ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Εθνικόν και Καποδιστριακόν
Πανεπιστήμιον Αθηνών
ΙΔΡΥΘΕΝ ΤΟ 1837

ΦΙΛΟΣΟΦΙΚΗ ΣΧΟΛΗ
ΤΜΗΜΑ ΦΙΛΟΛΟΓΙΑΣ
ΤΟΜΕΑΣ ΓΛΩΣΣΟΛΟΓΙΑΣ

Προσδιορισμός συγγραφικού προφίλ μέσω τεχνικών επεξεργασίας φυσικής γλώσσας

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Σοφία Μιχ. Γαγιάτσου

ΑΘΗΝΑ 2021

Προσδιορισμός συγγραφικού προφίλ μέσω τεχνικών επεξεργασίας φυσικής γλώσσας

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

της

Σοφίας Μιχ. Γαγιάτσου

Φιλολόγου - Γλωσσολόγου ΕΚΠΑ

ΕΠΙΒΛΕΠΩΝ: Μαρκόπουλος Γεώργιος, Αναπληρωτής Καθηγητής, Τμήμα Φιλολογίας, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών (ΕΚΠΑ)

ΤΡΙΜΕΛΗΣ ΣΥΜΒΟΥΛΕΥΤΙΚΗ ΕΠΙΤΡΟΠΗ

Μαρκόπουλος Γεώργιος, Αναπληρωτής Καθηγητής, Τμήμα Φιλολογίας ΕΚΠΑ
Μικρός Γεώργιος, Καθηγητής, College of Humanities and Social Sciences του Πανεπιστημίου Hamad Bin Khalifa, Κατάρ
Γούτσος Διονύσιος, Καθηγητής, Τμήμα Φιλολογίας ΕΚΠΑ

ΕΠΤΑΜΕΛΗΣ ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ

Μαρκόπουλος Γεώργιος, Αναπληρωτής Καθηγητής, Τμήμα Φιλολογίας ΕΚΠΑ
Μικρός Γεώργιος, Καθηγητής, College of Humanities and Social Sciences του Πανεπιστημίου Hamad Bin Khalifa, Κατάρ
Γούτσος Διονύσιος, Καθηγητής, Τμήμα Φιλολογίας ΕΚΠΑ
Ιακώβου Μαρία, Αναπληρώτρια Καθηγήτρια, Τμήμα Φιλολογίας ΕΚΠΑ
Μπέλλα Σπυριδούλα, Καθηγήτρια, Τμήμα Φιλολογίας ΕΚΠΑ
Λέγγερης Άγγελος, Επίκουρος Καθηγητής, Τμήμα Φιλολογίας ΕΚΠΑ
Παναρέτου Ελένη, Αναπληρώτρια Καθηγήτρια, Τμήμα Φιλολογίας ΕΚΠΑ

Η εκπόνηση της παρούσας διατριβής υποστηρίχθηκε οικονομικά από το Ίδρυμα Κρατικών Υποτροφιών (ΙΚΥ).

Στους αγαπημένους μου γονείς,
Μιχαήλ και Μαρία

Περίληψη

Η παρούσα διατριβή παρουσιάζει την ερευνητική προσπάθεια και τα αποτελέσματα που αυτή παρήγαγε αναφορικά με την αυτόματη αναγνώριση των χαρακτηριστικών της προσωπικότητας του συγγραφέα μέσω τεχνικών επεξεργασίας φυσικής γλώσσας. Συγκεκριμένα, εξετάζεται η υπόθεση ότι στοιχεία του χαρακτήρα ενός ατόμου μπορούν να προσδιοριστούν αυτόματα μέσω της γλώσσας που χρησιμοποιεί στο γραπτό του λόγο. Για το σκοπό αυτό αναπτύχθηκε ηλεκτρονικό σώμα κειμένων από εκθέσεις μαθητών Λυκείου στην Ελληνική γλώσσα. Στους συμμετέχοντες μαθητές χορηγήθηκαν δύο ερωτηματολόγια προσωπικότητας, το ένα βασισμένο στην τυπολογία του Carl Jung και το άλλο στο μοντέλο των Πέντε Παραγόντων (ψυχολογικών χαρακτηριστικών). Επιπλέον, καταγράφεται στην παρούσα διατριβή αναλυτική επισκόπηση της διεθνούς βιβλιογραφίας του εν λόγω ερευνητικού πεδίου, ώστε να μελετηθούν η μεθοδολογία, τα εργαλεία και τα αποτελέσματα των σχετικών ερευνών μέχρι σήμερα. Αξιοποιώντας τα σύγχρονα ερευνητικά πορίσματα, η διατριβή εξετάζει την αποτελεσματικότητα δεκάδων υφομετρικών χαρακτηριστικών για την πρόβλεψη της προσωπικότητας των μαθητών. Αυτά τα χαρακτηριστικά, αφού εξήχθησαν αυτόματα από τα ηλεκτρονικά κειμενικά σώματα των εκθέσεων με εργαλεία και πόρους επεξεργασίας φυσικής γλώσσας, τροφοδότησαν μια μεγάλη σειρά από αλγόριθμους μηχανικής μάθησης, των οποίων τα μοντέλα ελέγχθηκαν μετά ως προς την ακρίβεια της απόδοσής τους. Για το Ερωτηματολόγιο Τύπων Προσωπικότητας Myers-Briggs Type Indicator (MBTI) βρέθηκε ότι ο Naive Bayes αλγόριθμος αποδίδει το μεγαλύτερο συγκριτικά κατά μέσο όρο ποσοστό ακρίβειας, ανερχόμενο σε 76,5%, ενώ για την πρόβλεψη των χαρακτηριστικών προσωπικότητας βάσει του μοντέλου των Πέντε Παραγόντων, επικράτησε ο αλγόριθμος Generalized Linear Model με μέσο όρο ακρίβειας 72,2%. Από την διεξαχθείσα έρευνα προέκυψαν νέοι συνδυασμοί υφομετρικών χαρακτηριστικών και αντίστοιχες υπολογιστικές τεχνικές, που δίνουν ενδιαφέρουσες και ικανοποιητικές λύσεις στο πρόβλημα αυτόματης αναγνώρισης της προσωπικότητας του συγγραφέα για την Ελληνική γλώσσα, ενώ καταδείχθηκε (και για τα Ελληνικά) η βαρύνουσα αξία της χρήσης των υφομετρικών γλωσσολογικών χαρακτηριστικών στην

αντιμετώπιση των ερευνητικών προβλημάτων στο γενικότερο επιστημονικό και τεχνολογικό πεδίο της κατανόησης από τον υπολογιστή του ανθρώπινου γραπτού λόγου, με έμφαση στην επεξεργασία κειμένων φυσικής γλώσσας για αυτόματη εξόρυξη ιδιαίτερων χαρακτηριστικών του συγγραφέα τους.

Λέξεις κλειδιά: πρόβλεψη προσωπικότητας, τυπολογία Jung, μοντέλο Πέντε Παραγόντων, υπολογιστική υφομετρία, υφομετρικά χαρακτηριστικά, μηχανική μάθηση.

Abstract

This thesis presents the research and its outcomes regarding the automatic recognition of author's personality features based on natural language processing techniques. In particular, we examine the hypothesis that the elements of a person's character can be determined automatically through the language he/she uses in written speech. For this purpose, an electronic corpus of texts has been developed comprising essays written by high-school students in Modern Greek. Participant students were given two personality questionnaire, one based on the typology of Carl Jung and the other based on the Model of Five Factors (psychological traits). In addition, in this volume, a detailed overview is recorded of international literature of the research field in question, in order to study the methodology, the tools and the results of relevant research up to now. By utilizing modern research findings, the thesis examines the effectiveness of numerous stylometric features to predict the personality of the students. These features, after being automatically extracted from the electronic corpora of essays using tools and natural languages processing resources, fed a long series of machine learning algorithms, the models of which were later controlled for the accuracy of their efficiency. As regards the Personality Types Questionnaire (Myers-Briggs Type Indicator (MBTI)) we found that the Naive Bayes algorithm renders the highest average accuracy percentage, being 76.5%, while for the prediction of personality features based on the Five Factors model, the Generalized Linear Model algorithm prevailed by average accuracy 72.2%. From the research that we conducted new combinations of stylometric features resulted and corresponding computational techniques, giving interesting and satisfying solutions to the problem of author's personality automatic recognition for Greek, while the important value of the use of stylometric linguistic features was demonstrated (for Greek as well) in tackling the research problems in the general scientific and technological field of human written speech understanding by computers, emphasizing on the processing of natural language texts for automatic mining of distinct features of their author.

Keywords: personality prediction, Jung Typology, Big Five model, computational stylometry, stylometric features, machine learning.

Πίνακας Περιεχομένων

Ευρετήριο όρων

Ελληνικά	**Αγγλικά**
Ακρίβεια	Accuracy
Αλγόριθμοι Μηχανικής μάθησης	Machine learning algorithms
Ανάκληση	Recall
Αντίληψη	Perceiving
Δεκτικότητα στην εμπειρία	Openness to Experience
Διαίσθηση	Intuition
Διγράμματα χαρακτήρων	Character Bigrams
Δίλεκτα	Word Bigrams
Εξωστρέφεια	Extraversion
Εσωστρέφεια	Introversion
Ευσυνειδησία	Conscientiousness
Κρίση	Judging
Μέρη του λόγου	Part of Speech
Μοντέλο των Πέντε Παραγόντων	Big Five model
Νευρωτισμός	Neuroticism
Νόηση	Sensing
Ορθότητα	Precision
Πρόβλεψη προσωπικότητας	Personality prediction
Προσήνεια	Agreeableness
Σκέψη	Thinking
Συναίσθημα	Feeling
Τριγράμματα χαρακτήρων	Character Trigrams
Τρίλεκτα	Word Trigrams
Τυπολογία του Jung	Jung’s Typology
Υφομετρικά χαρακτηριστικά	Stylometric features

Ευρετήριο Εικόνων

Ευρετήριο Σχημάτων

Ευρετήριο Πινάκων

Πρόλογος

Είναι αδιαμφισβήτητο το ότι στις μέρες μας η καθημερινή ζωή, η λειτουργία και η εξέλιξη της κοινωνίας, σε ατομικό αλλά και σε συλλογικό επίπεδο, είναι όσο ποτέ άλλοτε συνδεδεμένη με υπολογιστικές συσκευές. Ο 20^{ος} αιώνας έχει παραστατικά χαρακτηριστεί ως αιώνας των ηλεκτρονικών υπολογιστών. Ο χαρακτηρισμός αυτός ισχύει περισσότερο εμφατικά σήμερα, στην αρχή της τρίτης δεκαετίας του νέου μεταβιομηχανικού και τεχνολογικού αιώνα. Πράγματι, οι εξελίξεις στην ευφυή εκμάθηση/αυτό-εκπαίδευση των ίδιων των υπολογιστικών μηχανών (Machine Learning), στην επιστήμη των Δεδομένων (Data Science) και στην Τεχνητή Νοημοσύνη (Artificial Intelligence) οδηγούν ήδη την ανθρωπότητα στην λεγόμενη 4^{η} βιομηχανική επανάσταση, στην διάρκεια της οποίας η επικοινωνία ανθρώπου-μηχανής αναμένεται να προσομοιάσει συγκλονιστικά στην επικοινωνία ανθρώπου με άνθρωπο. Σε αυτό θα συμβάλει, μεταξύ άλλων, η προσδοκώμενη ραγδαία πρόοδος στις γλωσσικές τεχνολογίες, άρα και στην κομβική συνιστώσα αυτών, την Υπολογιστική Γλωσσολογία. Δεν είναι λοιπόν τυχαία η κεντρική θέση που κατέχει αυτή η επιστημονική περιοχή, ούτε τυχαίος ο ρόλος που καλούνται να διαδραματίσουν οι ερευνητές του κλάδου.

Η επανάσταση που έφερε στην καθημερινή ζωή των κοινωνιών των ανεπτυγμένων κρατών η τεχνολογία των υπολογιστών δεν ήταν τελικά παρά το πρώτο βήμα· ο άνθρωπος κατασκεύασε υπολογιστικές μηχανές που υλοποιούν με επιτυχία μεν, επιφανειακά και εξωτερικά όμως στρώματα της λειτουργικότητας της ανθρώπινης νόησης. Οι εξελίξεις στο μέλλον θα είναι πολλές τόσο σε επίπεδο συνειδητοποίησης, οργάνωσης και μοντελοποίησης βαθύτερων νοητικών λειτουργιών, όσο και σε επίπεδο εξομοίωσης αντίστοιχων νοητικών δυνατοτήτων όπως η γνώση, η μάθηση, η πρόβλεψη με εξελιγμένες υπολογιστικές μονάδες. Το πεδίο είναι ανοικτό και πάρα πολλοί είναι εκείνοι που έχουν από καιρό αρχίσει να το εξερευνούν υπακούοντας συνειδητά ή ασυνείδητα στην πρόκληση αυτογνωσίας και εφευρετικότητας που ενέχει ένα τέτοιο εγχείρημα.

Λαμβάνοντας υπόψη μας τις παραπάνω διαπιστώσεις, παρουσιάζει ιδιαίτερο ενδιαφέρον η μελέτη των τρεχουσών εξελίξεων σε θέματα Ανάκτησης Πληροφορίας και Εξόρυξης Γνώσης, ειδικότερα δε σε ένα σύγχρονο πεδίο υπολογιστικής ανάλυσης

κειμένου που στοχεύει στην Αναγνώριση του Συγγραφέα (Authorship Identification) με έμφαση στον αυτόματο προσδιορισμό των ψυχολογικών χαρακτηριστικών του συγγραφέα, που αποτελεί και το αντικείμενο της παρούσας διατριβής. Η αναγνώριση της προσωπικότητας του συγγραφέα αποτελεί πεδίο έρευνας που προσελκύει έντονο ενδιαφέρον εξαιτίας της ανάγκης να αποκωδικοποιηθεί ο χαρακτήρας του γράφοντος, γνώση που αξιοποιείται από πληθώρα εφαρμογών.

Θέλουμε να πιστεύουμε ότι τα αποτελέσματα της εργασίας μας αυτής συμβάλλουν έστω και σε ένα μικρό βαθμό στην προώθηση της σχετικής έρευνας.

Στο σημείο αυτό, θα ήθελα να ευχαριστήσω τους κυρίους Γεώργιο Μαρκόπουλο, Αναπληρωτή Καθηγητή του τμήματος Φιλολογίας του ΕΚΠΑ και Γεώργιο Μικρό, Καθηγητή του College of Humanities and Social Sciences του Πανεπιστημίου Hamad Bin Khalifa του Κατάρ, που μου εμπιστεύθηκαν το συγκεκριμένο θέμα διδακτορικής διατριβής. Θα ήταν παράλειψη αν δεν εξέφραζα στον κ. Γ. Μαρκόπουλο, επιβλέποντα της παρούσας διατριβής τις θερμές μου ευχαριστίες για την αδιάκοπη ενθάρρυνσή του, το ειλικρινές ενδιαφέρον καθώς και την καθοριστική συμβολή του στην εκπόνηση της διατριβής. Ιδιαίτερα ευχαριστώ τον κ. Γ. Μικρό για την καθοδήγηση, τις συμβουλές και τις ουσιαστικές παρατηρήσεις του. Επίσης ευχαριστώ θερμά τον κ. Διονύσιο Γούτσο, Καθηγητή του τμήματος Φιλολογίας ΕΚΠΑ για την υποστήριξη και τις καίριες υποδείξεις του. Οι ευχαριστίες μου απευθύνονται και στα μέλη της επταμελούς εξεταστικής επιτροπής κ. Μαρία Ιακώβου, κ. Σπυριδούλα Μπέλλα, κ. Άγγελο Λέγγερη και κ. Ελένη Παναρέτου για το ενδιαφέρον, την προσοχή τους, τα ουσιώδη σχόλιά τους και τις συμβουλές τους.

Σημαντική υπήρξε και η συμβολή των διευθυντών των σχολείων, τα οποία συμμετείχαν στην έρευνα. Η επίτευξή της, βέβαια, κατέστη δυνατή χάρη στη βοήθεια των μαθητών και μαθητριών των Λυκείων που συνέβαλαν με προθυμία και ενδιαφέρον, γράφοντας εκθέσεις για τη δημιουργία του σώματος κειμένων, επάνω στο οποίο στηρίχθηκαν τα πειράματα. Τους ευχαριστώ θερμά.

Τέλος, θα ήθελα να ευχαριστήσω το Ίδρυμα Κρατικών Υποτροφιών (ΙΚΥ) που μου χορήγησε υποτροφία για την εκπόνηση της διατριβής μετά από επιτυχή συμμετοχή μου στον ειδικό διαγωνισμό.

Αθήνα, Φεβρουάριος 2021

Σοφία Γαγιάτσου
Φιλόλογος - Γλωσσολόγος ΕΚΠΑ

Κεφάλαιο 1

Εισαγωγή

Σε αυτό το κεφάλαιο περιγράφεται το αντικείμενο και οι στόχοι της παρούσας διατριβής και αναλύεται η διάρθρωση του κειμενικού περιεχομένου της σε επιμέρους κεφάλαια.

1.1 Αντικείμενο - Στόχοι

Στην παρούσα διατριβή εξετάζεται το θέμα της αυτόματης αναγνώρισης της προσωπικότητας του συγγραφέα με τεχνικές επεξεργασίας φυσικής γλώσσας. Η διατριβή αποσκοπεί στη γενικότερη κριτική θεώρηση του θέματος και στην καταγραφή των τελευταίων εξελίξεων στον ερευνητικό τομέα της αυτόματης αναγνώρισης ιδιοτήτων του συγγραφέα (Authorship Profiling) και εκτενώς της προσωπικότητας του συγγραφέα καθώς και στη δημιουργία ενός μοντέλου πρόβλεψης προσωπικότητας, που θα χαρακτηρίζεται από υψηλά επίπεδα ακρίβειας.

Συγκεκριμένα, γίνεται αναφορά στην έννοια της προσωπικότητας και στις σχετικές θεωρίες που έχουν διατυπωθεί. Ερευνάται η σχέση γλώσσας και προσωπικότητας και συζητείται η συμβολή του κλάδου της Ποσοτικής Γλωσσολογίας στην επίτευξη του στόχου. Περιγράφεται η δημιουργία σώματος κειμένων της Ελληνικής γλώσσας από μαθητές σχολείων Δευτεροβάθμιας Εκπαίδευσης, στο οποίο υλοποιούνται πειράματα για την αυτόματη αναγνώριση της προσωπικότητας των μαθητών με υφομετρικές μεθόδους. Επιπλέον, τονίζεται η ανάγκη της εξέλιξης των μεθόδων του τομέα αυτού με στόχο την αξιοποίηση των αποτελεσμάτων για σύγχρονες εφαρμογές σε πολλούς, διαφορετικούς μεταξύ τους, κλάδους.

Οι στόχοι της ερευνητικής προσπάθειας συνοψίζονται επιγραμματικά ως εξής:

- Επισκόπηση της διεθνούς βιβλιογραφίας της σχετικής με αλγορίθμους και τεχνικές για τον εντοπισμό χαρακτηριστικών του συγγραφέα με έμφαση στην προσωπικότητά του από κείμενα διαφόρων ειδών και μορφών.

- Δημιουργία σώματος κειμένων από εκθέσεις μαθητών Λυκείου στην Ελληνική γλώσσα.
- Καθορισμός της προσωπικότητας των συμμετεχόντων μαθητών με τη χορήγηση και συμπλήρωση ερωτηματολογίων προσωπικότητας.
- Προσδιορισμός των υφομετρικών χαρακτηριστικών που εξάγονται από το σώμα κειμένων και χρησιμοποιούνται αποτελεσματικά για την πρόβλεψη προσωπικότητας.
- Αξιολόγηση αλγορίθμων μηχανικής μάθησης ως προς την ικανότητά τους να προβλέπουν με σχετική ακρίβεια συγκεκριμένα χαρακτηριστικά προσωπικότητας ενός συγγραφέα με βάση γραπτές εκθέσεις του.

1.2 Δομή

Το περιεχόμενο της διατριβής έχει οργανωθεί σε οκτώ κεφάλαια, τα οποία παρουσιάζουν αναλυτικά τα επιμέρους θέματα της αναγνώρισης της προσωπικότητας του συγγραφέα που μελετήθηκαν στο πλαίσιο της συγκεκριμένης έρευνας.

Στο δεύτερο κεφάλαιο διασαφηνίζεται εννοιολογικά η προσωπικότητα και γίνεται αναφορά στις σχετικές θεωρητικές προσεγγίσεις. Στο τρίτο κεφάλαιο γίνεται η σύνδεση ανάμεσα στην επιστήμη της Υπολογιστικής Υφολογίας και στον αυτόματο εντοπισμό των χαρακτηριστικών του συγγραφέα ενός κειμένου με την παρουσίαση του θεωρητικού υπόβαθρου του επιστημονικού πεδίου, αλλά και με την επισήμανση ερευνών που έχουν γίνει μέχρι σήμερα. Το τέταρτο κεφάλαιο εστιάζει στον τομέα της αυτόματης αναγνώρισης της προσωπικότητας με αναλυτική βιβλιογραφική επισκόπηση και κατηγοριοποίησή των ερευνών που έχουν καταγραφεί μέχρι σήμερα με βάση το είδος του σώματος κειμένων που χρησιμοποίησαν. Το πέμπτο κεφάλαιο αναφέρεται στη μηχανική μάθηση και πώς αυτός ο κλάδος της τεχνητής νοημοσύνης αξιοποιείται για την πρόβλεψη της προσωπικότητας. Στο έκτο κεφάλαιο παρουσιάζεται η μεθοδολογία που ακολουθήθηκε στην παρούσα έρευνα τόσο για τη δημιουργία του ηλεκτρονικού σώματος κειμένων από εκθέσεις μαθητών στην Ελληνική γλώσσα και την συμπλήρωση ερωτηματολογίων προσωπικότητας από τους μαθητές όσο και για την εξαγωγή υφομετρικών χαρακτηριστικών από αυτό. Στο έβδομο κεφάλαιο εξετάζονται και αναλύονται τα αποτελέσματα της υλοποίησης των μοντέλων

πρόβλεψης προσωπικότητας με βάση τα δύο ερωτηματολόγια προσωπικότητας. Στο όγδοο και τελευταίο κεφάλαιο παρουσιάζονται τα συμπεράσματα που προκύπτουν από την διεξαχθείσα έρευνα, τα θέματα προς συζήτηση και η συνεισφορά της διατριβής στον τομέα της αναγνώρισης της προσωπικότητας από γραπτό λόγο. Τέλος, προδιαγράφονται οι τομείς και γίνονται προτάσεις για μεταδιδακτορική έρευνα, ενώ τονίζεται η σπουδαιότητα της αυτόματης ανίχνευσης στοιχείων της προσωπικότητας, με την οποία καταπιάνεται η παρούσα διατριβή, μέσω της παράθεσης ενδεικτικών χαρακτηριστικών δυνητικών περιοχών εφαρμογών της.

Κεφάλαιο 2

Προσωπικότητα: Ορισμοί και Θεωρίες

2.1 Ιστορία της Προσωπικότητας

Η έννοια της προσωπικότητας έχει απασχολήσει τον άνθρωπο πριν ακόμα αποτελέσει αντικείμενο της επιστήμης της ψυχολογίας, ήδη από την αρχαιότητα. Η φυσιολογία, η φιλοσοφία και η λογοτεχνία είναι μερικοί μόνο κλάδοι που ασχολήθηκαν συστηματικά με την προσωπικότητα. Ερωτήματα όπως τι είναι φυσιολογικό και τι όχι, καθώς και πώς και γιατί αλλάζουν οι άνθρωποι συνδέονται με τη μελέτη της προσωπικότητας.

Ο αρχαίος ιατρός Ιπποκράτης ο Κώος (460-370 π.Χ.), ο οποίος θεωρούσε πως αφετηρία της ιατρικής είναι η γνώση της φυσικής σύστασης του σώματος, μελέτησε ιδιαίτερα την ανθρώπινη φυσιολογία. Η Ιπποκρατική φυσιολογία διέπεται από τη θεωρία των χυμών. Ο Ιπποκράτης συστηματοποιώντας τις απόψεις των Πυθαγορείων για τους χυμούς και συνδυάζοντάς τις με τις απόψεις των προσωκρατικών φιλοσόφων για τα στοιχεία, ιδίως του Εμπεδοκλή, υποστήριξε πως η ανθρώπινη συμπεριφορά εδράζεται στη βιολογία και ότι οι πηγές της ατομικής διαφοροποίησης της προσωπικότητας καθώς και σχήματα μη φυσιολογικής συμπεριφοράς μπορεί να βρεθούν στην ανισορροπία μεταξύ των τεσσάρων χυμών του ανθρώπινου σώματος, του αίματος, της μέλαινας χολής, της κίτρινης χολής και του φλέγματος και των τεσσάρων στοιχείων, ξηρού, θερμού, ψυχρού και υγρού (Κιαπόκας, 1996: 97-98). Ο άνθρωπος απολαμβάνει τέλεια υγεία (ευκρασία), εφόσον όλα του τα στοιχεία, σωματικά και ψυχικά, βρίσκονται σε αρμονική σύνθεση.

Την αρχή αυτή ακολούθησαν και ο Πλάτωνας και ο Αριστοτέλης και σε αυτήν επανέρχεται σήμερα η ψυχοσωματική ιατρική. Η Ιπποκρατική σχολή διδάσκει επιπλέον ότι η ψυχή συμμετέχει σε όλες τις λειτουργίες του οργανισμού και υφίσταται τον αντίκτυπο από τις διαταραχές του. Σύμφωνα με τους Thetford & Walsh (1985) οι απόψεις του Ιπποκράτη έχουν σαφή βιολογική θεμελίωση.

Ακολουθούν ο Επίκουρος και ο Κικέρωνας, οι οποίοι υπήρξαν θεωρητικοί της προσωπικότητας. Πιστεύοντας λιγότερο σε ένα βιολογικό πρότυπο, προέκριναν μία ψυχολογική βάση για την προσωπικότητα (Thetford & Walsh, 1985).

Ιδιαίτερο ενδιαφέρον παρουσιάζουν οι τρόποι με τους οποίους η σύγχρονη ψυχολογία μελετά την προσωπικότητα, έννοια που αποτελεί κεντρικό σημείο στην ιστορία και εξέλιξη της επιστήμης της ψυχολογίας.

Στα τέλη του 19ου αιώνα και έως τα τέλη της δεκαετίας του 1970 κεντρικός άξονας είναι οι θεωρίες των χαρακτηριστικών (traits) και «η γενική υπόθεση ότι όλοι οι άνθρωποι έχουν μια εσωτερικά εδραιωμένη ικανότητα (προδιάθεση) να ανταποκρίνονται με συγκεκριμένους (παγιωμένους) τρόπους σε συγκεκριμένα ερεθίσματα» (Ποταμιάνος & Παπαστάμου, 2008: 13). Κοινό σημείο των θεωριών είναι η αντίληψη ότι η προσωπικότητα θεωρείται αναπόσπαστο στοιχείο της ανθρώπινης οντότητας. Τις θεωρίες αυτές θα δούμε αναλυτικότερα σε επόμενο υποκεφάλαιο.

Στις τελευταίες δεκαετίες του 20ου αιώνα επικρατεί η έννοια της ψυχολογιοποίησης, που έφερε στην επιφάνεια η κοινωνική ψυχολογία. Αναφέρεται στην απόδοση κοινωνικής συμπεριφοράς στα ψυχολογικά χαρακτηριστικά του ατόμου που την εκδηλώνει. Τέλος, οι άρρητες θεωρίες εντάσσονται και αυτές στην σύγχρονη προσέγγιση της προσωπικότητας και αναφέρονται σε γενικές πεποιθήσεις σχετικά με τη συχνότητα εμφάνισης και τη διακύμανση ενός γνωρίσματος στο γενικό πληθυσμό. Πρόκειται για μια διαδικασία κατηγοριοποίησης κατά την οποία καταλήγουμε σε συμπεράσματα για την προσωπικότητα κάποιου συσχετίζοντας στοιχεία του χαρακτήρα ή της συμπεριφοράς (Παπαστάμου, 1989).

2.2 Η έννοια της Προσωπικότητας

Σήμερα, παρόλο που όλοι χρησιμοποιούμε τον όρο «προσωπικότητα» και μάλιστα προβαίνουμε συχνά σε χαρακτηρισμούς της προσωπικότητας των ατόμων, ωστόσο θεωρείται ασαφής. Ο σχηματισμός της λέξης εντάσσεται στη λεξική διαδικασία της παραγωγής, από το επαυξημένο θέμα *προσωπικ-* (*προσωπ-* + *-ικ*) συν το παραγωγικό επίθημα *-οτητ* συν το κλιτικό επίθημα *-α* και στο Λεξικό της Νέας Ελληνικής Γλώσσας του Μπαμπινιώτη (2002) ορίζεται ως το σύνολο των φυσικών, πνευματικών, ψυχικών και κοινωνικών χαρακτηριστικών ενός ατόμου.

Σε αυτό το σημείο πρέπει να διευκρινίσουμε τη διαφορά ανάμεσα στους όρους «προσωπικότητα» και «χαρακτήρας», διότι δεν είναι συνώνυμοι. Ο χαρακτήρας θεωρείται ο σκελετός της προσωπικότητας. Κάθε γνώρισμα του χαρακτήρα ενός ανθρώπου είναι ταυτόχρονα και γνώρισμα της προσωπικότητας, αλλά δεν ισχύει το αντίστροφο. Ουσιαστικά, εδώ έγκειται και η διαφορά της προσωπικότητας από τον χαρακτήρα. Η προσωπικότητα περιλαμβάνει το χαρακτήρα.

