Predicting Personality Traits from Spontaneous Modern Greek Text: Overcoming the Barriers

Vasileios Komianos, Eleni Moustaka, Maria Andreou, Eirini Banou, Sofia Fanarioti, Katia L. Kermanidis

Predicting Personality Traits From Spontaneous Modern Greek Text: Overcoming the Barriers

Vasileios Komianos, Eleni Moustaka, Maria Andreou, Eirini Banou, Sofia Fanarioti and Katia Kermanidis

Ionian University, Department of Informatics, 7 Tsirigoti Square, Corfu

Abstract. The present work aims at identifying relations between the morphosyntactic and semantic properties of an author's writings and his/her personality traits. Machine learning schemata are used to classify an author according to the values of the Big Five traits, or predict their numerical value. Unlike related work, the current approach focuses on Modern Greek text, and makes use of limited data and resources, available at its disposal. Meta-learning and synthetic oversampling help overcome the small dataset and its imbalanced class distribution.

Keywords: Big Five Traits**,** Modern Greek, Machine Learning, Personality Prediction, Bagging, class imbalance, SMOTE

1 Introduction

There is a Greek expression that says "Show me your friend and I'll tell you who you are", and means that people tend to make friendship with persons who reflect to their personality. In a similar way, the present study investigates a new concept: "Show me what you are writing and I'll tell you who you are".

People, depending on their habits, differ from each other in the way they think, feel or act. These differences not only reflect in what people think, feel or do, but also in the way they say or write those thoughts, feelings and actions [1]. According to Tausczik and Pennebaker [2], the words we use in our daily life reflect who we are. Language is the common way to translate our thoughts and emotions in a form that people around us will understand. Language and words are the medium through which psychologists attempt to understand human beings [2].

The present work is based on the assumption that a person's writings may contain information, about his/her personality [3]. This information may concern the writer's gender [4], age [5], education, mental or physical health. While most of these research approaches deal with the English language, the present work aims at identifying characteristics related to this information in the Modern Greek language, i.e. at identifying writings' characteristics that are related to the personality's Big Five trait taxonomy [6].

The need for a consistent method to easily describe personality properties has led many researchers to efforts that involve categorizing many aspects of human

personality in a few distinct categories. A result of these efforts is the Big Five trait taxonomy [6]. The word 'Big' is used not to describe the importance of these five categories over others but to emphasize their broadness. According to the Big Five trait taxonomy there are five factors: extraversion, agreeableness, conscientiousness, neuroticism and openness. It has been quite useful because it offers a simple and concise description of a person's personality. Methods like this are often used by recruiters and human resource managers in order to help them find the appropriate employee for each job, in education by teachers trying to approach their students, even in advertising via profiling of the consumers for an individualized approach. One of the first efforts to correlate the Big Five factors with textual characteristics took place in 1999 by Pennebaker et al [7].

This paper describes an approach to automatically identify an author's Big Five traits, based on the properties of his/her text. The present work manages to address a series of interesting research challenges/innovations, namely:

- the Modern Greek language, which imposes certain idiosyncrasies, i.e. certain morphosyntactic features than need to be taken into account, not present in English
- the lack of sophisticated resources available in other languages, like LIWC
- the small amount of available data, compared to the data used in previous related work
- the imbalanced distribution of the class values (Big Five trait values) in the data.

The rest of this paper is organized as follows. Section 2 reviews related work for the task at hand. Section 3 describes in detail the proposed methodology. Section 4 gives an overview of the experimental setup. Section 5 presents the results from the research and the paper concludes in section 6.

2. Related Work

The authors in [8] and [9] discovered that the way people speak is related to their physical and mental health status. They developed a content-analysis method to track Freudian themes in text samples. Judges, then, evaluated each phrase to determine the degree it might reflect one or more themes related to anxiety (e.g., death, castration), hostility toward self or others, and various interpersonal and psychological topics.

There have been previous attempts to create software for this purpose [10], [11] but the most interesting was a word-based approach, the General Inquirer that has been developed by Stone et al [12]. The authors in [13] used word-based approaches on social sciences and psychology. Until then most studies had focused on psychiatry.

Pennebaker et al [7] used a word count program, named Linguistic Inquiry and Word Count (LIWC) [14]. LIWC is based on a dictionary used to find specific words in the text and counts their frequency of use. These words have specific attributes or may belong to special categories such as words describing happiness. Despite the appeal of computerized language measures they are still at early stages because they ignore context, irony, sarcasm, and idioms.

Mairesse et al [15] worked on the automatic recognition of personality traits. In their article, they reported experimental results for recognition of all Big Five traits, in both conversation and text. It was noted that for some personality factors, any type of

statistical model performs better than the baseline, but ranking models perform best overall.

Recently Roshchina et al [16] proposed TWIN (Tell me What I Need). This is a Personality-based Recommender System that analyzes the textual content of the reviews and estimates the personality of the user according to the Big Five model. The reviews were taken from the "Trip Advisor" site and they supposed that the similarity between people can be established by analyzing the content of the words they are using. They used 4 data mining algorithms: linear regression, M5 model tree, M5 regression tree and support vector machines for regression. They concluded that the M5 regression tree performs better than the other 3 algorithms.

3. Methodology

For the purpose of this work, 382 essays were written by 382 different authors, of various genders, ages and educational backgrounds as shown in Table 1. Regarding the educational background, 17 of the participants had graduated Primary school, 51 are Secondary school graduates, and 313 are higher education graduates. Five texts were excluded because of not meeting the criteria, i.e. they consisted only of emoticons or lyrics. The procedure was divided into two parts. During the first part the authors were asked to write spontaneously and continuously for 20 minutes about their thoughts and feelings. During the second part they had to complete 44 scaled statements that concern their perception about their self in a variety of situations. The main source of the questionnaire came from the research of John et al [6]. They used the specific questionnaire to conclude these five categories. The questionnaire was translated, and it was considered appropriate to include demographic information of the participants, while retaining their anonymity. In order to collect the data, two different questionnaire forms were created; an electronic form[1] and a handwritten form. Thereby, distant authors were given the opportunity to participate.

Table 1. Age and gender of the participants

Age Group	Male Participants	Female participants	Total
15-20	42	101	143
21-30	61	68	129
31-40	32	35	57
41-50	16	20	36
51-60	1	2	3
61-70	1	1	2
71+	2	0	2
Total	**155**	**227**	**382**

Since the collection of the essays was accomplished, the next step was to analyze the texts and create a dictionary containing the grammatically and semantically annotated words of the essay text. Text analysis software was developed specifically

[1]https://sites.google.com/site/dimsc2011/

for this purpose. To perform morphological tagging, a Part-of-Speech (POS) tagger was used [17]. The dictionary words belonged to 14 POS categories, as is shown in Figure 1. The first 10 represent the valid Modern Greek POS categories that the tagger can recognize and analyze. The last 4 (Misspelled, Number, Acronym/Abbreviation and Foreign Word) were added in order to categorize some words, that could not be analyzed and categorized by the POS tagger. The dictionary has 91 attributes for every word, representing the POS category and the conceptual or emotional category the word belongs to.

As this is the first time a research of this kind is being carried out for Modern Greek, there is no previous knowledge about the attributes that can reveal personality information. In this work the widest possible scope is covered, fifty-two of the attributes are related to part of speech, thirty-six to describe the conceptual/emotional category, one attribute states if the word belongs to a foreign language, one is used to state misspelled words and the last one for numbers. The attributes are shown in Tables 7, 8, 9 and 10. These attributes have values TRUE/FALSE depending on whether the word belongs to the corresponding category. In case the same word shows up with more than one POS, all instances are counted and the most frequent POS is being accepted. The dictionary consists of 8936 words.

Finally, the dictionary words had to be enriched with conceptual and emotional information, using an annotation tool developed for this purpose (Figure 1). This tool gave the language experts the ability to supplement the dictionary with the additional information related to concept or emotion, for every word of the dictionary the annotation tool provides a set of choices corresponding to the dictionary attributes that the user can choose. The software developed for this work, adapts the word count approach like LIWC in a less complicated way. Unlike LIWC it does not support Cognitive Processes features and Relativity of time.

Although the annotation tool was more than helpful for simplifying the semantic annotation of the text, the task was quite challenging. A team of four language experts manually tagged every word, into one of the special categories, shown in Figure 3. There are 22 categories and 13 subcategories in total, chosen, primarily, according to related research [15], and secondly, according to the idiosyncratic properties of Modern Greek. Unlike English, were verb morphology may be determined by other words surrounding it, in Modern Greek number and person of verb is determined by the verb itself.

The tagging process achieved an inter-annotator agreement of 91%. Explanatory discussions took place to clarify cases of disagreement. At the end of the annotation process, each word in the dictionary is tagged with its POS and belongs to none or more of the special categories.

After completing the dictionary, essays were analyzed by the text analysis tool, in order to calculate the characteristics we are interested in (i.e. the categories in Tables 5 - 11) and create the final learning datasets. A total of five datasets were formed – one for every Big Five trait. Each instance of the dataset represents one essay, and consists of the values corresponding to the text characteristics, plus the Big Five trait value resulting from the personality questionnaire.

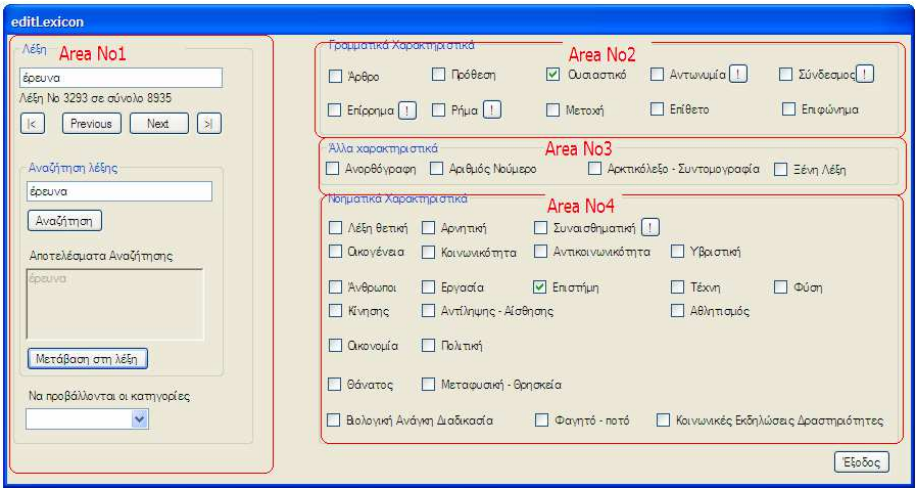

Figure 1. The annotation tool. In Area No1 the user can search words for editing. In Area No2 the user can choose the POS for the given word. In Area No3 there are checkboxes for words being misspelled, numbers, abbreviations or foreign. In Area No4 the user can identify a word as conceptual or emotional.

4. Experimental Setup

Two sets of experiments were run: one for regression, i.e. for the numerical prediction of the Big Five trait values, and one for binary classification, i.e. categorizing each essay into one of two binary classes (TRUE/FALSE) regarding the Big Five traits. The classifier employed was Quinlan's C4.5 [18]. For regression, the M5P regression tree algorithm was chosen. Trees are powerful predictors, and incorporate valuable and easily understandable knowledge related to the application domain.

When treating personality modeling as a regression problem, regression models can replicate the actual scalar values seen in the personality ratings. Data regression is more appropriate for fine-grained recognition. Regression trees tend to outperform more complex techniques, according to Oderlander and Nowson (2006).

The trait values resulting from the questionnaire are numerical, and, depending on their value, they indicate how well (or not) each trait describes the person. The greater the value of a trait, the better the trait describes the person; the lower the value, the less appropriately the trait describes a person. These values are calculated from the questionnaire answers [6], and constitute the regression class value.

For the classification experiments, the transformation of the numerical values into binary (TRUE/FALSE) has been performed by calculating (from the questionnaire) the lowest and highest value each trait could theoretically score, and taking the center value (threshold). "FALSE" was assigned to values lower than the threshold, and

"TRUE" was assigned to the ones greater than the threshold. The corresponding ranges are shown in the first three columns of Table 2.

Table 2. The Big Five Trait Taxonomy

Feature	Range of values	FALSE	TRUE	TRUE	FALSE
Extraversion	-10,…,+22	-10,…,+6	+7,…,+22	278	101
Agreeableness	-15,…,+21	-15,…+3	+4,…,+21	357	22
Conscientiousness	-15,…,+21	-15,…,+3	+4,…,+21	306	73
Neuroticism	-10,…,+22	-10,…+6	+7,…,+22	179	200
Openness	-2,…,+38	-2,…,+17	+18,…,+38	336	43

The personality trait values were thereby converted to binary variables according to the separation of values between TRUE and FALSE for every feature. After having converted the classes, it was noticed that the class distribution in the data was highly imbalanced for four Big Five traits. The number of instances belonging to each class is shown in the last two columns of Table 2.

The class imbalance tends to bias the classifier towards the majority class (TRUE), and consequently, classification accuracy for the minority class (FALSE) instances drops significantly. In order to reduce the influence of the imbalance between classes, the SMOTE oversampling technique was employed. SMOTE [20] increases the minority class instances by creating synthetic instances.

5. Results

Looking at Table 3, it is evident that predicting the numerical Big Five trait value is a difficult task. Stand-alone regression tree algorithms, like M5P, only marginally and sparsely manage to drop the relative absolute error rate (the ratio between the model's prediction error and the baseline error produced by returning the mean value of the big five trait values in the training set) below the 100% baseline. Taking into account the limited size of the dataset, the limited size of the essay lengths (the length varied between 1 and 664 words, with a mean value of 110.3 words and a standard deviation of 97.9 words), and the less sophisticated feature set, compared to datasets used in previous approaches (e.g. LIWC), these results are quite understandable.

To overcome the performance barriers imposed by the aforementioned problems, experiments with boosting were run, in an attempt to force the regression tree base learner to learn from previous mistakes, and improve its prediction accuracy in an iterative running scheme. Bagging [19] was chosen as the metalearning scheme and various sample bag sizes (100%, 80% and 60% of the initial training set size) were experimented with. Results improve slightly, especially for neuroticism and agreeableness. Comparing the results with the ones reported in [15] (the relative absolute error ranging from 93.3% for openness to 99.2% for extraversion), and taking into account that they use an essay corpus that consists of five times more essays, and, on average, seven times longer essays than the corpus used in the current approach, the regression error reported herein is quite noteworthy.

Table 4 shows the results of the second set of experiments, i.e. the classification tests. Using C4.5, precision and recall are unexpectedly high for the positive class (TRUE), and extremely low for the negative class (FALSE). This is attributed to the large degree of class imbalance in the data. The low number of negative instances, compared to that of the positive instances leads to the bias of the classifier towards assigning an instance to the positive class. SMOTE [20], a technique for increasing the number of the minority class instances in the data by synthetic oversampling, was chosen as an objective way to face the second barrier, namely the class imbalance problem. As a result, negative instances are doubled in number. Results improve significantly, reaching 60% precision and 57% recall for the negative class for openness. The only trait that is more uniformly distributed and not governed by class imbalance is neuroticism. Compared to related work, Mairesse et al. [15] report classification accuracy results between 51.1% and 62.1% (54.4% for decision trees). Evidently, binary classification, a more straightforward approach compared to the fine-grained nature of the regression task, leads to promising results.

Table 3. Regression (Relative Absolute Error)

Trait	M5P	Bagging (60%)	Bagging (80%)	Bagging (100%)
Agreeableness	99.79%	99.56%	99.38%	99.87%
Conscientiousness	100.79%	99.88%	99.18%	99.82%
Extraversion	99.91%	99.74%	100.00%	100.32%
Neuroticism	101.16%	99.45%	98.92%	99.04%
Openness	100.59%	100.38%	100.54%	100.29%

Table 4. Classification

	Classifier	Precision	Recall	Class
Agreeableness	J48	0.945	0.955	TRUE
		0.111	0.091	FALSE
	J48 SMOTE	0.934	0.944	TRUE
		0.5	0.455	FALSE
Conscientiousness	J48	0.803	0.905	TRUE
		0.147	0.068	FALSE
	J48 SMOTE	0.757	0.794	TRUE
		0.519	0.466	FALSE
Extraversion	J48	0.744	0.723	TRUE
		0.294	0.317	FALSE
	J48 SMOTE	0.658	0.637	TRUE
		0.521	0.545	FALSE
Neuroticism	J48	0.478	0.542	TRUE
		0.534	0.47	FALSE
	J48 SMOTE	0.721	0.729	TRUE
		0.505	0.495	FALSE
Openness	J48	0.892	0.914	TRUE
		0.171	0.14	FALSE
	J48 SMOTE	0.891	0.902	TRUE
		0.598	0.57	FALSE

Classification results revealed the most important features for each of the Big Five factors. Agreeableness, according to the decision trees, is affected by the presence of words related to conceptual category and belonging to the subcategories of politics, abusive, finance and the emotional category of pain/anguish. Conscientiousness appears to be dependent on conjunctions of separation and the length of sentences in text. Extraversion is affected by the presence of participles, by the presence of social words expressing exclusion and by emotional words expressing enthusiasm. Neuroticism is affected by verbs and adjectives, emotional words related to pain/anguish and emotional words expressing hope. Openness according to the results is affected by words related to sports, science, un/sociability, words related to human beings and by words expressing emotions.

6. Conclusion

An approach to automatically identify a writer's personality traits, based on the linguistic properties of his/her writings was described in the present work. Classification and regression experiments were run for predicting Big Five trait value. The approach focuses on Modern Greek text, and deals with several challenges, like the language itself, the small size of available data, the lack of sophisticated dictionaries and resources available in other languages. Bagging helps overcome the limited data, while oversampling manages to address the class imbalance problem. An interesting future research objective would be the creation and use of more sophisticated resources for the semantic annotation of words and phrases in the text. Syntactic information, i.e. phrase frequencies, structure and constituent orderings, could also prove quite useful. Other learning algorithms, not sensitive to limited and high-dimensional data are also worth exploring.

References

1. Yarkoni, T.: Personality in 100,000 Words: A large-scale analysis of personality and word use among bloggers, Journal of Research in Personality, Volume 44, Issue 3, June 2010, Pages 363-373, ISSN 0092-6566, 10.1016/j.jrp.2010.04.001, (2010)
2. Tausczik, Y.R., Pennebaker, J.W.: The Psychological Meaning of Words: LIWCand Computerized Text Analysis Methods. Journal of Language and Social Psychology 29(1) 24–54, (2010)
3. Argamon, S., Dhawle, S., Koppel, M., Pennebaker, J.W.: Lexical Predictors of Personality Type , In: Joint Annual Meeting of the Interface and the Classification Society of North America, (2005)
4. Argamon, S., Koppel, M., Fine, J., Shimon, A.R.,: Gender, genre and writing style in format written texts. Text 23: 321–346, (2003)
5. Argamon, S., Koppel, M., Pennebaker, J.W., Schler J.: Automatically Profiling the Author of an Anonymous Text, Magazine Communications of the ACM, (2009)
6. John, O.P, Srivastana, S.: The Big-Five Trait Taxonomy : History, Measurement, and Theoretical Perspectives,Berkeley, CA94720-1650, (1999)

7. Pennebaker, J. W., King, L. A., Linguistic Styles: Language Use as an Individual Difference, Journal of Personality and Social Psychology, Vol. 77, No 6, pp 1296-1312, (1999)
8. Graham, D. T., Stern,J. A., Winokur, G.: Experimental investigation of the specificity of attitude hypothesis in psychosomatic disease. Psychosomatic Medicine, Journal of Biobehavioral Medicine, (1958)
9. Gottschalk, L. A., Gleser, G.: The measurement of psychological states through the content analysis of verbal behavior, University of California Press, Berkeley, (1969)
10. Weintraub, W., Verbal behavior in everyday life. New York, NY, US: Springer Publishing Co. viii 200 (1989)
11. Stiles, W. B., Describing Talk: A Taxonomy of Verbal Response Modes, (1992) , Newbury Park, CA: Sage, 1992
12. Stone, P. J., Dunphy, D. C., Smith, M. S., and Ogilvie, D. G., The General Inquirer: A Computer Approach to Content Analysis. M.I.T. Press, Cambridge, Massachusetts, (1966)
13. Peterson,C., Ulrey, L. M., Can explanatory style be scored from TAT protocols? Personality and Social Psychology Bulletin, 20, pp. 102–106, (1994)
14. Pennebaker, J. W., & Francis, M. E. Linguistic Inquiry and Word Count: LIWC. Mahwah, NJ: Erlbaum, (1999)
15. Mairesse,F., Walker, M.A., Mehl, M.R., Moore, R.K.: Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. Journal of Artificial Intelligence Research 30. 457--500 (2007)
16. Roshchina, A. and Cardiff, J. and Rosso, P., A Comparative Evaluation of Personality Estimation Algorithms for the TWIN Recommender System. In: Proc. CIKM 3rd Int. Workshop on Search and Mining User-generated Contents, SMUC-2011
17. Koleli, E.: A new Greek parts of speech tagger, based on maximum entropy classifier, Thesis, Athens University of Economics, (2011).
18. Quinlan, J.R.: C4.5: Programs for Machine Learning. San Mateo, Calif.: Morgan Kaufmann, (1993)
19. Breiman, L.: Bagging Predictors, Kluwer Academic Publishers, Boston. Manufactured in The Netherlands, (1996)
20. Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, W. Philip Kegelmeyer: SMOTE: Synthetic Minority Over-sampling Technique, Journal of Artificial Intelligence Research 16 321–357, (2002)

Appendix:

Table 5. Modern Greek Parts of Speech

List of POS				
Article	Adjective	Adverb	Conjunction	Interjection
Noun	Pronoun	Verb	Participle	Preposition

Table 6. List of pronouns

List of Pronouns				
Personal	Definite	Possessive	Interrogative	Demonstrative

Referential	Indefinite	1st person	2nd person	3rd person
Singular	Plural	-	-	-

Table 7. List of verb features

List of Verbs features					
Past	Present	Future	Active	Passive	Imperative
1st person	2nd person	3rd person	Singular	Plural	-

Table 8. List of conjunctions

List of Conjunctions features						
Complex	Separation	Antithesis	Inference	Explanation	Special	Hesitance
Time	Cause	Hypothesis	Final	Result	Comparison	

Table 9. List of adverbs

List of Adverbs					
Time	Manner	Quantity	Affirmative	Negation	Hesitance

Table 10. List of conceptual words

Conceptual words				
Positive	Negative	Abusive	Sociability	Unsociability
Politics	Biological need/process	Metaphysics/Religion	Humans	-
Sports	Science	Family	Death	
Food/Beverage	Perception/Sense	Motion	Finance	
Labor	Nature	Social Events/Activities	Arts	

Table 11. List of emotional words

Emotional words				
Self-Confidence/Courage/Boldness	Joy	Enthusiasm	Love	Hope
Shyness	Desire	Anger/Rage	Pain/Anguish	-
Anxiety/Stress	Fear	Panic	Sadness	