Prediction of Authors' Personality Types and Traits in Modern Greek Essays Using Stylometric Features

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 study 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 two personality questionnaires, one based on the typology of Carl Jung and the other based on the Model of Five Factors. 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 in predicting the Jung's Typology types were obtained by the Naive Bayes algorithm. In contrast, for the prediction of personality features based on the Five Factors Model, the Generalized Linear Model (Binomial method) algorithm prevailed. 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. In the Big Five personality classification, the prediction accuracy reached 85.9% on Openness, 71.2% on Conscientiousness, 67.6% on Extraversion, 70.2% on Agreeableness, and 65.6% on Neuroticism. 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 author's personality prediction problem for Modern Greek.**

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

I. INTRODUCTION

Authorship identification represents one of the emerging text mining fields 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 group (Authorship Verification), and specifying the author's metadata such as demographic and psychological traits of the author (Authorship Profiling), including gender, age, personality [1], 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 think. The example of identifying students' personality that carry guns and participate in school shootings is typical [2]. CPP can highlight their psychological traits, which can be exploited in successfully identifying potential perpetrators.

Apart from the apparent contribution provided in Behavioural Psychology by connecting personality traits to human behaviour, CPP can also function in many other fields. For instance, companies utilize personality analysis of users/consumers in the marketing domain to adopt effective recruitment techniques and customer service techniques. Even in 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 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 [3]).

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 students' personality based on Jung's taxonomy and the model of Big-Five factor markers 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

This section presents the two personality questionnaires used to profile the writers (Carl Jung's and Isabel Briggs Myers' Personality Type Questionnaire and Big-Five Personality Test). Then, we review the findings of studies in the field of CPP from the text.

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

Research in the field of personality prediction uses Carl Jung's and Isabel Briggs Myers' personality type theory [4][5] or the Five-factor Model of Personality [6], which are the two most utilized personality models, to profile the participating authors. Therefore, the literature review presented in this section is referred to associated research, which involves the above-mentioned personality questionnaires since our students have been profiled with these tests.

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), which signifies the source and direction of a person's energy expression.
- their preference of one of the two functions of perception as Sensing (S) or Intuitive (N) represents how someone perceives information.
- their preference of one of the two functions of judging as Thinking (T) or Feeling (F), which describes how a person processes information.
- their orientation to the outer world as Judging (J) or Perceiving (P), which reflects how a person implements the information he/she has processed.

The Jung Typology Test classifies psychological personality differences in four dichotomies that 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. Big-Five Personality Test

One of the most widely accepted personality theories in psychology is the Five-Factor model. According to the Five-Factor Model of Personality, most human personality traits can be boiled down to five broad dimensions of personality, regardless of language or culture. There has been much research on how people describe others, and five major dimensions of human personality emerged. They are often referred to as the OCEAN model of personality because of the acronym from the names of the five dimensions. Openness to Experiences, Conscientiousness, Extraversion, Agreeableness, and Neuroticism are the five most essential personality traits [7]. More specifically:

- Openness to Experience

High scorers tend to be original, creative, curious, complex; Low scorers tend to be conventional, down to earth, have narrow interests, be uncreative.

- Conscientiousness

High scorers tend to be reliable, well-organized, self-disciplined, careful; Low scorers tend to be disorganized, undependable, negligent.

- Extraversion

High scorers tend to be sociable, friendly, fun-loving, talkative; Low scorers tend to be introverted, reserved, inhibited, quiet.

- Agreeableness

High scorers tend to be good-natured, sympathetic, forgiving, courteous; Low scorers tend to be critical, rude, harsh, callous.

- Neuroticism

High scorers tend to be nervous, high-strung, insecure, worrying; Low scorers tend to be calm, relaxed, secure, hardy.

C. *Personality and Language*

The way a person uses language as a communication code reveals much information for his/her personality. The selection of specific morphological, syntactic structures and lexical choices can indicate his/her age, gender, social class, and feelings. Moreover, we can understand whether the speaker or author of a text is extraverted, emotional, or distant. So, a critical element that needs to be examined is the relationship between personality and language.

In general, the dominant opinion is that personality affects and directs our behavior, thoughts, feelings, interpersonal relationships, and of course, language production. People speak and write in different ways, even if they want to express the same content. The language user chooses the appropriate level of speech depending on the specific instance of linguistic communication, thus shaping a personalized way of speaking or writing. Researchers in this field support that every human has a characteristic way of using the language, i.e., a kind of authorial fingerprint [8]. Since the idiolect is constructed through the selective use of specific linguistic elements and their differentiated usage frequency, we can infer that also a correlation between personality traits and language features, such as lexical categories, n-grams is evident.

