SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
HUMAN BEHAVIOUR
INFORMATION TECHNOLOGY

Received
23 January 2014

Accepted
4 April 2014

Published
23 April 2014

Correspondence and requests for materials should be addressed to Y.N. (yneuman@bgu.ac.il)

A Vectorial Semantics Approach to Personality Assessment

Yair Neuman & Yochai Cohen

Department of Education, Ben-Gurion University of the Negev, Beer-Sheva, Israel.

Personality assessment and, specifically, the assessment of personality disorders have traditionally been indifferent to computational models. Computational personality is a new field that involves the automatic classification of individuals' personality traits that can be compared against gold-standard labels. In this context, we introduce a new vectorial semantics approach to personality assessment, which involves the construction of vectors representing personality dimensions and disorders, and the automatic measurements of the similarity between these vectors and texts written by human subjects. We evaluated our approach by using a corpus of 2468 essays written by students who were also assessed through the five-factor personality model. To validate our approach, we measured the similarity between the essays and the personality vectors to produce personality disorder scores. These scores and their correspondence with the subjects' classification of the five personality factors reproduce patterns well-documented in the psychological literature. In addition, we show that, based on the personality vectors, we can predict each of the five personality factors with high accuracy.

The assessment and diagnosis of personality and, in particular, of *personality disorders* (PDs) is a common practice of psychologists and psychiatrists. This practice is guided mainly by classical psychometric tools such as questionnaires and relies heavily on human judgment and expertise.

Recently, there has been an increasing interest in introducing novel computational methods for the assessment of personality and related disorders[1–9]. These methods involve "personality recognition" tasks consisting of "automatic classification of authors' personality traits that can be compared against gold standard labels" (1, p. 1). The methods used for computing personality apply Natural Language Processing (NLP) techniques and Machine Learning (ML) algorithms and use various features (e.g., n-grams) and resources [e.g., Linguistic Inquiry and Word Count[10],] to classify the labeled data.

In this context, it was recently argued by Pianesi (7, p. 150) that "There is a tension between the invariance of personality traits and the natural variability of behavior in concrete situations that risks to seriously hamper current attempts at automatically predicting personality traits." In other words, while personality is sometimes portrayed as invariant (e.g., John is an extrovert), situational context generates variability that washes out the personality's alleged stability. For instance, while John may be a typical extrovert in certain situations, such as within his close family circle, he may behave like a typical introvert in other situations.

The implication of Pianesi's argument is that an attempt to identify a canonical set of features and parameters for personality recognition may work well in a given context but may hold no validity in a different one. Therefore, Pianesi's criticism encourages a more flexible and context-dependent approach to personality assessment.

One possible way to address this challenge, in the specific context of text analysis, is by a priori characterizing a set of linguistic cues describing certain personality dimensions and then measuring *the similarity of the text-in-context to each of these dimensions*.

Following this logic, in this paper, we present a novel approach to personality assessment, which is based on the idea of vectorial semantics models (VSM)[11–12], explained in the next section.

The idea of using a vectorial semantics approach to personality assessment emerged in the context of profiling political leaders for military intelligence, and was recently and successfully used in the context of forensic psychiatry for the screening of potential mass-murderers. This study represents the first time in which this approach is validated against the "big-five" personality factors and by using thousands of essays written by students.

Vectorial semantics and personality assessment

Vector space models of semantics, suggest that the meaning of a word and the concept it represents can be identified by analyzing words co-occuring with our target word in a given context. For instance, if we want to

Table 1 | Adjectives co-located with Suspicious and Vengeful

	Paranoid	Obsessive	Histrionic
Suspicious	3	3	3
Vengeful	6	3	2

understand the meaning of a *Paranoid*, we examine the adjectives co-located with it in texts. Using a corpus of the English language, we may find that the two adjectives most frequently co-located with Paranoid are: *Suspicious* (Frequency = 3) and *Vengeful* (Frequency = 6). This means that in the texts we have analyzed the word Paranoid is accompanied by the word Suspicious three times and by the word Vengeful six times. The shaded area in Table 1 presents these data.