Στον καθημερινό λόγο, ο όρος προσωπικότητα χρησιμοποιείται συνήθως με δύο διαφορετικούς τρόπους: αφενός για να τονίσουμε την ολοκλήρωση, τη συνέπεια και τη μοναδικότητα της ύπαρξης ενός συγκεκριμένου ατόμου και αφετέρου για να δώσουμε έμφαση σε διαστάσεις ομοιότητας και διαφοράς μεταξύ των ανθρώπων (Thomas, 1995: 373).

Ανατρέχοντας στη σχετική βιβλιογραφία διαπιστώνουμε ότι η διατύπωση και του επιστημονικού ορισμού παρουσιάζει δυσκολίες κυρίως λόγω των διαφορετικών απόψεων των θεωρητικών σε θέματα που αφορούν την προσωπικότητα και γι’ αυτό υπάρχει πλήθος ορισμών.

Σύμφωνα με τον Eysenck (1952) προσωπικότητα είναι η σταθερή και διαρκής οργάνωση των παρακάτω στοιχείων που καθορίζουν την προσαρμογή στο περιβάλλον: του χαρακτήρα, της συναισθηματικής ιδιοσυγκρασίας, της διανοητικής και φυσικής κατάστασης του ατόμου.

Ο Allport (1961) ύστερα από μελέτη πενήντα ορισμών ορίζει την προσωπικότητα ως τη δυναμική οργάνωση των ψυχοφυσικών συστημάτων του ατόμου τα οποία καθορίζουν την προσαρμογή του στο περιβάλλον, τη χαρακτηριστική συμπεριφορά και τη σκέψη του.

Κατά τον Hilgard (1962) προσωπικότητα είναι το μόρφωμα των ατομικών χαρακτηριστικών και του τρόπου συμπεριφοράς του ατόμου, που καθορίζει την προσαρμογή του στο περιβάλλον.

Ο Murphy (1966) θεωρεί ότι η προσωπικότητα αποτελείται από δομές επίκτητων προτιμήσεων και προσδοκιών σχετικά με τον τρόπο κάλυψης των βιολογικών αναγκών που η κάθε κοινωνία έχει επιβάλει.

Η προσωπικότητα αναφέρεται σε κάποιους σταθερούς, εσωτερικούς παράγοντες, οι οποίοι κάνουν τη συμπεριφορά ενός ατόμου να ακολουθεί μια συνέπεια σε διάφορες χρονικές περιόδους, αλλά και να είναι διαφορετική σε σχέση με την συμπεριφορά άλλων ανθρώπων, όταν αντιμετωπίζουν παρόμοιες καταστάσεις υποστηρίζει ο Child (1968).

Ως προσωπικότητα ορίζεται από τους Pervin & John (1997) η πολυσύνθετη οργάνωση αντιλήψεων, παρορμήσεων και συμπεριφορών που κατευθύνει και δίνει σχήμα στη ζωή ενός ατόμου.

Οι παραπάνω ορισμοί της προσωπικότητας συγκλίνουν στο ότι οι άνθρωποι έχουν την προδιάθεση να ανταποκρίνονται με διαφορετικούς τρόπους σε παρόμοια ερεθίσματα. Βασικός άξονας για αυτή την προσέγγιση του όρου είναι τα χαρακτηριστικά. Ο συνδυασμός των ιδιαίτερων διανοητικών, συναισθηματικών και σωματικών χαρακτηριστικών ενός ανθρώπου, χάρις στον οποίο (συνδυασμό) αυτός (ο άνθρωπος) κατέχει μια αναγνωρίσιμη και μοναδική ταυτότητα είναι αυτό που καθορίζει την προσωπικότητα του ανθρώπου. Τα ιδιάζοντα αυτά χαρακτηριστικά της προσωπικότητας ενός ανθρώπου αντανακλώνται κατά κανόνα και σε όλες γενικά τις εκδηλώσεις του: στις σκέψεις/τοποθετήσεις του, στον ψυχισμό του, στις καθημερινές πράξεις του και την εν γένει συμπεριφορά του προς τους άλλους, τον εαυτό του, το περιβάλλον. Η κατανόηση, επομένως, της προσωπικότητας ενός ανθρώπου σημαίνει την κατανόηση και την ερμηνεία εκείνων των εκδηλώσεών του που τον χαρακτηρίζουν και τον διαφοροποιούν κατά μοναδικό τρόπο από τους άλλους, με άλλα λόγια, εκείνων των εκδηλώσεών του που «υπογράφονται» από την δική του ιδιαίτερη προσωπικότητα και τα χαρακτηριστικά της.

2.3 Θεωρίες Προσωπικότητας

Ο επιστημονικός κλάδος που μελετάει την προσωπικότητα ονομάζεται Ψυχολογία της Προσωπικότητας και έχει στόχο την περιγραφή και την ερμηνεία των διαφορών μεταξύ των ατόμων καθώς και την αποτύπωση των ομοιοτήτων τους. Έχουν αναπτυχθεί διάφορες θεωρίες προσωπικότητας ανταγωνιστικές αλλά και συμπληρωματικές που στηρίζονται σε μια φιλοσοφική θεώρηση της ανθρώπινης φύσης με κοινό στοιχείο τη συμβολή τους στην κατανόηση της ανθρώπινης συμπεριφοράς.

Μεγάλη ώθηση στη μελέτη της προσωπικότητας του ανθρώπου έδωσαν οι θεωρητικές συλλήψεις και οι μελέτες σπουδαίων επιστημόνων της συμπεριφοράς του ανθρώπου με αποτέλεσμα να διαμορφωθούν στο χώρο της ψυχολογίας της προσωπικότητας πολλές διαφορετικές θεωρητικές κατευθύνσεις. Σύντομα θα

αναφερθούμε στις σύγχρονες αυτές θεωρητικές κατευθύνσεις της ψυχολογίας της προσωπικότητας που έχουν διατυπωθεί.

Η ψυχαναλυτική θεωρία του Freud (Freud, 1923) εισήγαγε το τοπογραφικό μοντέλο της ανθρώπινης προσωπικότητας. Σύμφωνα με το μοντέλο αυτό, η ψυχική ζωή κάθε ατόμου μπορεί να αναπαρασταθεί με τρία επίπεδα συνειδητότητας, το ασυνείδητο, το προσυνειδητό και το συνειδητό. Ύστερα από κάποιες αλλαγές στη θεωρία του ο Freud υποστήριξε ότι η συμπεριφορά του ατόμου είναι το αποτέλεσμα της αλληλεπίδρασης δυνάμεων που συνιστούν τη δομή της προσωπικότητας, η οποία αποτελείται από τρία μέρη: το Εκείνο (id), το Εγώ (ego) και το Υπερεγώ (super ego). Τα στοιχεία αυτά θα πρέπει να εκλαμβάνονται ως θεωρητικές κατασκευές.

Το Εκείνο βρίσκεται εξ ολοκλήρου στο ασυνείδητο επίπεδο, ενώ το Εγώ και το Υπερεγώ λειτουργούν σε όλα τα επίπεδα συνειδητότητας. Το Εκείνο περιλαμβάνει τα βιολογικά ένστικτα και λειτουργεί με βάση την αρχή της ευχαρίστησης. Το Εγώ περιλαμβάνει τις ψυχικές λειτουργίες (αντίληψη, νόηση, μνήμη, κρίση κλπ) με τις οποίες το άτομο αντιλαμβάνεται την πραγματικότητα και έχει το ρόλο του ρυθμιστή ανάμεσα στις παρορμήσεις του Εκείνο και στους περιορισμούς της κοινωνίας. Το Υπερεγώ περιλαμβάνει τους κανόνες συμπεριφοράς που επιβάλλει η κοινωνία στα μέλη της. Σε όλους υπάρχει μια σύγκρουση μεταξύ των επιπέδων καθώς και εντός του κάθε επιπέδου. Η προσωπικότητα είναι δυναμική και προσπαθεί να επιφέρει ισορροπία μεταξύ αντιτιθέμενων δυνάμεων. Σύμφωνα με τον Freud η προσωπικότητα διαμορφώνεται αμετάκλητα κατά την παιδική ηλικία.

Χαρακτηριστικό παράδειγμα διαφοροποίησης των αρχικών φροϋδικών θέσεων είναι η ατομική ψυχολογία του Adler, σύμφωνα με τον οποίο αυτό που χαρακτηρίζει το άτομο είναι η έμφυτη κοινωνικότητά του. Αυτό σημαίνει ότι, σε αντίθεση με τον Freud, πρώτος ο Adler τονίζει τις κοινωνικές παραμέτρους της προσωπικότητας. Η έννοια του τρόπου ή του πλάνου ζωής αποτελεί έννοια-κλειδί για την αντλεριανή θεωρία περί προσωπικότητας κατά την οποία βασικό κίνητρο των πράξεων του ανθρώπου είναι η ορμή για κοινωνική αναγνώριση και αποδοχή. Αποτυχία της προσπάθειας αυτής απολήγει στο συναίσθημα της μειωμένης αυτοεκτίμησης, στο σύμπλεγμα κατωτερότητας (Dreikurs, 1975).

Ο Adler διέκρινε τέσσερις τύπους προσωπικότητας ανάλογα με το βαθμό δραστηριοποίησης του ατόμου για την επίλυση των προβλημάτων της ζωής, τον αρχομανή, τον παρασιτικό, τον αποφεύγοντα και τον χρήσιμο τύπο. Ιδιαίτερη έμφαση έδωσε ο Adler στις επιδράσεις που ασκεί στη διαμόρφωση της προσωπικότητας του

ατόμου το ψυχολογικό κλίμα της οικογένειας και η δυναμική που αναπτύσσεται μεταξύ των μελών της.

Από τη θεωρία του Freud διαφοροποιήθηκε και ο Carl Jung (1954), του οποίου οι απόψεις για τις στάσεις και τις λειτουργίες της προσωπικότητας αποτελούν σημαντική προσφορά στην ψυχανάλυση. Πιο συγκεκριμένα, διέκρινε σε κάθε άτομο το συνειδητό (ego), το προσωπικό ασυνείδητο (personal unconscious) και το συλλογικό ασυνείδητο (collective unconscious).

Η προσωποκεντρική θεωρία του Rogers (1951) βασίζεται στην προσπάθεια του ατόμου για αυτοπραγμάτωση και στην έννοια της αυτοσυνέπειας και της συμφωνίας του εαυτού μας με την εμπειρία. Με τη θεωρία του ο Rogers έδωσε τη βασική κατεύθυνση στη φαινομενολογική θεωρία και υπήρξε ο κύριος εκφραστής του μοντέλου της εκπλήρωσης, σύμφωνα με το οποίο βασικό χαρακτηριστικό του ανθρώπου είναι η τάση του να εκπληρώνει τις δυνατότητές του.

Την προσέγγιση της προσωπικότητας από τη σκοπιά της θεωρίας της μάθησης επιχειρούν οι συμπεριφοριστές βασιζόμενοι στις απόψεις του Locke ότι το παιδί γεννιέται tabula rasa και ότι η συμπεριφορά του ατόμου είναι αποτέλεσμα της εμπειρίας. Ο φυσιοδίφης Pavlov (1927) μελέτησε και περιέγραψε πρώτος την κλασική εξαρτημένη μάθηση και ο Watson (1936) πρώτος αναφέρεται στη θεωρία της συμπεριφοράς ως μια κατεύθυνση της ψυχολογίας της προσωπικότητας.

Ο Skinner (1953), εκπρόσωπος της συντελεστικής μάθησης, θεωρεί ότι οι πολύπλοκες μορφές συμπεριφοράς που αποτελούν την προσωπικότητα του ανθρώπου μπορεί να περιγραφούν ως ένα σύστημα ενισχυτικών γεγονότων. Από τους ενισχυτές, που διαχωρίζονται σε θετικούς και αρνητικούς, ο Skinner προτάσσει τους θετικούς ως κίνητρο για την εκμάθηση κάποιας συμπεριφοράς.

Η κοινωνικογνωστική θεωρία για την προσωπικότητα των Bandura και Mischel αναφέρεται στις γνωστικές διεργασίες της ανθρώπινης συμπεριφοράς, η οποία εδράζει στην κοινωνική προέλευση του ατόμου. Έμφαση δίνεται στις επιδιώξεις, τους στόχους, τις ικανότητες και τις δεξιότητες επίλυσης προβλημάτων (Bandura, 1986). Η βάση αυτής της θεωρίας είναι η κοινωνική μάθηση κατά την οποία το άτομο μαθαίνει μέσω της μίμησης προτύπων.

Ο Kelly στη θεωρία των προσωπικών νοητικών κατασκευών του διατυπώνει την άποψη πως οι άνθρωποι αποδίδουν ζεύγη αντίθετων εννοιών στον κόσμο των αντικειμένων και των ανθρώπων που τους περιβάλλουν και καθοδηγούν τη συμπεριφορά τους. Έτσι, η συνέπεια στη συμπεριφορά ενός ατόμου είναι το άμεσο

αποτέλεσμα αυτού του συστήματος προσωπικών κατασκευών που χρησιμοποιεί, προκειμένου να ερμηνεύσει τον κόσμο και που αποτελεί ουσιαστικά την ίδια του την προσωπικότητα (Kelly, 1955: 50).

Κάθε θεωρία της προσωπικότητας πρεσβεύει έναν τρόπο ερμηνείας της ανθρώπινης συμπεριφοράς. Οι δύο βασικοί τρόποι ταξινόμησης και περιγραφής των στοιχείων της ανθρώπινης ψυχοσύνθεσης είναι τα χαρακτηριστικά της προσωπικότητας και οι ψυχολογικοί τύποι. Η προσωπικότητα, δηλαδή, μπορεί να περιγραφεί με χαρακτηριστικά αλλά και με διακριτές κατηγορίες που αποδίδουν τις διαστάσεις της προσωπικότητας. Η θεωρία των χαρακτηριστικών, όπως προαναφέραμε μέσω των ορισμών στο υποκεφάλαιο 2.2, βασίζεται σε ιδιότητες της συμπεριφοράς, οι οποίες παραμένουν σχετικά σταθερές. Η θεωρία των Τύπων, η οποία διατυπώθηκε από τον Jung, ταξινομεί τα άτομα σε κατηγορίες. Όσα ανήκουν στον ίδιο ψυχολογικό τύπο έχουν την ίδια ψυχοσύνθεση.

Από το σύνολο των θεωριών της προσωπικότητας, στις ενότητες που ακολουθούν θα δοθεί έμφαση στις δύο επικρατέστερες, την θεωρία των Τύπων του Carl Jung και το μοντέλο των Πέντε Παραγόντων, καθώς βάσει αυτών σχεδιάστηκαν τα αντίστοιχα ψυχομετρικά τεστ προσωπικότητας, που χορηγήθηκαν στους μαθητές Λυκείου που συμμετείχαν στην έρευνά μας.

2.3.1 Η Θεωρία των Τύπων του Carl Jung

Οι δύο βασικές συμπεριφορές που συνυπάρχουν σε κάθε άνθρωπο σύμφωνα με τον Carl Jung, Ελβετό ψυχίατρο, είναι η εξωστρέφεια και η εσωστρέφεια. Ο καθένας ακολουθεί μία από τις δύο κατευθύνσεις, αλλά και η άλλη παραμένει ως κομμάτι του εαυτού του.

Η συμπεριφορά τόσο των εξωστρεφών όσο και των εσωστρεφών τύπων διέπεται από τέσσερις λειτουργίες, τη Νόηση, το Συναίσθημα, την Αντίληψη και την Διαίσθηση. Μέσω της νόησης ο άνθρωπος προσπαθεί να κατανοήσει τη φύση του κόσμου και την εμπειρία με γνωστικές διεργασίες. Με το συναίσθημα, από την άλλη, αξιολογεί τα γεγονότα ανάλογα με την ευχαρίστηση ή τη δυσαρέσκεια που νιώθει.

Η αντίληψη σχετίζεται με την ερμηνεία του εξωτερικού κόσμου και των εσωτερικών καταστάσεων. Ο άνθρωπος δηλαδή προσλαμβάνει αλλά και αποδίδει

νόημα στην εμπειρία. Η διαίσθηση, τέλος, αναφέρεται στην αντίληψη που έχουμε μέσω του ασυνείδητου, το οποίο όμως, αντίθετα από τη θεωρία του Freud, είναι η αντικειμενική πραγματικότητα που συμπληρώνει το Εγώ.

Ο Jung διατύπωσε τη θεωρία των Τύπων, όπως ονομάζεται. Έχει τρεις άξονες και αντιπροσωπεύει τους κύριους τρόπους προσαρμογής και κατανόησης των συμπεριφεριολογικών, συναισθηματικών, γνωσιακών και εικονοπλαστικών προϊόντων της ψυχής (Κουτουβίδης, Μηνογιάννη & Βάρσου, 2004). Οι άξονες αυτοί είναι η βασική συμπεριφορά, Εξωστρέφεια (Extraversion) - Εσωστρέφεια (Introversion) και δύο λειτουργικές διαστάσεις, Νόηση (Sensing) - Διαίσθηση (Intuiting) και Σκέψη (Thinking) - Συναίσθημα (Feeling). Ο πρώτος δείχνει τον τρόπο με τον οποίο ενεργοποιείται το άτομο, ο δεύτερος φανερώνει τα σημεία στα οποία εστιάζει την προσοχή του και ο άξονας Σκέψη - Συναίσθημα εκφράζει τον τρόπο με τον οποίο το άτομο αποφασίζει.

Κάθε άνθρωπος, σύμφωνα με τον Jung, τυποποιείται βάσει των τριών παραπάνω αξόνων. Όλοι έχουν όλους τους δυνατούς συνδυασμούς, αλλά ο καθένας έχει ένα σύνολο λειτουργιών καλύτερα αναπτυγμένο από άλλες. Οι ενήλικες προσπαθούν να κατανοήσουν τις λιγότερο αναπτυγμένες λειτουργίες τους, οι οποίες παρουσιάζονται ως συμπλέγματα ξένα προς το Εγώ. Με αυτό το εγχείρημα επιτυγχάνεται η ολοκλήρωση της προσωπικότητας (Jung, 1954).

Ο αντιληπτικός άξονας Νόηση - Διαίσθηση αναφέρεται στον τρόπο με τον οποίο γίνεται αντιληπτή και εσωτερικεύεται η ψυχική πραγματικότητα. Κατά τον Jung Νόηση είναι η ψυχική λειτουργία που εσωτερικεύει το συγκεκριμένο στο εδώ και τώρα. Αντίθετα, η Διαίσθηση δείχνει έναν τρόπο σύλληψης της ψυχικής πραγματικότητας κατά τον οποίο εσωτερικεύονται σύνολα μάλλον παρά μέρη. Συλλαμβάνεται, δηλαδή, η συνολική εικόνα παρά οι λεπτομέρειες.

Η πολικότητα Σκέψη - Συναίσθημα έχει σχέση με την επεξεργασία της πληροφορίας και την κρίση. Όταν χρησιμοποιείται η Σκέψη, τα δεδομένα κρίνονται σύμφωνα με τη λογική. Το Συναίσθημα αναφέρεται στη διαμόρφωση κρίσεων μέσω μη λογικών διεργασιών, προσανατολισμένων στην αξία και στην εκτίμηση, ιδιαίτερα εφόσον αναφέρονται στις διαπροσωπικές σχέσεις.

Ο Jung συνδυάζοντας τους τύπους της προσωπικότητας με τις ψυχικές λειτουργίες κατέληξε σε οκτώ επιμέρους ψυχολογικούς τύπους που είναι: ο εξωστρεφής διανοητικός, ο εξωστρεφής συναισθηματικός, ο εξωστρεφής

αισθητηριακός, ο εξωστρεφής ενορατικός, ο εσωστρεφής διανοητικός, ο εσωστρεφής συναισθηματικός, ο εσωστρεφής αισθητηριακός και ο εσωστρεφής ενορατικός.

Πιο αναλυτικά, σύμφωνα με την Myers-Briggs (1962: 48), οι εξωστρεφείς προσανατολίζονται αρχικά προς το εξωτερικό αντικείμενο και έπειτα επιστρέφουν, για να συμπεριλάβουν και να προσαρμόσουν τον εαυτό τους στην κατανόηση της διαντίδρασης. Είναι πιθανό να διακρίνουν υποκειμενική δραστηριότητα μόνο υπό το φως της εξωτερικής πραγματικότητας και των σχέσεων αντικειμένου. Οι εξωστρεφείς άνθρωποι ενδιαφέρονται για κοινωνική δραστηριότητα, είναι επικοινωνιακοί και ανέμελοι.

Οι εσωστρεφείς, ωστόσο, προσανατολίζονται προς τον εσωτερικό κόσμο, ιδίως στον εσωτερικό κόσμο αντικειμένων ή συμπλεγμάτων και εικόνων. Η ενέργειά τους ρέει πρώτα εσωτερικά και έπειτα στην εξωτερική πραγματικότητα. Περιγράφονται ως παθητικοί, αντιδραστικοί και έχουν κλειστό χαρακτήρα.

Τα άτομα που προτιμούν να κατανοούν τα γεγονότα βάσει των αισθήσεών τους είναι ιδιαίτερα παρατηρητικά και ικανά να εντοπίζουν τις πραγματικές διαστάσεις των διαφόρων καταστάσεων. Σε αυτή την κατηγορία επικρατεί η Νόηση.

Όταν όμως η Διαίσθηση υπερισχύει, τότε το άτομο σχηματίζει στο νου του τη συνολική εικόνα και αναζητεί σχέσεις ανάμεσα σε γεγονότα. Πρόκειται για άτομο με ικανότητα στην επινόηση νέων τρόπων αντιμετώπισης καταστάσεων.

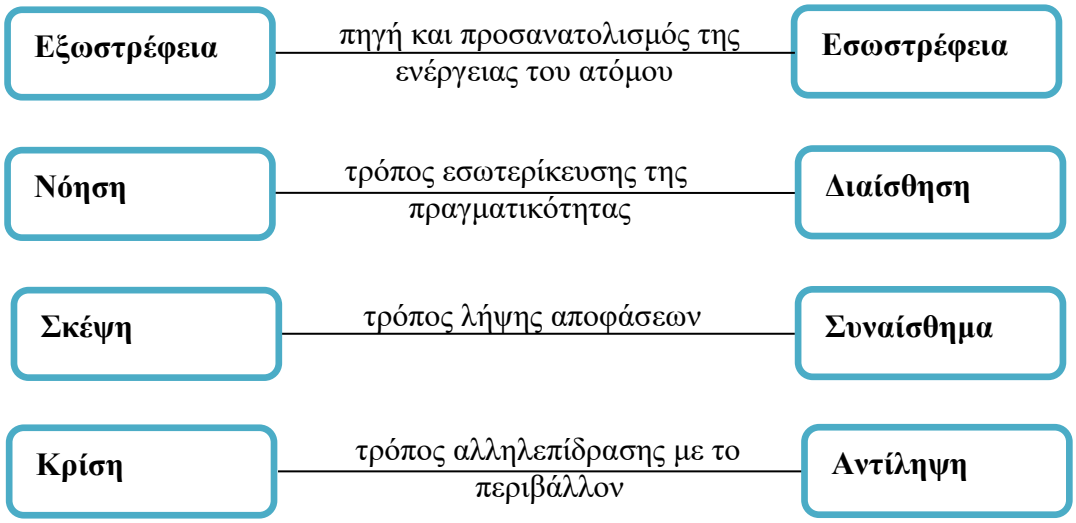
Όσοι χρησιμοποιούν τη Σκέψη για τη λήψη αποφάσεων στηρίζονται στη λογική. Αποστασιοποιούνται από το συναίσθημα και εξετάζουν με αντικειμενικότητα. Επιδιώκουν την κατανόηση των λαθών τους με στόχο την βέλτιστη επίλυση των προβλημάτων.

Αντίθετα, τα άτομα που χρησιμοποιούν το Συναίσθημα εμπλέκονται ψυχικά σε καταστάσεις και ταυτίζονται με συνανθρώπους τους. Δίνουν έμφαση στον άνθρωπο που αποτελεί και επίκεντρο των αποφάσεών τους.

Στον ψυχολογικό τύπο που η Κρίση χρησιμοποιείται στην αντίληψη των ερεθισμάτων δίνεται έμφαση στην πειθαρχία. Τα άτομα που εντάσσονται σε αυτό τον τύπο ακολουθούν πρόγραμμα, ολοκληρώνουν τους στόχους τους και επιζητούν να ελέγχουν τη ζωή τους.

Αντίθετα, τα άτομα με ιδιαίτερα αναπτυγμένη την Αντίληψη ζουν αυθόρμητα, αφού νιώθουν ότι περιορίζονται από προγράμματα. Εμπιστεύονται περισσότερο την επινοητικότητα τους και προσαρμόζονται ανάλογα με τις περιστάσεις.

Ένα από τα γνωστότερα ψυχολογικά τεστ βασισμένα στην παραπάνω τυπολογία, το Ερωτηματολόγιο Τύπων Προσωπικότητας Myers-Briggs Type Indicator (MBTI) που έχει σχεδιαστεί για να μετρά τις ατομικές διαφορές σύμφωνα με την τυπολογία του Jung, ενισχύει τις απόψεις του για το ότι τα άτομα διαφέρουν ανάλογα με τον τύπο στον οποίο ανήκουν, ανάλογα με το περιεχόμενο των ονείρων τους, διαφέρουν στο πώς βιώνουν τις εμπειρίες τους και πώς επεξεργάζονται τις αναμνήσεις τους. Το συγκεκριμένο τεστ αποτελεί ένα αυτό-περιγραφικό ερωτηματολόγιο και αφορά σε δεκαέξι τύπους προσωπικότητας, έναντι των οκτώ του Jung και έχει ευρεία αποδοχή (Hjelle & Ziegler, 1992: 177). Είναι ένα από τα δύο ερωτηματολόγια προσωπικότητας που χορηγήθηκαν στους μαθητές που συμμετείχαν στην έρευνα για την υλοποίηση της παρούσας διατριβής.



Σχήμα 1: Οι τέσσερις διπολικές διαστάσεις του ερωτηματολογίου MBTI.

2.3.2 Θεωρίες για τα χαρακτηριστικά της προσωπικότητας

Μια μεγάλη κατηγορία θεωριών δίνει έμφαση στα χαρακτηριστικά της προσωπικότητας. Οι θεωρητικοί της Προσωπικότητας υποστηρίζουν ότι οι άνθρωποι διαθέτουν ορισμένα χαρακτηριστικά γνωρίσματα (personality traits) που τους προδιαθέτουν να δρουν κατά συγκεκριμένους και συνεπείς τρόπους. Τα χαρακτηριστικά αυτά εξηγούν γιατί οι άνθρωποι διαφέρουν μεταξύ τους και συμπεριφέρονται ποικιλοτρόπως σε ίδιες ή παρόμοιες καταστάσεις. Οι πιο

αντιπροσωπευτικοί θεωρητικοί των χαρακτηριστικών είναι οι Allport, Cattell & Eysenck, οι οποίοι δίνουν έμφαση στις ατομικές διαφορές.

Για την ερμηνεία του όρου «χαρακτηριστικό» κατά τον Hirschberg (1978: 45) επικρατούν δύο απόψεις, η περιληπτική και η ερμηνεία της προδιάθεσης. Το χαρακτηριστικό σύμφωνα με την πρώτη ερμηνεία χρησιμοποιείται ως μέσο κατηγοριοποίησης παρόμοιων συμπεριφορών σε συμπεριφορικά σχήματα, ενώ αντίθετα η δεύτερη θεωρεί τα χαρακτηριστικά την αιτία που προϋπάρχει ως προδιάθεση στο άτομο και το οδηγεί σε συγκεκριμένες συμπεριφορές. Κάθε άνθρωπος έχει ένα μοναδικό σύνολο χαρακτηριστικών και επομένως μοναδική προσωπικότητα, που είναι διαχρονική και σταθερή. Βέβαια, η έννοια της σταθερότητας των χαρακτηριστικών αποτελεί αντικείμενο διαμάχης ανάμεσα στους θεωρητικούς, κυρίως για το αν αλλάζουν κάτω από διαφορετικές συνθήκες.

Η πρώτη προσπάθεια υπολογισμού και ταξινόμησης των χαρακτηριστικών της προσωπικότητας έγινε το 1884 από τον Galton, ο οποίος εφάρμοσε τη λεξικολογική υπόθεση στο πεδίο της ανθρώπινης προσωπικότητας. Η άποψη, δηλαδή, ότι τα σημαντικότερα στοιχεία της ανθρώπινης συμπεριφοράς αναφέρονται κωδικοποιημένα ως χαρακτηριστικά στις περισσότερες ή σε όλες τις γλώσσες του κόσμου, τον οδήγησε στην καταγραφή χιλίων επιθέτων από αγγλικό λεξικό που περιγράφουν χαρακτηριστικά (Galton, 1884.)

Ακολουθούν οι Allport & Odbert που πάλι με τη χρήση λεξικού δημιουργούν ένα κατάλογο δεκαοκτώ χιλιάδων χαρακτηριστικών, τον οποίο χρησιμοποίησε ο Cattell το 1943 για να τα ομαδοποιήσει και να καταλήξει σε δεκαέξι παράγοντες όπως θα δούμε παρακάτω.

Κατά τον Allport τα γνωρίσματα της προσωπικότητας είναι η πηγή της συνέπειας στην ανθρώπινη συμπεριφορά και ευθύνονται για τις διαφορές στον τρόπο αντίδρασης κάτω από τις ίδιες καταστάσεις. Εντοπίζονται έμμεσα μέσω της συχνότητας με την οποία ένα άτομο εκδηλώνει μία συγκεκριμένη μορφή συμπεριφοράς, καθώς και μέσω της έντασης της εκάστοτε προτιμώμενης αντίδρασης. Τα χαρακτηριστικά της προσωπικότητας κάθε ατόμου κατηγοριοποιούνται σε πρωτεύοντα (κυρίαρχη προδιάθεση), κεντρικά (γενικά) και δευτερεύοντα (στάση ζωής, προτιμήσεις). Τα χαρακτηριστικά αυτά δεν είναι πλήρως παρόντα κατά τη γέννηση, αλλά αναπτύσσονται ως αποτέλεσμα μάθησης μέσα σε ένα περίπλοκο περιβάλλον.