The above is confirmed by current research; [9] supported that language reveals each person's temperament and investigates how it is linked with his/her linguistic individuality. [10] emphasizes that all linguistic levels (phonology, morphology, syntax, semantics, pragmatics) affect the message recipient. Research shows that personality traits impact each person’s language production [11]. [12] points out that personality is projected through language, but that personality may also become perceivable to the recipient through language. Moreover, he mentions that different personality traits affect different levels of language production. [13] talks about the psychological aspect of language and focus on the choice of words by the language user as an indicative element of its character. Social psychologists have pointed out that the use of words, intonation, accent, and other language elements reveal their social, financial, and psychological position [14].

Although we perceive the importance of the connection between language and personality traits of the speaker or the author, the field has not been studied sufficiently, as most research focuses on verbal speech and the trait of Extraversion. According to [15], this is due to paralinguistic elements of the verbal speech, such as accent and intonation, as well as the fact that speech between family members and friends from a sociolinguistic point of view offers more useful linguistic data, since it is more spontaneous. Finally, Extraversion as a trait is more easily recognizable in somebody’s speech, and therefore in combination with the above, research has focused on identifying the language features denoting this speaker’s personality trait; as a result, it has been studied more than the other traits, both in the Five-Factor model as well as in Jung’s typology. Finally, the dominant language in written data is English; this makes the comparative study of findings in other languages more difficult.

Our research attempts to cover the void in this field by creating specialized corpora and utilizing natural language processing techniques in order to research all types and all traits of both personality theories, and thus showing that the relation between language and personality can be determined computationally.

D. *Personality Research from Text*

We briefly present previous research that involves either Jung Typology Test or Big-Five Personality Test in the author’s personality prediction task from the text.

1) *Jung Typology Test*

One of the first studies related to the author's personality prediction problem [16] 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 training machine learning algorithms, the experiments in personality prediction suggested that the personality dimensions Introverted-Extraverted and Intuitive-Sensing) can be predicted accurately.

CPP studies have also expanded to social media texts with an emphasis on Twitter. A study for predicting Twitter users’ personality type [17] 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 [18] 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 successfully distinguished 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 [19], 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 [20], 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 [21], researchers investigate the personality prediction of Twitter users in Japanese and conclude that the textual information of user behaviors is more valuable 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.

2) *Big-Five Personality Test*

Another study [22], which also treated personality prediction as a classification problem, has been conducted using student essays data. The corpus consisted of essays written by 198 psychology undergraduates over twenty minutes expressing thoughts and feelings. Each writer has been profiled by filling in a questionnaire testing the "Big Five" personality dimensions. The researchers focused on two of the Big Five traits, Extraversion and Neuroticism. Style and content features were extracted, and they concluded that style features provide a significant amount of information about personality.

In [23], authors developed classification, regression, and ranking models to recognize Big Five personality traits. They extracted a set of linguistic and psycholinguistic features from essays written by 2,479 psychology students, who were told to write whatever came through their minds for 20 minutes. The LIWC lexicon provided 88 word categories with syntactic and semantic information, while the Medical Research Council (MRC) Psycholinguistic Database [24] was used to extract 14 features. These features were used to train machine learning algorithms. The LIWC features outperformed the MRC features for every trait, and the LIWC features on their own always perform slightly better than the full feature set.

Using a publicly available dataset [11] consisting of essays, the authors of [25] developed a personality prediction model. They used psycholinguistic indices and language embeddings as features. Their results showed that language embeddings consistently outperform conventional psycholinguistic features.

In recent years, CPP studies have focused on corpora of social network data written in English and other languages. One of the most successful research initiatives in this area is the Author Profiling Task organised at PAN 2015. The specific task aimed to identify Twitter users' personality traits considering multilingual data (English, Spanish, Italian, and Dutch) [26].

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 in size (number of words per student) and students' demographics (gender and age).

The participating students of three different high schools were asked to write three essays to achieve our goal of collecting 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 volunteers were many more than 198, but their data have been ignored because they did not provide in their linguistic production the required text size. The mean length of the essays was 1,255 words. The topics, which were not given in advance, were related to the benefits 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 the personality types of students.