We can now consider Suspicious and Vengeful as two *dimensions* defining the semantic space of Paranoid. In this semantic space, the meaning of Paranoid is represented as a vector in a two-dimensional space defined by Suspicious and Vengeful, a point that is graphically represented in Figure 1.

Next, we find that *Obsessive* and *Histrionic* are two other words residing in this semantic space, as presented in Table 1.

Figure 1 is a graphical representation of Table 1. The X-axis signifies the dimension of Suspicious and the Y-axis signifies the Vengeful dimension. In this space, Paranoid is represented by the dashed vector whose coordinates are X = 3 and Y = 6. Along the same line, the bold vector represents Histrionic and the third vector Obsessive.

We can see that the vector of Histrionic is closer, and therefore more similar to the vector of Obsessive than to the vector of Paranoid. Measuring the similarity of vectors is a simple but rigorous procedure that relies on the cosine between them.

In reality, the situation is much more complex than described as each word is accompanied by many other words that co-occur with it in a linguistic corpus. Therefore, instead of a simple two-dimensional space, as appears in Figure 1, we have to deal with a high-dimensional space. In addition, in some cases we prefer to measure the similarity of texts comprising many words rather than the semantic similarity of isolated words only. However, beyond the above complexities, the basic idea of representing the meaning of words as a vector in a high-dimensional semantic space and measuring the similarity between words/texts by measuring the distance between the vectors has been proved to be extremely powerful and may be used for personality assessment.

For example, assume we intend to assess the personality of a certain individual by analyzing a sentence he published on Twitter (e.g., *I suspect that the CIA is responsible for this conspiracy*) and determine whether it is indicative of a paranoid personality disorder. The vectorial semantics approach would propose to represent the relevant words in the sentence as a vector in a high-dimensional semantic space and to measure the distance between the vector comprised of these words and the vector of words representing the paranoid personality disorder. The closer the vectors are the more similar the sentence is to the personality vector and, therefore, our confidence increases in the hypothesis that the sentence represents the paranoid personality.

The first step in the vectorial semantics approach to personality assessment is therefore to identify words that are the best representatives of a certain personality trait. Our basic assumption is that experts can characterize personality types and specifically PDs by using a minimal set of words that grasp the essence of the disorder. For instance, while describing a paranoid personality disorder, the adjective "suspicious" emerges as a prototypical keyword (e.g.,[13–15]). Using a set of adjectives that describes a PD, we may automatically analyze the dimensions of a given text by simply representing it as a vector and measuring its similarity to pre-defined vectors of PDs. The above assumption has been intensively discussed in personality research[16]. Moreover, the assumption has been specifically prominent in the study of the Big Five personality traits that have been assessed through trait-descriptive adjectives[17–19].

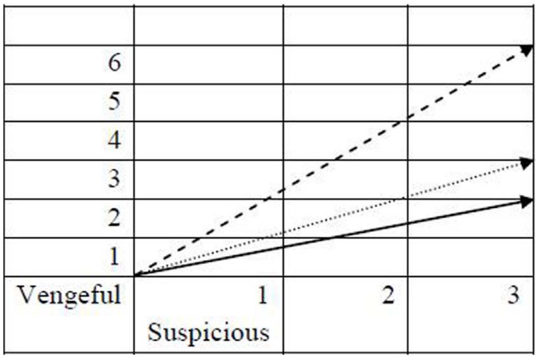

Figure 1 | A graphical representation of table 1.

The vectorial semantics approach to personality assessment can be summarized as follows: (1) Based on theoretical and/or empirical knowledge select a set of words that represent a psychological trait. (2) Represent this set as a vector. (3) Choose a text you would like to assess and represent its words as a vector. (4) Measure the similarity between the vectors. The similarity score is indicative of the degree in which the personality trait is represented in the text.

The Five Factor Model and PDs

The Big Five or the five-factor model of personality (FFM)[16,20,21] suggests that the taxonomy of personality can be described through five major traits. Extraversion (E) involves an "energetic approach" to the social and material world and includes traits such as sociability, activity, and positive emotionality. Agreeableness (A) involves a pro-social and communal orientation and includes traits such as altruism, tender-mindedness, trust, and modesty. Conscientiousness (C) describes socially prescribed impulse control and goal-directed behavior. Neuroticism (N) – sometimes referred to as its opposite pole, Emotional stability – involves negative emotionality and feeling anxious, sad, and tense. Openness to experience (O) describes the breadth, depth, and originality of the subjects' mental and experiential life.

The FFM has been of interest for studying various PDs and related behaviors (e.g. 22). Malouff et al.[23] found that mood disorders were associated mostly with a higher level of N and lower levels of E, A, and C. The study of[24] further supports these associations by showing that depressive disorders are associated with high levels of A and low levels of C. In a recent study[25], found that the presence and severity of depression among older adults was associated with a higher level of N and a lower level of E: Those who are neurotics and introverts are much more inclined toward depression.

In this study, we are interested in validating our vectorial semantics approach to personality assessment. *To test the validity of our approach, we draw on the comprehensive meta-analytic review of the five-factor model and personality disorders*[26]. This meta-analysis examines the relationships between each of the big-five personality dimensions and each of ten personality disorders.

We specifically draw on Table 7 in Saulsman and Page[26] (p. 1070). This table displays the binomial effect size for each personality disorder and five-factor personality dimension. The table shows "out of 100 people with a particular personality disorder, the number that would score high on a particular personality dimension and the number that would score low on the dimension". For instance, it was found that, out of 100 paranoids, we should expect 64 to score High on the Neuroticism dimension and 36 to score Low. It means

Table 2 | Binomial effect size for the personality disorders and five-factor personality dimension

Personality Disorder	N		E		A	
	High	Low	High	Low	High	Low
Paranoid	64	36			33	67
Schizoid			39	61		
Schizotypal	68	32	36	64	40	60
Histrionic			71	29	33	67
Narcissistic			60	40	37	63
Avoidant	75	25	28	72		
Dependent	71	29				

that a *significantly higher proportion of Paranoids would be among those who are classified as Neurotics*.

In general, the meta-analysis reveals three general patterns: (1) that most PDs have "positive associations with neuroticism", (2) that most PDs have "negative associations with agreeableness", and (3) that Extraversion is the more *mixed* and therefore "discriminating dimension"[26] with positive and negative associations with the PDs.

Saulsman and Page decided that "effect sizes $\geq$ .20 would be considered meaningful"[26]. Although this cut-point is arbitrary, it is justified on practical reasons, which are elaborated in the paper. Following this decision, we decided to report only the meaningful effect sizes found in the meta-analysis. In addition, we focus only on PDs that we were able to translate into vectors, and disorders such as "Borderline" that were difficult to conceptualize in terms of adjectives were not included in the study. Hence, Table 2 presents the meaningful effect sizes identified in the meta-analysis with seven PDs.

Research questions and hypotheses

Research question 1. In this study, we have defined PDs vectors and measured their similarity with essays written by students. The question is whether our results correspond with the three patterns identified in the meta-analysis[26] and reported on, above. If our vectorial semantics approach to personality assessment can validly measure the degree in which a PD is evident in a text, then the relation between our PDs scores and the five-factors should reproduce the above patterns. Therefore, we hypothesized that:

Hypothesis 1: Subjects classified as N (i.e., Neurotics) will score higher than subjects classified as non-N on the Paranoid, Schizotypal, Avoidant, and Dependent PDs scores.
Hypothesis 2: Subjects classified as non-A will score higher than subjects classified as A on the Paranoid, Schizotypal, and Narcissistic PDs.
Hypothesis 3: Non-Extraverts will score higher on Schizoid, Schizotypal and Avoidant PDs and will score lower on the Histrionic and Narcissistic PDs.

Research question 2. While our main aim is to provide minimal evidence to the validity of our approach, our second question is: can one classify the subjects into the different values of their personality factors (Neurotic vs. non-Neurotic) based on the vectorial analysis we have developed? This question concerns the pragmatic value of using the vectorial semantics approach for personality assessment.

To address this challenge, we constructed vectors representing the five personality factors (i.e., A, E, C, N, and O). Drawing on the list of adjectives provided by Trapnell and Wiggins[18], we constructed vectors of the personality dimensions along the lines used for constructing the PDs. For each dimension we constructed one vector representing the existence of the dimension (e.g., E), and another vector representing its opposite (e.g., non-E or E-negative). The vectors are:

Table 3 | Cross-tabulation of the predicted vs. observed value of the O

	Predicted	
Observed	N	Y
N	846	350
Y	551	721

'N' represents the subject being classified as 'non-O' and 'Y' that the subject has been classified as 'O'

E-POSITIVE: dominant, assertive, authoritarian, forceful, assured, confident, firm, persistent
E-NEGATIVE: nervous, modest, quiet, forceless, afraid, shy, calm, indecisive
A-POSITIVE: Tender, Gentle, Soft, Kind, Affectionate Helpful Sympathetic Friendly
A-NEGATIVE: Cruel Unfriendly Negative Mean Brutal Inconsiderate Insensitive Cold
C-POSITIVE: Organized Orderly Tidy Neat Efficient Persistent Systematic Straight Careful Reliable
C-NEGATIVE: Distracted Unreliable Incompetent Wild Inefficient Disloyal Chaotic Confused Messy Disorganized
N-POSITVE: Worried Stressed Anxious Nervous Fearful Touchy Guilty Insecure Restless Emotional
N-NEGATIVE: Balanced Stable Confident Fearless Calm Easy_going Relaxed Secure Comforted Peaceful
O-POSITIVE: Philosophical Abstract Imaginative Curious Reflective Literary Questioning Individualistic Unique Open
O-NEGATIVE: Narrow-minded Concrete Ordinary Incurious Thoughtless Ignorant Uneducated Common Conventional Restricted

The similarity of each vector to the essay was calculated using the same procedure described for each of the nine PDs. These similarity scores, which are meant to be indicative of the extent in which the personality factor is evident in the text, were titled the *Personality Factor Scores*.

Results

Research question 1. Regarding Hypothesis 1; it was found that, when the PD scores of Neurotics and non-Neurotics were compared, the Mean Rank of N is significantly higher for the Paranoid (z = 3.50, P = .000), Avoidant (z = 4.36, P = .000) and Dependent (z = 3.78, P = 001) PDs. The difference was not statistically significant for the Schizotypal PD, although N scored higher than non-N. These results largely confirm our hypothesis. Regarding Hypothesis 2; using the same procedure, it was found that non-As scored higher with regard to the Paranoid (z = 2.09, P = .05), Schizotypal (z = 4.23, P =.000), and Narcissitic (z = 2.84, P = .002) PDs. These results fully support the hypothesis. Regarding Hypothesis 3; the direction of the difference was in line with the meta-analysis results of[26] only for the Schizoid (z = 2.16, P =.016) PD where non-E scored higher than E.

Overall, there seems to be a good match between the patterns presented in the meta-analysis of[26] and our results. This match provides empirical support for the validity of our PDs measures. The results concerning the E factor were partially supported, although mixed results were clearly evident with regard to this personality dimension and associated PDs.

Research question 2. The classification model we have used classifies subjects into the value of each of the five factors by using the personality vectors automatically extracted from their essays. For

Table 4 | Cross-tabulation of the predicted vs. observed value of the C

	Predicted	
Observed	N	Y
N	651	563
Y	461	793

'N' represents the subject being classified as 'non-O' and 'Y' that the subject has been classified as 'O'

Table 6 | Cross-tabulation of the predicted vs. observed value of N

	Predicted	
Observed	N	Y
N	523	712
Y	305	928

'N' represents the subject being classified as 'non-O' and 'Y' that the subject has been classified as 'O'

example, the model classifies a certain subject as Neurotic or non-Neurotic based on his PDs Scores and Personality Factor Scores. Cross-tabulating the value of each of the five personality vectors (e.g., E vs. non-E) by their predicted values according to the classification procedure, the results were found statistically significant for all of the five factors. Tables 3–7 present the cross-tabulation for each of the big five personality factors. Table 8 presents the accuracy of the classification procedures and the χ^2 for each cross tabulation. All results were found statistically significant (P = .000). Across personality dimensions and on average, our analysis gained 60% accuracy, which is slightly higher than the average best performances reported by[27] (i.e., 57.09%).

The classification procedure includes a sub-procedure that allows us to rank the independent variables according to their normalized importance to the model. Here we can see in a descending order the three most important predictors that were found for each personality factor:

E: N-Negative, C-Positive, Paranoid
N: N-Positive, Depressive, Avoidant
A: C-Negative, N-Negative, O-Negative
C: N-Negative, Schizoid, Histrionic
O: O-Negative, Histrionic, Schizotypal

Discussion

The vectorial semantics approach merges the idea of VSM with personality assessment. To the best of our knowledge, this is a novel approach that is different from the other approaches used for automatic personality assessment. Using this approach for measuring the level of each PD in a text, our results largely agree with the most comprehensive meta-analysis that examined the relationships between the big five factors and PDs. While this agreement provides empirical support for our approach, it cannot be considered as a final validation but only as a first and primary step in providing empirical support for a minimal level of validity. Our secondary aim was to test our approach in a classification task. We must emphasize that, although we did not aim to compete with various ML approaches for classifying the "Big 5" of personality as done in a recent workshop[1], our approach has gained a higher level of accuracy when compared with the results gained elsewhere[27]. Fusing our approach with various ML algorithms described in (1) may improve current performance on this task. In addition, our analysis has been constrained by the binary nature of the dependent variables and the predictive strength of our approach may be increased by considering the Big-5 factor scores dimensionally i.e. as continous scores. Our approach may also have various practical applications. For instance, the approach may be used for automatically assessing the personality disorders of people who find it difficult to gain access to mental health services. These subjects may write a short essay that can then be analyzed through our approach and their PDs scores may be compared to a benchmark of essays written by healthy subjects. Subjects who score significantly higher on certain PDs may be advised to approach a mental health expert for an in-depth diagnosis. This form of automatic screening procedure, which has been proposed in various contexts such as in the automatic diagnosis of depression[28], will assist both patients and mental health authorities struggling with overload and lack of resources. However, the specific applications of the approach presented in this study must be better planned and elaborated. We may therefore conclude our paper with an invitation for further research and collaboration.

Methods

Data. We used the Essays dataset provided to the participants of the "Workshop on Computational Personality Recognition: Shard Task"[1]. As explained by the organizers of the workshop, this is a corpus of 2468 stream-of-consciousness texts that was produced by[29] and labeled with personality classes of the FFM.

The essays were written by students who were also assessed through a standard inventory of[30] for their score on each of the big five personality factors. The personality scores obtained from the Big5 test have been normalized by[27] and turned into nominal classes by[1] as this workshop focused on a classification task of the big five through a ML approach. The labels in the dataset are provided as categorical variables with a balanced frequency of around 50% in each category. The percentage of subjects in each personality dimension is presented in Table 9. It is important to remember that the Big-5 varibles are usually conceptualized in the literature as dimensional, as opposed to categorical (i.e., one is not "Open" or "Not-open," but instead falls somewhere along a continuum with regard to the personality characteristic). However, the nature of the data set we have used and the methodlogical approach necessitated a binary approach to these characteristics for purpose of the present study.

Defining the PDs vectors. We used the definition of Millon et al.[14] for PDs and extracted the first main adjectives[14] used to define each of nine PDs. The adjectives were checked against the Oxford English Dictionary and the Corpus of Contemporary American English[31]. In the case where an adjective has a synonym that perfectly preserves the original meaning while having a higher frequency in the population, the adjective has been replaced. In addition, two-word adjectives have been replaced with a one word description wherever possible. The following words were used to define Millon's PD Vectors:

Table 5 | Cross-tabulation of the predicted vs. observed value of the A

	Predicted	
Observed	N	Y
N	307	851
Y	171	1139

'N' represents the subject being classified as 'non-O' and 'Y' that the subject has been classified as 'O'

Table 7 | Cross-tabulation of the predicted vs. observed value of the E

	Predicted	
Observed	N	Y
N	405	786
Y	252	1025

'N' represents the subject being classified as 'non-O' and 'Y' that the subject has been classified as 'O'

Table 8 | Accuracy of the classification procedure

Personality Dimension	χ^2	Accuracy
N	57.35	59
E	64.25	58
O	92.05	64
A	62.22	59
C	11.35	59

1. Schizoid: indifferent, apathetic, remote, solitary
2. Depressive: sad, depressed, hopeless, gloomy, fatalistic
3. Avoidant: shy, reflective, embarrassed, anxious
4. Dependent: helpless, incapable, passive, immature
5. Histrionic: dramatic, seductive, shallow, hyperactive, vain
6. Narcissistic: selfish, arrogant, grandiose, indifferent
7. Compulsive: restrained, conscientious, respectful, rigid
8. Paranoid: cautious, defensive, distrustful, suspicious
9. Schizotypal: eccentric, alien, bizarre, absent

Preprocessing. Each essay was analyzed through a Part-of-Speech tagger[32] and only nouns, verbs, adjectives, and adverbs were processed for further analysis. Next, we measured the similarity between each of the essays and each of the PDs vectors by using the term-to-context matrix developed by Turney[33,34]. More specifically, we drew on the dual-space model developed by Turney[33,34]. This model allowed us to measure the semantic similarity of words and texts by using the "dual space", and examining the "domain" similarity of words (i.e., their "topic" similarity) and their "function" similarity (i.e., similarity of rule or usage). In this paper, we used the "mono" similarity matrix[33] that combines the two measures. By using this matrix, we measured the similarity between the vector of each essay and each of the personality factors. The result is a similarity score. The higher the score the closer the essay is to the personality vector. Next we classified each similarity score as follows: For each similarity measure (e.g., the similarity between the essay and the Paranoid vector) we found the .75 percentile, i.e., the score above which only 25% of the subjects scored. Essays that scored higher than the .75 percentile score were classified as "1" and the rest as "0". The output of this procedure was, that for each essay/subject, we had nine binary scores indicating whether the essay/subject could be characterized by the PD or not. These scores are titled the *PDs scores*.

Procedure. To test hypotheses 1–3, we used the Mann-Whitney U Test with the personality vector (e.g., E versus non-E) as the grouping variable, and the PDs scores as the test variables. For example, we compared the PDs scores of Neurotics vs. non-Neurotics. Given the ordinal scale of the test variables and the unique test ranking procedure, we considered it appropriate for our data analysis. Moreover, in the analysis, we applied a Monte Carlo Estimate to gain an unbiased estimate of the exact level of significance, by repeatedly sampling from our dataset. Specifically, we used a *Monte Carlo procedure with 10,000 samples and a 95% confidence interval*. All significance tests used in this study were one-tailed.

For the classification task, we used a tree-based classification model. As explained by Cosma Shalizi, "The basic idea is very simple. We want to predict a response or class Y from inputs X1;X2; : : :Xp. We do this by growing a binary tree. At each internal 1 node in the tree, we apply a test to one of the inputs. Depending on the outcome of the test, we go to either the left or the right sub-branch of the tree. Eventually we come to a leaf node, where we make a prediction. This prediction aggregates or averages all the training data points that reach that leaf" [www.stat.cmu.edu/~cshalizi/350/lectures/22/lecture-22.pdf]. A layman-friendly explanation of decision tree-based classification appears in Wikipedia [http://en.wikipedia.org/wiki/Classification_and_regression_tree]. Specifically, we used the Classification and Regression Tree (CRT) model, which splits the data into segments that are as homogeneous as possible with respect to the dependent variable. This is a common machine-learning approach to classification[35], which is detailed in[36]. We used the factor value (e.g., Neurotic vs. non-Neurotic) as a dependent variable, and the nine PDs scores and the ten five factor scores gained through our vectorial analysis as independent continuous variables. For example, we attempted to classify our subjects into Neurotics and non-Neurotics by using the above set of vectors. Overall, we ran the procedure five times, once for each of the five personality factors. For improving the validity of our results, we have used a ten-fold cross validation procedure also known as *rotation estimation*. One round of cross-validation involves partitioning a sample of data into complementary subsets, performing the analysis on one subset (called the *training set*), and validating the analysis on the other subset (called the *validation set* or *testing set*). The non-expert may find an easy explanation of this procedure in Wikipedia. The classification procedure produced a predicted value for each subject, for instance whether s/he is Neurotic or not. These predictions were cross-tabbed against the real observed score, and a Chi Test was used to measure the significance of our classification model.

Table 9 | Percentage of subjects in each of the personality dimensions (N = 2468)

	Personality Dimensions				
	N	E	O	A	C
0	50	48	49	47	49
1	50	52	51	53	51

1. Celli, F., Pianesi, F., Stillwell, D. & Kosinski, M. The Workshop on Computational Personality Recognition: Shared Task. AAAI Technical Report WS-13-01, Boston, USA: AAAI Press (2013 July 11).
2. Dixon, N. You are what you tweet. *Interactions* **20**, 48–52 (2013).
3. Farnadi, G., Zoghbi, S., Moens, M. F. & De Cock, M. Recognizing personality traits using Facebook status updates. Paper presented at the Proceedings of the Workshop on Computational Personality Recognition (WCPR 13).The 7th international AAAI conference on weblogs and social media (ICWSM13): Boston USA: AAAI Press (2013 July 11) [http://clic.cimec.unitn.it/fabio/wcpr13/farnadi_wcpr13.pdf].
4. He, Q., Veldkamp, B. P. & de Vries, T. Screening for posttraumatic stress disorder using verbal features in self narratives: A text mining approach. *Psychiat. Res.* **198**, 441–447 (2012).
5. Kosinski, M., Stillwell, D. & Graepel, T. Private traits and attributes are predictable from digital records of human behavior. *P. Natl. Acad. Sci. USA.* **110**, 5802–5805 (2013).
6. Neuman, Y., Assaf, D. & Cohen, Y. Automatic identification of themes in small group dynamics through the analysis of network motifs. *Bull.Men. Cli.* **76**, 53–68 (2012).
7. Pianesi, F. Searching for Personality [Social Sciences]. *IEEE Signal Proc. Mag.* **30**, 146–158 (2013).
8. Qiu, L., Lin, H., Ramsay, J. & Yang, F. You are what you tweet: Personality expression and perception on Twitter. *J. Res. Pers.* **46**, 710–718 (2012).
9. Zhou, M. X., Wang, F., Zimmerman, T., Yang, H., Haber, E. & Gou, L. Computational discovery of personal traits from social multimedia. Paper presented at the IEEE International Conference on Multimedia and Expo Workshops (ICMEW) San Jose, California, USA (2013, July 15–19).
10. 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).
11. Clark, S. Vector space models of lexical meaning. Handbook of Contemporary Semantics–second edition [Lappin, S. & Fox C. (eds)] (Wiley-Blackwell, New York) In press, http://www.cl.cam.ac.uk/~sc609/pubs/sem_handbook.pdf.
12. Turney, P. D. & Pantel, P. From frequency to meaning: Vector space models of semantics. *J. Artif. Intell. Res.* **37**, 141–188 (2010).
13. McWilliams, N. *Psychoanalytic diagnosis* (Guilford Press, New York, 1994).
14. Millon, T., Millon, C. M. & Meagher S. [Brief Description of the Fourteen Personality Disorders of *DSM-III, DSM-III-R*, and *DSM-IV*.] *Personality Disorders in Modern Life* (John Wiley & Sons Inc., New Jersey, 2004).
15. Freeman, D. Suspicious minds: the psychology of persecutory delusions. *Clin. Psychol. Rev.* **27**, 425–457 (2007).
16. John, O. P. & Srivastava, S. [The Big Five trait taxonomy: History, measurement, and theoretical perspectives] [102–138] [Pervin L. A. & John O. P. (eds.)] *Handbook of Personality: Theory and Research* (University of California Press, Berkeley, CA. 1999).
17. Goldberg, L. R. The development of markers for the Big-Five factor structure. *Psychol. Assessment* **4**, 26. (1992).
18. Trapnell, P. D. & Wiggins, J. S. Extension of the interpersonal adjective scales to include the Big Five dimensions of personality. *J. Pers. Soc. Psychol.* **59**, 781 (1990).
19. Wiggins, J. S., Trapnell, P. & Phillips, N. Psychometric and geometric characteristics of the Revised Interpersonal Adjective Scales (IAS-R). *Multivar. Behav. Res.* **23**, 517–530 (1988).
20. McCrae, R. R. & John, O. P. An introduction to the five-factor model and its applications. *J. Pers.* **60**, 175–215 (1992).
21. McCrae, R. R. & Costa Jr, P. T. *Personality in adulthood: A five-factor theory perspective.* (Guilford Press, New York, 2012).
22. McCann, S. J. Suicide, big five personality factors, and depression at the American state level. *Arch. Suicide Res.* **14**, 368–374 (2010).
23. Malouff, J. M., Thorsteinsson, E. B. & Schutte, N. S. The relationship between the five-factor model of personality and symptoms of clinical disorders: a meta-analysis. *J. Psychopathol. Behav.* **27**, 101–114 (2005).
24. Kotov, R., Gamez, W., Schmidt, F. & Watson, D. Linking "big" personality traits to anxiety, depressive, and substance use disorders: A meta-analysis. *Psychol. Bull.* **136**, 768 (2010).
25. Koorevaar, A. M. L. *et al.* Big Five personality and depression diagnosis, severity and age of onset in older adults. *J. Affect. Disorders* **151**, 178–185 (2013).
26. Saulsman, L. M. & Page, A. C. The five-factor model and personality disorder empirical literature: A meta-analytic review. *Clin Psychol Rev.* **23**, 1055–1085 (2004).
27. 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).

28. Neuman, Y. *et al.* Proactive screening for depression through automatic and metaphorical text analysis. *Artif. Intell. Med.* **56**, 19–25 (2012).
29. Pennebaker, J. W. & King, L. A. Linguistic styles: language use as an individual difference. *J. Pers. Soc. Psychol.* **77**, 1296 (1999).
30. John, O. P., Donahue, E. M. & Kentle, R. L. *The Big Five Inventory—Versions 4a and 4b.* (Technical Report) (Berkeley: Institute for Personality and Social Research, University of California, 1991).
31. Davies, M. The 385+ million word corpus of contemporary American English (1990–2008+) *Design, architecture, and linguistic insights. Int. J. Corp. Ling.* **14**, 159–190 (2009).
32. Toutanova, K., Klein, D., Manning, C. D. & Singer, Y. (2003, May). Feature-rich part-of-speech tagging with a cyclic dependency network. Paper presented at the the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology Association for Computational Linguistics (HLT-NAACL) Edmonton, Alberta, Canada. Stroudsburg, PA, USA: Association for Computational Linguistics (2003, May).
33. Turney, P., Neuman, Y., Assaf, D. & Cohen, Y. Literal and metaphorical sense identification through concrete and abstract context. Paper presented at the Conference on Empirical Methods in Natural Language Processing. Edinburgh, Scotland, UK. (2011, July 27–31).
34. Turney, Peter D. Domain and function: A dual-space model of semantic relations and compositions. *arXiv preprint arXiv*: (2013).
35. Loh, W. Y. Classification and regression trees. *WIREs DMKD.* **1**, 14–23 (2011).
36. Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. *Classification and regression trees* (Chapman & Hall/CRC, New York, 1984).

Author Contributions

YN is responsible for all aspects to exclude the programming part which is the contribution of YC.

Additional information

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Neuman, Y. & Cohen, Y. A Vectorial Semantics Approach to Personality Assessment. *Sci. Rep.* **4**, 4761; DOI:10.1038/srep04761 (2014).