Ο Cattell χρησιμοποιώντας τη μέθοδο της παραγοντικής ανάλυσης για την ανακάλυψη του τρόπου οργάνωσης των χαρακτηριστικών της προσωπικότητας

κατέληξε σε δεκαέξι παράγοντες, που χωρίζονται σε δύο κατηγορίες, τα πηγαία ή βασικά χαρακτηριστικά, τα οποία είναι υποκείμενες δομές και αποτελούν τη βάση της προσωπικότητας και τα επιφανειακά χαρακτηριστικά, τα οποία είναι σύνολα συμπεριφορών και αποτελούνται από σειρά στοιχείων. Ως αποτέλεσμα αυτού του θεωρητικού μοντέλου, κατασκευάστηκε το Ερωτηματολόγιο Προσωπικότητας των 16 Παραγόντων (Personality Factor Questionnaire: 16PF). Οι παράγοντες, τους οποίους μετρά το ερωτηματολόγιο, καλύπτουν ένα ευρύ πεδίο από χαρακτηριστικά της ανθρώπινης προσωπικότητας, ιδιαίτερα από το πεδίο των ικανοτήτων και της ιδιοσυγκρασίας (Cattell, 1970). Η ανάλυση της θεωρίας του Cattell αποτέλεσε τη βάση για την ερευνητική ανάπτυξη του μοντέλου των Πέντε Παραγόντων των McCrae & Costa (1996), η ανάλυση του οποίου ακολουθεί.

Πίνακας 1: Οι 16 παράγοντες για την προσωπικότητα Cattell, R. B. (1957).

1. Συγκρατημένος	Εξωστρεφής
2. Λιγότερο Έξυπνος	Περισσότερο Έξυπνος
3. Σταθερός, Ισχυρό Εγώ	Συναισθηματικός/Νευρωτισμός
4. Ταπεινός	Διεκδικητικός
5. Σοβαρός	Εύθυμος
6. Συμφεροντολόγος	Ευσυνείδητος
7. Ντροπαλός	Ριψοκίνδυνος
8. Επίμονος	Διαλλακτικός
9. Γεμάτος Εμπιστοσύνη	Καχύποπτος
10. Πρακτικός	Επινοητικός
11. Ευθύς	Πονηρός
12. Ήρεμος	Ενεργητικός
13. Συντηρητικός	Τολμηρός
14. Εξαρτημένος από την ομάδα	Αυτάρκης
15. Απείθαρχος	Πειθαρχημένος
16. Χαλαρός	Σφιγμένος

Ο Eysenck στη θεωρία των Τριών Παραγόντων διαφοροποιεί την έννοια του «χαρακτηριστικού» από την έννοια του «τύπου» (Pervin & John, 1997). Το χαρακτηριστικό αναφέρεται σε μια ενότητα σχετικών συμπεριφορών οι οποίες εμφανίζονται μαζί. Ο τύπος είναι ευρύτερη έννοια που περιλαμβάνει συσχετιζόμενα χαρακτηριστικά. Κατά τον Eysenck, λοιπόν, οι συμπεριφορές που είναι σχετικές μεταξύ τους συνιστούν ένα χαρακτηριστικό και ανήκουν στατιστικά στον ίδιο παράγοντα. Στόχος της τυπολογίας είναι η ερμηνεία των αλληλοσυσχετίσεων που υπάρχουν στα χαρακτηριστικά της προσωπικότητας.

Σύμφωνα με τη θεωρία του Eysenck, γνωστή με το αρκτικόλεξο P-E-N, οι καταγεγραμμένες στο DNA κληρονομικές καταβολές ασκούν επίδραση στις νευροφυσιολογικές λειτουργίες, που με τη σειρά τους επιδρούν στις εκδηλώσεις της ανθρώπινης συμπεριφοράς, η οποία μπορεί να αποτιμηθεί στη βάση τριών διαστάσεων της προσωπικότητας, της Εξωστρέφειας (Extraversion-E), του Νευρωτισμού (Neuroticism-N) και του Ψυχωτισμού (Psychoticism-P). Οι δυο πρώτες διαστάσεις περιγράφηκαν για πρώτη φορά το 1947, ενώ η τρίτη προστέθηκε τη δεκαετία του 1970 (Eysenck, 1992).

Ο Eysenck αποδίδει στις δύο διπολικές διαστάσεις της προσωπικότητας, στην Εσωστρέφεια-Εξωστρέφεια και στο Νευρωτισμό, χαρακτηριστικά των τύπων της ιδιοσυγκρασίας που είχε ήδη αναπτύξει ο Ιπποκράτης. Ο Κώος ιατρός διέκρινε τέσσερις τύπους ιδιοσυγκρασίας, τον αιματώδη, τον μελαγχολικό, τον χολερικό και φλεγματικό βάσει των τεσσάρων σωματικών χυμών. Ο Eysenck απέδωσε στον αιματώδη τύπο χαμηλά ποσοστά σε νευρωτισμό και υψηλά σε εξωστρέφεια, στον μελαγχολικό τύπο υψηλά ποσοστά σε νευρωτισμό και χαμηλά σε εξωστρέφεια, στον χολερικό τύπο υψηλά ποσοστά σε νευρωτισμό και εξωστρέφεια, στον φλεγματικό χαμηλά ποσοστά σε νευρωτισμό και σε εξωστρέφεια.

2.3.2.1 Το μοντέλο των Πέντε Παραγόντων

Η προσέγγιση των γνωρισμάτων έχει αποδειχθεί ιδιαίτερα διαφωτιστική, καθώς αποδίδει στην απόκτηση συγκεκριμένων γνωρισμάτων της προσωπικότητας την συνέπεια στην συμπεριφορά. Ωστόσο, τα προηγούμενα μοντέλα των Cattell και

Eysenck δεν βρήκαν την αναμενόμενη ανταπόκριση, διότι το πρώτο θεωρήθηκε δύσχρηστο λόγω των δεκαέξι παραγόντων και το δεύτερο ανεπαρκές.

Σήμερα, οι θεωρίες περί χαρακτηριστικών φαίνεται να συγκλίνουν στο μοντέλο των Πέντε Παραγόντων ή Διαστάσεων (Five-factor Model) ως επαρκές για την χαρτογράφηση των ατομικών διαφορών. Σύμφωνα με αυτό, η ανθρώπινη προσωπικότητα στο γενικότερο επίπεδο ανάλυσής της ρυθμίζεται από πέντε διαπολιτισμικά και γενετικά προκαθορισμένους παράγοντες που είναι η Δεκτικότητα στην εμπειρία (Openness to Experience-O), η Ευσυνειδησία (Conscientiousness-C), η Εξωστρέφεια (Extraversion-E), η Προσήνεια (Agreeableness-A) και ο Νευρωτισμός (Neuroticism-N). Τα αρχικά των παραγόντων σχηματίζουν τη λέξη OCEAN στα Αγγλικά.

Τα τελευταία εξήντα περίπου χρόνια το μοντέλο των Πέντε Παραγόντων έχει καθιερωθεί στην Ψυχολογία και η έρευνα έχει τεκμηριώσει τη σχέση ανάμεσα στα χαρακτηριστικά του μοντέλου και σε γλωσσικά στοιχεία (Pennebaker & King, 1999; Oberlander & Gill, 2006).

Βέβαια, το μοντέλο σε καμία περίπτωση δεν υποστηρίζει ότι οι διαφορές στην προσωπικότητα οφείλονται αποκλειστικά στους πέντε παράγοντες, αλλά ότι από ένα πλήθος χαρακτηριστικών οι παράγοντες αυτοί επικρατούν με υψηλά ποσοστά. Έτσι, το μοντέλο είναι γνωστό και ως Μεγάλη Πεντάδα (Big Five), διότι σε κάθε παράγοντα περιλαμβάνονται πολλά χαρακτηριστικά, τα οποία θα δούμε παρακάτω. (Pervin & John, 1997).

Η επιστημονική τεκμηρίωση του μοντέλου έχει αξιολογηθεί με βάση τέσσερα κριτήρια που πρότειναν οι Paul Costa & Robert McCrae (1985). Κάθε παράγοντας για να θεωρείται σημαντικός πρέπει να παραμένει σταθερός για μεγάλο χρονικό διάστημα και να εμφανίζει ίδια αποτελέσματα σε διαφορετικές έρευνες, τα χαρακτηριστικά του να επιβεβαιώνονται όχι μόνο από επιστημονικές έρευνες αλλά και στις διαπροσωπικές σχέσεις. Ως τρίτο κριτήριο ορίζεται η εμφάνιση του παράγοντα σε διάφορα πολιτισμικά πλαίσια και ως τέταρτο να έχει βιολογικό υπόβαθρο (Τσαούσης, 1999).

Είναι αρκετές οι διαπολιτισμικές μελέτες που επιβεβαιώνουν την εγκυρότητα του μοντέλου σε ποικίλα πολιτισμικά πλαίσια όχι μόνο ευρωπαϊκά αλλά και ασιατικά (Ισραήλ, Φιλιππίνες, Κίνα, Ιαπωνία, Ινδία). Σημαντικό είναι να αναφέρουμε ότι το μοντέλο των Πέντε Παραγόντων έχει εφαρμοστεί για την περιγραφή της προσωπικότητας και στην Ελλάδα και έχει επιβεβαιωθεί η εγκυρότητά του (Παυλόπουλος & Μπεζεβέγκης, 1999). Συγχρόνως, η εκδήλωση νευρωτισμού, η

εξωστρέφεια και η εσωστρέφεια αποδεικνύεται πως έχουν βιολογικό υπόστρωμα (Eysenck, 1990).

Υπάρχουν ωστόσο και κάποιοι ψυχολόγοι που ασκούν αρνητική κριτική στο μοντέλο με το επιχείρημα ότι η ανθρώπινη προσωπικότητα είναι τόση σύνθετη που δεν μπορεί να περιγραφεί μόνο με πέντε διαστάσεις και κάποιοι άλλοι που υποστηρίζουν ότι αρκούν οι τρεις παράγοντες π.χ. του Eysenck.

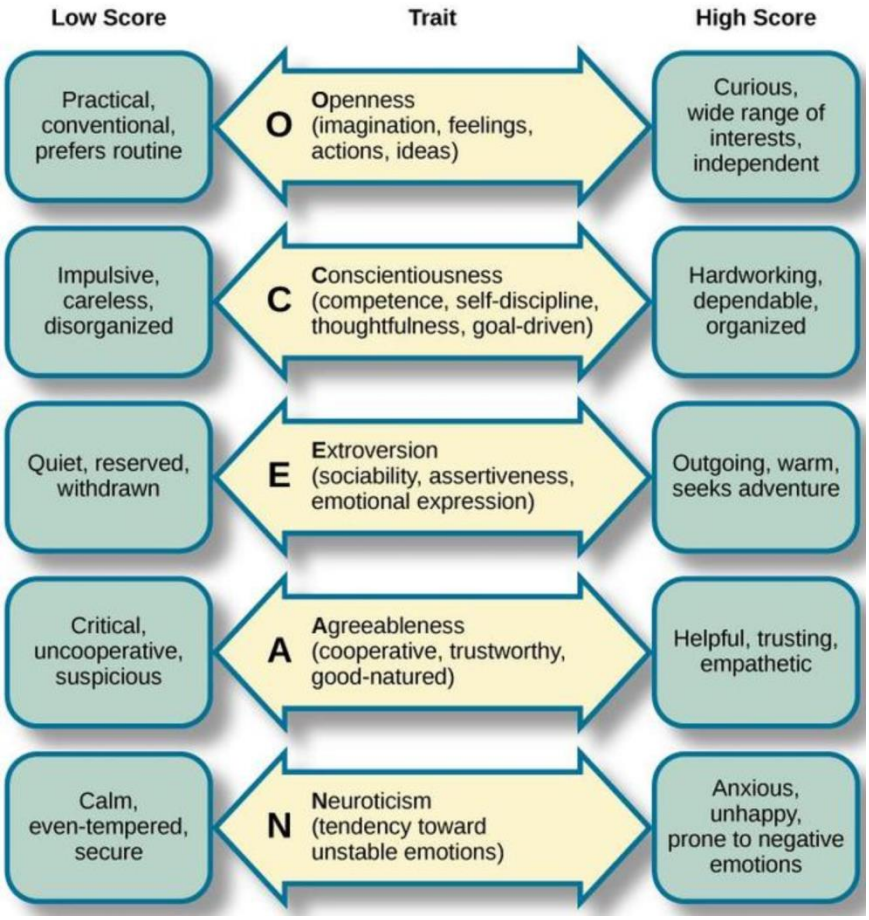
Όπως προαναφέραμε, ως βάση του μοντέλου των Πέντε Παραγόντων θεωρείται η λεξικολογική υπόθεση και η συνακόλουθη δημιουργία ταξινομιών που βασίζονται σε λεξικά, αλλά περισσότερο επηρέασε τη θεωρία των χαρακτηριστικών της προσωπικότητας η μελέτη των Costa & McCrae. Οι προαναφερθέντες συνέθεσαν ερωτηματολόγια για να μελετήσουν τη δομή της προσωπικότητας σε αντίθεση με τους προηγούμενους μελετητές που δημιούργησαν καταλόγους επιθέτων.

Οι ανωτέρω ερευνητές τεκμηρίωσαν το μοντέλο των Πέντε Παραγόντων στα εξής ερωτηματολόγια: Ερωτηματολόγιο προσωπικότητας του Eysenck (1985), Πολυδιάστατο Ερωτηματολόγιο Προσωπικότητας της Μινεσσότα (1986), Ταξινόμηση Q της Καλιφόρνιας (1986), Ερωτηματολόγιο Καταστάσεων-Χαρακτηριστικών Προσωπικότητας (1987), Έντυπο Έρευνας Προσωπικότητας του Jackson (1988), Αναθεωρημένες Διαπροσωπικές Κλίμακες Επιθέτων (1989), Τυπολογικό Δείκτη Myers-Briggs (1989) και στο NEO Ερωτηματολόγιο Πέντε-Παραγόντων (1989).

Το 1992 κατασκεύασαν το NEO-Αναθεωρημένο Ερωτηματολόγιο της Προσωπικότητας (NEO-Personality Inventory Revised/ NEO-PI-R) για τη μέτρηση των Πέντε Παραγόντων. Στο πρώτο στάδιο χρησιμοποίησαν τους παράγοντες νευρωτισμό, εξωστρέφεια και δεκτικότητα στην εμπειρία και στο επόμενο στάδιο αξιοποιώντας σχετικές έρευνες πρόσθεσαν την προσήνεια και την ευσυνειδησία. Τέλος στο τρίτο στάδιο ο κάθε παράγοντας διαφοροποιήθηκε σε έξι χαρακτηριστικά, τα οποία μετρώνται από οκτώ ερωτήσεις το καθένα. Με τις μετέπειτα έρευνές τους απέδειξαν πως το νέο μοντέλο που είχε προκύψει μπορούσε να καλύψει πάρα πολλά θεωρητικά πλαίσια και τεκμηρίωσαν την ύπαρξή του σε συγκεκριμένα ερωτηματολόγια άλλων μελετητών.

Ένα άλλο θέμα που προκύπτει είναι η ονοματοδοσία των παραγόντων. Οι θεωρητικοί των χαρακτηριστικών της προσωπικότητας αναφέρονται στους ίδιους παράγοντες με παρεμφερείς όρους. Ο νευρωτισμός των Costa & McCrae απαντά και ως συναισθηματική σταθερότητα, η εξωστρέφεια ως ορμητικότητα (Costa & McCrae, 1985; Digman, 1988), ενώ αντί για προσήνεια συναντάμε τους όρους φιλικότητα,

φιλική προσαρμογή-εχθρική προσαρμογή και προσήνεια-ανταγωνισμός (Digman & Takemoto-Chock, 1981). Ο παράγοντας που ονομάστηκε ευσυνειδησία λέγεται και συνέπεια ή επιθυμία (Costa & McCrae 1985; Goldberg, 1990) και τέλος η δεκτικότητα στην εμπειρία από άλλους αναφέρεται ως κουλτούρα ή διανόηση (Costa & McCrae, 1985; Goldberg, 1990). Πάντως, κοινά αποδεκτοί είναι οι όροι που διατύπωσαν οι Costa & McCrae το 1985 (Pervin & John, 1997).



Σχήμα 2: Αντιστοίχιση του μοντέλου των Πέντε Παραγόντων σε χαρακτηριστικά (Goldberg, 1990).

Πιο αναλυτικά, για κάθε παράγοντα και τις υποκλίμακές του ισχύουν τα ακόλουθα (Piedmont, 1998): Η Δεκτικότητα στην εμπειρία έχει σχέση με την αναζήτηση δραστηριότητας και νέων εμπειριών και τα χαρακτηριστικά είναι: φαντασία, αισθητική, συναισθήματα, ενεργητικότητα, ιδέες, αξίες. Όσοι ανήκουν σε αυτή την κατηγορία έχουν πολλά ενδιαφέροντα, είναι έντονα συναισθηματικά άτομα και ο τρόπος σκέψης τους χαρακτηρίζεται από ανεξαρτησία και δημιουργικότητα.

Η Ευσυνειδησία μετρά το βαθμό οργανωτικότητας του ατόμου και την επιμονή του για την επίτευξη στόχων. Οι επιμέρους έννοιες που τον προσδιορίζουν είναι: ικανότητα, επιμέλεια, συνέπεια στις αρχές, φιλοδοξία, επιμονή, περίσκεψη. Τα ευσυνείδητα άτομα χαρακτηρίζονται από οργάνωση και τάξη στη ζωή τους όλα γίνονται βάσει προγράμματος. Είναι φιλόδοξα, πειθαρχημένα, διακρίνονται από σοβαρότητα και σπάνια αθετούν τις υποσχέσεις τους.

Η Εξωστρέφεια αξιολογεί το επίπεδο της κοινωνικής αλληλεπίδρασης και εκφράζει την ένταση των συναισθημάτων και την ικανότητα του ατόμου να νιώσει ευχαρίστηση. Τα χαρακτηριστικά είναι: εγκαρδιότητα, κοινωνικότητα, σιγουριά, δραστηριότητα, αναζήτηση συγκινήσεων, θετικά συναισθήματα. Τα εξωστρεφή άτομα είναι δραστήρια, ενθουσιάζονται και αγαπούν τη ζωή, είναι ομιλητικά και αισιόδοξα.

Η Προσήνεια δηλώνει την ποιότητα στην διαπροσωπική συμπεριφορά και τις απόψεις του ατόμου προς τους άλλους και προσδιορίζεται από: εμπιστοσύνη, ευθύτητα, αλτρουισμό, συμβιβαστικότητα, μετριοφροσύνη, ευαισθητοποίηση. Άτομα με υψηλό σκορ στην προσήνεια αγαπούν και ενδιαφέρονται για τον συνάνθρωπο. Είναι ευαίσθητα, καλοπροαίρετα, εμπιστεύονται εύκολα τους άλλους και προσπαθούν να αποφεύγουν τις συγκρούσεις.

Ο Νευρωτισμός ελέγχει την προσαρμοστικότητα του ατόμου και τη συναισθηματική του σταθερότητα και περιλαμβάνει τα εξής επιμέρους χαρακτηριστικά: άγχος, επιθετικότητα, θλίψη, ντροπαλότητα, παρορμητικότητα, ευαισθησία. Τα άτομα με υψηλό σκορ σε αυτό τον παράγοντα διακατέχονται από διαρκές άγχος. Βιώνουν φοβίες, θλίψη και χάνουν εύκολα την ψυχραιμία τους.

Συνοψίζοντας, το μοντέλο των Πέντε Παραγόντων έχει πρακτικές εφαρμογές σε πολλά πεδία μερικά από τα οποία είναι η κλινική ψυχολογία και ψυχοπαθολογία, ο επαγγελματικός προσανατολισμός, η εκπαιδευτική ψυχολογία. Όπως προέκυψε από την επισκόπηση της βιβλιογραφίας, παρά τις όποιες αντιρρήσεις έχουν διατυπωθεί αποτελεί ένα από τα πιο τεκμηριωμένα μοντέλα στον τομέα της προσωπικότητας καθώς προσφέρει αποτελέσματα με αξιοπιστία και εγκυρότητα. Άλλωστε αυτό καταδεικνύεται και από τη μακρόχρονη ιστορία και χρήση του σε διεθνή κλίμακα.

Λόγω της σπουδαιότητας του μοντέλου, αλλά και της ευρείας χρήσης του στη βιβλιογραφία για την αυτόματη πρόβλεψη της προσωπικότητας-χρησιμοποιείται από την πλειοψηφία των ερευνητών του συγκεκριμένου πεδίου- χορηγήσαμε στους συμμετέχοντες στην έρευνά μας ψυχολογικό τεστ βασισμένο στο μοντέλο των Πέντε Παραγόντων.

Ολοκληρώνοντας το κεφάλαιο της σύντομης επισκόπησης των θεωριών της προσωπικότητας, ανεξάρτητα από τα πλεονεκτήματα και τα μειονεκτήματα της κάθε μιας, αυτό που έχει ιδιαίτερη αξία σύμφωνα με τους Pervin & John (1997: 31) είναι πως όλες προσέφεραν πολλά πιθανά κομμάτια επίλυσης του περίπλοκου παζλ που ονομάζεται προσωπικότητα.

2.4 Προσωπικότητα και Γλώσσα

Ο τρόπος με τον οποίο χρησιμοποιεί ένα άτομο τη γλώσσα ως κώδικα επικοινωνίας αποκαλύπτει πολλές πληροφορίες για την προσωπικότητά του. Η επιλογή συγκεκριμένων μορφολογικών, συντακτικών δομών και λεξιλογικών επιλογών μπορεί να είναι ενδεικτική της ηλικίας, του φύλου, της κοινωνικής θέσης, των συναισθημάτων του. Για το λόγο αυτό, άλλωστε, αντιλαμβανόμαστε αν ο ομιλητής ή συγγραφέας ενός κειμένου είναι εξωστρεφής, συναισθηματικά φορτισμένος, απόμακρος ή νευρωτικός. Ένα βασικό στοιχείο, λοιπόν, που πρέπει να εξεταστεί είναι η σχέση προσωπικότητας και γλώσσας.

Γενικότερα, η επικρατούσα θεώρηση είναι ότι η προσωπικότητα επηρεάζει και κατευθύνει τη συμπεριφορά μας, τις σκέψεις, τα συναισθήματά μας, τις διαπροσωπικές σχέσεις και βέβαια τη γλωσσική παραγωγή. Οι άνθρωποι μιλούν και γράφουν με διαφορετικό μεταξύ τους τρόπο ακόμα και αν θέλουν να εκφράσουν το ίδιο περιεχόμενο. Ο χρήστης της γλώσσας επιλέγει το κατάλληλο επίπεδο λόγου ανάλογα με τη συγκεκριμένη περίσταση γλωσσικής επικοινωνίας διαμορφώνοντας έναν εξατομικευμένο τρόπο ομιλίας ή γραφής, στον οποίο όμως καθοριστικής σημασίας είναι η προσωπικότητα. Οι ερευνητές του τομέα υποστηρίζουν ότι ο κάθε άνθρωπος έχει ένα χαρακτηριστικό τρόπο χρήσης της γλώσσας, ένα είδος συγγραφικού αποτυπώματος (Juola, 2008). Οι ψυχολόγοι έχουν δείξει τη συσχέτιση μεταξύ χαρακτηριστικών της προσωπικότητας και γλωσσικών στοιχείων, όπως λεξικών κατηγοριών, ν-γραμμάτων (ακολουθία από ν-πλήθος χαρακτήρων ή λέξεων).

Τα παραπάνω επιβεβαιώνονται και ερευνητικά· ο Sanford (1942) υποστήριξε πως η γλώσσα αποκαλύπτει την ιδιοσυγκρασία του καθενός και μελέτησε τη γλωσσολογική του ατομικότητα. Ο Bradac (1990) τονίζει ότι όλα τα επίπεδα της γλώσσας (φωνολογία, μορφολογία, σύνταξη, σημασιολογία, πραγματολογία)

επηρεάζουν τον δέκτη του μηνύματος. Έρευνες δείχνουν πως τα χαρακτηριστικά της προσωπικότητας επιδρούν στη γλωσσική παραγωγή του καθενός (Pennebaker & King, 1999). Ο Gill (2003) τονίζει ότι η προσωπικότητα προβάλλεται μέσω της γλώσσας αλλά και ότι η προσωπικότητα μπορεί να γίνει αντιληπτή στον δέκτη μέσω της γλώσσας. Επιπλέον, αναφέρει ότι διαφορετικά χαρακτηριστικά της προσωπικότητας επηρεάζουν διαφορετικά επίπεδα της γλωσσικής παραγωγής. Οι Pennebaker, Mehl & Niederhoffer (2003) αναφέρονται στην ψυχολογική πλευρά της γλώσσας και επικεντρώνουν το ενδιαφέρον τους στην επιλογή λέξεων από τον χρήστη της γλώσσας ως ενδεικτικό στοιχείο του χαρακτήρα του. Η γλώσσα θα μπορούσε να διαγιγνώσκει την ψυχολογική κατάσταση ενός ατόμου.

Κοινωνικοί ψυχολόγοι έχουν τονίσει πως η χρήση των λέξεων, ο επιτονισμός, η προφορά και άλλα γλωσσικά στοιχεία αποκαλύπτουν την κοινωνική, οικονομική και ψυχολογική τους θέση (Pennebaker & Stone 2003). Άλλωστε, κάθε δημιουργία κειμένου αποκαλύπτει και ενσωματώνει στοιχεία της προσωπικής και κοινωνικοπολιτισμικής μας ταυτότητας (Γεωργακοπούλου & Γούτσος, 2008: 191). Η ποικιλότητα που εκφράζεται στη γλώσσα επηρεάζεται από χαρακτηριστικά του ομιλητή ή συγγραφέα. Σύμφωνα με τον Περήφανο (2019) η ιδιόλεκτος μπορεί να θεωρηθεί ως το χαρακτηριστικό αποτύπωμα της χρήσης της γλώσσας ενός ατόμου, καθώς οι μεμονωμένοι ομιλητές ή συγγραφείς διαθέτουν τα δικά τους ιδιαίτερα γλωσσικά δομικά σχήματα και προτιμήσεις. Από την έρευνά του προκύπτει ότι η ανίχνευση της ιδιολέκτου μέσω υπολογιστικής προσέγγισης μπορεί να χρησιμοποιηθεί για την ομαδοποίηση συγγραφέων με βάση το ύφος, παρέχοντας εργαλεία για επίλυση προβλημάτων συσχέτισης κειμενικού ύφους και προσωπικότητας.

Παρόλο που αντιλαμβανόμαστε τη σπουδαιότητα της σύνδεσης γλώσσας και χαρακτηριστικών της προσωπικότητας του ομιλητή ή συγγραφέα, δεν είναι αρκετές οι έρευνες που έχουν γίνει σε αυτόν τον τομέα. Οι περισσότερες από αυτές αφορούν στον προφορικό λόγο και μάλιστα στο χαρακτηριστικό της Εξωστρέφειας. Αυτό οφείλεται σύμφωνα με τους Oberlander & Gill (2006) στα παραγλωσσικά στοιχεία του προφορικού λόγου, όπως η προφορά και ο επιτονισμός καθώς και στο ότι από κοινωνιογλωσσολογική άποψη ο προφορικός λόγος ανάμεσα σε πρόσωπα της οικογένειας και φίλους προσφέρει πιο αποκαλυπτικά δεδομένα προς επεξεργασία, αφού είναι πιο αυθόρμητος. Τέλος, η εξωστρέφεια ως χαρακτηριστικό είναι πιο εύκολα αναγνωρίσιμο στο λόγο κάποιου και έτσι σε συνδυασμό με τα παραπάνω οι έρευνες επικεντρώθηκαν στον προσδιορισμό των γλωσσικών χαρακτηριστικών που

υποδηλώνουν την εξωστρέφεια του ομιλητή με αποτέλεσμα να έχει μελετηθεί περισσότερο από τα άλλα χαρακτηριστικά τόσο του μοντέλου των Πέντε Παραγόντων όσο και της τυπολογίας των Jung and Briggs-Myers.

Το κενό που υπάρχει σε αυτόν τον τομέα επιχειρεί να καλύψει η παρούσα διατριβή δημιουργώντας σώμα κειμένων και αξιοποιώντας τεχνικές επεξεργασίας φυσικής γλώσσας, ώστε να ερευνηθούν όλα τα χαρακτηριστικά και όλοι οι τύποι και των δύο θεωριών της προσωπικότητας και έτσι να φανεί ότι η σχέση γλώσσας και προσωπικότητας είναι δυνατό να προσδιοριστεί υπολογιστικά.

Κεφάλαιο 3

Υπολογιστική Υφολογία

3.1 Ποσοτική Γλωσσολογία

Η Ποσοτική Γλωσσολογία (Quantitative Linguistics) είναι κλάδος της Γλωσσολογίας που ασχολείται με την ποσοτική ανάλυση της γλωσσικής δομής και τη γλωσσολογική ερμηνεία της. Η ποσοτική ανάλυση χρησιμοποιείται για να ολοκληρωθεί η ποιοτική ανάλυση που διεξάγεται από τη θεωρητική γλωσσολογία. Η χρήση ποσοτικών μεθόδων λειτουργεί συμπληρωματικά με τις ποιοτικές θεωρήσεις ως προς την κατανόηση του γλωσσικού φαινομένου (Μικρός, 2015b: 2).