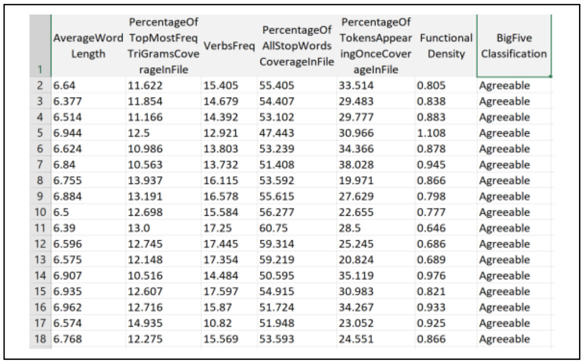
A. *Approach*

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.) [27]-[29]. In a top-down approach, instead, specialized dictionaries with custom entries are used to check the potential correlation with personality traits [30]-[32]. 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 part of a broader feature set 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 (Figure 1) of the preprocessing phase (matrix of stylometric features) was submitted to the data mining platform Rapidminer [33]. The text preprocessing pipeline was initially applied to the original texts of the students. However, we observed that various language errors were scattered across all linguistic levels and inserted significant bias in the modeling process negatively affecting the prediction results. Therefore, the essays were corrected manually without loss of information on the morphosyntactic level.



1	AverageWord Length	PercentageOf TopMostFreq TriGramsCoverageInFile	VerbsFreq	PercentageOf AllStopWords CoverageInFile	PercentageOf TokensAppearingOnceCoverageInFile	Functional Density	BigFive Classification
2	6.64	11.622	15.405	55.405	33.514	0.805	Agreeable
3	6.377	11.854	14.679	54.407	29.483	0.838	Agreeable
4	6.514	11.166	14.392	53.102	29.777	0.883	Agreeable
5	6.944	12.5	12.921	47.443	30.966	1.108	Agreeable
6	6.624	10.986	13.803	53.239	34.366	0.878	Agreeable
7	6.84	10.563	13.732	51.408	38.028	0.945	Agreeable
8	6.755	13.937	16.115	53.592	19.971	0.866	Agreeable
9	6.884	13.191	16.578	55.615	27.629	0.798	Agreeable
10	6.5	12.698	15.584	56.277	22.655	0.777	Agreeable
11	6.39	13.0	17.25	60.75	28.5	0.646	Agreeable
12	6.596	12.745	17.445	59.314	25.245	0.686	Agreeable
13	6.575	12.148	17.354	59.219	20.824	0.689	Agreeable
14	6.907	10.516	14.484	50.595	35.119	0.976	Agreeable
15	6.935	12.607	17.597	54.915	30.983	0.821	Agreeable
16	6.962	12.716	15.87	51.724	34.267	0.933	Agreeable
17	6.574	14.935	10.82	51.948	23.052	0.925	Agreeable
18	6.768	12.275	15.569	53.593	24.551	0.866	Agreeable

Figure 1. Output of the preprocessing phase.

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 [34] and gender identification [35], and we tested them for author personality prediction as well. A list of the stylometric features extracted from the textual data is reported in Table I.

C. Classification Algorithms

In this project, the problem of predicting the personality type and personality traits 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 and on the other hand, the Five Factors of personality, **O**penness to Experience, **C**onscientiousness, **E**xtraversion, **A**greeableness, and **N**euroticism. The extracted stylometric features matched the texts whose authors clearly belonged to a positive or negative category to have a valid prediction.

TABLE I. STYLOMETRIC FEATURES EXTRACTED FROM CORPUS

1. Frequency of Verbs	13. Frequency of Active Voice Verbs	25. Functional Density
2. Frequency of Nouns	14. Frequency of Passive Voice Verbs	26. Average Word Length
3. Frequency of Adjectives	15. Percentage of all Stop Words	27. Average Sentence Length
4. Frequency of Articles	16. Percentage of Top Most Frequent Tokens	28. Percentage of Top Most Frequent Word Bigrams
5. Frequency of Pronouns	17. Percentage of Top Most Frequent Non Stop Words	29. Percentage of Top Most Frequent Word Trigrams
6. Frequency of Adverbs	18. Percentage of Bottom Least Frequent Tokens	30. Percentage of Top Most Frequent Character Bigrams
7. Frequency of Prepositions	19. Percentage of Bottom Least Frequent Non Stop Words	31. Percentage of Top Most Frequent Character Trigrams
8. Frequency of Conjunctions	20. Number of Single Non Stop Words per all Words Occurrences	32. Percentage of 100 Most Frequent Words
9. Frequency of Personal Pronouns	21. Percentage of Tokens Appearing Once	33. Percentage of 100 Most Frequent Word Bigrams
10. Frequency of Coordinative Conjunctions	22. Percentage of Tokens Appearing Twice	34. Percentage of 100 Most Frequent Character Bigrams
11. Frequency of Subordinative Conjunctions	23. Ratio of Twice over Once Appearing Tokens	35. Percentage of 100 Most Frequent Character Trigrams
12. Frequency of Personal And Possessive Pronouns	24. Percentage of all Non Stop Words	

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 (Binomial Method), 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 the students' essays as training data. Their personality type and traits had been defined before using the appropriate psychometric questionnaires.

V. RESULTS

This section presents the results of the procedure that we followed to automatically classify the students' essays based on the personality type and personality traits defined by the personality questionnaires they filled in. From the nine algorithms trained in the textual data, we present the evaluation metrics of the most effective algorithm (Table II and Table III) along with the corresponding weights that positively affected the prediction of the personality type and traits depending on the psychological theory used.

A. Jung Typology Test

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% [36]. A more detailed list of evaluation metrics (accuracy, precision, and recall) is reported in Table II.

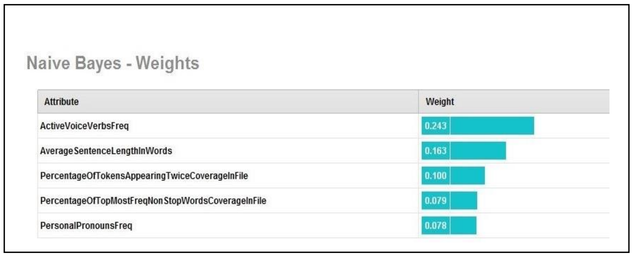
TABLE II. NAIVE BAYES MODEL PERFORMANCE

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 (Binomial Method), 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 71.7% by Generalized Linear Model (Binomial Method) and Logistic Regression. For the Feeling type, the Decision Tree algorithm exhibits the second-best performance with 63.2%. Random Forest is in the third position with 63.1%. 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 2), 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.

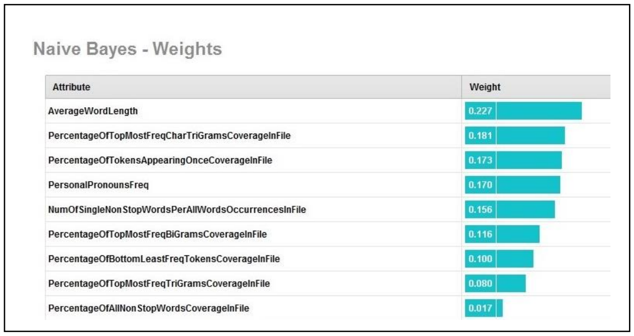


Naive Bayes - Weights

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

Figure 2. Weights for Extraversion.

Figure 3 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.

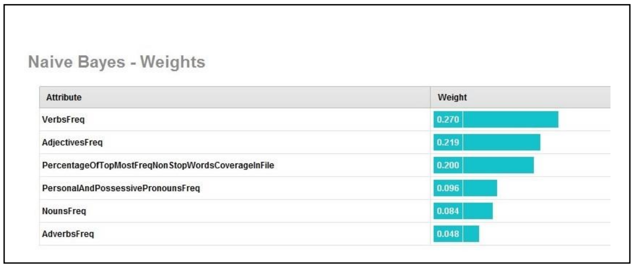


Naive Bayes - Weights

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 3. 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 4).

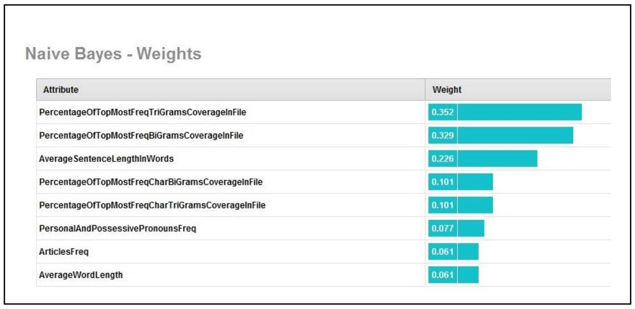


Naive Bayes - Weights

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

Figure 4. Weights for Feeling.

Finally, in Figure 5, 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
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 5. Weights for Judging.

B. Big-Five Personality Test

Regarding the prediction of all Big Five personality traits, the algorithm with the best results was the Generalized Linear Model (Binomial Method). The accuracy rate revealed a range from 65.6%% to 85.9%, with an average of 72.1%. Openness to Experience was predicted with 85.9%, Conscientiousness with 71.2%, Extraversion with 67.6%, Agreeableness with 70.2%, and the trait of Neuroticism with 65.6% [36]. It clearly emerges that Openness to Experience is the easiest trait that can be predicted from the textual data, followed by Conscientiousness and Agreeableness. Table III reports a more detailed list of evaluation metrics (accuracy, precision, and recall).

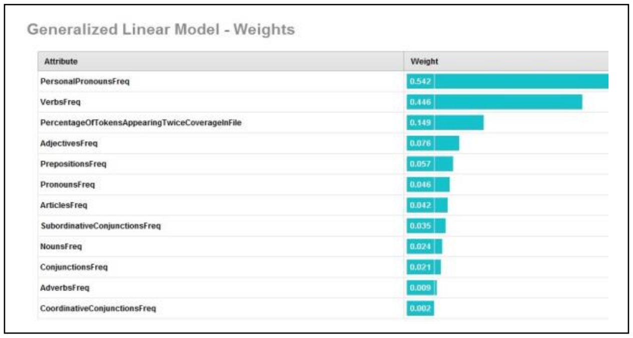
TABLE III. GENERALIZED LINEAR MODEL PERFORMANCE (BINOMIAL METHOD)

Personality Trait	Generalized Linear Model (Binomial Method)		
	Accuracy	*Precision*	*Recall*
Openness	85.9%	85.4%	100%
Conscientiousness	71.2%	68.6%	80.0%
Extraversion	67.6%	66.7%	86.7%
Agreeableness	70.2%	67.9%	98.7%
Neuroticism	65.6%	64.8%	71.9%

In terms of classification accuracy, the next best algorithms trained in the corpus produced the following results: Regarding Openness to Experience, Logistic Regression achieved 85.2%, Fast Large Margin, Decision Tree, Random Forest and Support Vector Machine 81.4% and Deep Learning 80.4%. The algorithms with the best performance for the trait of Conscientiousness were Naive Bayes, Gradient Boosted Trees, and Random Forest with calculated accuracies of 61.8%, 57.2%, and 56.0%, respectively. Extraversion was predicted with 65.1% by Fast Large Margin algorithm, 60.0% by Decision Tree and Random Forest, and 57.5% Deep Learning and Support Vector Machine. For Agreeableness, Random Forest algorithm exhibits the second-best performance with 69.1%, Deep Learning, Decision Tree, Gradient Boosted Trees, and Support Vector Machine being in the third position with 62.1%. The algorithms with the best performance for the trait of Neuroticism were Deep Learning, Naive Bayes, and Logistic Regression with calculated accuracies of 59.2%, 58.0%, and 57.6%, respectively.

In the following paragraphs, we present the weights we extracted from the Generalized Linear Model (Binomial Method) that measure the importance of each stylometric feature to the classification decisions of the algorithm for each personality trait separately with the aim to classify the essays of the students in personality traits 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 Openness to Experience (Figure 6), the use of personal pronouns had a significant impact. In addition, the use of verbs, dis legomena, adjectives, prepositions, pronouns, articles, subordinative conjunctions, nouns, conjunctions, adverbs, and coordinative conjunctions complete the list with the twelve most important stylometric features.



Generalized Linear Model - Weights

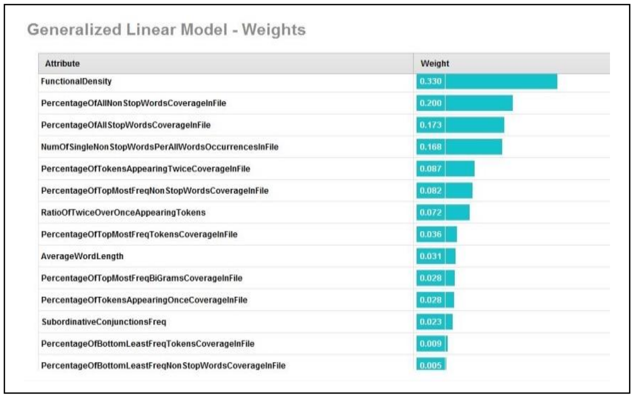
Attribute	Weight
PersonalPronounsFreq	0.542
VerbsFreq	0.446
PercentageOfTokensAppearingTwiceCoverageInFile	0.149
AdjectivesFreq	0.076
PrepositionsFreq	0.057
PronounsFreq	0.046
ArticlesFreq	0.042
SubordinativeConjunctionsFreq	0.035
NounsFreq	0.024
ConjunctionsFreq	0.021
AdverbsFreq	0.009
CoordinativeConjunctionsFreq	0.002

Figure 6. Weights for Openness to Experience

The stylometric features that contributed to the prediction of Conscientiousness were in descending order: functional density, non stop words, stop words, dis legomena, ratio of twice over once appearing tokens, the top most frequent tokens, the average word length, the top most frequent word bigrams, the hapax legomena, subordinative conjunctions, the bottom least frequent tokens, and the bottom least frequent non stop words (Figure 7).

Figure 8 depicts the prediction ability of the stylometric features for Extraversion used by the algorithm. The average sentence length had the most significant impact. The features that follow are ratio of twice over once appearing tokens, personal and possessive pronouns, the top most frequent word bigrams, adverbs, the bottom least frequent non stop words, conjunctions, prepositions, and dis legomena.

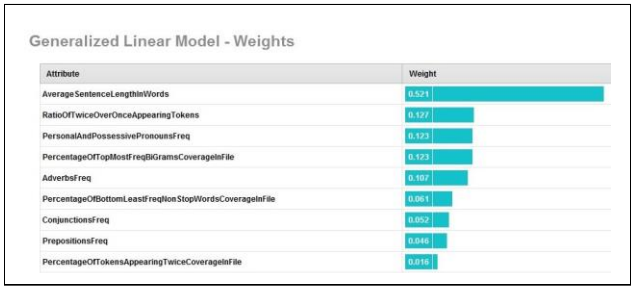
The stylometric features that affected the result of the classification of the essays in terms of Agreeableness are the verbs, ratio of twice over once appearing tokens, dis legomena, the use of verb types in active voice, personal and possessive pronouns, the top most frequent word bigrams, the average word length in characters, the average sentence length in words, prepositions, the top most frequent tokens, and the top most frequent character trigrams (Figure 9).



Generalized Linear Model - Weights

Attribute	Weight
FunctionalDensity	0.330
PercentageOfAllNonStopWordsCoverageInFile	0.200
PercentageOfAllStopWordsCoverageInFile	0.173
NumOfSingleNonStopWordsPerAllWordsOccurrencesInFile	0.168
PercentageOfTokensAppearingTwiceCoverageInFile	0.087
PercentageOfTopMostFreqNonStopWordsCoverageInFile	0.082
RatioOfTwiceOverOnceAppearingTokens	0.072
PercentageOfTopMostFreqTokensCoverageInFile	0.036
AverageWordLength	0.031
PercentageOfTopMostFreqBiGramsCoverageInFile	0.028
PercentageOfTokensAppearingOnceCoverageInFile	0.028
SubordinativeConjunctionsFreq	0.023
PercentageOfBottomLeastFreqTokensCoverageInFile	0.009
PercentageOfBottomLeastFreqNonStopWordsCoverageInFile	0.005

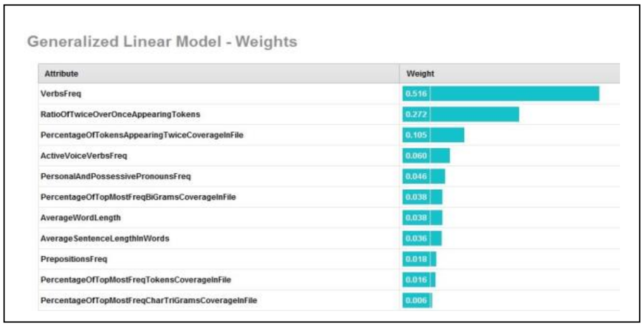
Figure 7. Weights for Conscientiousness



Generalized Linear Model - Weights

Attribute	Weight
AverageSentenceLengthInWords	0.521
RatioOfTwiceOverOnceAppearingTokens	0.127
PersonalAndPossessivePronounsFreq	0.123
PercentageOfTopMostFreqBiGramsCoverageInFile	0.123
AdverbsFreq	0.107
PercentageOfBottomLeastFreqNonStopWordsCoverageInFile	0.061
ConjunctionsFreq	0.052
PrepositionsFreq	0.046
PercentageOfTokensAppearingTwiceCoverageInFile	0.016

Figure 8. Weights for Extraversion

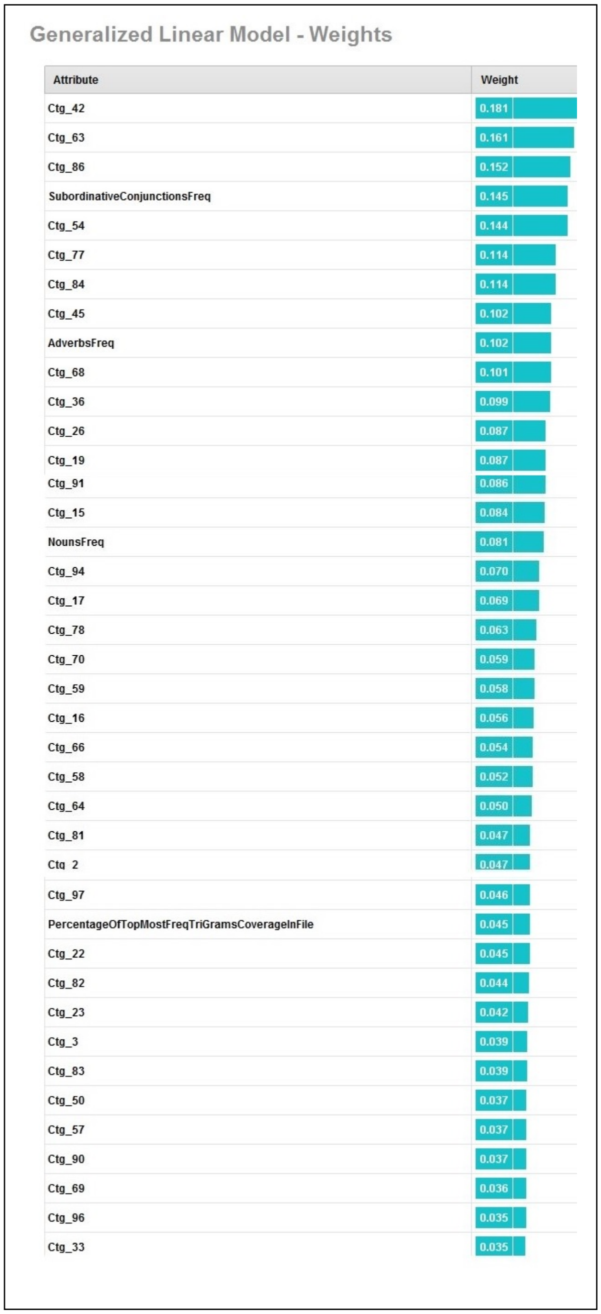


Generalized Linear Model - Weights

Attribute	Weight
VerbsFreq	0.516
RatioOfTwiceOverOnceAppearingTokens	0.272
PercentageOfTokensAppearingTwiceCoverageInFile	0.105
ActiveVoiceVerbsFreq	0.060
PersonalAndPossessivePronounsFreq	0.046
PercentageOfTopMostFreqBiGramsCoverageInFile	0.038
AverageWordLength	0.038
AverageSentenceLengthInWords	0.036
PrepositionsFreq	0.018
PercentageOfTopMostFreqTokensCoverageInFile	0.016
PercentageOfTopMostFreqCharTriGramsCoverageInFile	0.006

Figure 9. Weights for Agreeableness

Finally, in Figure 10, the stylometric features that contributed to the prediction of Neuroticism are many of the 100 most frequent character trigrams, which were extracted from the whole corpus in contrast to the other features extracted from the subcorpora depending on the personality trait. Additionally, personality prediction was affected by subordinative conjunctions, adverbs, nouns, and the top most frequent word trigrams.



Generalized Linear Model - Weights

Attribute	Weight
Ctg_42	0.181
Ctg_63	0.161
Ctg_86	0.152
SubordinativeConjunctionsFreq	0.145
Ctg_54	0.144
Ctg_77	0.114
Ctg_84	0.114
Ctg_45	0.102
AdverbsFreq	0.102
Ctg_68	0.101
Ctg_36	0.099
Ctg_26	0.087
Ctg_19	0.087
Ctg_91	0.086
Ctg_15	0.084
NounsFreq	0.081
Ctg_94	0.070
Ctg_17	0.069
Ctg_78	0.063
Ctg_70	0.059
Ctg_59	0.058
Ctg_16	0.056
Ctg_66	0.054
Ctg_58	0.052
Ctg_64	0.050
Ctg_81	0.047
Ctg_2	0.047
Ctg_97	0.046
PercentageOfTopMostFreqTriGramsCoverageInFile	0.045
Ctg_22	0.045
Ctg_82	0.044
Ctg_23	0.042
Ctg_3	0.039
Ctg_83	0.039
Ctg_50	0.037
Ctg_57	0.037
Ctg_90	0.037
Ctg_69	0.036
Ctg_96	0.035
Ctg_33	0.035

Figure 10. Weights for Neuroticism

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

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 [8]. On the other hand, research on textual data from essays using the Big Five model achieved an average accuracy of 60.6% [25], while we got 72.2%. The other studies mentioned [9]-[13][26] implemented machine learning techniques in textual data that were retrieved from social media. Therefore, their results can't be directly compared since they 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.

It is essential, of course, to emphasize that in the research field in which this research belongs, there are no reference data measuring and comparing the performance of different personality traits prediction methods objectively. None of the existing research uses comparable methods that have been applied to identical or comparable sets of textual data in the same language. Therefore, the percentages of accuracy from literature involve research with textual data, but not those of students but those of adults, written under other circumstances and in a different language; and, of course, with other features, not always stylometric.

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.

A basic target for continuing the research work in CPP is the investigation of new traits but also testing more stylometric features. In this study, we utilized only linguistic stylometric features. In this direction, future research will employ experimentation with new linguistic features or features already examined in the literature, such as content features, psycholinguistic, and syntactic features. We plan to localize well-known psychometric lexicons in Modern Greek (e.g., LIWC) and use them to complement our feature sets. In addition, we need to select features depending on the corpus, as, for instance, a person writes differently in a school essay and differently on social media.

We have ascertained the need to develop high-volume representative data since this constitutes a prerequisite for any relevant research. We need to create specialized corpora of Greek education texts by means of the used algorithmic methods, as well as the respective reference corpora to review the performance of the methods. Moreover, we plan to increase the size of the used corpus with additional students' essays. To draw more reliable conclusions, the growth of the corpus needs to ensure a balance between textual genres and different personality profiles. To achieve this, the corpus could be enriched with essays of different topics and textual genres and sufficient data for every psychological type.

References

[1] S. Gagiatsou, G. Markopoulos, and G. Mikros, "Using Stylometric Features to Predict Author Personality Type in Modern Greek Essays" The Fifteenth International Conference on Digital Society (ICDS 2021), Jul. 2021, pp. 34-39, ISBN: 978-1-61208-869-3

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

[3] 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: November, 2021]

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

[7] The Big Five Project Personality Test https://www.outofservice.com/bigfive. [retrieved: November, 2021].

[8] P. Juola, "Authorship Attribution," Foundations and Trends in Information Retrieval, vol. 1, no. 3, pp. 233-334, 2008.

[9] F. H. Sanford, "Speech and Personality: A Comparative Case Study," Journal of Personality, vol. 10, pp. 169-198, 1942.

[10] J. J. Bradac, "Language Attitudes and Impression Formation", In H. Giles and W.P. Robinson (eds.), Handbook of Language and Social Psychology, pp. 387-412, 1990.

[11] J. W. Pennebaker and L. A. King, "Linguistic Styles: Language Use as an Individual Difference," Journal of Personality and Social Psychology, vol. 77, pp. 1296-1312, 1999.

[12] A. J. Gill, "Personality and Language: The Projection and Perception of Personality in Computer-mediated Communication", Ph.D. Thesis, University of Edinburgh: Scotland, 2003.

[13] J. W. Pennebaker, M. R. Mehl, and K. G. Niederhoffer, "Psychological Aspects of Natural Language Use: Our Words, our Selves," Annual Review of Psychology, vol. 54, pp. 547-577, 2003.

[14] J. W. Pennebaker and L. D. Stone, "Words of Wisdom: Language Use over the Life Span," Journal of Personality and Social Psychology, vol. 85, no. 2, pp. 291-301, 2003.

[15] J. Oberlander and A. J. Gill, "Language with Character: A Stratified Corpus Comparison of Individual Differences in e-mail Communication," Discourse Processes, vol. 42, pp. 239-270, 2006.

[16] 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.

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

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] S. Argamon, M. Koppel, J. W. Pennebaker, and J.Schler, “Automatically Profiling the Author of an Anonymous Text,” Communications of the Association for Computing Machinery, vol. 52, no. 2, pp. 119-123, 2009.

[23] F. Mairesse and M. A. Walker, “Words Mark the Nerds: Computational Models of Personality Recognition through Language” The 28th Annual Conference of the Cognitive Science Society, Jul. 2006, pp. 543-548.

[24] M. Wilson, “MRC Psycholinguistic Database: Machine Usable Dictionary, Version 2.00,” Behavioural Research Methods, Instruments and Computers, vol. 20, pp. 6-11, 1988.

[25] Y. Mehta et al., “Bottom-up and top-down: Predicting Personality with Psycholinguistic and Language Model Features” 20th IEEE International Conference on Data Mining (ICDM), Nov. 2020, pp. 1184-1189.

[26] F. Rangel et al., “Overview of the 3rd Author Profiling Pask at PAN 2015” In L. Cappellato, N. Ferro, J.Gareth, and E. San Juan (Eds), CLEF 2015 Evaluation Labs and Workshop - Working Notes Papers, Sep. 2015.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

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

[34] 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.

[35] 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.

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