H ($H_{I,T}$).
4) Choose $(5 \times p)/N_V$ densest rows of all subgroups, where N_V is the number of themes. Create $N_{\neg I}$ groups that include one subgroup from each theme where $N_{\neg I}$ is the number of backgrounds related latent topics. Then assign the mean of each of these groups to each background related latent topic rows of H which are $H_{\neg I,T}$.

bCool exploits the structure of our proposed SS-NMF model to initialize the excitation matrix. In simple terms, we use document vectors that belong to a specific theme to initialize it's $H_{I,T}$, while we use all documents to initialize $H_{\neg I,T}$. We use five times more document vectors to initialize background related latent topic rows than theme related latent topic rows to preserve the background related latent topics.

V. Scoring of Documents and Terms

This section describes the scoring of documents and the calculation of term scores.

A. Scoring Documents

After training the model, we obtained the excitation matrices, H, for each theme. The tf vectors of the documents in the training corpus are computed. The template vectors for each theme are computed using pretrained excitation matrices. The entries corresponding to the theme related topics, W_I, are normalized by the entry of background related topic, $W_{\neg I}$. The maximum value of theme related topics is the score for the theme. For example, the $3 + 1$ sized vector, $W = [0.4, 0.1, 0.5, 0.3]$ for the *Benevolence* theme is obtained. The first three entries correspond to W_I, and the last entry corresponds to $W_{\neg I}$. is the score of the test document for the The *Benevolence* score for the test document is 62.5 (the highest score in the normalized vector $[57, 25, 62.5]$).

B. Theme Term Scores

Our model uses terms to associate themes with documents. The training of the model forms an H matrix of term-topic distributions for each theme. A classic, simpler NMF model, to rank the significance of the terms for each theme, would use the normalized H matrix for each latent topic.

$$score\,(\tau_i) = \hat{h}_{i,\tau}, \quad i \in \{I \cup \neg I\} \tag{3}$$

where $\hat{h}$ is the normalized h.

Figure 3 shows the topmost five terms of three topics of different BHV's acquired using Equation 3. However, our model offers further information due to its semi-supervised nature and latent topics for each theme. Using these advantages, we present two new word scoring schemes which offer more tailored words for each latent topic.

Benevolence:	Hedonism:	Power:
good	happiness	authority
evil	pleasure	power
justice	social	bill
pardon	desire	social
trust	life	state

Fig. 3. The topmost 5 terms of example topics for three BHVs.

1) Direct Term Score (DTS): Direct term score exploits the latent topic structure of the model to come up with different term scores and orders for each document. The H matrix includes different term distributions for each theme related latent topics as well as the background. For example, if the *Power* BHV in the H matrix has three latent topics, our model learns three different concepts that leads three different term score distributions for *Power*.

To calculate the DTS of a term, instead of dealing with all latent topics like in Equation 3 we only use theme related latent topics. Most importantly, we put W_I to use to obtain scores for all terms under each themes' latent topics for each document that can be compared with each other. DTS enables words under high scored theme related latent topics to get higher scores than others.

$$DTS_{\nu,i,\tau} = w_{\nu,i} h_{i,\tau}, \quad i \in I \tag{4}$$

2) Purity Term Score (PTS): The aforementioned method provides a ranking of the terms in the documents for every theme. However, since our model uses tf for the data matrix, the terms with higher tf tend to have higher scores even though they are not the best representative of themes. There is nothing wrong here, but one may want to see terms that are both significant and specific to a theme. Thus, we propose the *Purity* concept that attempts to emphasize theme-specific terms by decreasing the scores of terms with hight tf values that are not specific to some themes.

The *purity* of the terms in documents for a theme is their ratio of their DTS to the background term score (BTS). The product of the purity of terms with their DTS is considered as the purity term scores (PTS).

$$BTS_{\nu,i,\tau} = w_{\nu,i} h_{i,\tau}, \quad i \in \neg I \tag{5}$$

$$Purity_{\nu,i,\tau} = \frac{DTS_{\nu,i,\tau}}{DTS_{\nu,i,\tau} + BTS_{\nu,i,\tau}} \tag{6}$$

$$PTS_{\nu,i,\tau} = DTS_{\nu,i,\tau} \times Purity_{\nu,i,\tau} \tag{7}$$

By using this method, the score of a term increases relatively if it mostly appears in a specific theme. Whereas the score decreases relatively if the term is in the background corpus. Consequently, theme-specific terms which are also not very rare are carried to the top of the score list.

VI. Evaluation

In this section, we evaluate our model in two different aspects: (1) the robustness of the training procedure via leave-one-out cross-validation (LOOCV), and (2) the quality of our scoring approach.

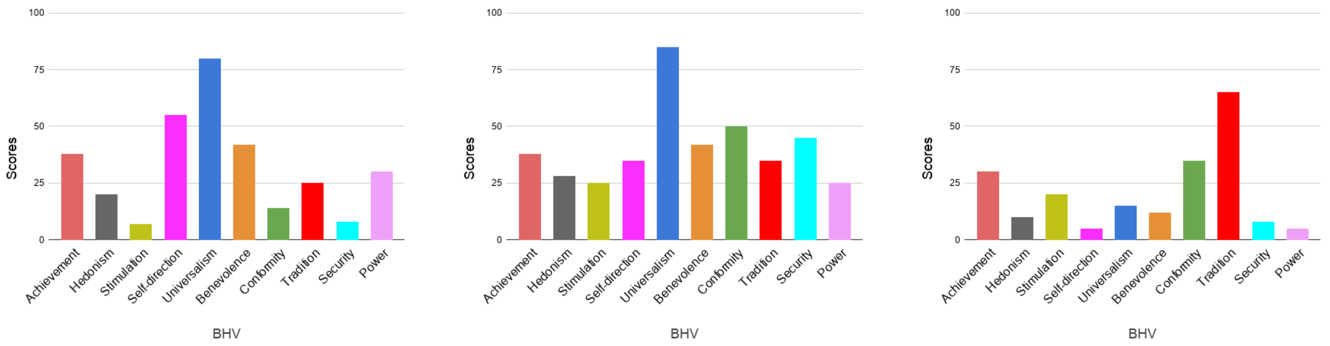
A. Robustness of Training

We create a training set consisting of 434 documents collected from Wikipedia, which are labelled as described in section III-A. Robustness refers to the repeatability of the experiments. Since NMF minimizes a non-convex cost, results using the same dataset with different initializations may differ, making interpretation hard.

In LOOCV, we remove one of the documents from the training set and train a model in the absence of this document. Then, the document that was removed is used to test the model. This procedure is repeated for each document. With this evaluation method, we aim to evaluate both the robustness of our procedure, as well as the quality of our training set.

1) Scoring Method: Formally, we compute our score as $\sum_{i,j} t(i)S(i,j)m(j)$ where t is a $0-1$ vector that is one only for the index of the correct label, and m are the predictions of our model. S is a scoring matrix. In a standard classification problem, typically a diagonal scoring matrix S is used as we wish to maximize the probability of a correct classification. However, in our problem, a document may not be exclusive to a single theme. Thus, we modified our scoring matrix by taking into account the relationship between BHVs in the Schwartz's circular structure. That is, we expect the score of the labelled BHV to be the highest, its adjacent and bipolar BHVs to be relatively high, and other BHVs to be low. Here, the intuition is that a document is likely to have the main theme and some supporting themes. Furthermore, it is common to refer to opposing themes for the purpose of comparing and contrasting. Supporting and opposing themes respectively correspond to the adjacent and bipolar themes in the Schwarts BHV structure. Table II shows the weights used in our scoring matrix S for each theme.

Figure 4 shows sample scenarios and their scores for the *universalism* BHV. **(a)** shows an optimal case with a high score. In this scenario, *universalism* and its neighbour and bipolar BHV's have relatively higher scores. In scenario **(b)**, *universalism* has the highest score, but others are nearly identical scores. In scenario **(c)** the score is negative. Although



(a) Score: 1010 (b) Score: 410 (c) Score: -347

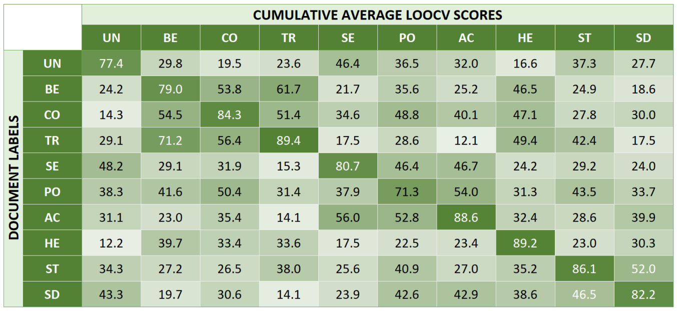
Fig. 4. Examples of alternative BHV scores a document labeled as *universalism* may take, in decreasing order of representation quality from (a) to (c).

TABLE II
WEIGHTS OF LOOCV MATRIX

BHV	UN	BE	CO	TR	SE	PO	AC	HE	ST	SD
UN	9	4	−5	−5	−5	4	4	−5	−5	4
BE	3	9	3	3	−6	3	3	−6	−6	−6
CO	−6	3	9	3	3	−6	−6	−6	3	3
TR	−6	3	3	9	3	−6	−6	−6	3	3
SE	−6	−6	3	3	9	3	−6	−6	3	3
PO	4	4	−5	−5	4	9	4	−5	−5	−5
AC	4	4	−5	−5	−5	4	9	4	−5	−5
HE	−3	−3	−3	−3	−3	−3	6	9	6	−3
ST	−6	−6	3	3	3	−6	−6	3	9	3
SD	3	−6	3	3	3	−6	−6	−6	3	9

UN: Universalism, BE: Benevolence, CO: Conformity, TR: Tradition, SE: Security, PO: Power, AC: Achievement, HE: Hedonism, ST: Stimulation, SD: Self-direction

TABLE III
AVERAGE LOOCV SCORES OF DOCUMENTS GROUPED BY THEIR THEMES.



		CUMULATIVE AVERAGE LOOCV SCORES									
		UN	BE	CO	TR	SE	PO	AC	HE	ST	SD
DOCUMENT LABELS	UN	77.4	29.8	19.5	23.6	46.4	36.5	32.0	16.6	37.3	27.7
	BE	24.2	79.0	53.8	61.7	21.7	35.6	25.2	46.5	24.9	18.6
	CO	14.3	54.5	84.3	51.4	34.6	48.8	40.1	47.1	27.8	30.0
	TR	29.1	71.2	56.4	89.4	17.5	28.6	12.1	49.4	42.4	17.5
	SE	48.2	29.1	31.9	15.3	80.7	46.4	46.7	24.2	29.2	24.0
	PO	38.3	41.6	50.4	31.4	37.9	71.3	54.0	31.3	43.5	33.7
	AC	31.1	23.0	35.4	14.1	56.0	52.8	88.6	32.4	28.6	39.9
	HE	12.2	39.7	33.4	33.6	17.5	22.5	23.4	89.2	23.0	30.3
	ST	34.3	27.2	26.5	38.0	25.6	40.9	27.0	35.2	86.1	52.0
	SD	43.3	19.7	30.6	14.1	23.9	42.6	42.9	38.6	46.5	82.2

UN: Universalism, BE: Benevolence, CO: Conformity, TR: Tradition, SE: Security, PO: Power, AC: Achievement, HE: Hedonism, ST: Stimulation, SD: Self-direction

the main theme is *universalism*, other BHVs have higher scores. This results in a decrease in the score.

2) *Results:* We have evaluated our approach via extensive simulations. The performance of the trained model in predicting the BHV scores of documents is examined using LOOCV. Table III shows the results of the average LOOCV scores for each BHV. The scores on the diagonal of the table are the highest scores for each row which indicates that the model can successfully assign the highest score to the correct theme. Neighbor and bipolar BHVs are also relatively higher than other BHVs.

We analyzed the performances of two initialization procedures, *random* and deterministic *bCool*. We carry out 10 independent training runs and we compute the average sample variance of the LOOCV scores. In general, we observe that the *bCool* method obtains very close results to the *random* case. One of the important points is that while using *random* initialization, we may get different results at each time. In contrast, the deterministic initialization removes this undesired variation at the expense of getting suboptimal solutions. In addition to this, the result indicates that the variance of *random* initialization is 40 times more than *bCool*. Thus, we prefer to use our proposed deterministic *bCool* initialization procedure.

In order to compare initialization procedures, we group the documents based on their themes and get mean through all themes. Then, we calculate scores for each theme using Equation 3. Table IV demonstrates the results of comparisons for *random* and *bCool* initialization procedures by showing the percentage increase/decrease of the score for each theme. In this table, we can see that *random* and *bCool* initialization procedures got better scores at different themes, but this is an expected behaviour because of the variance of the random method. But, on the average, the score is close to zero.

TABLE IV
THE (%) IMPACT OF INITIALIZING WITH *bCool* INSTEAD OF RANDOM

Universalism	Benevolence	Conformity	Tradition	Security
-6.49	1.48	-20.50	-7.95	-10.97
Power	**Achievement**	**Hedonism**	**Stimulation**	**Self-direction**
8.39	5.41	2.75	1.18	-17.88

Table V shows the scoring results for using *bCool* initialization procedure. By considering our scoring metric and sample scenarios in Figure 4, we obtain satisfactory results for each theme by using *bCool* initialization procedure. These results point out that our approach predicts almost correctly of each theme.

TABLE V
THE SCORES FOR *bCool* INITIALIZATION

Universalism	Benevolence	Conformity	Tradition	Security
484.02	641.77	452.08	704.51	277.22
Power	**Achievement**	**Hedonism**	**Stimulation**	**Self-direction**
377.61	485.87	513.37	531.08	352.10

B. Anomaly Detection for Training Documents

As discussed in section III, the training data was semi-automatically collected using qualitatively selected seed documents. The collected documents were expected to be related to themes corresponding to their seeds, thus resulting in a set of documents appropriate for training our model. However, we observed that some documents relate more strongly to unintended themes, which could negatively impact the training. To detect documents labelled as such, the documents were filtered using the LOOCV scores. Documents whose LOOCV scores are less than zero are labelled as an anomaly and eliminated from the training setlist. In this manner, we eliminated 13 universalism, 4 benevolence, 7 conformity, 5 security, 5 power, 3 achievement, 4 hedonism and 6 self-direction documents from the training set.

A model was trained based on this new training set using *bCool* initialization and evaluated with LOOCV. The results of this new cleaned model were compared with the model that used the original document set. To perform a fair comparison, we only considered the scores of documents that belong to the eliminated dataset. Table VI shows that eliminating documents resulted in an overall increase in scores. However, continuously eliminating documents in this manner could end up with a very small training set in hand, which does not sufficiently represent the themes.

TABLE VI
THE DIFFERENCE IN % PERCENTAGE OF THE SCORES OBTAINED USING THE *bCool Cleaned* AND *bCool* METHODS.

Universalism	Benevolence	Conformity	Tradition	Security
15.49	-0.26	3.09	13.66	3.25
Power	**Achievement**	**Hedonism**	**Stimulation**	**Self-direction**
4.17	28.39	9.48	11.51	-6.96

C. Evaluation Through Big Five Personality Traits

While there are many studies on BHV that report results based on surveys and questionnaires, there are no labelled document sets based on BHV to evaluate our approach. Although BHV aims to capture human values, not traits, a relation among FFM traits and BHV have been proposed [25].

We evaluated our model using PAN-AP-2015 corpus in conjunction with these relations, which are shown in Table VII. Note that only four of the five traits are taken into consideration since Neuroticism was eliminated as no association with any values were identified.

PAN-AP-2015 corpus [15] consists of FFM traits for users whose tweets are in English, Spanish, Italian, and Dutch. These users also took the BFI-10 online test [26] that scores personality traits according to FFM that is normalized between -0.5 and 0.5.

TABLE VII
THE CONVERSION MATRIX THAT RELATES FFM TRAITS TO BHV.

	Empirical Correlations [25]			
BHV	**E**	**A**	**C**	**O**
Universalism	-0.07	0.15	-0.17	0.47
Benevolence	0.01	0.45	0.04	-0.06
Conformity	-0.13	0.20	0.16	-0.34
Tradition	-0.29	0.36	-0.10	-0.29
Security	-0.11	0.06	0.22	-0.29
Power	0.13	-0.45	0.05	-0.38
Achievement	0.31	-0.41	0.22	-0.06
Hedonism	0.18	-0.34	-0.05	0.07
Stimulation	0.26	-0.26	-0.24	0.33
Self-direction	0.10	-0.25	-0.01	0.48

E: Extroversion, A: Agreeableness, C: Conscientiousness, O: Openness

In order to gain insight into how our model performs on the PAN-AP-2015 corpus, we extracted BHV scores for the English tweets using the pretrained SS-NMF (Section III). The BHVs predicted for documents are mapped to corresponding FFM traits (Table VII) in order to render our output comparable.

A document consists of the concatenation of all the tweets of an individual. A vector of size 10 is obtained for each document to represent the weight of its relationships to the BHVs. The prediction of test data is represented with a matrix, $R \in \mathcal{R}^{N\times 10} \geq 0$, of size $N \times 10$.

$$Y = RC \tag{8}$$

$$\hat{Y}_i = \frac{Y_i - \sum_{j\in C_{i,j}<0} C_{i,j}}{\sum_{j\in C_{i,j}>=0} C_{i,j} - \sum_{j\in C_{i,j}<0} C_{i,j}} - 0.5 \tag{9}$$

The number of individuals in the test set is denoted with N. The conversion matrix, $C \in \mathcal{R}^{10\times 4}$, consists of values corresponding the correlation between a BHV and a FFM trait as shown in Table VII. The conversion matrix, C, is the empirical correlations found in [25]. The product of R and C yields the unscaled predictions of FFM traits, Y. The scaled predictions for each trait, $i \in \{0, 1, 2, 3\}$, are denoted by the vector $\hat{Y}_i$. The scaling function is shown in Equation 9, which scales the raw predictions in the range of $[-0.5, 0.5]$.

The Root Mean Squared Error (RMSE) of our results are shown in Table VIII along with the RMSE of the random predictions and the mean of RMSE of the best performers in [15]. The RMSE's of random predictions are the mean of ten trials with predictions from a uniform distribution between -0.5 and 0.5. It is worth to notice that our experiments do not use training data of the PAN-AP-2015 corpus. Even though supervised methods perform better, our semi-supervised method's performance is much better than a random prediction, and it is corpus independent.

TABLE VIII
RESULTS OF MAPPING BHVS OBTAINED USING OUR MODEL FOR THE DOCUMENTS IN THE PAN-AP-2015, WHICH ARE MAPPED TO FFM .

	E	**A**	**C**	**O**
RMSE of Random Prediction	0.3746	0.3591	0.3653	0.4157
RMSE of our Experiment	0.2294	0.2468	0.2456	0.3606
The Mean of Best Performers' RMSE's in [15]	0.1330	0.1406	0.1507	0.1358

E: Extroversion, A: Agreeableness, C: Conscientiousness, O: Openness

VII. DISCUSSION AND CONCLUSIONS

Understanding the personalities of individuals from written text is a challenging task in the pipeline of human level sentiment analysis. Our work aims to model the word usage in order to score documents with a novel model. In this respect, our approach does not directly applicable to profiling individual but we believe that a consistent scoring procedure for scoring documents is an important first step.

The training corpus we have collected and used was not homogeneous among the BHV. One of the next steps would be enlarging the training corpus with a balanced distribution.

Evaluation is very important but difficult to do entirely data-driven. We first carried out LOOCV to validate our model and got promising results. We also used LOOCV to find anomaly documents that represent different themes than the intended ones. Then apart form LOOCV in order to evaluate our model with ground truth values, we tested the PAN-AP-2015 corpus, which has Big Five scores. Since we trained our model with BHV training corpus, we had to convert the BHV scores to Big Five scores. Even though this not the best way to evaluate, the correlation between the two personality traits allowed us to arrive at a scoring method. Our findings should be revisited once a direct method for evaluation can be carried out.

In contrast to standard topic models that use a bag-of-words representation, we believe that using explicit models for synonyms and antonyms are very important to verify the BHV for each document. In order for a document to converge to its BHV, the antonyms of the words in that document must also exist. The antonyms are a part of context but cause confusion as to which value the document belongs to. This provides competition between the BHVs that their higher order group belong to the bipolar dimension. This point is also an important future study.

Source codes for this study is available online on the GitHub page: https://github.com/suyunu/semi-supervised-nmf.

ACKNOWLEDGMENT

This work has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement no 687941 as well as the Turkish Ministry of Development under the TAM Project no DPT2007K120610.

REFERENCES

[1] E. Cambria, S. Poria, A. Gelbukh, & M. Thelwall. (2017). Sentiment analysis is a big suitcase. IEEE Intelligent Systems, 32(6), 74-80.

[2] W. Bilsky and S. H. Schwartz, (1994). Values and personality. European journal of personality, 8(3), 163-181.

[3] S. H. Schwartz, (1994). Are there universal aspects in the structure and contents of human values?. Journal of social issues, 50(4), 19-45.

[4] S. Battiston & H. Zeytinoglu, (2016). Social Intelligence Networks. A Novel Framework for On-Line Social Platforms.

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

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

[7] T. Yarkoni, (2010). Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. Journal of research in personality, 44(3), 363-373.

[8] Y. R. Tausczik & J. W. Pennebaker, (2010). The psychological meaning of words: LIWC and computerized text analysis methods. Journal of language and social psychology, 29(1), 24-54.

[9] S. H. Schwartz, (2012). An overview of the Schwartz theory of basic values. Online readings in Psychology and Culture, 2(1), 11.

[10] P. T. Costa, & R. R. McCrae, (2008). The revised neo personality inventory (neo-pi-r). The SAGE handbook of personality theory and assessment, 2(2), 179-198.

[11] S. Argamon, S. Dhawle, M. Koppel, & J. Pennebaker, (2005). Lexical predictors of personality type.

[12] J. Oberlander & S. Nowson, (2006, July). Whose thumb is it anyway?: classifying author personality from weblog text. In Proceedings of the COLING/ACL on Main conference poster sessions (pp. 627-634). Association for Computational Linguistics.

[13] F. Celli, F. Pianesi, D. Stillwell, & M. Kosinski, (2013, June). Workshop on computational personality recognition (shared task). In Proceedings of the Workshop on Computational Personality Recognition.

[14] F. Celli, B. Lepri, J. I. Biel, D.Gatica-Perez, G. Riccardi & F. Pianesi (2014, November). The workshop on computational personality recognition 2014. In Proceedings of the 22nd ACM international conference on Multimedia (pp. 1245-1246). ACM.

[15] F. M. Rangel Pardo, F. Celli, P. Rosso, M. Potthast, B.Stein & W. Daelemans (2015). Overview of the 3rd Author Profiling Task at PAN 2015. In CLEF 2015 Evaluation Labs and Workshop Working Notes Papers (pp. 1-8).

[16] N. Majumder, S. Poria, A.Gelbukh & E. Cambria, (2017). Deep learning-based document modeling for personality detection from text. IEEE Intelligent Systems, 32(2), 74-79.

[17] D. Xue, L. Wu, Z. Hong, S. Guo, L. Gao, Z. Wu, ... and J. Sun. (2018). Deep learning-based personality recognition from text posts of online social networks. Applied Intelligence, 1-15.

[18] M. Kosinski, D. Stillwell, & T. Graepel (2013). Private traits and attributes are predictable from digital records of human behavior. Proceedings of the National Academy of Sciences, 201218772.

[19] G. Carducci, G. Rizzo, D. Monti, E. Palumbo, & M. Morisio (2018). TwitPersonality: Computing Personality Traits from Tweets Using Word Embeddings and Supervised Learning. Information, 9(5), 127.

[20] D. D. Lee & H. S. Seung (1999). Learning the parts of objects by non-negative matrix factorization. Nature, 401(6755), 788.

[21] C. C. Aggarwal & C. Zhai (2012). A survey of text clustering algorithms. In Mining text data (pp. 77-128). Springer, Boston, MA.

[22] K. Stevens, P. Kegelmeyer, D. Andrzejewski & D. Buttler (2012, July). Exploring topic coherence over many models and many topics. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (pp. 952-961). Association for Computational Linguistics.

[23] C. Ding, X. He & H. D. Simon (2005, April). On the equivalence of nonnegative matrix factorization and spectral clustering. In Proceedings of the 2005 SIAM International Conference on Data Mining (pp. 606-610). Society for Industrial and Applied Mathematics.

[24] A. N. Langville, C. D. Meyer, R. Albright, J. Cox & D. Duling. (2006, August). Initializations for the nonnegative matrix factorization. In Proceedings of the twelfth ACM SIGKDD international conference on knowledge discovery and data mining (pp. 23-26).

[25] S. Roccas, L. Sagiv, S. H. Schwartz, & A. Knafo, (2002). The big five personality factors and personal values. Personality and social psychology bulletin, 28(6), 789-801.

[26] B. Rammstedt & O. P. John,(2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. Journal of research in Personality, 41(1), 203-212.