Η ποσοτική αντιμετώπιση της γλωσσικής χρήσης άρχισε να εφαρμόζεται στην Ελληνική γλώσσα στη δεκαετία του 1980 και συνεχώς διευρύνεται καθώς αυξάνεται η δημιουργία ηλεκτρονικών σωμάτων κειμένων χωρίς τα οποία δεν θα υπήρχε ο κλάδος (Mikros, 2005).

Ανάλογα με το γλωσσικό επίπεδο όπου χρησιμοποιούνται οι ποσοτικές μέθοδοι διακρίνονται οι κλάδοι της Ποσοτικής Γλωσσολογίας, ένας από τους οποίους είναι η Υφομετρία (Stylometry) ή αλλιώς Υπολογιστική Υφολογία, που αναφέρεται στην κειμενική ανάλυση και ασχολείται με την ποσοτική επεξεργασία του ύφους των κειμένων.

Η Υφομετρία είναι ένας διεπιστημονικός κλάδος που εξετάζει τον τρόπο γραφής των κειμένων και το πώς αυτός συνδέεται με την ταυτότητα του συγγραφέα τους ή με άλλα χαρακτηριστικά του όπως το φύλο, η ηλικία, το ψυχολογικό του προφίλ. Ο συγκεκριμένος κλάδος απαιτεί την συνεργασία των επιστημών της Γλωσσολογίας, της Επεξεργασίας Φυσικής Γλώσσας, της Λογοτεχνικής Ανάλυσης, της Στατιστικής, της Ανάκτησης Πληροφορίας και κλάδου της Τεχνητής Νοημοσύνης, της Μηχανικής Μάθησης.

Βασικές εφαρμογές του κλάδου της Υφομετρίας είναι ο εντοπισμός της πατρότητας ενός κειμένου, όταν διεκδικείται από δύο ή περισσότερους πιθανούς συγγραφείς καθώς και η αυτόματη πρόβλεψη των δημογραφικών και ψυχολογικών

χαρακτηριστικών του συγγραφέα. Στα πλαίσια της Υφομετρίας μια σειρά από μετρήσεις σε κείμενα όπως το μέσο μήκος λέξης, το μέσο μήκος πρότασης, η συχνότητα χρήσης γραμμάτων και σημείων στίξης έχει φανεί ότι σχετίζονται με το προσωπικό ύφος των συγγραφέων. Αξιοποιώντας τις μεθόδους της Υφομετρίας γίνεται εφικτή και μάλιστα με ικανοποιητικά αποτελέσματα η έρευνα στο ευρύτερο πεδίο της αυτόματης αναγνώρισης συγγραφέα.

3.1.1 Ύφος

Στο σημείο αυτό κρίνεται απαραίτητο να αναφερθούμε στο «ύφος», αφού πρόκειται για έννοια που έχει σημαντική θέση στην παρούσα διατριβή δεδομένου ότι η μέθοδος της έρευνάς μας είναι η Υφομετρική προσέγγιση, κατά την οποία ποσοτικοποιείται το ύφος των συγγραφέων του σώματος κειμένων που δημιουργήθηκε. Η έννοια του ύφους πρέπει να προσδιοριστεί και λόγω της πολυσημίας που τη συνοδεύει, καθώς παρόλο που μελετάται εδώ και πολλούς αιώνες, οι ορισμοί ποικίλλουν ανάλογα με την εποχή, την περιοχή και το επιστημονικό πεδίο που τη χρησιμοποιεί. Κοινό πάντως χαρακτηριστικό όλων των προσεγγίσεων είναι ότι το ύφος εξατομικεύει τον τρόπο ομιλίας ή γραφής.

Πριν δούμε πώς η Υφομετρία αντιλαμβάνεται το ύφος θα παραθέσουμε τους γενικούς ορισμούς που επικρατούν. Στο φιλολογικό πεδίο, στο Λεξικό Μπαμπινιώτη (2002) το λήμμα «ύφος» ορίζεται ως ο ιδιαίτερος, εξατομικευμένος τρόπος με τον οποίο χρησιμοποιεί καθένας τη γλώσσα σύμφωνα με τις επιλογές του από το σύστημα (λεξιλογικό, συντακτικό κ.λπ.) της γλώσσας. Σύμφωνα με το Λεξικό Τριανταφυλλίδη (1998) «ύφος» είναι ο ιδιαίτερος τρόπος με τον οποίο διατυπώνει κάποιος τα διανοήματά του στο γραπτό ή στον προφορικό λόγο, η συνειδητή επιλογή ορισμένων επαναλαμβανόμενων, κατά κανόνα, δομικών σχημάτων που απαρτίζουν ένα ιδιαίτερο γλωσσικό σύστημα, στο οποίο δίνεται ιδιαίτερη έμφαση στη μορφή.

Για την Υφομετρία το ύφος είναι γλωσσική κατηγορία και χαρακτηρίζει κάθε είδος λόγου. Προσδιορίζεται από μορφές των βασικών γλωσσικών επιπέδων που συνδέονται με τη γλωσσική παραγωγή και δεν ελέγχονται συνειδητά από τον ομιλητή ή συγγραφέα. Πρόκειται για τα γλωσσικά χαρακτηριστικά που εμφανίζονται αυθόρμητα στο λόγο και δεν αποτελούν συνειδητές επιλογές του συγγραφέα κατά τη

διαδικασία της γραφής, όπως το μέγεθος των προτάσεων, το μέγεθος των λέξεων, το ποσοστό των λέξεων που εμφανίζονται μόνο μια φορά στο κείμενο (Μικρός, 2013: 290). Αυτά τα χαρακτηριστικά αξιοποιούνται από την Υφομετρία για εξαγωγή μεταδεδομένων σε σχέση με το κείμενο (π.χ. θέμα, είδος) αλλά και τον συγγραφέα (π.χ. φύλο, ηλικία, προσωπικότητα) .

Οι Herrmann, van Dalen-Oskam & Schöch (2015) προτείνουν έναν νέο, ευρύτερο και πιο αφηρημένο ορισμό που προκύπτει από τη μελέτη του ύφους σε τρεις διαφορετικές παραδόσεις, της Γερμανικής, της Ολλανδικής και της Γαλλικής γλώσσας και φιλολογίας: «ύφος είναι η ιδιότητα των κειμένων που αποτελείται από ένα σύνολο τυπικών χαρακτηριστικών τα οποία μπορούν να παρατηρηθούν ποσοτικά ή ποιοτικά». Με τον όρο σύνολο αναφέρονται στον συνδυασμό πολλών πιθανών χαρακτηριστικών από διάφορα γλωσσικά επίπεδα. Τυπικά χαρακτηριστικά ονομάζουν τα γλωσσολογικά χαρακτηριστικά σε επίπεδο φωνητικό, λεξιλογικό, συντακτικό και σημασιολογικό αλλά και χαρακτηριστικά πέρα από την πρόταση, όπως η αφηγηματική οπτική ή η κειμενική μακροδομή. Με την έρευνά τους επιδιώκουν να συνδυάζουν γλωσσολογία, φιλολογικές μελέτες και την επιστήμη των υπολογιστών.

Όσον αφορά τον στόχο της Υφομετρίας, αυτός είναι να εντοπίσει στο λόγο κάποιου τις δομές που συνδέονται με τη διαδικασία της γλωσσικής παραγωγής και γραφής. Αυτές οι δομές, που είναι αόρατες στον αναγνώστη, αποτελούν ασυνείδητους αλλά μετρήσιμους δείκτες ατομικού ύφους, δείκτες της ιδιολέκτου του καθενός (Πολίτου-Μαρμαρινού, Μικρός & Δημητρούλια, 2011). Οι δείκτες αυτοί συνθέτουν το υφομετρικό «αποτύπωμα» ή «γονιδίωμα» του κάθε συγγραφέα (Μικρός, 2015a: 8-9).

Στη βιβλιογραφία (van Halteren et al., 2005) γίνεται λόγος για την «υπόθεση του ανθρώπινου υφομετρικού γονιδιώματος» (human stylome hypothesis), δηλαδή για την άποψη ότι το ύφος του καθενός είναι μοναδικό, όπως ακριβώς το δακτυλικό του αποτύπωμα ή το γονιδίωμά του. Ωστόσο, η υφομετρική μας ταυτότητα δεν ταυτίζεται απόλυτα όπως η βιολογική. Ο παραπάνω λοιπόν παραλληλισμός δεν σημαίνει ότι ένας συγγραφέας διακρίνεται απόλυτα από έναν άλλο βάσει συγκεκριμένων γλωσσικών χαρακτηριστικών. Πάντως θεωρείται πως το ύφος του κάθε συγγραφέα είναι μοναδικό.

Το ύφος, λοιπόν, ενός κειμένου αφού αναλύεται, όπως αναφέραμε, σε ποσοτικά στοιχεία μπορεί να περιγραφεί με μετρήσιμο τρόπο. Με τη χρήση στατιστικών μεθόδων η Υφομετρία αναλύει ένα κείμενο όχι με τη φιλολογική έννοια, αλλά με την μετατροπή των κειμενικών δεδομένων σε αριθμούς παρέχει μια ποσοτική ανάλυση.

3.1.2 Ιστορική αναδρομή υφομετρικών μελετών

Στο υποκεφάλαιο αυτό σύντομα θα αναφερθούμε σε βασικές μελέτες της υφομετρίας, που αποτέλεσαν σταθμούς στην εξέλιξή της. Η ιστορία της υφομετρίας έχει τις ρίζες της στην Αναγέννηση. Τότε έγιναν επιτυχημένες προσπάθειες υφομετρικής ανάλυσης με στόχο την ανίχνευση της πατρότητας κειμένων.

Χαρακτηριστική είναι η περίφημη αποκάλυψη της πλαστότητας της δωρεάς του δυτικού τμήματος της Ρωμαϊκής Αυτοκρατορίας από τον Μέγα Κωνσταντίνο στον Πάπα Σίλβεστρο. Ο Ιταλός ουμανιστής Lorenzo Valla το 1439 στο έργο του "De falso credita et ementita Constantini Donatione declamatio" εξέτασε τα υφολογικά χαρακτηριστικά της γλώσσας του κειμένου " Constitutum Constantini" και απέδειξε ότι δεν ήταν έργο του Μεγάλου Κωνσταντίνου, αλλά είχε γραφτεί αργότερα, τον 8ο αιώνα μ.Χ.

Στα μέσα του 18ου αιώνα ξεκίνησαν οι έρευνες σχετικά με το ύφος του Σαίξπηρ. Ο λόγος ήταν οι έριδες γύρω από την πατρότητα σαιξπηρικών κειμένων, η γνησιότητα των οποίων ελέγχεται και με υφομετρικά κριτήρια.

Ένα από τα βασικότερα ερευνητικά προβλήματα των κλασικών σπουδών τον 19ο αιώνα ήταν η χρονολόγηση των Πλατωνικών Διαλόγων. Στο έργο "Principes de stylométrie" ο Πολωνός φιλόσοφος Lutoslawski κωδικοποιεί τις αρχές της υφομετρίας και προσεγγίζει το θέμα με υφομετρικές τεχνικές.

Στα τέλη του 19ου αιώνα και ενώ η πατρότητα των έργων του Σαίξπηρ αποτελούσε πρόβλημα υφομετρικής ανάλυσης, ο Mendenhall εξέτασε το μήκος των λέξεων ως κειμενικό χαρακτηριστικό για την εξαγωγή υφομετρικών διαφορών (Mendenhall, 1887). Η ιδέα είχε διατυπωθεί για πρώτη φορά από τον Βρετανό μαθηματικό Augustus De Morgan, με τις μελέτες όμως του φυσικού Thomas Corwin Mendenhall το μήκος λέξεων έγινε γνωστό ως δείκτης συγγραφικής πατρότητας.

Ένας άλλος ερευνητής που συνέβαλε στην επιστήμη της υφομετρίας είναι ο William Benjamin Smith, ο οποίος με το ψευδώνυμο Conrad Mascol για πρώτη φορά μέτρησε το μέσο μήκος πρότασης, τη χρήση σημείων στίξης και τη συχνότητα των λειτουργικών λέξεων (Mascol, 1888: 454), για να δείξει πως δεν ανήκαν όλες οι επιστολές του Αποστόλου Παύλου στον ίδιο.

Στον 20ο αιώνα η εξέλιξη της υφομετρίας ήταν ταχύτατη. Ο Αμερικανός γλωσσολόγος George Kingsley Zipf (1935) διατύπωσε τον νόμο που έχει πάρει το

όνομά του, ο οποίος ορίζει ότι η συχνότητα οποιουδήποτε γλωσσικού τύπου φυσικής γλώσσας είναι αντιστρόφως ανάλογη με την κατάταξή της στη λίστα λεξιλογικής συχνότητας. Έτσι, σε ένα ηλεκτρονικό σώμα κειμένων παρατηρούμε, όντως, πως ο πιο συχνός λεξικός τύπος εμφανίζεται δυο φορές περίπου πιο συχνός από την συχνότητα του δεύτερου πιο συχνού τύπου κ.ο.κ.

Το πιο γνωστό πρόβλημα συγγραφικής πατρότητας που προέκυψε τον 20ο αιώνα ήταν τα Ομοσπονδιακά Κείμενα των ΗΠΑ, τα οποία γράφτηκαν μεταξύ 1787 και 1788 από τρεις μετέπειτα ηγετικές μορφές Alexander Hamilton, John Jay και James Madison, με στόχο την επικύρωση του Συντάγματος των ΗΠΑ από τους πολίτες. Τα συνολικά 85 άρθρα (77 αρχικά τα οποία είχαν δημοσιευθεί σε τρεις εφημερίδες της Νέας Υόρκης και 8 που γράφτηκαν αργότερα) δημοσιεύθηκαν όλα μαζί σε δίτομο βιβλίο με τίτλο "The Federalists: collection of essays, written in favour of the New Constitution, as agreed upon by the Federal Convention, September 17, 1787". Για 12 από αυτά τα κείμενα υπάρχει αμφισβήτηση της πατρότητάς τους. Οι μελετητές που ασχολήθηκαν με τη διερεύνηση της πατρότητας των αμφισβητούμενων κειμένων ήταν οι Mosteller & Wallace (1984). Αρχικά, το 1941, οπότε ξεκίνησαν την έρευνα, χρησιμοποίησαν το μέγεθος των προτάσεων ως υφομετρικό δείκτη, αλλά αφού δεν λειτούργησε διακριτικά, βασίστηκαν σε λεξιλογικές μεταβλητές και μάλιστα σε λειτουργικές λέξεις και αξιοποιώντας το θεώρημα του Bayes απέδωσαν τα κείμενα στον συγγραφέα τους.

Η χρήση του μήκους της πρότασης ως υφολογικού χαρακτηριστικού του συγγραφέα επανέρχεται στην έρευνα από τον Udny Yule (1939). Διαπιστώνει βέβαια τη δυσκολία στον χωρισμό ενός κειμένου σε προτάσεις, αλλά προβαίνει στον έλεγχο της αξιοπιστίας της πρότασης ως δείκτη συγγραφικού ύφους θεωρώντας πως μπορεί να αξιοποιηθεί. Για τον σκοπό αυτό επέλεξε έργα των Bacon, Coleridge, Lamb και Macaulay και όντως συμπέρανε ότι το προτασιακό μέγεθος είναι χαρακτηριστικό του συγγραφέα.

Σημαντική είναι και η προσφορά του John Burrows (1987), αφού για πρώτη φορά χρησιμοποιεί πολυπαραγοντικές στατιστικές μεθόδους σε συνδυασμό με την συχνότητα των πιο συχνών λέξεων για την υφομετρική απόδοση πατρότητας σε συγγραφείς. Από τον Burrows εισήχθη η χρήση της Ανάλυσης Κύριων Συνιστωσών και σε αυτό το εργαλείο βασίστηκε η έρευνα κατά τη δεκαετία του 1990. Εφαρμογή έγινε στο χώρο της λογοτεχνίας και συγκεκριμένα στους ήρωες των διαλόγων των διηγημάτων της συγγραφέως Jane Austen. Την ίδια τεχνική ακολουθεί και σε διήγημα

του Henry Fielding, για να αποδώσει πατρότητα και συγχρόνως να προσδιορίσει το φύλο του συγγραφέα.

Έπονται πολλές μελέτες που, όπως προείπαμε, χρησιμοποίησαν την Ανάλυσης Κύριων Συνιστωσών στις πιο κοινόχρηστες λέξεις. Χαρακτηριστικές είναι αυτές των Holmes & Forsyth (1995) για την εκ νέου ανάλυση των Ομοσπονδιακών Κειμένων, των Craig & Kinney (2009) για έργα του Σαίξπηρ και των Somers & Tweedie (2003) για το έργο «Η Αλίκη στη χώρα των θαυμάτων» και την μίμησή του από άλλο δημιουργό. Σε κάποιες μελέτες η Ανάλυση Κύριων Συνιστωσών χρησιμοποιεί εκτός από τις συχνότερες λέξεις και άλλους υφομετρικούς δείκτες (Baayen, van Halteren, & Tweedie, 1996). Οι Forstall & Scheirer (2010) συνδυάζουν την Ανάλυση Κύριων Συνιστωσών με τεχνικές μηχανικής μάθησης. Τέλος, πρέπει να αναφέρουμε και μια πιο σύγχρονη τεχνική στο χώρο του εντοπισμού του συγγραφέα, τα αποτυπώματα γραπτού λόγου (Writeprints), που αξιοποιεί σύνθετες μεθόδους (Abbasi & Chen, 2008).

Είναι σημαντικό να αναφέρουμε πως η Υφομετρία σημείωσε και αποτυχίες με αποτέλεσμα να αμφισβητηθεί η αξιοπιστία της. Η πρώτη πιο χαρακτηριστική περίπτωση είναι τα γραφήματα Cusum ή QSUM (Μικρός, 2015a: 45-46), τα οποία χρησιμοποιήθηκαν στις δεκαετίες του 1980 και 1990 για να δείξουν αν ένα κείμενο ανήκει εξ ολοκλήρου σε έναν συγγραφέα ή έχει παρέμβει και άλλος. Πρόκειται για μια στατιστική μέθοδο κατά την οποία παρακολουθείται ο μέσος όρος μιας μέτρησης και οι αλλαγές που αυτός υφίσταται κατά την εξέλιξη ενός φαινομένου. Ήταν τόσο εύκολος ο υπολογισμός των διαγραμμάτων και η ερμηνεία τους, που η μέθοδος έγινε δημοφιλής ακόμα και ως δικανική τεχνική. Ωστόσο, η τεχνική παρουσίαζε έλλειψη θεωρητικού υπόβαθρου και αντικειμενικότητας και συγχρόνως αμφισβητήθηκε η ακρίβεια και η εγκυρότητά της. Η διαμάχη έφτασε στο αποκορύφωμα όταν ο άνθρωπος που είχε παρουσιάσει σε έκδοσή του την τελική μορφή της μεθόδου, ο Andrew Morton, σε εκπομπή της βρετανικής τηλεόρασης όχι μόνο δεν κατάφερε να αναλύσει τα κείμενα, αλλά δεν διέκρινε αυτά που ανήκαν σε καταδικασμένους εγκληματίες από αυτά του υπουργού δικαιοσύνης της Αγγλίας.

Μια δεύτερη αποτυχία είναι η απόδοση στον Σαίξπηρ του ποιήματος "A Funeral Elegy" από τον γλωσσολόγο Foster στα τέλη του 1980 (Μικρός, 2015a: 51-52). Η φήμη του μεγάλωσε όταν αποκάλυψε ότι ο ανώνυμος συγγραφέας του έργου "Primary Colors" ήταν ο δημοσιογράφος Joe Klein, ο οποίος αργότερα το παραδέχτηκε. Έτσι, ο Foster χαρακτηρίστηκε ως ο καλύτερος "Literary Detective". Όμως, το 1996 οι Elliot & Valenza με υφομετρικές μελέτες τους απέδειξαν εκτός των

άλλων ότι το ποίημα "A Funeral Elegy" δεν ανήκει στον Σαίξπηρ. Η μεταξύ τους επιστημονική διαμάχη τερματίστηκε το 2002 με την υποχώρηση του Foster.

Σε θεωρητικό πλαίσιο, κατά τον Daelemans (2013), η Υφομετρία πρέπει να εξεταστεί στο πλαίσιο της επεξεργασίας φυσικής γλώσσας ως ένα από τα τρία επίπεδα κατανόησης κειμένου. Τα τρία είδη της γνώσης που μπορούν να εξαχθούν από ένα κείμενο είναι η αντικειμενική (όπως ιδέες, αιτιακές και χρονικές σχέσεις), η υποκειμενική (απόψεις και συναισθήματα) και η μεταγνώση. Η Υφομετρία ανήκει στην τελευταία κατηγορία. Πρόκειται δηλαδή για γνώση γύρω από το ίδιο το κείμενο και κυρίως σχετικά με τον συγγραφέα του.

Σήμερα, που η υφομετρική ανάλυση επικουρείται από τους ηλεκτρονικούς υπολογιστές, οι οποίοι διαθέτουν τεράστιες δυνατότητες επεξεργασίας σωμάτων κειμένων και αξιοποιούν μεθόδους τεχνητής νοημοσύνης και περίπλοκες στατιστικές τεχνικές, οι εφαρμογές του κλάδου είναι εξαιρετικά χρήσιμες και ποικίλες.

Τα αποτελέσματα της Υφομετρίας ενδιαφέρουν όχι μόνο τους ειδικούς στην στατιστική, αλλά μελετητές της λογοτεχνίας, οι οποίοι μέσω της τεχνητής νοημοσύνης, που ανοίγει ένα τεράστιο εύρος δυνατοτήτων, διερευνούν και επιλύουν θέματα λογοτεχνικής πατρότητας, καθηγητές που ζητούν σημάδια λογοκλοπής, δημοσιογράφους για την εξακρίβωση ενός εγγράφου, ανακριτές ενός εγκλήματος, δικηγόρους για μια επίμαχη διαθήκη. Οι Υφομετρικές προσεγγίσεις εφαρμόζονται στη λογοτεχνική έρευνα με κυριότερους στόχους τον προσδιορισμό του χρόνου συγγραφής ενός έργου, τον προσδιορισμό της σχέσης ύφους και λογοτεχνικού είδους ή θέματος. Η Υφομετρία, επίσης, αξιοποιείται στον εντοπισμό δημιουργών κακόβουλων προγραμμάτων, σε ζητήματα πνευματικής ιδιοκτησίας, μπορεί να έχει εφαρμογές στην ψυχογλωσσολογία, στην κοινωνιογλωσσολογία, στην εγκληματολογία (π.χ. πατρότητα τρομοκρατικών μηνυμάτων, διαδικτυακός εκβιασμός) και στην ιατρική, στον τομέα της διάγνωσης (π.χ. εντοπισμός σχιζοφρένειας και αλτσχάιμερ). Ο πλέον βέβαια αποτελεσματικός τομέας εφαρμογής είναι η αναγνώριση της πατρότητας ανώνυμων ή αμφισβητούμενης πατρότητας κειμένων.

3.2 Αυτόματος Εντοπισμός Συγγραφέα

3.2.1 Υφομετρική απόδοση της συγγραφικής πατρότητας

Οι εξελίξεις στην υπολογιστική γλωσσολογία τα τελευταία χρόνια επιτρέπουν ανώτερου τύπου αναλύσεις στα κείμενα. Σε αυτό έχει συμβάλει και η ωρίμανση των εργαλείων όπως οι λημματοποιητές, οι μορφολογικοί και συντακτικοί αναλυτές, που έχουν διευρύνει το πεδίο εφαρμογών της Ανάκτησης Πληροφορίας (Information Retrieval). Ένας νέος και σύγχρονος τύπος ανάλυσης κειμένου που συναντάται στη βιβλιογραφία και εντάσσεται στην παραπάνω κατηγορία εφαρμογών στοχεύει στην αυτόματη Αναγνώριση του Συγγραφέα (Authorship Identification). Ο τομέας αυτός με δεδομένα τα παραπάνω επιστημονικά επιτεύγματα σημειώνει πρόοδο και στην αξιοπιστία των μεθόδων και στην αποτελεσματικότητα των τεχνικών.

Θέματα πατρότητας κειμένων τίθενται από τότε που υπάρχουν κείμενα. Η δυνατότητα, όμως, να απαντάμε στα θέματα αυτά, όχι με τον παραδοσιακό τρόπο, δόθηκε από την ανάπτυξη της στατιστικής και της πληροφορικής. Η βασική ιδέα στον τομέα της αναγνώρισης του συγγραφέα, η οποία επιτυγχάνεται πλέον με υφομετρικές τεχνικές, είναι η δυνατότητα διάκρισης κειμένων που ανήκουν σε διαφορετικούς συγγραφείς μετρώντας κάποια κειμενικά χαρακτηριστικά. Μπορούμε, δηλαδή, να μετρήσουμε τη συχνότητα γλωσσικών χαρακτηριστικών σε κείμενα ενός συγγραφέα και χρησιμοποιώντας τις μετρήσεις αυτές να συνδέσουμε στατιστικά ένα ανώνυμο κείμενο με τον πραγματικό συγγραφέα του.

Όπως είδαμε στο προηγούμενο κεφάλαιο, οι προσπάθειες ποσοτικοποίησης του ύφους ενός κειμένου ανάγονται πολύ πίσω στο χρόνο. Από το πιο γνωστό πρόβλημα συγγραφικής πατρότητας, τα Ομοσπονδιακά Κείμενα των ΗΠΑ μέχρι και τα τέλη της δεκαετίας του 1990 η έρευνα επικεντρώθηκε στον καθορισμό των χαρακτηριστικών εκείνων που θα μπορούσαν να αποτελέσουν υφολογικούς δείκτες. Κατά τον Rudman (1997: 360) είχαν προταθεί ως τότε περίπου χίλιοι διαφορετικοί δείκτες. Χαρακτηριστικό είναι ότι οι έρευνες βοηθούνταν από τους υπολογιστές, αφού δεν υπήρχε σύστημα αναγνώρισης συγγραφέα πλήρως αυτοματοποιημένο.

Με την ραγδαία όμως έκρηξη κειμενικής παραγωγής στο διαδίκτυο η ανάγκη για αποτελεσματική οργάνωση, ακριβή ανάκτηση και κυρίως για τον προσδιορισμό της πατρότητας καθίσταται αναντίρρητη. Έτσι, η αναγνώριση του συγγραφέα εξελίσσεται στο πλαίσιο της δικαστικής γλωσσολογίας, αφού η ανωνυμία του διαδικτύου οδηγεί

συχνά στην κάλυψη νομικών παραβιάσεων αλλά και ποινικών αδικημάτων. Η έρευνα επικεντρώθηκε σε κείμενα ηλεκτρονικού ταχυδρομείου, αναρτήσεις σε ιστολόγια, μηνύματα σε ιστότοπους συζητήσεων λόγω των προβληματισμών που προέκυψαν σχετικά με την πατρότητά τους. Η διερεύνηση της συγγραφικής πατρότητας απαιτείται σε περιπτώσεις πνευματικών δικαιωμάτων, επιστολών τρομοκρατικής οργάνωσης, απειλητικών μηνυμάτων, επιστολών αυτοκτονίας, κακόβουλων προγραμμάτων.

3.2.2 Προβλήματα συγγραφικής απόδοσης

Ο ευρύτερος τομέας της αυτόματης αναγνώρισης του συγγραφέα υποδιαιρείται σε επιμέρους κλάδους/προβλήματα (Μικρός, 2015a: 7-8):

α) Στην απόδοση ενός κειμένου σε συγκεκριμένο συγγραφέα από ένα πεπερασμένο σύνολο συγγραφέων (Authorship Attribution). Στον κλάδο αυτό προσπαθούμε να εντοπίσουμε την πατρότητα ενός ή περισσότερων ανώνυμων κειμένων μέσα από συγκεκριμένη λίστα υποψήφιων συγγραφέων, των οποίων έχουμε δείγμα γραφής. Είμαστε, δηλαδή, σίγουροι ότι ένας από τους συγγραφείς της λίστας μας είναι αυτός που έχει γράψει το κείμενο.

Η πιο σημαντική μελέτη απόδοσης συγγραφικής πατρότητας του 20^{ου} αιώνα είναι τα Ομοσπονδιακά Κείμενα των ΗΠΑ. Δώδεκα από τα συνολικά ογδόντα πέντε κείμενα είναι αμφισβητούμενης πατρότητας τη διερεύνηση της οποίας έκαναν, όπως αναφέρθηκε παραπάνω, οι Mosteller και Wallace με υφομετρική ανάλυση. Το θέμα των κειμένων, το πλήθος των υποψήφιων συγγραφέων καθώς και ο όγκος των κειμενικών δεδομένων φαίνεται ότι επηρεάζουν την επιτυχία της απόδοσης ενός κειμένου σε συγκεκριμένο συγγραφέα (Luyckx, 2010). Για τις τελευταίες εξελίξεις στον τομέα του Authorship Attribution παραπέμπουμε στο διεθνή διαγωνισμό PAN 2019 (Kestemont et al., 2019), στον οποίο υπήρχε η εξής ιδιαιτερότητα σε σχέση με προηγούμενες χρονιές: τα δεδομένα ελέγχου (test texts) περιείχαν και δείγματα κειμένων από συγγραφείς που δεν ανήκαν στη λίστα των υποψηφίων.

β) Στην απόδοση κειμένου σε συγγραφέα που δεν ανήκει σε κλειστό σύνολο - Πιστοποίηση συγγραφέα (Authorship Verification). Για την επίλυση αυτού του προβλήματος πρέπει να εντοπίσουμε την πατρότητα ενός ή περισσότερων ανώνυμων κειμένων μέσα από μια ανοικτή λίστα συγγραφέων των οποίων διαθέτουμε δείγμα

γραφής, αλλά δεν γνωρίζουμε αν είναι υποψήφιοι συγγραφείς. Επομένως, ο συγγραφέας μπορεί να είναι οποιοσδήποτε και έτσι το έργο είναι δυσκολότερο. Παρόλο που η ερευνητική δραστηριότητα είναι σημαντική στον τομέα αυτό, η ακρίβεια αναγνώρισης υπολείπεται των αποτελεσμάτων που επιτυγχάνονται στις έρευνες απόδοσης πατρότητας (van Halteren, 2004; Koppel, Schler & Argamon, 2009).

γ) Στον καθορισμό δημογραφικών και ψυχολογικών χαρακτηριστικών του συγγραφέα (Authorship Profiling), που περιλαμβάνει την κατηγοριοποίηση του συγγραφέα με κριτήριο το φύλο (Μικρός 2015a: 178-191), την ηλικία, την προσωπικότητα. Έτσι, γίνονται προσπάθειες αυτόματης πρόβλεψης του φύλου, της ηλικίας και της ψυχολογικής κατάστασης του συγγραφέα. Γενικότερα, σε αυτή την κατηγορία εντάσσεται ο εντοπισμός οποιασδήποτε ιδιότητας του συγγραφέα ενός κειμένου. Το χαρακτηριστικό που είναι αντικείμενο της παρούσας διατριβής είναι η προσωπικότητα του συγγραφέα.

δ) Στον αυτόματο εντοπισμό λογοκλοπής (Gollub et al., 2013) και κακόβουλης τροποποίησης του περιεχομένου των ιστοσελίδων συνεργατικών ψηφιακών μέσων. Μελετάμε, δηλαδή, την κανονικότητα του υφομετρικού προφίλ ενός κειμένου και τη χρήση ποσοτικών μεθόδων για την αξιολόγησή του. Το θέμα της αλλαγής συγγραφικού στυλ (Data change detection) αποτέλεσε αντικείμενο του διεθνούς διαγωνισμού PAN 2019 (Zangerle et al., 2019), όπου μελετήθηκε από τις ερευνητικές ομάδες εάν ένα κείμενο είναι γραμμένο από έναν ή περισσότερους συγγραφείς και στην περίπτωση που οι συγγραφείς είναι όντως περισσότεροι πόσοι ακριβώς είναι.

Όλοι οι παραπάνω τομείς αποτελούν αντικείμενο έρευνας με αρκετά καλά αποτελέσματα κυρίως στην απόδοση ενός κειμένου σε συγκεκριμένο συγγραφέα από ένα πεπερασμένο σύνολο συγγραφέων. Εάν ο συγγραφέας δεν ανήκει σε κλειστό σύνολο ή δεν ανήκει στη λίστα μας ή ακόμα όταν το πλήθος των υποψηφίων είναι μεγάλος ή το δείγμα γραφής είναι μικρό, τα αποτελέσματα δεν είναι ικανοποιητικά.

3.2.3 Υφομετρική απόδοση συγγραφέα στην Ελληνική γλώσσα

Μελετώντας τη βιβλιογραφία για την αυτόματη αναγνώριση συγγραφέα στα Ελληνικά διαπιστώνουμε ότι η πρώτη προσπάθεια υφομετρικής ανάλυσης έγινε από τον Μικρό (2006, 2007) σε κείμενα που ανακτήθηκαν από το Διαδίκτυο και

συγκεκριμένα 1.200 δημοσιευμένα άρθρα στην εφημερίδα «Τα Νέα» γραμμένα από τέσσερις διαφορετικούς συγγραφείς. Στόχος της έρευνας ήταν να μελετηθεί μια νέα μέθοδος για την επιλογή λέξεων που χαρακτηρίζουν ένα συγγραφέα και οδηγούν στην αυτόματη αναγνώρισή του. Το στατιστικό μοντέλο που χρησιμοποιήθηκε ήταν η Διακριτική Ανάλυση (Discriminant Function Analysis) με 80 λέξεις ενδεικτικές του συγγραφέα σε συνδυασμό με υφομετρικά και ψυχογλωσσολογικά χαρακτηριστικά. Η ακρίβεια απόδοσης της συγγραφικής πατρότητας έφτασε το 92,45%.

Ενδεικτικό παράδειγμα διερεύνησης πατρότητας κειμένων ιστορικού και φιλολογικού ενδιαφέροντος αποτελεί η μελέτη που απέδωσε ανώνυμες μεταφράσεις του 19ου αιώνα στον Αλέξανδρο Παπαδιαμάντη (Πολίτου-Μαρμαρινού κ.ά, 2011).

Οι Mikros & Perifanos (2013) ανέπτυξαν το πρώτο σώμα κειμένων από το Twitter αποτελούμενο από 12.973 αναρτήσεις. Οι χρήστες που επιλέχθηκαν ήταν δέκα δημοφιλή άτομα με έντονη δραστηριότητα στο συγκεκριμένο μέσο κοινωνικής δικτύωσης. Ως χαρακτηριστικά εξήχθησαν τα 1.000 πιο συχνά διγράμματα και τριγράμματα χαρακτήρων και λέξεων με αποτέλεσμα την Υφομετρική Ανάλυση με ν-γράμματα αυξανομένου μήκους και επιπέδου, τα οποία αποτελούν μια συνεχόμενη ακολουθία ν τεμαχίων ενός κειμένου και υπολογίζονται ανεξαρτήτως γλώσσας. Χρησιμοποιήθηκε ο αλγόριθμος ταξινόμησης Multi-class support vector και από το σώμα κειμένων δημιουργήθηκαν τέσσερις υποομάδες με κείμενα των 25, 50, 75 και 100 λέξεων το καθένα. Η μέγιστη ακρίβεια του συστήματος ήταν 95,1% με 10-πτυχη διασταυρούμενη επικύρωση στο σετ των δεδομένων με τις 100 λέξεις. Από την έρευνα προέκυψε ότι η ακρίβεια επηρεάζεται από το μέγεθος του κειμένου.

Ο Mikros (2013a) παρουσίασε τη δημιουργία σώματος κειμένων από ιστολόγια στην Ελληνική γλώσσα με στόχο την απόδοση συγγραφικής πατρότητας. Πρόκειται για αναρτήσεις από 10 ιστολόγια ανδρών και 10 γυναικών (406.460 λέξεις) με κοινό θέμα. Χρησιμοποιώντας αλγόριθμο μηχανικής μάθησης και υφομετρικά χαρακτηριστικά έφτασε σε ακρίβεια 85,4%.

Από τον Μικρό (2015a: 122-143) διεξήχθη πείραμα υφομετρικής απόδοσης συγγραφέα σε ανώνυμα κείμενα που ανήκουν στον έναν μεταξύ δύο πιθανών συγγραφέων. Το σώμα κειμένων αποτελούμενο από τριακόσια (300) άρθρα ειδησεογραφικού χαρακτήρα με πολιτική θεματολογία αντλήθηκε από την εφημερίδα «Τα Νέα». Επιλέχθηκαν υφομετρικά χαρακτηριστικά και χρησιμοποιήθηκε ως αλγόριθμος ταξινόμησης η λογιστική παλινδρόμηση. Η απόδοση της συγγραφικής πατρότητας ενός ανώνυμου κειμένου επιτεύχθηκε με ακρίβεια 97,7%.

Πιο απαιτητικό είναι το πρόβλημα υφομετρικής απόδοσης συγγραφέα με περισσότερους από δύο υποψήφιους. Η έρευνα του Μικρού (2015a: 145-170) καλύπτει την υφομετρική απόδοση αγνώστων κειμένων σε ομάδα 20 πιθανών συγγραφέων. Το σώμα κειμένων στο οποίο εφαρμόστηκε ο αλγόριθμος Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines, SVM) περιλαμβάνει αναρτήσεις ιστολογίων. Από τα κείμενα αυτά εξήχθησαν 1.553 υφομετρικά χαρακτηριστικά και η ακρίβεια της απόδοσης έφτασε στο 88%.

Οι Mikros & Markopoulos (2017) σε έρευνα που πραγματοποίησαν σε 250 αναρτήσεις πέντε ελληνικών ιστολογίων διαφορετικών συγγραφέων ανέπτυξαν μοντέλο απόδοσης συγγραφέα του οποίου η ακρίβεια έφτασε στο 84,8%. Τα υφομετρικά χαρακτηριστικά που επιλέχθηκαν για την ταξινόμηση από τον αλγόριθμο Naive Bayes ήταν τα ν-γράμματα λέξεων. Από τα αποτελέσματα προέκυψε ότι οι πολυλεκτικές ακολουθίες μεταβλητού μήκους ως δείκτες συνέβαλαν, ώστε ο αλγόριθμος να πετύχει πιο υψηλό ποσοστό ακρίβειας σε σχέση με τα διγράμματα και τριγράμματα λέξεων.

Σε διατριβή για την υπολογιστική αναπαράσταση της ιδιολέκτου (Περήφανος, 2019) ανιχνεύεται σε σώμα κειμένων του Twitter στα Ελληνικά το υφολογικό αποτύπωμα των 4.949 χρηστών, το οποίο μπορεί να χρησιμοποιηθεί στη συσταδοποίηση ιδιολεκτικής ομοιότητας και να εφαρμοστεί αποτελεσματικά στην αναγνώριση συγγραφέα.

3.3 Αυτόματος εντοπισμός χαρακτηριστικών του συγγραφέα

Στο προηγούμενο υποκεφάλαιο αναφέραμε πως ένα από τα προβλήματα της υφομετρικής ανάλυσης που εμπίπτει στον ερευνητικό τομέα του εντοπισμού ενός συγγραφέα είναι η ιχνογράφηση των χαρακτηριστικών του (Authorship Profiling). Όταν μιλάμε ή γράφουμε επικεντρωνόμαστε στο θέμα μας και την πληροφορία που θέλουμε να μεταδώσουμε, αλλά ταυτόχρονα σε ένα ασυνείδητο επίπεδο μεταδίδουμε πληροφορίες για το φύλο μας, την ηλικία μας, την προσωπικότητά μας ακόμα και για την πολιτική μας ιδεολογία. Συχνά μας ενδιαφέρει να προσδιορίσουμε τέτοιου είδους κοινωνιολογικά χαρακτηριστικά όπως είναι το φύλο, η ηλικία, η μόρφωση, η κοινωνική

τάξη καθώς και ψυχολογικούς παράγοντες που περιλαμβάνουν την προσωπικότητα του συγγραφέα, τη νοητική του κατάσταση ή το αν είναι φυσικός ομιλητής ή όχι. Και όταν χρησιμοποιούμε τον όρο «συγγραφέας» εννοούμε το άτομο που παράγει λόγο είτε προφορικό είτε γραπτό οποιουδήποτε είδους (π.χ. μια έκθεση ιδεών, ένα άρθρο, ένα σημείωμα ακόμα και αναρτήσεις στο διαδίκτυο ή και κώδικα προγράμματος υπολογιστή).

Η ψυχολογία υποστηρίζει ότι συγκεκριμένα χαρακτηριστικά της προσωπικότητας συνδέονται με τη γλωσσική συμπεριφορά του ατόμου και αυτή η σχέση μπορεί να μοντελοποιηθεί με τεχνικές επεξεργασίας φυσικής γλώσσας που βασίζονται στη στατιστική. Αυτό αποτελεί βασικό ερευνητικό πρόβλημα της Υφομετρίας, δηλαδή να περιγράψει και να ερμηνεύσει τη σχέση ανάμεσα στην προσωπικότητα του συγγραφέα και στο ύφος της γραφής του, έναν μοναδικό τρόπο να χρησιμοποιεί την γλώσσα, τον ιδιοσυγκρασιακό τρόπο με τον οποίο εκφράζεται.

Η γλώσσα ως προϊόν του ανθρώπινου νου αναμένεται να παρέχει πληροφορίες για τη νοητική κατάσταση του ατόμου που την παρήγαγε. Χαρακτηριστική είναι η περίπτωση της αξιοποίησης γλωσσικών δεικτών στις ιατρικές διαγνώσεις, όπως για παράδειγμα η μελέτη της γλωσσικής παραγωγής των σχιζοφρενών που ανάγεται στον 19ο αιώνα (Covington et al., 2005). Έτσι, έρευνες δείχνουν πως σε τέτοιες καταστάσεις η χρήση απλοϊκών συντακτικών δομών μπορεί να είναι ένδειξη γενικευμένου γνωστικού ελλείμματος, ενώ λάθη μορφολογικά είναι σπάνια στη σχιζοφρένεια.

Άλλες μελέτες παρουσιάζουν γλωσσικά χαρακτηριστικά που είναι ενδεικτικά συγκεκριμένης συναισθηματικής κατάστασης. Για παράδειγμα άτομα αγχωμένα και μελαγχολικά χρησιμοποιούν συχνότερα αντωνυμίες πρώτου ενικού προσώπου (Rude, Gortner & Pennebaker, 2004), μειώνουν την χρήση άρθρων και αυξάνουν την συχνότητα στην χρήση βοηθητικών ρημάτων σε χρόνο ενεστώτα (Pennebaker & Lay, 2002). Άλλοι ερευνητές του ίδιου τομέα (Newman et al., 2003) εξετάζουν το ψέμα σε σχέση με το γλωσσικό ύφος και καταλήγουν στο συμπέρασμα ότι όσοι λένε ψέματα, σε αντίθεση με τους ειλικρινείς, αναφέρονται λιγότερο στον εαυτό τους, χρησιμοποιούν συχνότερα ρήματα κίνησης και περισσότερες αρνητικές λέξεις. Οι χαρακτηριστικές αυτές περιπτώσεις μελετών που έχουν πραγματοποιηθεί αφορούν γενικότερα στο προφίλ του συγγραφέα και αποδεικνύουν ότι το ύφος με το οποίο είναι γραμμένο ένα κείμενο παρέχει σημαντικές πληροφορίες για την ψυχοσύνθεση του συγγραφέα.

3.3.1 Μελέτες εντοπισμού χαρακτηριστικών του συγγραφέα

Το πρόβλημα που επισημαίνεται από τους επιστήμονες του χώρου είναι ο περιορισμένος αριθμός σωμάτων κειμένων για τη διεξαγωγή ερευνών (Verhoeven & Daelemans, 2014). Για το λόγο αυτό οι έρευνες τα τελευταία χρόνια αξιοποιούν κείμενα από τα μέσα κοινωνικής δικτύωσης που συνεχώς αυξάνονται. Παράδειγμα τέτοιας ενδιαφέρουσας έρευνας αποτελεί πρόσφατος διεθνής διαγωνισμός (Wiegmann, Stein & Potthast, 2019), που έθεσε ως στόχο στους συμμετέχοντες τον αυτόματο προσδιορισμό τεσσάρων δημογραφικών στοιχείων, του φύλου, του έτους γέννησης, της έκτασης της φήμης και του επαγγέλματος 48.335 διάσημων προσώπων από τις αναρτήσεις τους στο Twitter.

Στόχος αυτού του κεφαλαίου είναι να παρουσιάσει μια επισκόπηση στις έρευνες που έχουν γίνει στον τομέα της αυτόματης πρόβλεψης χαρακτηριστικών του συγγραφέα και συγκεκριμένα του φύλου, της ηλικίας, της μητρικής γλώσσας και της γλωσσικής ποικιλίας, καθώς τα συγκεκριμένα δημογραφικά χαρακτηριστικά αποτελούν αντικείμενο πρόβλεψης συχνότερα, σύμφωνα με τη βιβλιογραφία.

3.3.1.1 Απόδοση του φύλου του συγγραφέα

Ο κλάδος της Γλωσσολογίας που ασχολήθηκε με την επίδραση του φύλου στη γλωσσική παραγωγή είναι η κοινωνιογλωσσολογία. Βέβαια, ως τη δεκαετία του 1990 οι έρευνες περιορίζονταν στον προφορικό λόγο. Τα τελευταία, όμως χρόνια με την αύξηση των ηλεκτρονικών σωμάτων κειμένων και λογισμικών μηχανικής μάθησης, όπως έχει αναφερθεί και παραπάνω, γίνονται προσπάθειες αυτόματης κατηγοριοποίησης κειμένων βάσει του φύλου του συγγραφέα. Το βασικό ερώτημα σε αυτό το πρόβλημα είναι εάν τα κείμενα ανδρών και γυναικών διαφοροποιούνται με συστηματικό τρόπο.

Η γλωσσική παραγωγή σε αλληλεπίδραση με το φύλο αποτελεί διεπιστημονικό τομέα έρευνας, αφού ασχολούνται με αυτόν επιστήμονες από διαφορετικούς κλάδους όπως κοινωνιογλωσσολόγοι, νευροφυσιολόγοι ανθρωπολόγοι. Έρευνες νευροβιολογίας έδειξαν ότι η διαφορά ανάμεσα στα δύο φύλα έγκειται στην ανατομία του εγκεφάλου. Υπάρχουν, δηλαδή, ασυμμετρίες τόσο στη βιολογική δομή όσο και στη

λειτουργία του εγκεφάλου. Συγκεκριμένα, οι γυναίκες χρησιμοποιούν συγχρόνως και τα δύο ημισφαίρια του εγκεφάλου κατά τη γλωσσική επεξεργασία, σε αντίθεση με τους άνδρες που χρησιμοποιούν μόνο το αριστερό (Shaywitz et al., 1995). Οι ανατομικές διαφοροποιήσεις και η διαλειτουργικότητα των εγκεφαλικών ημισφαιρίων κατά την επεξεργασία γλωσσικών δεδομένων και τη γλωσσική παραγωγή λειτουργούν υπέρ των γυναικών που είναι πιο ικανές στη γλωσσική χρήση από τους άνδρες. Η υπεροχή αυτή έχει πιστοποιηθεί διαχρονικά και διαπολιτισμικά, αφού ισχύει ανεξαρτήτως της εθνικότητας και του πολιτισμικού υπόβαθρου των ομιλητών.

Πιο ειδικά, οι έμφυλες διαφορές στη γλωσσική επεξεργασία εντοπίστηκαν και σε παιδιά και εφήβους μεταξύ 9 και 15 ετών (Burman, Bitan & Booth, 2008). Κατά τη διάρκεια των γλωσσικών δραστηριοτήτων που ανατέθηκαν παρατηρήθηκε στατιστικά μεγαλύτερη δραστηριότητα στην περιοχή του εγκεφάλου που ελέγχει τη γλώσσα στα κορίτσια απ’ ότι στα αγόρια, γεγονός που οδήγησε σε καλύτερα αποτελέσματα τα κορίτσια.

Επίσης σε εφήβους έχει μελετηθεί η επίδραση του φύλου στη γλωσσική ανάπτυξη αποκλειστικά στο συντακτικό επίπεδο. Τα αποτελέσματα έδειξαν ότι οι δείκτες της συντακτικής πολυπλοκότητας είναι πιο αυξημένοι στα κορίτσια σε σχέση με τα συνομήλικά τους αγόρια (Cornett, 2014).

Κοινωνιογλωσσολογικά υπάρχουν επίσης διαφορές εξαιτίας διαφορετικών επιδράσεων από την κοινωνική δομή. Οι άνδρες συνήθως χρησιμοποιούν πιο συχνά κοινωνικά στιγματισμένους τύπους και υιοθετούν γλωσσικά στοιχεία χαμηλότερου κοινωνικού κύρους, ενώ οι γυναίκες υιοθετούν γλωσσικούς τύπους υψηλού κύρους και προτιμούν τον κοινωνικά καταξιωμένο γλωσσικό κώδικα (Μικρός, 2015a:176).

Ενδιαφέρον παρουσιάζει η ιστορική αναδρομή που επιχειρούν οι Hall, Borba & Hiramoto (2021) στο θεωρητικό πεδίο που αφορά στη σχέση γλώσσας και φύλου. Σε άρθρο του Abdalgane (2019) καταγράφονται συγκεκριμένες διαφορές σε γλωσσικά επίπεδα μεταξύ ανδρών και γυναικών που μιλούν την Αγγλική γλώσσα.

Διαφορές εντοπίζονται στη γλωσσική έκφραση των δύο φύλων στο λεξιλόγιο, στις συντακτικές επιλογές και σε πολλά γλωσσικά χαρακτηριστικά που είναι υποσυνείδητα ως προς τη λειτουργία τους, όπως στο μήκος λέξεων και προτάσεων, στην συχνότητα ακολουθιών χαρακτήρων και μερών του λόγου. Πρόκειται για υφομετρικές μεταβλητές, η χρήση των οποίων αποκλείει την πιθανότητα αξιοποίησης γλωσσικών χαρακτηριστικών που αντιπροσωπεύουν συνειδητές επιλογές των

συγγραφέων που επιστρατεύονται ανάλογα με την επικοινωνιακή περίσταση (Μικρός, 2015a: 191).

Οι παραπάνω θεωρητικές προσεγγίσεις επιβεβαιώνονται και από την εφαρμογή υφομετρικών τεχνικών σε σώματα κειμένων. Από την ανάλυση κειμένων (Tausczik & Pennebaker, 2010) προκύπτει πως η μεγαλύτερη διαφορά ανάμεσα στη γλώσσα των δύο φύλων είναι η συχνή χρήση κοινωνικών αναφορών από τις γυναίκες και σύνθετου λόγου από τους άνδρες. Οι άνδρες χρησιμοποιούν περισσότερο μεγάλες λέξεις, άρθρα και προθέσεις, ενώ οι γυναίκες αντωνυμίες κυρίως πρώτου ενικού και τρίτου προσώπου.

Μια από τις πρώτες έρευνες που αξιοποίησαν υφομετρικά χαρακτηριστικά είναι των Koppel, Argamon & Shimoni (2002). Σε υποσύνολο του British National Corpus πέτυχαν ακρίβεια πρόβλεψης φύλου 79,5% σε λογοτεχνικά και 82,6% σε μη λογοτεχνικά κείμενα. Σημαντικό είναι πως κάποια χαρακτηριστικά που είχαν μελετηθεί παλαιότερα λειτούργησαν ως δείκτες του φύλου και σε αυτή την έρευνα: οι γυναίκες χρησιμοποιούν συχνότερα αντωνυμίες ενώ οι άνδρες οριστικά άρθρα. Γενικότερα, βρέθηκε πως υπάρχει συσχέτιση μεταξύ υφολογικών επιλογών ανδρών-γυναικών και των χαρακτηριστικών των λογοτεχνικών ή μη κειμένων (Argamon et al., 2003).

Την επόμενη χρονιά ο Corney (2003) μετρώντας υφομετρικά χαρακτηριστικά σε μηνύματα ηλεκτρονικού ταχυδρομείου έφτασε σε ακρίβεια πρόβλεψης του φύλου του αποστολέα το 70,1%. Τα πιο σημαντικά χαρακτηριστικά αποδείχθηκαν οι συχνότερες λειτουργικές λέξεις και ο μέσος όρος μήκους λέξης και πρότασης.

Μια ακόμη έρευνα στο χώρο των διαδικτυακών κειμένων που εξετάζει συγχρόνως την πρόβλεψη φύλου και ηλικίας του συγγραφέα έγινε από τους Schler et al. (2006) σε κείμενα από ιστολόγια (37.478 αναρτήσεις). Χρησιμοποιήθηκαν 1.502 χαρακτηριστικά και ο αλγόριθμος μηχανικής μάθησης Multi-Class Real Winnow και η ακρίβεια της πρόβλεψης του φύλου ήταν στο 80,1%. Αξίζει να σημειωθεί ότι σημαντικό ρόλο στη διάκριση του φύλου είχαν τα χαρακτηριστικά που ήταν σημασιολογικά ουδέτερα (π.χ. συχνές λειτουργικές λέξεις και τα μέρη του λόγου) παρά τη μεγάλη διαφοροποίηση των δύο φύλων στη χρήση στερεότυπων λέξεων περιεχομένου.

Στο χώρο της λογοτεχνίας σε μια μελέτη (Hota et al., 2006) εξετάστηκε κατά πόσο ο Σαίξπηρ, ένας άνδρας συγγραφέας, προσέγγισε τα χαρακτηριστικά της γυναικείας ομιλίας στους διαλόγους 34 έργων του. Το ποσοστό ακρίβειας, που κυμάνθηκε μεταξύ 60% και 75% ανάλογα με τα χαρακτηριστικά, ερμηνεύεται από τους

ερευνητές ως ένδειξη της μη επιτυχημένης προσπάθειας του Σαίξπηρ να προσεγγίσει πλήρως τον γυναικείο λόγο.

Σε άλλη μελέτη πρόβλεψης φύλου και ηλικίας (Peersman, Daelemans & Van Vaerenbergh, 2011) αναλύθηκαν κείμενα από μια βελγική online πλατφόρμα κοινωνικής δικτύωσης. Η συλλογή περιλαμβάνει 1.537.283 αναρτήσεις σε Φλαμανδικά Ολλανδικά. Στην έρευνα αυτή το φύλο χρησιμοποιείται ως ένα επιπλέον χαρακτηριστικό για να επιτευχθούν καλύτερα αποτελέσματα στην πρόβλεψη της ηλικίας.

Στο πλαίσιο του διεθνούς διαγωνισμού PAN 2014 παρουσιάστηκε (Marquardt et al., 2014) έρευνα για την πρόβλεψη ηλικίας και φύλου από κείμενα κοινωνικών δικτύων στην Αγγλική και Ισπανική γλώσσα. Εξετάστηκαν δυο διαφορετικά μοντέλα και τα αποτελέσματα για την ακρίβεια στην πρόβλεψη φύλου ανάλογα με τη γλώσσα και το είδος κοινωνικού δικτύου κυμάνθηκαν μεταξύ 54,22% και 80,68%.

Στο πλαίσιο του 3^{ου} διεθνούς διαγωνισμού απόδοσης χαρακτηριστικών σε συγγραφέα PAN 2015 το υψηλότερο ποσοστό ακρίβειας στον εντοπισμό του φύλου επιτεύχθηκε στην Ολλανδική και στην Ισπανική γλώσσα με πάνω από 95%.

Πρέπει να τονίσουμε πως όλες οι παραπάνω μελέτες εφάρμοσαν υφομετρικές τεχνικές σε κείμενα το λιγότερο 250 λέξεων το καθένα, η έκταση των οποίων είναι γνωστό πόσο επηρεάζει τα αποτελέσματα. Για παράδειγμα οι Zhang & Zhang (2010) σε πολύ σύντομες αναρτήσεις ιστολογίων (15 tokens per segment, 10.000 segments) έφτασαν σε ακρίβεια πρόβλεψης φύλου μόλις 72,1%.

Στον διεθνή διαγωνισμό PAN 2016 είκοσι δύο συμμετέχουσες ερευνητικές ομάδες αναγνώρισαν το φύλο του συγγραφέα κειμένων Αγγλικής, Ισπανικής και Ολλανδικής γλώσσας. Το καλύτερο αποτέλεσμα (ακρίβεια: 0.7564) για τα Αγγλικά έφεραν οι τεχνικές των Modaresi, Liebeck & Conrad (2016) με μοντέλο λογιστικής παλινδρόμησης με συνδυασμό υφομετρικών και λεξικών χαρακτηριστικών.

Ακολούθως, στο διαγωνισμό PAN 2017, στο εργαστήριο για το συγγραφικό προφίλ είκοσι δύο συμμετέχουσες ομάδες αξιολογήθηκαν για την πρόβλεψη του φύλου του συγγραφέα σε ένα σώμα κειμένων από το Twitter που συλλέχθηκε σε τέσσερις διαφορετικές γλώσσες: στα Αγγλικά, Αραβικά, Ισπανικά και Πορτογαλικά. Τα καλύτερα αποτελέσματα επιτεύχθηκαν με μέγιστη ακρίβεια 87% για την Πορτογαλική γλώσσα, η οποία σημειωτέον μαζί με την Αραβική για πρώτη φορά στην ιστορία του διαγωνισμού αποτέλεσαν μέρος του σώματος κειμένων. Το χαμηλότερο ποσοστό ακρίβειας σημειώθηκε για τα Αραβικά (72,10%).

Αξίζει να παραθέσουμε την προσπάθεια για χρήση περιορισμένου αριθμού χαρακτηριστικών για την κατηγοριοποίηση αναρτήσεων ιστολογίου βάσει του φύλου του συγγραφέα. Σε ανακοίνωση των Soler & Wanner (2014) ενημερωνόμαστε για την χρήση 67 συντακτικών χαρακτηριστικών σε συνδυασμό με 16 χαρακτηριστικά άλλων κατηγοριών και με αυτό τον τρόπο ξεπέρασαν την απόδοση άλλων μοντέλων που είχαν στη διάθεσή τους πολλά περισσότερα χαρακτηριστικά (ακρίβεια 82,72%).

Άλλη έρευνα (Verhoeven, Daelemans & Plank 2016a, 2016b) αφορά στη δημιουργία ενός πολύγλωσσου σώματος κειμένων από το Twitter με το όνομα TwiSty. Πρόκειται για αναρτήσεις στα Ολλανδικά, Γερμανικά, Γαλλικά, Ιταλικά, Πορτογαλικά και Ισπανικά 18.168 συγγραφέων, των οποίων το φύλο και η προσωπικότητα αποτελεί στόχο πρόβλεψης των ερευνητών. Σημαντικό είναι πως μεγάλη πρόοδος σημειώθηκε στην πρόβλεψη του φύλου στα Ολλανδικά με F-score 82,61.

Επίσης το 2016, οι Soler & Wanner (2016) παρουσίασαν μια προσέγγιση αναγνώρισης του φύλου του συγγραφέα με ημι-εποπτευόμενη μάθηση (semi-supervised). Είναι σημαντικό ότι χρησιμοποίησαν μικρό αριθμό κατηγοριοποιημένων δεδομένων και μεγαλύτερο μη κατηγοριοποιημένων δεδομένων για την κατηγοριοποίηση των κειμένων, επιλογή που βελτίωσε την ακρίβεια του μοντέλου.

Το φύλο του συγγραφέα μελετήθηκε και σε έρευνα που διεξήχθη για την Πορτογαλική γλώσσα με δεδομένα από το Facebook της Βραζιλίας (Hsieh, Dias & Paraboni, 2018). Η καλύτερη ταξινόμηση έγινε με τη χρήση μοντέλου βασισμένου στον αριθμό λέξεων, ενώ η χειρότερη με τη χρήση ψυχογλωσσολογικού μοντέλου, γεγονός το οποίο δείχνει σύμφωνα με τους ερευνητές ότι τα ψυχογλωσσολογικά χαρακτηριστικά λειτουργούν αποτελεσματικότερα στην πρόβλεψη της προσωπικότητας και του συναισθήματος και όχι του φύλου ή της ηλικίας. Τα χαρακτηριστικά αυτά προέκυψαν από το LIWC (Pennebaker et al., 2015), πρόγραμμα κειμενικής ανάλυσης με ενσωματωμένα λεξικά των οποίων τα λήμματα είναι ψυχολογικού και κοινωνικού περιεχομένου και πρόκειται για κατηγορίες λέξεων όπως π.χ. θυμός, οικογένεια, πλούτος.

Η ακόλουθη έρευνα (Dell'Orletta & Nissim, 2018) αφορά στην πρόβλεψη του φύλου των συγγραφέων κειμένων της Ιταλικής γλώσσας που εντάσσονται σε πέντε διαφορετικά είδη (Twitter, YouTube, εκθέσεις μαθητών των δύο πρώτων χρόνων της δευτεροβάθμιας εκπαίδευσης, άρθρα εφημερίδων και προσωπικά ημερολόγια). Πιο αποτελεσματικά προβλέπεται το φύλο από τα προσωπικά ημερολόγια ενώ λιγότερο από τα δημοσιογραφικά κείμενα με τη χρήση τεχνικών μηχανικής μάθησης αλλά και

νευρωνικών δικτύων, δηλαδή συστημάτων επεξεργασίας με ικανότητα αποθήκευσης και ανάκλησης γνώσης προερχόμενης από εμπειρία τα οποία υιοθετούν το δομικό πρότυπο του ανθρώπινου εγκεφάλου.

Εκτός από τις μελέτες για την πρόβλεψη του φύλου στις γλώσσες που αναφέρθηκαν, έχει πραγματοποιηθεί έρευνα και για τα βιετναμέζικα από αναρτήσεις ιστολογίων (Dang, Giang & Son, 2009) με ακρίβεια 83,3%.

Όσον αφορά τα Ελληνικά, ο Μικρός (2009, Mikros 2013b) στην έρευνά του ανέπτυξε ηλεκτρονικό σώμα κειμένων αποτελούμενο από 700 κείμενα δημοσιευμένα στην εφημερίδα «Ελευθεροτυπία» στη διάρκεια ενός έτους ισομερώς κατανεμημένα σε άνδρες και γυναίκες συγγραφείς συνολικά 479.439 λέξεων. Με την κατάλληλη επιλογή υφομετρικών μεταβλητών, 54 από τις έξι ομάδες του γλωσσικού πλούτου, μήκους λέξης και πρότασης, συχνοτήτων γραμμάτων και Μερών του Λόγου και τέλος συχνών λειτουργικών λέξεων, έγινε η εκπαίδευση ενός τεχνητού νευρωνικού δικτύου. Η ακρίβεια σωστής πρόβλεψης του φύλου του συγγραφέα είναι πάνω από 80%, αποτέλεσμα που κατατάσσει το εν λόγω μοντέλο νευρωνικού δικτύου στα πιο ακριβή της βιβλιογραφίας.

Τα τελευταία χρόνια η γλωσσολογική κοινότητα έχει στραφεί στη μελέτη κειμένων από τα κοινωνικά δίκτυα εξαιτίας του ενδιαφέροντος που παρουσιάζει η γλωσσική παραγωγή σε αυτά. Στην ουσία πρόκειται για ένα νέο κειμενικό είδος. Ειδικότερα, η γλώσσα των ιστολογίων έχει υβριδική δομή, δεδομένου ότι περιέχει στοιχεία και μονολόγου και διαλόγου. Ο Mikros (2013a) εξέτασε παράλληλα με την απόδοση συγγραφικής πατρότητας και την απόδοση του φύλου σε ένα σώμα κειμένων που δημιουργήθηκε από 100 ελληνικά ιστολόγια. Για την απόδοση του φύλου του συγγραφέα αξιοποιήθηκαν αναρτήσεις από 10 ιστολόγια ανδρών και 10 γυναικών (406.460 λέξεις) με κοινό θέμα. Χρησιμοποιώντας αλγόριθμο μηχανικής μάθησης και υφομετρικά χαρακτηριστικά (λεξιλογικός πλούτος, μήκος λέξης, συχνότητα γραμμάτων και ν-γράμματα χαρακτήρων και λέξεων) η πρόβλεψη του φύλου έφτασε σε ακρίβεια το 82,6%.

Κλείνοντας την ενότητα της αυτόματης πρόβλεψης του φύλου του συγγραφέα, πρέπει να αναφερθούμε σε μια νέα ερευνητική πρόκληση που εξετάσθηκε από τον διεθνή διαγωνισμό PAN 2019 (Rangel & Rosso, 2019). Πρόκειται για την απόδοση ενός κειμένου σε άνθρωπο ή σε αυτοματοποιημένο πρόγραμμα (bot), το οποίο στοχεύει να επηρεάσει εμπορικά, πολιτικά, ιδεολογικά την κοινή γνώμη. Στην περίπτωση που

έχει γραφτεί από άνθρωπο πρέπει να προβλεφθεί το φύλο του. Το συγκεκριμένο εγχείρημα εφαρμόστηκε σε αγγλικές και ισπανικές αναρτήσεις του Twitter.

3.3.1.2 Απόδοση της ηλικίας του συγγραφέα

Ένα σημαντικό πρόβλημα στον τομέα της αναγνώρισης των χαρακτηριστικών ενός συγγραφέα είναι η σχέση ανάμεσα στην γλωσσική παραγωγή και την ηλικία του. Όπως και στο φύλο, διαφορετικές εποχές, πολιτισμοί, θρησκείες αντιμετωπίζουν διαφορετικά το κάθε ηλικιακό στάδιο και αυτό προβάλλεται και στη γλώσσα. Με κοινωνιογλωσσολογικούς όρους (Κακριδή-Φερράρι, 2005) οι διαφοροποιήσεις στη γλώσσα των ατόμων που εξαρτώνται από την ηλικία τους συνιστούν το φαινόμενο της ηλικιακής διαβάθμισης, κατά την οποία λόγω εξελικτικών ή κοινωνικών παραγόντων συγκεκριμένα γλωσσικά στοιχεία ή εκφραστικοί τρόποι συνδέονται με συγκεκριμένες ηλικιακές ομάδες.

Αν το άτομο μεγαλώνοντας δεν εγκαταλείψει αυτά τα στοιχεία τότε πρόκειται για γλωσσική μεταβολή και όχι για ηλικιακή διαβάθμιση. Έτσι, αν μια ηλικιακή ομάδα υιοθετήσει διαφοροποιήσεις στη γλώσσα και τις διατηρήσει και στα ηλικιακά στάδια που ακολουθούν, τότε η χρήση αυτών των διαφοροποιήσεων επεκτείνεται. Βέβαια, είναι δύσκολο να καταλήξουμε στο συμπέρασμα πως η ηλικία ως ανεξάρτητη κοινωνική μεταβλητή δημιουργεί συγκεκριμένες διαφοροποιήσεις στη γλώσσα, καθώς συχνά οι διαφοροποιήσεις συνδέονται και με άλλους παράγοντες τόσο του κοινωνικού μακροεπιπέδου όσο και του μικροεπιπέδου της ατομικής ζωής με τους οποίους αλληλεπιδρά ο παράγοντας ηλικία.

Η κοινωνιογλωσσολογική βιβλιογραφία έχει εκτενώς εμπλουτισθεί με θέματα που αφορούν το ηλικιακό στάδιο της εφηβείας. Η γλώσσα των νέων είναι μια γλωσσική ποικιλία που διαφοροποιείται αρκετά από την καθιερωμένη γλώσσα και χαρακτηρίζεται από ιδιαίτερη δημιουργικότητα. Η έμφαση στη γλώσσα των εφήβων και η διάκρισή της από τη γλώσσα των ενηλίκων έχει δοθεί και στην επιστήμη της υφομετρίας. Οι περισσότερες, μάλιστα, έρευνες παρουσιάζουν ικανοποιητικά αποτελέσματα στον αυτόματο εντοπισμό εφήβων συγγραφέων.

Σύμφωνα με δύο μελέτες που πραγματοποιήθηκαν από τους Pennebaker & Stone (2003) σε λογοτέχνες και σε μη λογοτέχνες φάνηκε πως η γλωσσική παραγωγή

επηρεάζεται από την ηλικία. Συγκεκριμένα, βρέθηκε ότι όσο αυξάνεται η ηλικία τα άτομα χρησιμοποιούν περισσότερο θετικές παρά αρνητικές λέξεις, λιγότερες αυτοαναφορές, συχνότερα ρήματα σε μελλοντικούς χρόνους και ο λόγος τους παρουσιάζει μεγαλύτερη γνωστική πολυπλοκότητα. Εντοπίζονται, επομένως, μεταβολές στη γλώσσα καθώς τα άτομα μεγαλώνουν, οι οποίες υποδηλώνονται με γλωσσικά στοιχεία.

Η δυνατότητα πρόβλεψης ηλικίας και φύλου των συγγραφέων έχει μελετηθεί σε σώμα κειμένων αποτελούμενο από 37.478 αναρτήσεις ιστολογίων (Schler et al., 2006). Οι τρεις ηλικιακές κατηγορίες που εξετάστηκαν ήταν 13-17, 23-27 και 33-42. Τα αποτελέσματα έδειξαν πως με τη σύγχρονη χρήση υφομετρικών χαρακτηριστικών και χαρακτηριστικών περιεχομένου η πρώτη ηλικιακή κατηγορία διακρίθηκε από την τρίτη με ακρίβεια πάνω από 96%. Πάντως, συνολικά η ακρίβεια της πρόβλεψης ήταν 76,2% και τονίζεται ότι περισσότερο αύξησαν το ποσοστό αυτό χαρακτηριστικά περιεχομένου (λίστες λέξεων).

Οι Burger & Henderson (2006) δημιούργησαν ένα σώμα κειμένων από 100.000 αναρτήσεις ιστολογίων και 87.883 διαφορετικούς συγγραφείς και επιδίωξαν μελετώντας τα κείμενα και τις πληροφορίες που παρείχαν οι συγγραφείς για τον εαυτό τους να εντοπίσουν χαρακτηριστικά για την πρόβλεψη της ηλικίας. Ο καλύτερος ταξινομητής τους μείωσε το λάθος σε ελάχιστο μόνο ποσοστό.

Χρησιμοποιώντας το σώμα κειμένων από ιστολόγια που βρίσκεται στην ιστοσελίδα του Koppel (Schler et al., 2006) και με χαρακτηριστικά το μήκος πρότασης και λέξεις που δεν υπάρχουν σε λεξικό (non-dictionary) οι Goswami, Sarkar & Rustagi (2009) έφτασαν σε ακρίβεια 89,68% για την κατάταξη στην ηλικιακή κατηγορία 13-17 ή 33-42. Συνολικά για όλες τις ηλικιακές ομάδες και με την προσθήκη του χαρακτηριστικού των λέξεων περιεχομένου στις λέξεις εκτός λεξικού έφτασαν στο 80,32%. Με τον όρο λέξεις που δεν καταγράφονται στο λεξικό εννοούν την αργκό, συντμήσεις, λέξεις με λάθη που ωστόσο χρησιμοποιούνται τακτικά στα μέσα κοινωνικής δικτύωσης.

Σε έρευνα που αφορά αποκλειστικά την πρόβλεψη ηλικίας συγγραφέων σε διαδικτυακή συνομιλία (online chat) 160.740 αναρτήσεων οι Tam & Martell (2009) κάνοντας μετρήσεις με διάφορα μοντέλα και υφομετρικά χαρακτηριστικά κατόρθωσαν να πετύχουν 0,996 f-score στη διάκριση ανάμεσα σε εφήβους και ενήλικες. Αυτό το αποτέλεσμα επιτεύχθηκε με τη χρήση Μηχανών Διανυσμάτων Υποστήριξης και με χαρακτηριστικό τα τριγράμματα λέξεων. Στόχος της έρευνας ήταν η συμβολή στη

δημιουργία ενός συστήματος αυτόματης αναγνώρισης ενηλίκων που συνομιλούν διαδικτυακά με εφήβους, ώστε να αποφευχθούν φαινόμενα παρενόχλησης.

Σε άλλη μελέτη πρόβλεψης φύλου και ηλικίας (Peersman et al., 2011), η οποία αναφέρθηκε και στο προηγούμενο υποκεφάλαιο, αναλύθηκαν κείμενα από βελγική online πλατφόρμα κοινωνικής δικτύωσης με αναρτήσεις σε Φλαμανδικά Ολλανδικά. Η πιο ισχυρή κατηγορία χαρακτηριστικών για την πρόβλεψη της ηλικίας ήταν τα ν-γράμματα λέξεων και ανάλογα με την ηλικιακή ομάδα επιτεύχθηκε ακρίβεια από 71,3% έως 88,2%.

Σε εργασία των Rosenthal & McKeown (2011) ελέγχθηκε αν οι λέξεις, οι υφολογικές επιλογές και η συμπεριφορά του χρήστη στο ιστολόγιό του μπορούν να χρησιμοποιηθούν για την πρόβλεψη της ηλικιακής του κατηγορίας. Σε ένα σώμα κειμένων 24.500 αναρτήσεων από το LiveJournal (υπηρεσία κοινωνικής δικτύωσης) έφτασαν σε ακρίβεια 81,57%.

Ακόμη μια έρευνα με βασικό στοιχείο τη σχέση ηλικίας και γλώσσας πραγματοποιήθηκε από τους Nguyen, Smith & Rosé (2011). Η πρωτοτυπία έγκειται στο σώμα κειμένων που αποτελείται από τρία διαφορετικά είδη: αναρτήσεις ιστολογίων, καταγεγραμμένες τηλεφωνικές κλήσεις και αναρτήσεις από forum. Με τη χρήση της εξίσωσης γραμμικής παλινδρόμησης το καλύτερο αποτέλεσμα που πέτυχαν ήταν δείκτης συσχέτισης r 0,742.

Με τη χρήση του λογισμικού ανάλυσης κειμένου Linguistic Inquiry and Word Count (LIWC) και με τη μέθοδο της διαφορικής γλωσσικής ανάλυσης (differential language analysis, DLA) ερευνητική ομάδα (Schwartz et al., 2013) μελέτησε την προσωπικότητα, το φύλο και την ηλικία στα κοινωνικά δίκτυα. Το αποτέλεσμα για την εξαρτημένη μεταβλητή «ηλικία» ήταν 0,84.

Στο πλαίσιο του διεθνούς διαγωνισμού PAN 2014 μια ακόμη ερευνητική ομάδα (Marquardt et al., 2014) παρουσίασε ένα σύνολο χαρακτηριστικών για την αύξηση της ακρίβειας στην πρόβλεψη ηλικίας και φύλου από κείμενα κοινωνικών δικτύων στην Αγγλική και Ισπανική γλώσσα. Εξετάστηκαν δυο διαφορετικά μοντέλα και τα αποτελέσματα για την ακρίβεια στην πρόβλεψη ηλικίας ανάλογα με τη γλώσσα και το είδος κοινωνικού δικτύου κυμάνθηκαν μεταξύ 24% και 48,31%.

Στον διαγωνισμό PAN 2016 με τη χρήση Μηχανών Διανυσμάτων Υποστήριξης και συνδυασμό υφομετρικών χαρακτηριστικών οι κάτοχοι της πρώτης θέσης πρόβλεψαν την ηλικία του συγγραφέα αγγλικών κειμένων με ακρίβεια 58,97% (Rangel et al., 2016).

Η ηλικία του συγγραφέα μελετήθηκε στην έρευνα που αναφέραμε για την πρόβλεψη του φύλου για την Πορτογαλική γλώσσα με δεδομένα από το Facebook της Βραζιλίας (Hsieh, Dias & Paraboni, 2018). Η καλύτερη ταξινόμηση έγινε με τη χρήση μοντέλου βασισμένου στον αριθμό λέξεων.

3.3.1.3 Απόδοση της μητρικής γλώσσας του συγγραφέα

Σύμφωνα με τη βιβλιογραφία, για τα λάθη που κάνει ένας συγγραφέας σε οποιοδήποτε γλωσσικό επίπεδο ευθύνονται οι διαφορές ανάμεσα στη μητρική του γλώσσα και στην ξένη γλώσσα στην οποία γράφει. Με βάση αυτά τα δεδομένα ο αυτόματος εντοπισμός της μητρικής γλώσσας του συγγραφέα ενός κειμένου που γράφει σε γλώσσα διαφορετική από τη μητρική του έχει αρκετές αναφορές. Ο τομέας παρουσιάζει ενδιαφέρον, καθώς συμβάλλει και στη μελέτη της διαδικασίας κατάκτησης της δεύτερης γλώσσας.

Σε σχετική έρευνα (Argamon et al., 2009) χρησιμοποιήθηκε το International Corpus of Learner English (ICLE), το οποίο δημιουργήθηκε για την μελέτη γραπτού λόγου που παρήχθη από μη φυσικούς ομιλητές της Αγγλικής γλώσσας. Οι συγγραφείς ήταν φοιτητές της Αγγλικής ως δεύτερης γλώσσας από την Βουλγαρία, την Γαλλία, την Ισπανία, την Ρωσία και την Τσεχία και ο στόχος της έρευνας να εντοπιστεί η μητρική γλώσσα του συγγραφέα του αγγλικού κειμένου. Η ακρίβεια έφτασε το 82,3% με τη χρήση χαρακτηριστικών περιεχομένου και αυτό διότι φάνηκε ότι μη φυσικοί ομιλητές της Αγγλικής χρησιμοποιούν αναλόγως με τη μητρική τους γλώσσα συγκεκριμένες λέξεις πιο συχνά από άλλες.

Οι Wong & Dras (2009) πειραματίστηκαν στις παραπάνω πέντε γλώσσες και επιπλέον στα Κινεζικά και Ιαπωνικά και πέτυχαν ακρίβεια 73,71% με τον συνδυασμό όλων των λεξικών χαρακτηριστικών. Όταν πρόσθεσαν στα χαρακτηριστικά και τα συντακτικά λάθη η ακρίβεια δεν βελτιώθηκε.

Σε παλαιότερη έρευνα (Koppel, Schler & Zigdon, 2005) είχε προκύψει ακρίβεια 80% για τον εντοπισμό της μητρικής γλώσσας χρησιμοποιώντας υφομετρικά χαρακτηριστικά και λάθη των συγγραφέων κατά το πέρασμα από τη μια γλώσσα στην άλλη.

Οι Yu, Mei & Zhai (2005) με την εφαρμογή ανάλυσης σε ν-γράμματα λέξεων διέκριναν φυσικούς ομιλητές της Αγγλικής και της Κινεζικής γλώσσας σε ακαδημαϊκό επίπεδο. Το σώμα κειμένων συστάθηκε από πτυχιακές εργασίες και διδακτορικά καθώς και από επιστημονικά άρθρα. Δημιούργησαν μια λίστα λέξεων αποτελούμενη από 626 λειτουργικές λέξεις και κάποιες λέξεις περιεχομένου με στόχο την παραγωγή συχνοτήτων λέξεων. Το πιο χαρακτηριστικό παράδειγμα στην ανάλυση του σώματος κειμένων ήταν η χρήση της λέξης "never", η οποία απαντά στο 75% των κειμένων των φυσικών ομιλητών της Αγγλικής ενώ μόνο στο 20% των Κινέζων.

3.3.1.4 Απόδοση της γλωσσικής ποικιλίας του συγγραφέα

Ιδιαιτέρως ενδιαφέρει τους ερευνητές, εκτός από τον εντοπισμό της μητρικής γλώσσας του συγγραφέα ενός κειμένου, η αυτόματη αναγνώριση της διαλέκτου ή του ιδιώματος. Πολλές διαλεκτολογικές διαφοροποιήσεις μπορούν να παρέχουν πληροφορίες για το άτομο που έγραψε ένα δεδομένο κείμενο. Για παράδειγμα ο συλλαβισμός της λέξης "colour" απαντά στο Ηνωμένο Βασίλειο αλλά όχι στις Ηνωμένες Πολιτείες. Υπ' αυτή την έννοια τέτοιου είδους μελέτες έχουν πραγματοποιηθεί για διαλέκτους της Αγγλικής, Ισπανικής, Πορτογαλικής, Ρουμανικής, Σλαβικής, Περσικής, Ινδονησιακής, Κινεζικής και Αραβικής γλώσσας. Οι προσεγγίσεις του προβλήματος βασίστηκαν κυρίως σε ν-γράμματα και ταξινομητές όπως Naive Bayes μοντέλα και Μηχανές Διανυσμάτων Υποστήριξης.

Δύο διαγωνισμοί του 2017 είχαν ως αντικείμενο τη διάκριση κειμένων βάσει γλωσσικής ποικιλίας. Το 4ο Θεματικό Εργαστήριο στην Επεξεργασία Φυσικής Γλώσσας (DSL Task)[1] και ο διεθνής διαγωνισμός ιχνογράφησης χαρακτηριστικών του συγγραφέα PAN 2017. Στον πρώτο τα δεδομένα αποτελούνταν από κείμενα εφημερίδων και οι γλώσσες ήταν ομαδοποιημένες ανάλογα με τις μεταξύ τους ομοιότητες, ενώ στον δεύτερο τα κείμενα αντλήθηκαν από το Twitter στα Αγγλικά, Αραβικά, Ισπανικά και Πορτογαλικά. Αξίζει να σημειωθεί ότι οι αραβικές διάλεκτοι μελετώνται εκτεταμένα και μάλιστα μια υποενότητα του DSL Task ήταν εξ ολοκλήρου αφιερωμένη στον εντοπισμό αραβικών διαλέκτων.

[1] http://ttg.uni-saarland.de/vardial2017/sharedtask2017.html

Όσον αφορά στο διαγωνισμό PAN 2017 (Rangel et al., 2017), οι γλωσσικές ποικιλίες κυμαίνονταν από δύο για τα Πορτογαλικά (Βραζιλίας και Πορτογαλίας) μέχρι επτά για τα Ισπανικά (Αργεντινής, Βενεζουέλας, Ισπανίας, Κολομβίας, Μεξικό, Περού, Χιλής). Τα καλύτερα αποτελέσματα στην αυτόματη αναγνώριση της γλωσσικής ποικιλίας επιτεύχθηκαν με τη χρήση ν-γραμμάτων χαρακτήρων και λέξεων και Μηχανών Διανυσμάτων Υποστήριξης για τα Πορτογαλικά με μέγιστο ποσοστό ακρίβειας 98,50% και ακολουθούν τα Ισπανικά με 96,21%. Από τις προσεγγίσεις των ερευνητικών ομάδων που πέτυχαν υψηλά ποσοστά ακρίβειας φάνηκε ότι η πρόβλεψη του συγκεκριμένου τομέα εξαρτάται σαφώς από τη χρήση των λέξεων.

Ενδιαφέρον επίσης συμπέρασμα που προκύπτει από την έρευνα είναι η επίδραση του φύλου του συγγραφέα στην αναγνώριση της γλωσσικής ποικιλίας που χρησιμοποιεί. Συγκεκριμένα, στα Αραβικά και Πορτογαλικά η διαφορά του φύλου είναι στατιστικά σημαντική, δηλαδή σε κείμενα γυναικών η πρόβλεψη της γλωσσικής ποικιλίας ήταν ευκολότερη. Επιπλέον μελετήθηκε η δυσκολία της αναγνώρισης του φύλου σε σχέση με τη γλωσσική ποικιλία. Για παράδειγμα, στις περισσότερες αραβικές διαλέκτους και στα Πορτογαλικά το θηλυκό είναι πιο εύκολο να αναγνωρισθεί.

Κεφάλαιο 4

Αυτόματη αναγνώριση της προσωπικότητας του συγγραφέα

4.1 Εισαγωγή

Η γλώσσα ως κώδικας επικοινωνίας είναι δηλωτική της διαφορετικότητας του κάθε ατόμου. Η μελέτη γλωσσικών χαρακτηριστικών μπορεί να οδηγήσει σε επιτυχείς προβλέψεις για τον χαρακτήρα του ατόμου που παράγει λόγο. Το θέμα της αναγνώρισης της προσωπικότητας (Computational Personality Recognition) του ομιλούντος ή του γράφοντος μέσω τεχνικών επεξεργασίας φυσικής γλώσσας αποτελεί ένα σχετικά νέο ερευνητικό πεδίο με πολλές εφαρμογές.

Ιδιαίτερη έμφαση έχει δοθεί σε κείμενα που παρουσιάζουν εγκληματολογικό ενδιαφέρον (Neuman, 2016), τα οποία μπορούν να αναλυθούν και στη συνέχεια να εξαχθούν συμπεράσματα ως προς την προσωπικότητα του συγγραφέα τους. Χαρακτηριστικό είναι το παράδειγμα αυτόματου προσδιορισμού της προσωπικότητας των μαθητών που οπλοφορούν και σκοτώνουν στο σχολικό χώρο (Neuman et al., 2015). Επισημαίνονται τα ψυχολογικά τους χαρακτηριστικά, τα οποία μπορούν να αξιοποιηθούν και στον επιτυχή εντοπισμό των εν δυνάμει δραστών.

Η αυτόματη αναγνώριση προσωπικότητας εκτός από την προφανή συνεισφορά που μπορεί να παρέχει στον τομέα της ψυχολογίας με τη σύνδεση των χαρακτηριστικών προσωπικότητας και της ανθρώπινης συμπεριφοράς, μπορεί να λειτουργήσει σε πολλά ακόμα πεδία.

Στον τομέα του μάρκετινγκ, η ανάλυση της προσωπικότητας των χρηστών-καταναλωτών αξιοποιείται από τις επιχειρήσεις για την υιοθέτηση πιο αποτελεσματικών τεχνικών προσέλκυσης αλλά και εξυπηρέτησης των πελατών.

Ακόμα και στον επαγγελματικό τομέα και μάλιστα στο πεδίο της διαχείρισης ανθρώπινων πόρων η πρόβλεψη της προσωπικότητας μπορεί να επηρεάσει ή να διευκολύνει την επιλογή και να κρίνει την καταλληλότητα, για παράδειγμα υποψηφίων για μια θέση εργασίας (job screening).

Επιπλέον, βάσει της προσωπικότητας του χρήστη μπορούν να προσαρμοστούν διαλογικά συστήματα, ώστε να είναι πλησιέστερα στην ιδιοσυγκρασία του και επομένως η διάδραση αποτελεσματικότερη και πιο ικανοποιητική για τον χρήστη.

Μπορεί ακόμα να διερευνηθεί η συναισθηματική φόρτιση του κειμένου και να προκύψει η στάση για το θέμα την οποία διατηρεί ο συγγραφέας π.χ. η αξιολόγηση μιας κριτικής για ένα εστιατόριο ως θετικής ή αρνητικής (polarity detection).

Ένας άλλος τομέας ανάλυσης στον οποίο χρησιμοποιείται η αυτόματη αναγνώριση της προσωπικότητας είναι η εκπαίδευση. Αναλύοντας το λόγο των μαθητών θα μπορούσαν να αναδειχθούν ταλαντούχοι μαθητές ή μαθητές με δυσκολίες και έτσι να προσαρμοστεί η διδασκαλία, με στόχο να απευθύνεται στο κατάλληλο για τον καθένα ή για κάθε ομάδα γνωστικό επίπεδο. Ακόμα πιο εξειδικευμένα, μπορούμε να μιλήσουμε για την ανάλυση της αναγνωστικής δυσκολίας των κειμένων και την αυτόματη κατηγοριοποίησή τους σε επίπεδα δυσκολίας, ανάλογα με το εκπαιδευτικό επίπεδο των μαθητών στους οποίους απευθύνονται (Μικρός, 2013).

Βέβαια, ο ερευνητής του συγκεκριμένου τομέα, για να εφαρμόσει αποτελεσματικά την πρόβλεψη προσωπικότητας, καλείται να αντιμετωπίσει βασικά προβλήματα, εκ των οποίων πρωταρχικό είναι η δυσκολία συλλογής δεδομένων επισημειωμένων ως προς την προσωπικότητα του συγγραφέα τους. Η δεύτερη δυσκολία έγκειται στο θέμα της γλώσσας των δεδομένων. Τα περισσότερα εργαλεία υποστηρίζουν την Αγγλική με αποτέλεσμα ο εντοπισμός χαρακτηριστικών σε άλλες γλώσσες να είναι πιο απαιτητικός. Επίσης πρέπει να ληφθεί υπ' όψιν ότι κάθε κειμενικό είδος από το οποίο αποτελούνται τα δεδομένα ίσως έχει και διαφορετική μέθοδο επεξεργασίας. Για παράδειγμα η γλώσσα των κοινωνικών δικτύων έχει άλλα χαρακτηριστικά από τη γλώσσα επιστημονικών άρθρων ή των εκθέσεων των μαθητών της δευτεροβάθμιας εκπαίδευσης και επομένως η προσέγγιση θα είναι διαφορετική. Επιπλέον, ο τρόπος αξιολόγησης των μοντέλων αναγνώρισης πραγματοποιείται με διαφορετικό τρόπο σχεδόν από κάθε ερευνητή με αποτέλεσμα τη δυσχέρεια στον προσδιορισμό του υψηλότερου επιπέδου τεχνικής.

Ανάλογα με την εφαρμογή και τις δυνατότητες προσαρμογής του κάθε συστήματος χρησιμοποιούνται διαφορετικά μοντέλα προσωπικότητας. Δηλαδή, η μοντελοποίηση της προσωπικότητας είτε μπορεί να αντιμετωπίζεται ως ένα πρόβλημα ταξινόμησης (Argamon et al., 2007; Oberlander & Nowson 2006) είτε μπορεί να γίνει μέσω των τιμών που προκύπτουν από τα ψυχομετρικά τεστ που συμπληρώνουν οι ίδιοι

οι συμμετέχοντες ή από τις αναφορές των παρατηρητών για την προσωπικότητα των συμμετεχόντων.

4.2 Δύο διαφορετικές προσεγγίσεις για την αναγνώριση προσωπικότητας

Στη βιβλιογραφία καταγράφονται δύο διαφορετικές προσεγγίσεις για την αυτόματη αναγνώριση της προσωπικότητας του συγγραφέα: μία ανοδική (bottom-up) και μία καθοδική (top-down).

Στην πρώτη περίπτωση, οι ερευνητές (Oberlander & Nowson, 2006; Iacobelli et al., 2011; Bachrach et al., 2012) ξεκινούν από τα δεδομένα και αναζητούν γλωσσολογικά στοιχεία, τα οποία σχετίζονται με τα χαρακτηριστικά προσωπικότητας. Πρόκειται για τεχνικές ανοικτού λεξιλογίου (open-vocabulary), δηλαδή εξαγωγή γλωσσολογικών χαρακτηριστικών από τα σώματα κειμένων και διαφορική γλωσσική ανάλυση (Differential Language Analysis, DLA).

Αντίθετα, στην καθοδική προσέγγιση (Argamon et al., 2005; Mairesse et al., 2007; Golbeck, Robles & Turner, 2011) αξιοποιούνται λεξικά, για να ελέγξουν πιθανή συσχέτιση ανάμεσα σε αυτά και στα χαρακτηριστικά προσωπικότητας. Πρόκειται για τεχνικές κλειστού λεξιλογίου (closed-vocabulary approach), δηλαδή επιλέγονται εργαλεία που εξαρτώνται από τη γλώσσα του κειμένου και μέσω αυτών επιλέγονται τα χαρακτηριστικά που θα εξαχθούν από τα σώματα κειμένων και θα τροφοδοτήσουν το μοντέλο πρόβλεψης προσωπικότητας. Στη βασισμένη σε λεξικό προσέγγιση υποτίθεται ότι η προσωπικότητα του συγγραφέα μπορεί να προκύψει από τις επισημειωμένες λέξεις που χρησιμοποιεί. Πρέπει να σημειωθεί ότι για την Ελληνική γλώσσα δεν έχει αναπτυχθεί, απ' όσο γνωρίζουμε, κάποιο αντίστοιχο λεξικό.

Στη συνέχεια παραθέτουμε μια λίστα με τέτοιου είδους εργαλεία που απαντούν στη βιβλιογραφία και έχουν εφαρμοστεί για την αναγνώριση προσωπικότητας.

- **LIWC L**inguistic **I**nquiry and **W**ord **C**ount[2] (Pennebaker, Boyd, Jordan & Blackburn, 2015)

[2] https://liwc.wpengine.com/

Πρόκειται για πρόγραμμα κειμενικής ανάλυσης με ενσωματωμένα λεξικά των οποίων τα λήμματα είναι ψυχολογικού και κοινωνικού περιεχομένου. Στην τελευταία του έκδοση παράγει 90 περίπου διαφορετικά χαρακτηριστικά των κειμενικών δεδομένων, τα οποία κατατάσσονται σε 4 ομάδες. Περιληπτικά χαρακτηριστικά γλώσσας (π.χ. συναισθηματικός τόνος), 3 κατηγορίες γενικής περιγραφής (π.χ. πλήθος λέξεων ανά πρόταση, ποσοστό λέξεων με περισσότερα από έξι γράμματα), 21 γλωσσολογικές διαδικασίες (π.χ. το ποσοστό των λέξεων που είναι λειτουργικές, βοηθητικά ρήματα, επιρρήματα), 41 λεξικές κατηγορίες που εκφράζουν ψυχολογικές διαδικασίες (π.χ. συναίσθημα, αντίληψη, κοινωνικότητα), 6 κατηγορίες με προσωπικά θέματα (π.χ. επάγγελμα, χρήμα, ελεύθερος χρόνος), 5 δείκτες που υποδηλώνουν προφορικότητα της γλώσσας (π.χ. υβριστικές λέξεις) και 12 κατηγορίες σημείων στίξης (π.χ. περίοδοι, ημιπερίοδοι, απόστροφοι). Οι παραπάνω κατηγορίες σχετίζονται με συναισθηματικές και γνωσιακές καταστάσεις του συγγραφέα. Το πρόγραμμα δεν υποστηρίζει την Ελληνική γλώσσα.

- **MRC** The Medical Research Council Psycholinguistic Database Machine Usable Dictionary (Wilson, 1988)

Είναι βάση δεδομένων με γλωσσολογικά και ψυχολογικά χαρακτηριστικά. Είναι σχεδιασμένη για χρήση από ψυχογλωσσολόγους και για ερευνητές στο πεδίο της επεξεργασίας φυσικής γλώσσας. Διαφέρει από άλλα υπολογιστικά λεξικά όχι μόνο διότι παρέχει συντακτικές πληροφορίες αλλά επιπλέον και ψυχολογικά δεδομένα για κάθε εγγραφή. Για τις 150.837 λήμματα που περιέχει, παρέχει στατιστικά στοιχεία για 26 διαφορετικές γλωσσολογικές ιδιότητες π.χ. μέρη του λόγου, πλήθος γραμμάτων, φωνημάτων και συλλαβών των λέξεων, ομόηχες λέξεις.

- **NRC**[3] National Research Council Word-Emotion Association Lexicon (Mohammad & Turney, 2013)

Λεξικό που περιέχει 14.182 αγγλικές λέξεις επισημειωμένες με οκτώ βασικά συναισθήματα (θυμό, φόβο, ελπίδα, εμπιστοσύνη, έκπληξη, θλίψη, χαρά και αποστροφή) και δύο απόψεις (αρνητική και θετική). Κάθε λέξη μπορεί να

[3] https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

συσχετίζεται με κανένα, ένα ή περισσότερα συναισθήματα. Αυτόματη μετάφραση είναι διαθέσιμη και για άλλες γλώσσες, όχι όμως την Ελληνική.

- **SentiStrength**[4] (Thelwall et al., 2010)

Πρόγραμμα αυτόματης ανάλυσης άποψης, που βασίζεται σε λίστες θετικών και αρνητικών όρων. Σε κάθε σύντομο κείμενο αποδίδει θετικό, αρνητικό ή ουδέτερο χαρακτηρισμό σε μια κλίμακα από 1 έως 5. Υπάρχει επιλογή για την μορφή των αποτελεσμάτων, μπορεί δηλ. να δίνει μόνο τη μία διάσταση, τις δύο ή και τις τρεις.

- **SenticNet** [5]

Λεξικός πόρος που παρέχει σημασιολογική πληροφορία, κατηγοριοποίηση συναισθημάτων και πολικότητα για 100.000 έννοιες της Αγγλικής γλώσσας.

- **SPLICE** Structured Programming for Linguistic Cue Extraction [6]

Είναι ένα γλωσσολογικό εργαλείο ανάλυσης κειμένου σε μορφή διαδικτυακής υπηρεσίας.

Συνοψίζοντας, η ανοδική προσέγγιση φαίνεται να είναι αποτελεσματικότερη, αλλά ενέχει τον κίνδυνο της υπερπροσαρμογής του μοντέλου στα δεδομένα (overfitting), δηλαδή ο αλγόριθμος έχει προσαρμοστεί τόσο σε ένα σύνολο δεδομένων με αποτέλεσμα να χάνει την προβλεπτική του ικανότητα όταν εφαρμόζεται σε νέα δεδομένα και για να αποφευχθεί πρέπει να εφαρμόζεται σε μεγάλα σώματα κειμένων. Από την άλλη, η καθοδική προσέγγιση είναι πιο ανθεκτική (robust), δηλαδή πραγματοποιούνται ορθές προβλέψεις ακόμα και όταν τα δεδομένα χαρακτηρίζονται από προβλήματα, αλλά επιδέχεται λιγότερες βελτιώσεις. Επιπλέον, η μέθοδος που βασίζεται στη χρήση κλειστού λεξιλογίου περιορίζει την εφαρμογή της στην ανάλυση κειμένων γραμμένων στη γλώσσα του λεξικού και η απόδοσή της εξαρτάται και από την πληρότητα και την ποιότητα του λεξικού. Ένας περιορισμός που πρέπει να ληφθεί υπ' όψιν όταν χρησιμοποιείται η μέθοδος με το λεξικό σε πλατφόρμες κοινωνικής δικτύωσης είναι η δομή και η ποιότητα του λόγου. Οι χρήστες γράφουν συχνά με λάθη σε κάθε γλωσσικό επίπεδο, με συντμήσεις και λέξεις ή φράσεις που απαντούν μόνο σε

[4] http://sentistrength.wlv.ac.uk/
[5] https://sentic.net/
[6] splice.cmi.arizona.edu/

αυτό το περιβάλλον, πραγματικότητα που απαιτεί περαιτέρω βελτίωση των μεθόδων που βασίζονται στη χρήση κλειστού λεξικού.

Από την άλλη πλευρά, οι τεχνικές ανοικτού λεξιλογίου με τη χρήση μηχανικής μάθησης πετυχαίνουν καλύτερα αποτελέσματα, αλλά χρειάζεται εκπαίδευση που μπορεί να είναι αρκετά χρονοβόρα.

Για τους παραπάνω λόγους, η επιστημονική έρευνα έχει στραφεί σε υβριδικές μεθόδους που συνδυάζουν τη χρήση λεξικού με μηχανική μάθηση, ώστε να επωφεληθούν από τα πλεονεκτήματα και των δύο προσεγγίσεων, δηλαδή της ταχύτητας και της ακρίβειας αντίστοιχα. Από τον Celli (2012), για παράδειγμα, προτάθηκε μια συνδυαστική μέθοδος, χρήσιμη για domain adaptation, δηλαδή για προσαρμογή μοντέλου (μηχανικής μάθησης) σε σύνολο δεδομένων διαφορετικό αλλά παρεμφερές με το σύνολο δεδομένων εκπαίδευσης (εκμάθησης).

4.3 Μελέτες αυτόματης αναγνώρισης της προσωπικότητας

Στο υποκεφάλαιο που ακολουθεί γίνεται αναφορά στις έρευνες που έχουν πραγματοποιηθεί με αντικείμενο την αυτόματη αναγνώριση της προσωπικότητας των συγγραφέων κειμένου γραπτού (essay) και προφορικού λόγου, χρηστών κινητού τηλεφώνου, αποστολέων μηνυμάτων μέσω ηλεκτρονικού ταχυδρομείου, χρηστών ιστοσελίδων, βιντεο-ιστοσελίδων καθώς και ατόμων που διατηρούν λογαριασμό στο Facebook και στο Twitter.

Πάντως, σύμφωνα με μελέτη (Hinds & Joinson, 2019) η αυτόματη πρόβλεψη προσωπικότητας με υπολογιστικά μοντέλα είναι πιο αποτελεσματική από την πρόβλεψη που επιχειρήθηκε από ερευνητές μέσω της ερμηνείας συμπεριφορών χωρίς τη χρήση μηχανικών τεχνικών.

4.3.1 Αυτόματη αναγνώριση προσωπικότητας από κείμενο

Η πρώτη αναφορά έχει σχέση με τα σώματα κειμένων, απαραίτητα για οποιαδήποτε ανάλυση. Ο Coniam (2004) σε μελέτη του παρουσιάζει τον τρόπο δημιουργίας σώματος κειμένων από ακαδημαϊκά άρθρα δύο επιστημόνων του τομέα

της εφαρμοσμένης γλωσσολογίας και της γλωσσικής διδασκαλίας της Αγγλικής. Παρόλο που θα θεωρούσε κανείς ότι το υλικό αυτό έχει ομοιομορφία και είναι απρόσωπο, ωστόσο εντοπίζεται διαφορά στο ύφος. Ως χαρακτηριστικά της γλωσσικής ανάλυσης χρησιμοποιήθηκαν οι πιο συχνές λέξεις και ν-γράμματα τεσσάρων λέξεων και όχι λιγότερων, για να αποφευχθούν συνδυασμοί λειτουργικών λέξεων. Όσον αφορά την προσωπικότητα των συγγραφέων δόθηκε έμφαση στην απρόσωπη σύνταξη, η οποία δημιουργεί απόσταση και στη χρήση του πρώτου προσώπου, που ενισχύει την επαφή με τον αναγνώστη.

Κατά τη μελέτη γραπτών εκθέσεων που γράφτηκαν από φοιτητές της Ψυχολογίας (Pennebaker & King, 1999) διαπιστώθηκε η συσχέτιση ανάμεσα σε γλωσσικά χαρακτηριστικά και τα χαρακτηριστικά προσωπικότητας του μοντέλου των πέντε παραγόντων. Οι εξωστρεφείς χρησιμοποίησαν περισσότερες λέξεις κοινωνικού περιεχομένου, αυτοαναφορές και λέξεις δηλωτικές θετικών συναισθημάτων. Οι νευρωτικοί φοιτητές έγραφαν κυρίως σε πρώτο ενικό πρόσωπο και οι λέξεις που δήλωναν συναισθήματα είχαν αρνητικό περιεχόμενο. Όσοι ήταν δεκτικοί σε εμπειρίες, χρησιμοποίησαν περισσότερα άρθρα και μεγαλύτερες λέξεις και λιγότερες λέξεις σε πρώτο ενικό και χρόνο ενεστώτα. Άτομα με υψηλό σκορ στην προσήνεια βρέθηκε ότι χρησιμοποιούν περισσότερο το πρώτο ενικό πρόσωπο και λέξεις που εκφράζουν θετικά συναισθήματα και λιγότερα άρθρα. Τέλος, σε συμμετέχοντες με αυξημένο το χαρακτηριστικό της ευσυνειδησίας φάνηκε ότι ο λόγος τους συσχετίζεται θετικά με τη χρήση θετικών συναισθηματικά λέξεων και λιγότερων αρνήσεων.

Μια από τις πρώτες προσπάθειες που πραγματοποιήθηκαν (Argamon et al., 2005; Argamon et al., 2007) για την αυτόματη πρόβλεψη του ψυχολογικού τύπου των συγγραφέων με τεχνικές μηχανικής μάθησης αφορούσε σε κείμενα με ύφος οικείο και καθημερινό. Η έρευνα κινείται στο πλαίσιο της ψυχολογίας της γλώσσας αλλά και της υφομετρίας. Επεξεργάστηκαν 1.157 εκθέσεις αυθόρμητες και εκθέσεις στις οποίες ανέλυαν τον χαρακτήρα τους 1.106 φοιτητές Ψυχολογίας του πανεπιστημίου του Τέξας, Austin, που γράφτηκαν μεταξύ των ετών 1997 και 2003. Από το ερωτηματολόγιο προσωπικότητας βασισμένο στο μοντέλο των Πέντε Παραγόντων που συμπλήρωσαν οι φοιτητές, αξιοποιήθηκαν τα αποτελέσματα για την Εξωστρέφεια και τον Νευρωτισμό.

Τα χαρακτηριστικά που εξήχθησαν από τα κείμενα, για να χρησιμοποιηθούν στην αυτόματη αναγνώριση της προσωπικότητας των φοιτητών με υφομετρική ταξινόμηση, περιλάμβαναν 675 λειτουργικές λέξεις. Επιπλέον, κατασκευάζοντας ένα

λεξικό βάσει της Συστημικής Λειτουργικής Γραμματικής (Systemic Functional Grammar), προέκυψαν τρεις κατηγορίες γλωσσολογικών χαρακτηριστικών: συνδετικές λέξεις-φράσεις (Conjunction), δείκτες τροπικότητας (Modality) και αξιολογικά επίθετα και τροποποιητές (modifiers) με θετικό ή αρνητικό προσανατολισμό (Attitude and Orientation). Και οι δυο ομάδες χαρακτηριστικών είναι ανεξάρτητες περιεχομένου. Για την δημιουργία του μοντέλου αναγνώρισης χρησιμοποιήθηκε ο αλγόριθμος μηχανών υποστήριξης SMO με τον οποίο για τον Νευρωτισμό επιτεύχθηκε ακρίβεια 57%.

Η προαναφερθείσα έρευνα ακολούθησε μια ανοδική (bottom-up) προσέγγιση, ενώ η επίσης βασική μελέτη στον τομέα της αυτόματης αναγνώρισης της προσωπικότητας των Mairesse & Walker (2006) (Mairesse et al., 2007), έχει καθοδική (top-down) προσέγγιση. Σε αντίθεση με άλλες που αναφέρουν μόνο αποτελέσματα ταξινόμησης, σε αυτήν οι ερευνητές εφάρμοσαν μοντέλα ταξινόμησης (classification), παλινδρόμησης (regression) και ιεράρχησης (ranking) για κάθε χαρακτηριστικό του μοντέλου των Πέντε Παραγόντων. Πραγματοποίησαν μια σειρά από πειράματα σε ένα σώμα κειμένων 2.479 εκθέσεων, οι οποίες συντάχθηκαν από φοιτητές ψυχολογίας σε είκοσι λεπτά γράφοντας ό,τι σκεφτόντουσαν σε αυτή τη διάρκεια, για να διαπιστώσουν εάν αυτόματα εκπαιδευμένα μοντέλα μπορούν να αναγνωρίσουν την προσωπικότητα άγνωστων ατόμων. Συγχρόνως, αξιοποιήθηκε και ένα μικρότερο σώμα κειμένων αποτελούμενο από ηχογραφημένες συζητήσεις, στο οποίο αναφερόμαστε στο επόμενο υποκεφάλαιο με θέμα την αναγνώριση προσωπικότητας από προφορικό λόγο.

Εξήχθησαν από 88 κατηγορίες λέξεων από το LIWC, γλωσσολογικά χαρακτηριστικά τα οποία περιλαμβάνουν συντακτική και σημασιολογική πληροφορία. Επιπλέον με τη χρήση της ψυχογλωσσολογικής βάσης MRC (Wilson, 1988) προέκυψαν 14 ψυχογλωσσολογικά χαρακτηριστικά. Για την αξιολόγηση των μοντέλων ταξινόμησης έγινε χρήση έξι αλγορίθμων: αλγόριθμος δέντρων αποφάσεων C4.5, αλγόριθμος του Πλησιέστερου Γείτονα J48, Naive Bayes, Ripper, Adaboost και αλγόριθμος Μηχανών Διανυσμάτων Υποστήριξης. Ενδιαφέρον είναι το αποτέλεσμα που αφορά στη Δεκτικότητα στην εμπειρία, καθώς πέντε στους έξι αλγόριθμους ξεπέρασαν το σημείο αναφοράς (ακρίβεια 62,1%) και τέσσερις παρουσιάζουν την καλύτερη επίδοση γι' αυτό το χαρακτηριστικό. Η ευσυνειδησία είναι το δυσκολότερο χαρακτηριστικό για μοντελοποίηση. Τα καλύτερα αποτελέσματα ταξινόμησης επιτεύχθηκαν με τη χρήση του Naive Bayes αλγόριθμου (73,2% σωστές ταξινομήσεις στην Εξωστρέφεια). Τέλος, αξιολογήθηκε ο τρόπος με τον οποίο κάθε σύνολο

χαρακτηριστικών συνέβαλε στο τελικό αποτέλεσμα και φάνηκε πως τα χαρακτηριστικά από το LIWC ήταν πιο αποδοτικά σε σχέση με αυτά του MRC για όλους τους παράγοντες του μοντέλου των Πέντε Παραγόντων και περισσότερο για την Δεκτικότητα στην εμπειρία.

Αποκλειστικά για τη μελέτη της Εξωστρέφειας η ίδια ομάδα δημιούργησε μοντέλο παραγωγής λόγου (Mairesse & Walker, 2007) βασισμένο σε δεδομένα της ψυχογλωσσολογικής έρευνας. Αξιοποίησαν στοιχεία γλωσσικά που εντοπίσθηκαν στο λόγο των εξωστρεφών, όπως ευρεία θεματολογία, πολλές αυτοαναφορές, λίγα άρθρα, λίγες λέξεις ανά πρόταση, πολλά ρήματα, επιρρήματα και αντωνυμίες, λίγες αρνήσεις, πολλές λέξεις δηλωτικές θετικού συναισθήματος.

Με στόχο την πρόβλεψη της προσωπικότητας του συγγραφέα από κείμενο εργάστηκαν και οι Luyckx & Daelemans (2008a, 2008b) στο πανεπιστήμιο Antwerp του Βελγίου. Η πρωτοτυπία τους έγκειται στην δημιουργία σώματος κειμένων (“Personae”) 200.000 λέξεων στην Ολλανδική γλώσσα και όχι στην Αγγλική ζητώντας από 145 φοιτητές να γράψουν για ένα θέμα που δεν σχετίζεται με την προσωπικότητα, αλλά με την περιγραφή και τον σχολιασμό ενός ντοκιμαντέρ για την τεχνητή ζωή. Επίσης σημαντική είναι η επιλογή όχι του μοντέλου των Πέντε Παραγόντων για το ψυχομετρικό τεστ στο οποίο υποβλήθηκαν οι συμμετέχοντες, αλλά του Ερωτηματολογίου Τύπων Προσωπικότητας Myers-Briggs Type Indicator (MBTI). Η μεθοδολογία της συγκεκριμένης έρευνας αντιμετωπίζει το πρόβλημα ως θέμα ταξινόμησης κειμένου, επομένως, χρησιμοποίησαν κείμενα εκπαίδευσης, αυτόματη εξαγωγή χαρακτηριστικών και εκπαίδευση αλγορίθμου μηχανικής μάθησης. Χρησιμοποιήθηκαν κυρίως συντακτικά χαρακτηριστικά ως πιο αξιόπιστα, καθώς δεν ελέγχονται συνειδητά από τους συγγραφείς. Η εξαγωγή τους έγινε αφού είχε προηγηθεί επιφανειακή συντακτική ανάλυση (Memory-Based Shallow Parser). Επιπλέον, όμως έγινε χρήση και λεξικών χαρακτηριστικών με τη μορφή ν-γραμμάτων και λειτουργικών λέξεων και ν-γράμματα με λεπτομερειακή (fine-grained) και αδρή (coarse-grained) ανάλυση μερών του λόγου. Τα αποτελέσματα βασίστηκαν σε 10-πτυχη διασταυρούμενη επικύρωση (tenfold cross validation) με εκμάθηση βασισμένη σε μνήμη (memory-based learning). Όσον αφορά τους τύπους Εξωστρέφεια-Εσωστρέφεια και Διαίσθηση-Νόηση τα αποτελέσματα ακρίβειας είναι 65,5% και 62% αντίστοιχα.

Έρευνα πραγματοποιήθηκε στον ίδιο τομέα και για τα Κορεάτικα (Lee et al., 2007). Ογδόντα Κορεάτες φοιτητές συμπλήρωσαν δύο ερωτηματολόγια προσωπικότητας, ένα με βάση το μοντέλο των Πέντε Παραγόντων και το

Ερωτηματολόγιο Τύπων Προσωπικότητας Myers-Briggs Type Indicator (MBTI) και τους ζητήθηκε να γράψουν αυθόρμητα ένα κείμενο με όποιο θέμα επιθυμούσαν εντός είκοσι λεπτών, το οποίο αναλύθηκε από την κορεατική έκδοση του LIWC. Στόχος ήταν να ελεγχθεί η σχέση ανάμεσα στα χαρακτηριστικά της προσωπικότητας και στις 86 από τις μεταβλητές του κορεατικού LIWC, στόχος που επιβεβαιώθηκε. Μεταξύ αυτών των χαρακτηριστικών ήταν τόσο γλωσσολογικά (π.χ. μέρη του λόγου) όσο και ψυχολογικά χαρακτηριστικά (π.χ. λέξεις δηλωτικές αρνητικών ή θετικών συναισθημάτων). Σημαντική ήταν η διαπίστωση της στενής σχέσης ανάμεσα στην Εξωστρέφεια και των δύο ερωτηματολογίων και την χρήση ρημάτων.

Στο άρθρο του Wright (2012) διαβάζουμε για την έρευνα που πραγματοποιήθηκε σε εκθέσεις 2593 φοιτητών του πανεπιστημίου του Texas, οι οποίες συντάχθηκαν αυθόρμητα εντός είκοσι λεπτών. Χρησιμοποιήθηκαν δύο χαρακτηριστικά, η προσωπική αντωνυμία του πρώτου προσώπου (“I”) σε όλες της τις μορφές και διγράμματα λέξεων (“have to”). Με 10-πτυχη διασταυρούμενη επικύρωση η ακρίβεια έφτασε το 51,3%. Ο συγγραφέας του άρθρου καταλήγει στο συμπέρασμα ότι απαιτούνται χαρακτηριστικά υψηλότερου δομικού επιπέδου, ώστε να επιτρέπουν μεγαλύτερη δυνατότητα πρόβλεψης.

Η μελέτη του Celli (2012) είχε στόχο την αναγνώριση προσωπικότητας με υβριδική μέθοδο. Το σώμα κειμένων αποτελείται από εκθέσεις 2.500 ατόμων στα Αγγλικά και αναρτήσεις 23 χρηστών του Facebook και σύντομα κείμενα στα Ιταλικά. Το σύστημα δέχεται τέσσερα διαφορετικά είδη χαρακτηριστικών: ψυχολογικά (βάσει του MRC), γλωσσολογικά (λέξεις και ν-γράμματα που συσχετίζονται με χαμηλά ή υψηλά σκορ στα χαρακτηριστικά προσωπικότητας), διαγλωσσικά (όλα τα ανεξάρτητα γλώσσας από το LIWC και το MRC, π.χ. σημεία στίξης, αριθμοί, συχνότητα επαναλήψεων) και ψυχογλωσσολογικά (βάσει του LIWC, π.χ. λέξεις που σχετίζονται με άγχος, θυμό, συναισθήματα, καθώς και αντωνυμίες, αρνητικά μόρια και άλλους γλωσσολογικούς δείκτες).

Αξίζει να σημειωθεί ότι τα καλύτερα αποτελέσματα επιτεύχθηκαν με τη χρήση των διαγλωσσικών χαρακτηριστικών. Στο πλαίσιο της ίδιας μελέτης πραγματοποιήθηκε ένα πείραμα κατά το οποίο ζήτησαν από ανθρώπους να διαβάσουν το σώμα κειμένων και να απαντήσουν το ερωτηματολόγιο προσωπικότητας για τον συγγραφέα των κειμένων. Προκαλεί εντύπωση το αποτέλεσμα της πρόβλεψης της Εξωστρέφειας· βρέθηκε με μεγάλη επιτυχία αυτό το χαρακτηριστικό αλλά όχι τα υπόλοιπα από τους ανθρώπους, αντίθετα από το σύστημα, το οποίο είχε την χειρότερη

παρουσίαση για την Εξωστρέφεια. Αυτό υποδηλώνει πως η Εξωστρέφεια είναι θέμα περισσότερο σημασιολογίας. Σε πρόσφατη έρευνα (Bowden-Green, Hinds & Joinson, 2020) που διενεργεί βιβλιογραφική επισκόπηση της συσχέτισης της Εξωστρέφειας με τα μέσα κοινωνικής δικτύωσης διαπιστώνεται ότι ένας από τους δείκτες είναι οι λέξεις με κοινωνικό σημασιολογικό φορτίο, με θετικό περιεχόμενο.

Σε συνέχεια της προαναφερθείσας έρευνας δημιουργήθηκε ένα μη εποπτευόμενης μάθησης (supervised learning) σύστημα αναγνώρισης προσωπικότητας, διαδικτυακά προσβάσιμο (Celli & Poesio, 2014), το οποίο έχει ελεγχθεί στα Αγγλικά και στα Ιταλικά με f=0,68. Για αυτό το αποτέλεσμα χρησιμοποιείται ένα υποσύνολο του παραπάνω σώματος κειμένων και μόνο τα διαγλωσσικά χαρακτηριστικά, που είναι ανεξάρτητα γλώσσας.

Οι Mohammad & Kiritchenko (2013) μελέτησαν την επίδραση λέξεων που εκφράζουν συγκεκριμένα συναισθήματα στην αυτόματη πρόβλεψη της προσωπικότητας του συγγραφέα ενός κειμένου. Με δεδομένα των Pennebaker & King (1999), εκπαίδευσαν ταξινομητές Μηχανών Διανυσμάτων Υποστήριξης για κάθε χαρακτηριστικό του μοντέλου των Πέντε Παραγόντων και διαπίστωσαν ότι η χρήση των χαρακτηριστικών που εκφράζουν συναίσθημα αυξάνει το ποσοστό ακρίβειας της πρόβλεψης.

Επιπλέον, μελετήθηκε το θέμα των γλωσσικών χαρακτηριστικών που χρησιμοποιούνται για την αναγνώριση προσωπικότητας σε σχέση με το είδος των κειμένων (Wright & Chin, 2016). Το σύνολο, δηλαδή, των χαρακτηριστικών που εξήχθησαν για την πρόβλεψη της προσωπικότητας π.χ. από εκθέσεις φοιτητών είναι πιθανό να μην είναι αποτελεσματικό και σε σώματα κειμένων από αναρτήσεις στο διαδίκτυο. Δεδομένου ότι η χρήση των λέξεων εξαρτάται από το περικείμενο, μελέτησαν ν-γράμματα λέξεων, ν-γράμματα μερών του λόγου και υβριδικά ν-γράμματα που συνδυάζουν λέξεις και μέρη του λόγου, σε σώμα κειμένων από forums και από εκθέσεις φοιτητών. Τελικά, τα χαρακτηριστικά έδειξαν σύνδεση με την Ευσυνειδησία.

Μια άλλη προσέγγιση του θέματος της αυτόματης αναγνώρισης της προσωπικότητας και συγκεκριμένα για την ύπαρξη ή απουσία (δυαδικός ταξινομητής) των χαρακτηριστικών του μοντέλου των Πέντε Παραγόντων είναι η βαθιά μάθηση (Deep Learning) χρησιμοποιώντας συνελικτικά νευρωνικά δίκτυα (convolutional neural networks), ένα για κάθε παράγοντα του Big Five (Majumder et al., 2017). Χρησιμοποιήθηκαν τα πιο συχνά μονογράμματα, διγράμματα και τριγράμματα λέξεων

και ανά λέξη σημασιολογικά χαρακτηριστικά, τα οποία συνδυάζονται σε μια αναπαράσταση μεταβλητού μήκους του κειμένου. Όταν αυτή η αναπαράσταση εισαχθεί στο νευρωνικό δίκτυο, ενώνονται οι λέξεις σε ν-γράμματα, τα ν-γράμματα σε προτάσεις και οι προτάσεις σε ένα αρχείο.

Τα αποτελέσματα συνδυάστηκαν με όλα τα χαρακτηριστικά των Mairesse et al., (2007), που είχαν ήδη εξαχθεί και έτσι προκύπτει η αναπαράσταση του κάθε αρχείου. Η τεχνική βελτιώνεται με την αποβολή των προτάσεων του κειμένου που είναι συναισθηματικά ουδέτερες. Η μέθοδός τους, εφαρμοσμένη σε γραπτά κείμενα-εκθέσεις, είναι η αποτελεσματικότερη και μάλιστα υπερτερεί για όλα τα χαρακτηριστικά.

Σε έρευνα πρόβλεψης προσωπικότητας μελετήθηκε η Σουηδική γλώσσα με την ανάκτηση κειμένων που ήταν αναρτημένα σε σουηδικά forums και ιστοσελίδες ειδήσεων (Akrami et al., 2019). Η ιδιαιτερότητα της έρευνας έγκειται, αρχικά, στην αντιμετώπιση των Πέντε Παραγόντων με μορφή κλίμακας από -3 έως 3 για το κάθε χαρακτηριστικό, αντί για τη μεθοδολογία της δυαδικής ταξινόμησης που ακολουθούν οι περισσότεροι ερευνητές. Βέβαια, έγιναν πειράματα και για δυαδική ταξινόμηση αποκλειστικά για τη σύγκριση των αποτελεσμάτων με άλλες έρευνες. Επιπλέον, τα κείμενα επισημειώθηκαν από φοιτητές ψυχολογίας με τη χρήση διαδικτυακού εργαλείου και όχι από τους ίδιους τους συντάκτες των κειμένων με τη συμπλήρωση ερωτηματολογίου προσωπικότητας. Ως χαρακτηριστικά εξήχθησαν διγράμματα λέξεων και χαρακτήρων και τετραγράμματα χαρακτήρων.

Χρησιμοποιώντας ένα ήδη υπάρχον και προαναφερθέν σώμα κειμένων (Pennebaker & King, 1999) και με την εξαγωγή ψυχογλωσσολογικών χαρακτηριστικών (LIWC, SenticNet, NRC) οι Mehta et al. (2020) μελέτησαν την προσωπικότητα των συγγραφέων των κειμένων. Συμπέραναν ότι τα αποτελέσματα βελτιώνονται συνδυάζοντας τεχνικές βαθιάς μάθησης.

Για την Ελληνική γλώσσα έχει γίνει μια προσπάθεια αυτόματης αναγνώρισης των πέντε στοιχείων της προσωπικότητας 73 φοιτητών πληροφορικής από κείμενα που γράφτηκαν αυθόρμητα χωρίς συγκεκριμένο θέμα στη διάρκεια 15-20 λεπτών (Kermanidis, 2012). Ο μέσος όρος σε λέξεις της έκτασης των κειμένων ήταν 172,1. Όσον αφορά τα γλωσσικά χαρακτηριστικά ήταν χαμηλού επιπέδου (low level) γλωσσολογικά χαρακτηριστικά (μορφολογικά και λέξεις που εκφράζουν συναίσθημα). Για την αυτόματη κατηγοριοποίηση εφαρμόστηκαν Μηχανές Διανυσμάτων

Υποστήριξης. Τα αποτελέσματα οδήγησαν στο συμπέρασμα ότι μπορεί να γίνει πρόβλεψη της προσωπικότητας

Άλλη μια έρευνα για την Ελληνική γλώσσα πραγματοποιήθηκε (Komianos et. al., 2012) σε περισσότερα δεδομένα από την προηγούμενη με στόχο τη συσχέτιση ανάμεσα σε μορφοσυντακτικές και σημασιολογικές δομές της γραφής ενός ατόμου και σε στοιχεία του χαρακτήρα του βάσει του μοντέλου των Πέντε Παραγόντων. Από τους συμμετέχοντες συγκεντρώθηκαν 382 κείμενα (μέσος όρος λέξεων: 110,3) γραμμένα αυθόρμητα σε χρόνο είκοσι λεπτών που εξέφραζαν σκέψεις και συναισθήματα. Το σώμα κειμένων επισημειώθηκε με μορφολογική και σημασιολογική πληροφορία (π.χ. μέρη του λόγου, ξένες λέξεις, ορθογραφικά λάθη, λέξεις που εκφράζουν συναισθήματα θετικά ή αρνητικά). Οι ερευνητές πραγματοποίησαν δύο σετ πειραμάτων: Το πρώτο για παλινδρόμηση (αλγόριθμος δένδρων παλινδρόμησης M5P), δηλαδή, για τιμές κάθε χαρακτηριστικού του μοντέλου των Πέντε Παραγόντων και το δεύτερο για δυαδική ταξινόμηση (ταξινομητής Quinlan's C4.5), δηλαδή, για ένταξη ενός κειμένου σε συγκεκριμένο χαρακτηριστικό του συγγραφέα ή όχι.

Στο υποκεφάλαιο αυτό παρουσιάσαμε τις μελέτες που έχουν πραγματοποιηθεί μέχρι και το 2020 στο ερευνητικό πεδίο της αυτόματης αναγνώρισης χαρακτηριστικών της προσωπικότητας του συγγραφέα κειμένων με τεχνικές μηχανικής μάθησης. Οι μελέτες, που ξεκίνησαν πριν από δεκαπέντε χρόνια (2005), αφορούν κυρίως στην Αγγλική γλώσσα στην οποία αναπτύχθηκαν σώματα κειμένων κυρίως από φοιτητές. Για τον έλεγχο των αποτελεσμάτων αξιοποιήθηκε στην πλειοψηφία των περιπτώσεων το τεστ προσωπικότητας βασισμένο στο μοντέλο των Πέντε Παραγόντων. Τα χαρακτηριστικά που χρησιμοποιήθηκαν για τον προσδιορισμό της προσωπικότητας ήταν οι λειτουργικές λέξεις, ν-γράμματα και λεξικά με ψυχογλωσσολογικό περιεχόμενο. Έρευνες στα Ελληνικά έγιναν δύο το 2012, αλλά είναι πολύ περιορισμένης εμβέλειας. Τα ποσοστά ακρίβειας των μοντέλων είναι χαμηλά, σύμφωνα με τους ερευνητές, εξαιτίας περιορισμένων χαρακτηριστικών και του σώματος κειμένων που αποτελείται από μικρό αριθμό λέξεων ανά συγγραφέα.

4.3.2 Αυτόματη αναγνώριση προσωπικότητας από προφορικό λόγο

Έρευνες που αφορούν στην επεξεργασία προφορικού λόγου με στόχο την πρόβλεψη της προσωπικότητας του ομιλούντος έχουν πραγματοποιηθεί με τη χρήση διάφορων χαρακτηριστικών. Θα αναφερθούμε σε αυτές που αξιοποίησαν λεκτικά χαρακτηριστικά και όχι μη λεκτικά, όπως παραγλωσσικά, ακουστικά και οπτικοακουστικά στοιχεία (Pianesi et al., 2008; Mohammadi & Vinciarelli, 2012; Ivanov et al., 2011; Staiano et al., 2011; Alam & Riccardi, 2013).

Σε εξήντα οκτώ διαφόρων εθνικοτήτων φοιτητές αμερικανικού πανεπιστημίου εφαρμόστηκε ένα πείραμα (Cohen et al., 2008), κατά το οποίο τους ζητήθηκε να μιλούν για όποιο θέμα επιθυμούσαν για τρία λεπτά και να ηχογραφούνται. Ακολούθησε μετεγγραφή και γλωσσική ανάλυση του υλικού μέσω του LIWC. Η διαφορά από άλλες έρευνες έγκειται στη μη χρήση του μοντέλου των Πέντε Παραγόντων για την ψυχολογική περιγραφή των συμμετεχόντων. Αντί αυτού χρησιμοποίησαν ένα νευροβιολογικό μοντέλο που ελέγχει τη συστολή (behavioral inhibition) και ένα μοντέλο που ελέγχει τα χαρακτηριστικά που επηρεάζονται από θετικά ή αρνητικά συναισθήματα.

Οι Mehl, Gosling & Pennebaker (2006), αφού χορήγησαν το ερωτηματολόγιο του μοντέλου των Πέντε Παραγόντων σε ενενήντα έξι φοιτητές ψυχολογίας, τους εφοδίασαν με τον ηλεκτρονικό καταγραφέα EAR (Electronically Activated Recorder) και κατέγραψαν ό,τι έλεγαν στη διάρκεια δύο εικοσιτετραώρων. Στη συνέχεια έγινε μετεγγραφή και γλωσσολογική ανάλυση με το LIWC, για να προβλέψουν το ψυχολογικό τους προφίλ. Η έρευνα είχε επιπλέον στόχο να τονίσει πόσο σημαντική είναι η παρατήρηση των συμμετεχόντων στο φυσικό τους περιβάλλον, γιατί μόνο έτσι μπορεί να εκφραστεί η προσωπικότητά τους.

Οι Mairesse et al. (2007) εξετάζουν το πρόβλημα χρησιμοποιώντας ηχητικά δεδομένα από το σώμα κειμένων που αναφέρθηκε στην προηγούμενη παράγραφο (96.468 λέξεις). Γίνεται αυτόματη εξαγωγή των χαρακτηριστικών χρησιμοποιώντας το LIWC για λεξικά στοιχεία και το MRC για ψυχογλωσσολογικά. Παράλληλα εξετάζεται και η προσωδία (π.χ. ένταση και τόνος φωνής) αλλά και το είδος της έκφρασης (π.χ. προσταγή, προτροπή, ερώτηση, ισχυρισμός), καθώς φάνηκε ότι το επίπεδο της πρωτοβουλίας του ομιλητή και ο τύπος των εκφράσεων που χρησιμοποιεί είναι

στοιχεία ενδεικτικά κάποιων χαρακτηριστικών της προσωπικότητας. Εφάρμοσαν μοντέλα ταξινόμησης (classification), παλινδρόμησης (regression) και ιεράρχησης (ranking) για κάθε χαρακτηριστικό του μοντέλου των Πέντε Παραγόντων από τα οποία η τελευταία κατηγορία εμφάνισε τα καλύτερα αποτελέσματα συνολικά και είναι πιο ακριβής σε σχέση με τους ταξινομητές πολλαπλών κατηγοριών (multi-class).

Για την αξιολόγηση του χαρακτήρα των συμμετεχόντων συμπλήρωσαν οι ίδιοι ερωτηματολόγια προσωπικότητας, αλλά συγχρόνως αξιολογήθηκαν και από κριτές. Παρατηρήθηκε ότι με την χρήση των ερωτηματολογίων τα αποτελέσματα πρόβλεψης είναι χειρότερα από αυτά που προέκυψαν με την χρήση της αξιολόγησης των κριτών. Πραγματοποιήθηκε και σύγκριση όλων των χαρακτηριστικών που εξήχθησαν και από το γραπτό σώμα κειμένων και από το προφορικό, για να ελεγχθεί ο τρόπος με τον οποίο κάθε κατηγορία χαρακτηριστικών συμβάλλει στο αποτέλεσμα της πρόβλεψης. Προέκυψε ότι για την Εξωστρέφεια με τον συνδυασμό γλωσσικών, ψυχογλωσσολογικών και προσωδιακών χαρακτηριστικών η ακρίβεια φτάνει στο 73% και μάλιστα φαίνεται ότι η προσωδία επηρεάζει μόνο την Εξωστρέφεια.

Για τον αυτόματο εντοπισμό του χαρακτηριστικού της Εξωστρέφειας πραγματοποιήθηκε μια έρευνα (Nowson & Gill, 2014) με ενενήντα έξι συμμετέχοντες των οποίων οι διάλογοι ηχογραφήθηκαν με τον ηλεκτρονικό καταγραφέα EAR και μεταγράφηκαν χειρωνακτικά. Τα κειμενικά χαρακτηριστικά εξήχθησαν με το λογισμικό LIWC. Τα αποτελέσματα της ταξινόμησης συγκρίθηκαν με αυτά από ένα σώμα κειμένων που προήλθε από βίντεο-ιστολόγιο και η ακρίβεια ήταν μεγαλύτερη στους διαλόγους.

4.3.3 Αυτόματη αναγνώριση προσωπικότητας από τη χρήση κινητού τηλεφώνου

Στο κεφάλαιο αυτό θα ανατρέξουμε στις μελέτες που αφορούν βέβαια την αναγνώριση προσωπικότητας αλλά από πραγματικά δεδομένα που προέρχονται από κινητά τηλέφωνα. Πρόσφατα, κατέστη δυνατή η πρόβλεψη της προσωπικότητας του χρήστη ενός κινητού τηλεφώνου από τον τρόπο με τον οποίο χρησιμοποιεί το τηλέφωνό του. Σε μια εποχή που οι συνδρομές της κινητής τηλεφωνίας παγκοσμίως ξεπερνούν τα επτά δισεκατομμύρια και η πρόσβαση στα δεδομένα των τηλεφώνων

είναι εφικτή, ο συγκεκριμένος ερευνητικός τομέας καθίσταται εξαιρετικά ενδιαφέρων με εφαρμογές στο εμπόριο, τη διαφήμιση αλλά και στις υπολογιστικές και κοινωνικές επιστήμες.

Ο Holtgraves Thomas (2011) μελέτησε μηνύματα κινητών τηλεφώνων για να εξετάσει τη σχέση ανάμεσα στη γλώσσα των μηνυμάτων και τους τύπους της προσωπικότητας σύμφωνα με το μοντέλο των Πέντε Παραγόντων. Στο πανεπιστήμιο Ball State διακόσιοι είκοσι τέσσερις πρωτοετείς φοιτητές της Ψυχολογίας στο πλαίσιο των μαθημάτων συμμετείχαν στο πείραμα με τίτλο «Έρευνα κινητού τηλεφώνου». Αφού συμπλήρωσαν ερωτηματολόγιο προσωπικότητας, τους ζητήθηκε να καταγράψουν τα τελευταία είκοσι μηνύματα που είχαν στείλει από το κινητό τους τηλέφωνο και επιπλέον την ημερομηνία και ώρα που στάλθηκαν, αν ήταν άλλοι γύρω τους όταν έγραφαν τα μηνύματα, το μέρος από όπου τα έστειλαν και κάποια ακόμα στοιχεία που αφορούσαν τους παραλήπτες. Τα κείμενα των μηνυμάτων αναλύθηκαν με το LIWC. Στα αποτελέσματα υπάρχει κάποια διαφοροποίηση σε σχέση με έρευνες που αφορούν γενικά στην επεξεργασία κειμένου, ίσως λόγω της ιδιαιτερότητας των μηνυμάτων. Για παράδειγμα, η Εξωστρέφεια συνδέεται με τη συχνότητα αποστολής μηνυμάτων παρά με τον αριθμό των λέξεων. Το φύλο του συγγραφέα καθώς και η σχέση του με τον παραλήπτη, όπως αναμενόταν, επηρέασε τη γλώσσα των μηνυμάτων.

Οι έρευνες που περιγράφονται στις επόμενες παραγράφους έχουν στόχο την πρόβλεψη της προσωπικότητας, αλλά με δεδομένα μη γλωσσικά, καθώς από τη βιβλιογραφία προκύπτει πως αξιοποιούνται χαρακτηριστικά που έχουν σχέση με τη χρήση του τηλεφώνου π.χ. διάρκεια κλήσεων και εφαρμογών.

Σε έρευνα των Chittaranjan, Blom & Gatica-Perez (2012) συλλέχθηκαν δεδομένα από 117 συμμετέχοντες στην Ελβετία, οι οποίοι είχαν στη διάθεσή τους συγκεκριμένο μοντέλο κινητού τηλεφώνου για 17 μήνες. Τα χαρακτηριστικά που χρησιμοποιήθηκαν προέρχονταν από στοιχεία όχι μόνο για τις κλήσεις και τα μηνύματα αλλά και τις εφαρμογές, Bluetooth και προφίλ. Οι χρήστες των τηλεφώνων συμπλήρωσαν ένα διαδικτυακό ερωτηματολόγιο προσωπικότητας βασισμένο στο μοντέλο των Πέντε Παραγόντων. Στο υλικό έγινε επεξεργασία με δυο στατιστικές τεχνικές, τη συσχέτιση και την πολλαπλή παλινδρόμηση. Υπολόγισαν τον δείκτη συσχέτισης Pearson ανάμεσα σε κάθε χαρακτηριστικό του μοντέλου των Πέντε Παραγόντων και στα χαρακτηριστικά που επιλέχθηκαν από τα δεδομένα του τηλεφώνου. Έτσι, για παράδειγμα, η Εξωστρέφεια βρέθηκε να συσχετίζεται θετικά με τη χρήση εφαρμογών γραφείου και εφαρμογών ημερολογίου. Μέσω γραμμικής

παλινδρόμησης ελέγχθηκε η σχέση ανάμεσα σε εξαρτημένες και ανεξάρτητες μεταβλητές, από όπου βρέθηκε ενδεικτικά ότι το Διαδίκτυο χρησιμοποιήθηκε περισσότερο από Εσωστρεφείς χρήστες του τηλεφώνου. Επιπλέον, οι ερευνητές κατασκεύασαν ένα σύστημα ταξινόμησης των χρηστών βασισμένο στην προσωπικότητά τους. Το πλεονέκτημα της μεθοδολογίας είναι ότι προσαρμόζεται εύκολα σε μεγάλο πλήθος δεδομένων και χαρακτηριστικών.

Οι de Oliveira et al. (2011) σε 39 άτομα από το Μεξικό χορήγησαν το ερωτηματολόγιο προσωπικότητας βασισμένο στο μοντέλο των Πέντε Παραγόντων. Για την εξαγωγή των κατάλληλων χαρακτηριστικών χρησιμοποιήθηκαν Μηχανές Διανυσμάτων Υποστήριξης. Κατέληξαν σε 474 μεταβλητές από τα αρχεία καταγραφής κλήσεων και σε 9 δομικά χαρακτηριστικά των κοινωνικών δικτύων ανάμεσα στα οποία και το πλήθος των επαφών. Για την Εξωστρέφεια, την Προσήνεια και την Δεκτικότητα στην εμπειρία σημείωσαν σημαντική πρόοδο.

Στον ίδιο ερευνητικό τομέα κινήθηκαν οι Staiano et al. (2012), οι οποίοι συνέλεξαν δεδομένα οκτώ εβδομάδων από 53 χρήστες κινητών Android, αποτελούμενα από στοιχεία των κλήσεων, στοιχεία εγγύτητας (κοντινά τηλέφωνα και συσκευές bluethooth) και στοιχεία προσωπικότητας των συμμετεχόντων καθώς και των μεταξύ τους σχέσεων. Με τις παραπάνω πληροφορίες δημιουργήθηκαν δίκτυα στις άκρες των οποίων τοποθετήθηκαν ως βάρη τα αποτελέσματα. Σε δίκτυα (network-level features) οργάνωσαν και τα εξαχθέντα χαρακτηριστικά, στοιχείο που διαφοροποιεί την συγκεκριμένη εργασία από άλλες. Φάνηκε από τα αποτελέσματα ότι αυτού του είδους τα χαρακτηριστικά σημείωσαν μεγαλύτερη επιτυχία από τα actor-based μοντέλα (π.χ. πλήθος και διάρκεια κλήσεων) των άλλων ερευνών.

Άλλη ερευνητική ομάδα (de Montjoye et al., 2013) συγκέντρωσε τα δεδομένα (κλήσεις, μηνύματα) από 69 ανθρώπους του πανεπιστημιακού χώρου (MIT), αφού τους εξόπλισε με ένα κινητό τηλέφωνο και συμπλήρωσαν ερωτηματολόγιο προσωπικότητας. Δημιούργησαν δείκτες που θα μπορούσαν να εντοπίσουν τις διαφορές ανάμεσα στα χαρακτηριστικά του μοντέλου των Πέντε Παραγόντων και κατηγοριοποιούνται σε: βασική χρήση του κινητού (πλήθος κλήσεων και μηνυμάτων), ενεργή συμπεριφορά του χρήστη (πλήθος εξερχόμενων κλήσεων, χρόνος απάντησης μηνύματος), τοποθεσία (πλήθος περιοχών από τις οποίες έγινε κλήση), κανονικότητα (μεσοδιάστημα κλήσεων και μηνυμάτων) και ποικιλία (λόγος αριθμού αλληλεπίδρασης προς αριθμό επαφών). Είναι αξιοσημείωτο ότι η Εξωστρέφεια και ο Νευρωτισμός ήταν τα χαρακτηριστικά με την καλύτερη πρόβλεψη. Ο μέσος όρος ακρίβειας στην οποία

έφτασε το σύστημα, το οποίο χρησιμοποίησε Μηχανές Διανυσμάτων Υποστήριξης, είναι 42%, με υψηλότερο ποσοστό (63%) στην πρόβλεψη του Νευρωτισμού.

Μια από τις πιο πρόσφατες μελέτες (Rüegger et al., 2020) που αναγνωρίζει με τεχνικές μηχανικής μάθησης τα χαρακτηριστικά της προσωπικότητας βάσει του μοντέλου των Πέντε Παραγόντων αξιοποιεί στοιχεία από τη χρήση του κινητού τηλεφώνου, όπως π.χ. η διάρκεια και συχνότητα των κλήσεων και οι επαφές. Προκύπτει ότι μόνο για το χαρακτηριστικό της Εξωστρέφειας μπορεί να γίνει ακριβής πρόβλεψη από τα δεδομένα του κινητού τηλεφώνου.

4.3.4 Αυτόματη αναγνώριση προσωπικότητας από μηνύματα ηλεκτρονικού ταχυδρομείου

Ακολουθούν οι εργασίες που αφορούν στην πρόβλεψη προσωπικότητας από μηνύματα ηλεκτρονικού ταχυδρομείου. Αυτό που πρέπει να ληφθεί υπόψη στη μελέτη ψηφιακών κειμένων που προέρχονται από επικοινωνία μέσω ηλεκτρονικού ταχυδρομείου είναι τα ιδιαίτερα χαρακτηριστικά αυτού του κειμενικού είδους. Συνήθως, πρόκειται για σύντομα κείμενα που δεν ακολουθούν πάντα τους καθορισμένους γραμματικούς και συντακτικούς κανόνες και παρατηρούνται λάθη. Δεν έχουν μόνο στοιχεία του προφορικού αλλά και του γραπτού λόγου, χρησιμοποιούνται συντμήσεις και το ύφος ποικίλλει ανάλογα με τον παραλήπτη. Επομένως, η σκιαγράφηση συγγραφικού προφίλ μέσω μηνυμάτων ηλεκτρονικού ταχυδρομείου δεν είναι εύκολη.

Οι Gill & Oberlander (2002) ζήτησαν από 105 φοιτητές ή πρόσφατα πτυχιούχους με μητρική γλώσσα την Αγγλική να γράψουν από δύο μηνύματα, τα οποία να απευθύνονται σε ένα φίλο που είχαν καιρό να δουν. Στο πρώτο έπρεπε να περιγράψουν πώς πέρασε η εβδομάδα τους και στο άλλο πώς σχεδίαζαν να περάσουν την επόμενη. Για τη συγγραφή κάθε μηνύματος έπρεπε να διαθέσουν 10 περίπου λεπτά. Το σώμα κειμένων που συγκεντρώθηκε αποτελείται από 65.000 λέξεις. Συγχρόνως οι συμμετέχοντες συμπλήρωσαν διαδικτυακά δημογραφικό ερωτηματολόγιο και το Ερωτηματολόγιο προσωπικότητας του Eysenck.

Συνδυάστηκαν τεχνικές ψυχογλωσσολογίας και στατιστικής ανάλυσης για τη μελέτη της εξωστρέφειας στα μηνύματα ηλεκτρονικού ταχυδρομείου. Αναζητήθηκαν

χαρακτηριστικές λέξεις και ακολουθίες λέξεων. Στο πρώτο πείραμα αναλύθηκε το σώμα κειμένων με το λογισμικό LIWC και την ψυχογλωσσολογική βάση MRC. Και με τις δυο παραπάνω αναλύσεις που βασίζονται σε λεξικά βρέθηκε πως ελάχιστα χαρακτηριστικά διακρίνουν τα κείμενα των εξωστρεφών από αυτά των εσωστρεφών συγγραφέων τους.

Στη δεύτερη δοκιμασία αξιοποιήθηκαν ανοδικές (bottom up) στατιστικές αναλύσεις κειμένων και συγκεκριμένα τα διγράμματα λέξεων: διγράμματα που απαντούν σε κείμενα και εξωστρεφών και εσωστρεφών, διγράμματα που απαντούν μόνο στα κείμενα των εξωστρεφών και μόνο των εσωστρεφών. Οι ερευνητές χώρισαν τα διγράμματα που θεώρησαν πως θα διέκριναν την προσωπικότητα του συγγραφέα σε οκτώ κατηγορίες: επιφανειακή συναίσθηση, ποσοτικοποίηση, κοινωνικοί μηχανισμοί, αυτο/ετεροαναφορές, σθένος, ικανότητα, τροπικότητα και σχεδιασμός μηνύματος/ έκφραση. Τελικά, επιβεβαίωσαν προηγούμενες μελέτες και παρουσίασαν κάποια νέα δεδομένα, όπως ότι οι εξωστρεφείς έχουν μειωμένη πυκνότητα λόγου και οι εσωστρεφείς προτιμούν αριθμούς και ποσοτικοποιήσεις (Oberlander & Gill, 2006).

Ως προς το θέμα της γλωσσικής παραγωγής και της σχέσης της με την προσωπικότητα, οι Oberlander & Gill (2004b) με το σώμα κειμένων των 65.000 λέξεων από τα μηνύματα ηλεκτρονικού ταχυδρομείου προσπάθησαν να μελετήσουν τη δημιουργία συστήματος για την αυτόματη παραγωγή φυσικής γλώσσας.

Οι ίδιοι ερευνητές (Oberlander & Gill, 2004a) χρησιμοποιώντας το σώμα κειμένων αποτελούμενο από μηνύματα ηλεκτρονικού ταχυδρομείου, που περιγράψαμε στο προηγούμενο υποκεφάλαιο, επιχείρησαν να επιβεβαιώσουν ότι η χρήση των μερών του λόγου επηρεάζει την πρόβλεψη της προσωπικότητας και μάλιστα την εξωστρέφεια και τον νευρωτισμό. Συγκεκριμένα, επιδίωξαν να βρουν συχνότητες και ακολουθίες μερών του λόγου. Αφού προηγήθηκε μορφολογική επισημείωση του σώματος κειμένων, ακολούθησε ανάλυση ν-γραμμάτων λέξεων από ν=1 έως ν=5. Σημαντικό ήταν το αποτέλεσμα της ανάλυσης μονογραμμάτων κατά την οποία φάνηκε ότι τα μέρη του λόγου συμβάλλουν στην πρόβλεψη του νευρωτικού αλλά όχι του εξωστρεφή συγγραφέα.

Από τους Gill & French (2007) ελέγχθηκε η πιθανότητα να αξιοποιηθούν τεχνικές κειμενικής συνεμφάνισης (text co-occurrence) στην αυτόματη πρόβλεψη της προσωπικότητας συγγραφέων ηλεκτρονικών μηνυμάτων. Πρόκειται για δημιουργία διανυσματικών αναπαραστάσεων, όπου κάθε λέξη αντιστοιχίζεται σε ένα διάνυσμα και παρέχονται πληροφορίες για την συνύπαρξη λέξεων στο κείμενο. Στο πρώτο πείραμα

χρησιμοποίησαν ζεύγη λέξεων που σχετίζονται με υψηλού ή χαμηλού βαθμού Εξωστρέφεια ή Νευρωτισμό και στο δεύτερο πείραμα άντλησαν λέξεις από κείμενα συγγραφέων που οι ίδιοι χαρακτηρίστηκαν υψηλού ή χαμηλού βαθμού εξωστρεφείς ή νευρωτικοί. Η μέθοδος απέτυχε και στις δύο περιπτώσεις να εντοπίσει συσχετισμό ανάμεσα στις λέξεις-κλειδιά και στα κείμενα προς εξέταση, διότι κατά τη γνώμη των ερευνητών τα κείμενα ήταν σύντομα και επιπλέον οι συγκεκριμένες τεχνικές δεν έχουν ακόμα υψηλό επίπεδο αναπαραστάσεων της προσωπικότητας σε σχέση με το πώς οι άνθρωποι αποτιμούν την προσωπικότητα.

Μια ενδιαφέρουσα έρευνα των Gill, