Using Stylometric Features to Predict Author Personality Type in Modern Greek Essays

Gagiatsou Sofia
Department of Linguistics, School of Philosophy
National and Kapodistrian University of Athens
Athens, Greece
e-mail: sgagiats@phil.uoa.gr

Markopoulos Georgios
Department of Linguistics, School of Philosophy
National and Kapodistrian University of Athens
Athens, Greece
e-mail: gmarkop@phil.uoa.gr

Mikros George
College of Humanities and Social Sciences
Hamad Bin Khalifa University
Doha, Qatar
e-mail: gmikros@hbku.edu.qa

***Abstract*—We present a research focused on the prediction of the author's personality based on natural language processing techniques applied to essays written in Modern Greek by high-school students. Each writer has been profiled by filling in the Jung Typology Test. In addition, personality prediction is being discussed under the general research framework of author profiling by examining the effectiveness of several stylometric features to predict students' personality types. The feature set we employed was a combination of the word and sentence length, the most frequent part-of-speech tags, most frequent character/word bigrams and trigrams, most frequent words, as well as hapax/dis legomena. Since personality prediction represents a complex multidimensional research problem, we applied various machine learning algorithms to optimize our model's performance after extracting the stylometric features. We compared nine machine learning algorithms and ranked them according to their cross-validated accuracy. The best results were obtained by the Naive Bayes algorithm. According to the personality classification based on the Jung Typology Test, the author's personality prediction accuracy reached 80.7% on Extraversion, 79.9% on Intuition, 68.8% on Feeling, 75.7% on Judging, according to the personality classification. The reported results show a competitive approach to the personality prediction problem. Furthermore, our research revealed new combinations of stylometric features and corresponding computational techniques, giving interesting and satisfying solutions to the problem of the author's personality prediction for Modern Greek.**

Keywords-Author profiling; stylometry; Personality prediction; Jung Typology Test; corpus processing; computational stylistics; machine learning.

I. INTRODUCTION

Authorship identification represents one of the emerging text mining fields that stands at the intersection of Machine Learning, Information Retrieval, and Natural Language Processing. Under the stylometric framework, the author's identity is a multidimensional construct based mainly on writing patterns scattered across multiple linguistic levels and expressed quantitatively. The specific research domain splits into three subdomains: attributing a text to a particular author among a finite set of authors (Authorship Attribution), attributing a text to an author that does not belong to a closed set (Authorship Verification), and specifying the author's metadata such as demographic and psychological traits of the author (Authorship Profiling), including gender, age, personality, etc.

Language as a communication mechanism denotes the diversity of every individual. Therefore, the quantitative study of linguistic features can lead to predictions regarding the individual's character. The subject of Computational Personality Prediction (CPP) through natural language processing techniques constitutes a relatively new research field with many applications.

One critical application domain of this field is Forensic Linguistics. Criminals can be identified by the way they write. Moreover, conclusions can be drawn regarding their personalities and the way they are thinking. The example of identifying students' personality that carry guns and participate in school shootings is typical [1]. CPP can highlight their psychological traits, which afterward can be exploited in the successful identification of potential perpetrators.

Apart from the obvious contribution that provides in Behavioural Psychology by connecting personality traits to human behaviour, CPP can function in many other fields as well. For instance, in the marketing domain, personality analysis of users/consumers is utilized by companies to adopt effective recruitment techniques and customer service techniques. Even in the field of human resources management, predicting the personality can affect or facilitate the selection and determine the eligibility of candidates for a particular job. Moreover, based on the user's personality, dialogic systems can be customized and brought closer to users' temperament making interaction more effective and satisfying.

Another vital analysis domain where automatic personality prediction is used is education. For example, by analyzing students' writings, talented students or students with difficulties could be recognized and thus receive adaptive teaching, addressing the appropriate cognitive level for each one or each group.

One of the most crucial issues in CPP research is developing appropriate linguistic resources enriched with the author's personality metadata. Unfortunately, these resources are challenging to create due to the increased level of manual interaction with the authors and the various privacy and ethical considerations linked with administering psychometric questionnaires to many individuals.

Another issue is that most Natural Language Processing (NLP) tools specialized in psychometric text profiling support only English. Therefore, research in other languages should be done by developing specialized dictionaries and other supporting linguistics resources from scratch (see, for example, the case of Linguistic Inquiry and Word Count-LIWC [2]).

To cover the above-mentioned research gaps, we performed the first CPP study in Modern Greek focused on high-school students. For this reason, we developed a model for predicting the personality of students based on Jung's taxonomy by analyzing their term-essays and applying various machine learning methods to rich document representation based on several stylometric features.

The rest of this paper is organized as follows. In Section II we provide an overview of previous work on personality prediction. Section III describes our researching methods. In Section IV we present the research results. We summarize our findings and discuss future work in Section V.

II. Literature Review

In this section we present the personality questionnaire used to profile the writers. Then, we review the findings of studies in the field of CPP from text.

A. *Carl Jung's and Isabel Briggs Myers' Personality Type Questionnaire*

Research in the field of personality prediction uses the Five-factor Model of Personality [3] or the Carl Jung's and Isabel Briggs Myers' personality type theory [4][5] to profile the participating authors. Therefore, the literature review presented in this section is referred to associated research, which involves the Jung Typology Test since our students have been profiled with the above-mentioned personality questionnaire.

According to Jung's theory of psychological types [4] people can be characterized by

- their preference of general attitude as Extraverted (E) or Introverted (I)
- their preference of one of the two functions of perception as Sensing (S) or Intuitive (N)
- their preference of one of the two functions of judging as Thinking (T) or Feeling (F)
- their orientation to the outer world as Judging (J) or Perceiving (P).

The Jung Typology Test classifies psychological differences of personality in four dichotomies which yield 16 different combinations or personality types. Each personality type can be assigned a 4-letter acronym of the corresponding combination of preferences: ESTJ, ISTJ, ENTJ, INTJ, ESTP, ISTP, ENTP, INTP, ESFJ, ISFJ, ENFJ, INFJ, ESFP, ISFP, ENFP, INFP.

B. *Personality Research from Text*

One of the first studies related to the author's personality prediction problem [6] defined the research problem as a text categorization task. They developed a corpus consisted of essays written in Dutch by 145 students (BA level). By selecting syntactic features and by training machine learning algorithms, the experiments in personality prediction suggested that the personality dimensions Introverted-Extraverted and iNtuitive-Sensing) can be predicted fairly accurately.

CPP studies have also expanded to social media texts with an emphasis on Twitter. A study for predicting Twitter users' personality type [7] showed that the classifier's performance on training data was quite good. Still, the classifier failed to achieve satisfying results for the test data. Another study [8] describes a logistic regression classifier's training process to predict each of the four dimensions of Jung Typology. Their results showed that linguistic features are the most predictive features. Although they succeeded in distinguishing between the personality dimensions Introverted-Extraverted and Feeling-Thinking, the other two dimensions were hard to predict.

In a study of a multilingual corpus of tweets [9], based on six languages (Dutch, German, French, Italian, Portuguese, and Spanish), the researchers extracted the most frequent word and character n-grams. Their results confirmed the findings of the previous work in that particular personality distinctions could be predicted from social media data with success. In another study focused on tweets [10], the researchers used a Naive Bayes classifier achieving 80% accuracy for Introverted-Extraverted and 60% for the other dimensions.

CPP has also being applied to languages with a different graphemic organization compared to Western languages. For example, in [11], researchers investigate the personality prediction of Twitter users in Japanese and conclude that the textual information of user behaviors is more useful than the users' cooccurrence behavior information such as the likes.

In this study, the problem of author personality prediction was treated as a set of binary classification tasks using Support Vector Machines.

III. Corpus

To test our research hypothesis, that is, whether it is possible to detect personality traits of the authors of written Modern Greek texts, it is necessary to have a corpus of Modern Greek texts and at the same time to connect each author of these texts to a psychological profile. Due to the lack of such material, the first step was to collect primary textual data from native speakers of Modern Greek. In particular, the corpus that we developed consists of essays of 198 high school students and comprises 250.000 words in total. It is balanced both in size (number of words per student) and in students' demographics (gender and age).

The participating students of three different high schools were asked to write three essays to achieve our goal, which was to collect at least 1,000 words from each student. The task was voluntary, lasted three school years, and the writing was held in the classroom. The experiment was repeated three times at different periods. The authors had to write spontaneously and continuously for 60 minutes an essay. The topics, which were not given in advance, were related to the benefit of art, the role of school in raising environmental awareness, and fighting against child labor. Finally, since the provided texts were handwritten, we had to digitize them by manually typing all of them.

IV. Methodology

The following section describes the approach used to predict personality types of students.

A. *Approach*

In the literature, two approaches stand out for an automatic author's personality prediction. In a bottom-up approach, personality labels are predicted from linguistic features that are being extracted from the corpora used using standard NLP document representations (e.g., Bag-of-Words - BoW models, etc.) [12]-[14]. In a top-down approach, instead, specialized dictionaries with custom entries are used to check the potential correlation with personality traits [15]-[17]. Both approaches have advantages, as well as restrictions. Therefore, modern techniques are oriented towards hybrid methods that combine the use of a dictionary with extended document representations trained on machine learning algorithms to exploit the best from both approaches, i.e., speed and precision, respectively. In this study, we followed the bottom-up approach, which among other benefits explained above, is also language-independent.

B. *Feature extraction*

The features used in our research can be considered as part of a broader feature set that has been characterized as stylometric, i.e., models quantitatively the text's style. The linguistic features that have been used previously as stylometric indices are numerous. They increase continuously and belong to the whole range of linguistic levels. Stylometric features are compact, information-rich signaling linguistic devices. They are correlated with many different textual functions and carry multilevel information related to both the author's identity and his/her metadata. In CPP, stylometric features can unchain the hidden link between linguistic production and its correlation with specific personality types. This is because our personality traits are defining and be defined by our socio-cognitive and psychological conditions. In that sense, aspects of our linguistic behavior reflect these personality traits indirectly and amplify them using identity perceptions.

We processed the corpus with natural language processing tools during the pre-processing phase, i.e., tokenizer, lemmatizer, and POS tagger. The output of the preprocessing phase (matrix of stylometric features) was submitted to the data mining platform Rapidminer [18]. The text preprocessing pipeline was initially applied to the original texts of the students. However, we observed that various language errors scattered across all linguistic levels inserted significant bias in the modeling process and negatively affected the prediction results. Therefore, the essays were corrected manually without loss of information on the morphosyntactic level.

We designed and ran multiple experiments in order to extract and quantify many different subsets of stylometric features from the corpus. We extracted the most frequent character bigrams and trigrams, words bigrams, and trigrams, mean word and sentence length, the occurrence frequency of content and functional words, the most and less frequent words, the occurrence frequency of parts of speech, as well as hapax and dis legomena. These features have been proven effective in the field of authorship attribution [19] and gender identification [20], and we tested them for author personality prediction as well.

C. *Classification Algorithms*

In this project, the problem of predicting the personality type was treated as a binary classification task among the four dimensions of personality, **E**xtraversion-**I**ntroversion, **S**ensing-i**N**tuition, **T**hinking-**F**eeling and **J**udging-**P**erceiving. To have a valid prediction, the extracted stylometric features matched the texts whose authors clearly belonged to a positive or negative category.

Since personality detection presents a complex classification task, we decided to use several different machine learning algorithms to find the best approach in terms of model performance. We compared nine machine learning methods, i.e., Naive Bayes, Generalized Linear Model, Logistic Regression, Fast Large Margin, Deep Learning, Decision Trees, Random Forest, Gradient Boosted Trees, Support Vector Machines and we ranked them according to their cross-validated accuracy (10-fold). We evaluated the machine learning algorithms in terms of their predictive ability using as training data the students' essays.

Their personality type had been defined before using the appropriate psychometric questionnaire.

V. Results

In this section, we present the results of the procedure that we followed to automatically classify the students' essays based on the personality type defined by the personality questionnaire they filled in. From the nine algorithms that were trained in the textual data, we present the evaluation metrics of the most effective algorithm (Table I) along with the corresponding weights that positively affected the prediction of the personality type.

Regarding the prediction of all personality types of Jung's typology, the algorithm with the best results was Naive Bayes. The accuracy rate revealed a range from 68.8% to 80.7%, with an average of 76.5%. Extraversion type was predicted with 80.7%, the Intuition type with 79.9%, the Feeling with 68.8%, and the Judging type with 75.7% [21]. A more detailed list of evaluation metrics (accuracy, precision, and recall) is reported in Table I.

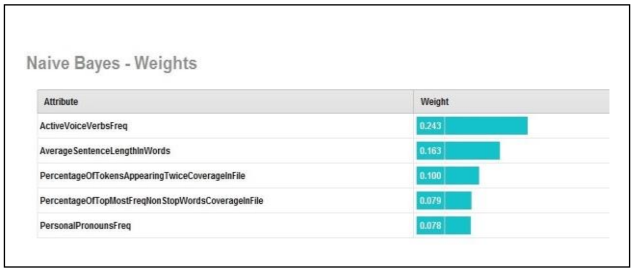
TABLE I. Naive Bayes Model Performance

Personality Type	**Naive Bayes Classifier**		
	Accuracy	***Precision***	***Recall***
Extraversion	80.7%	80.5%	100%
Intuition	79.9%	81.3%	92.6%
Feeling	68.8%	67.7%	96.7%
Judging	75.7%	76.2%	95.2%

The remaining algorithms that were trained in the corpus produced the following results in terms of classification accuracy: Regarding the Extraversion type, the Generalized Linear Model, Logistic Regression, Fast Large Margin, Deep Learning, Decision Trees, Random Forest and Gradient Boosted Trees algorithms have the same percentage of accuracy being 80.0% and the Support Vector Machine algorithm has 79.0%. The Intuition type was predicted with 75.0% by Gradient Boosted Trees algorithm, with 71.9% by Deep Learning and with 71.7% by Generalized Linear Model and Logistic Regression. For the Feeling type, the Decision Tree algorithm exhibits the second best performance with 63.2%, Random Forest being in third position with 63.1% and the next best result was 63% using Gradient Boosted Trees. The algorithms with the best performance for the Judging type were Support Vector Machine, Fast Large Margin and Deep Learning with calculated accuracies of 71.1%, 71.0% and 70.3% respectively.

The study aimed to classify the essays of the students in personality types by using stylometric indices. Therefore, we had to check whether and which of these features are the most useful and contribute to the prediction accuracy of the algorithm. For this reason, we extracted the weights from the Naive Bayes model that measure the importance of each stylometric feature to the classification decisions of the algorithm for each personality type separately.

For Extraversion (Figure 1), the use of verb types in active voice had a significant impact. In addition, the mean length of the sentence in words of all sentences, the words that occur only twice in one text, the most frequent content words, and finally the personal pronouns complete the list with the five most important stylometric features.

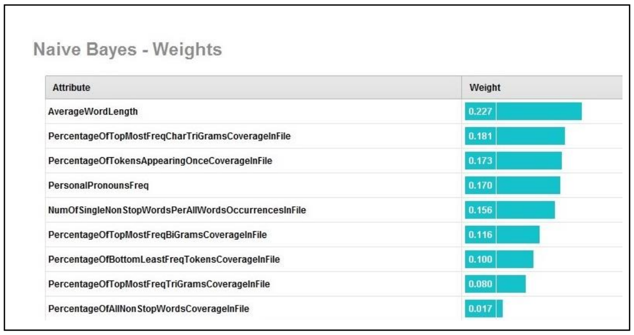


Naive Bayes - Weights

Attribute	Weight
ActiveVoiceVerbsFreq	0.243
AverageSentenceLengthInWords	0.163
PercentageOfTokensAppearingTwiceCoverageInFile	0.100
PercentageOfTopMostFreqNonStopWordsCoverageInFile	0.079
PersonalPronounsFreq	0.078

Figure 1. Weights for Extraversion.

Figure 2 depicts the prediction ability of the stylometric features for Intuition used by the algorithm. The word's mean length in characters had the most significant impact. The features that follow are the most frequent trigrams of characters, the hapax legomena, the personal pronouns, the content words, the most frequent word bigrams, the rarest words, the most frequent word trigrams, and all content words.



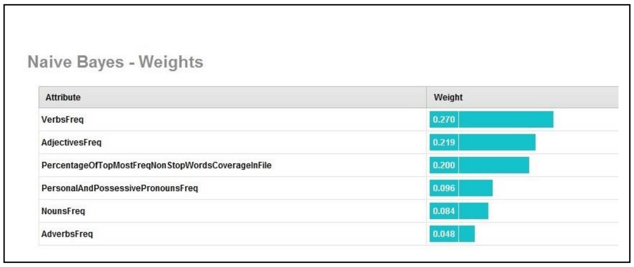
Naive Bayes - Weights

Attribute	Weight
AverageWordLength	0.227
PercentageOfTopMostFreqCharTriGramsCoverageInFile	0.181
PercentageOfTokensAppearingOnceCoverageInFile	0.173
PersonalPronounsFreq	0.170
NumOfSingleNonStopWordsPerAllWordsOccurrencesInFile	0.156
PercentageOfTopMostFreqBiGramsCoverageInFile	0.116
PercentageOfBottomLeastFreqTokensCoverageInFile	0.100
PercentageOfTopMostFreqTriGramsCoverageInFile	0.080
PercentageOfAllNonStopWordsCoverageInFile	0.017

Figure 2. Weights for Intuition.

The stylometric features that affected the result of the classification of the essays in terms of Feeling are the verbs, the adjectives, the most frequent content words, the personal and the possessive pronouns, the nouns, and the adverbs (Figure 3).

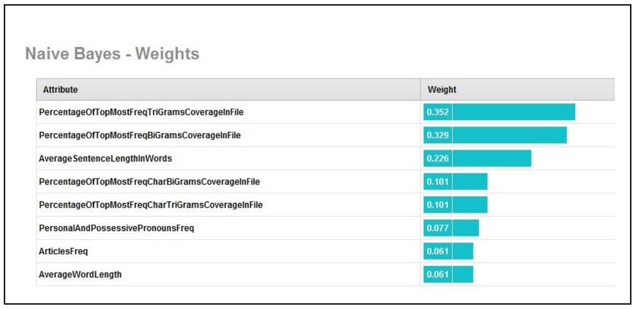
Finally, in Figure 4, the eight stylometric features that contributed to the prediction of the Judging type were in descending order: The most common word trigrams, the most common word bigrams, the mean length of the sentence in words, the most common character bigrams and the most common character trigrams with the same percentage, the personal and possessive pronouns, the articles, and the mean length of the word in characters.



Naive Bayes - Weights

Attribute	Weight
VerbsFreq	0.270
AdjectivesFreq	0.219
PercentageOfTopMostFreqNonStopWordsCoverageInFile	0.200
PersonalAndPossessivePronounsFreq	0.096
NounsFreq	0.084
AdverbsFreq	0.048

Figure 3. Weights for Feeling.



Naive Bayes - Weights

Attribute	Weight
PercentageOfTopMostFreqTriGramsCoverageInFile	0.352
PercentageOfTopMostFreqBiGramsCoverageInFile	0.329
AverageSentenceLengthInWords	0.226
PercentageOfTopMostFreqCharBiGramsCoverageInFile	0.101
PercentageOfTopMostFreqCharTriGramsCoverageInFile	0.101
PersonalAndPossessivePronounsFreq	0.077
ArticlesFreq	0.061
AverageWordLength	0.061

Figure 4. Weights for Judging.

The most important features extracted from the model vary considerably for each personality type. Therefore, we can infer that each type is based on different combination of linguistic features and these subsets are different between the different personality types.

It also becomes clear that the predictive accuracy of the proposed classification model is high compared to the existing literature on the field of personality prediction. Regarding Jung's Typology Test, we got an average accuracy of 76.5%, compared to the 68.62% reported for Dutch [6]. The other studies mentioned [7]-[11] implemented machine learning techniques in textual data that were retrieved from social media. Therefore, their results involve research with textual data from adults written under different circumstances and in a different language.

VI. Conclusion and Future Work

To summarize, in this paper, we presented the results of our research in the field of personality prediction. We applied CPP for the first time in texts written by high-school students, making our dataset unique. Our results confirmed our initial research hypothesis that stylometric features could be used as reliable prediction indices for the author's psychological profile. Our findings further support the latent link of personality traits with a wide array of linguistic behaviour aspects. Different personality types correlate with different stylometric features that belong to different linguistic levels. Therefore, the personality prediction through text demands a highly dynamic feature set to capture the widest possible spectrum of linguistic structures.

To this direction, future research will employ experimentation with more linguistic features. Furthermore, we plan to localize in Modern Greek well-known psychometric lexicons (e.g., LIWC) and use them complementing our feature sets.

References

[1] Y. Neuman, D. Assaf, Y. Cohen, and L. J. Knoll, "Profiling school shooters: Automatic text-based analysis," Front. Psychiatry, vol.6, p.86, 2015.

[2] J.W. Pennebaker, R.L. Boyd, K. Jordan, and K. Blackburn, "The Development and Psychometric Properties of LIWC2015". Austin, TX: University of Texas at Austin. (www.LIWC.net), 2015. [retrieved: June, 2021]

[3] Jr. P. T. Costa and R. R. McCrae, "NEO-PI-R: Professional Manual," Odessa, Fla.: Psychological Assessment Resources, 1993.

[4] C. G. Jung, "Psychological Types," Princeton, New Jersey: Princeton University Press, 1971.

[5] I. Briggs Myers and P. B. Myers, "Gifts Differing: Understanding Personality Type," Mountain View, CA: Davies-Black Publishing, 1980.

[6] K. Luyckx and W. Daelemans, "Personae: A corpus for Author and Personality Prediction from Text" The Sixth International Language Resources and Evaluation Conference (LREC 2008), 28-30 May 2008, pp. 2981-2987.

[7] D. Brinks and H. White, "Detection of Myers - Briggs Type Indicator via Text based Computer-mediated Communication," CS 229 Machine Learning Projects, Stanford, 2012.

[8] B. Plank and D. Hovy, "Personality Traits on Twitter-or-how to get 1,500 Personality Tests in a Week" The Sixth Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Association for Computational Linguistics, 2015, pp. 92-98.

[9] B.Verhoeven, W. Daelemans, and B. Plank, "Creating TwiSty: Corpus Development and Statistics," Computational Linguistics and Psycholinguistics Research Center CLiPS Technical Report Series, University of Antwerp, Belgium, CTRS-006, 2016.

[10] L. C. Lukito, A. Erwin, J. Purnama, and W. Danoekoesoemo, "Social Media User Personality Classification using Computational Linguistic" The Eighth International Conference on Information Technology and Electrical Engineering, Oct. 2016, pp. 1-6.

[11] K. Yamada, R. Sasano, and K. Takeda, "Incorporating Textual Information on User Behavior for Personality Prediction" The 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, Jul.-Aug. 2019, pp. 177-182.

[12] J. Oberlander and S. Nowson, "Whose Thumb is it anyway? Classifying Author Personality from Weblog Text" The 44th Annual Meeting of the Association for Computational Linguistics ACL, Jul. 2006, pp. 627-634.

[13] F. Iacobelli, A. J. Gill, S. Nowson, and J. Oberlander, "Large Scale Personality Classification of Bloggers" The Fourth International Conference on Affective Computing and Intelligent Interaction, 2011, Heidelberg: Springer-Verlag, pp. 568-577.

[14] Y. Bachrach, M. Kosinski, T. Graepel, P. Kohli, and D. Stillwell, "Personality and Patterns of Facebook Usage" The Fourth Annual ACM Web Science Conference, 2012, pp. 36-45.

[15] S. Argamon, S. Dhawle, M. Koppel, and J. W. Pennebaker, “Lexical Predictors of Personality Type” The Joint Annual Meeting of the Interface and the Classification Society of North America, 2005, pp. 1-16.

[16] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore, “Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text,” Journal of Artificial Intelligence Research, vol. 30, pp. 457-500, 2007.

[17] J. Golbeck, C. Robles, and K. Turner, “Predicting Personality with Social Media” The 2011 Annual Conference Extended Abstracts on Human Factors in Computing Systems, pp. 253-262.

[18] I. Mierswa and R. Klinkenberg. Rapidminer Studio (9.1) [Data science, machine learning, predictive analytics]. (https://rapidminer.com). [retrieved: June, 2021].

[19] G. K. Mikros and G. Markopoulos, “Using Multiword Sequences as Features in Authorship Attribution: Experiments based on Greek Blog Texts,” In A. Christofidou (Ed.), Aspects of Corpus Linguistics: Principles, applications and challenges Vol. 14, pp. 56-67, 2017. Athens: Academy of Athens: Research Center for Scientific Terms and Neologisms.

[20] G. K. Mikros, “Authorship Attribution and Gender Identification in Greek Blogs,” In I. Obradović, E. Kelih & R. Köhler (Eds.), Selected papers of the VIIIth International Conference on Quantitative Linguistics (QUALICO), Apr. 2013, Belgrade: Academic Mind, pp. 21-32.

[21] S. Gagiatsou, “Automatic author profiling based on natural language processing techniques”, Ph.D. Thesis, National and Kapodistrian University of Athens: Greece, 2021.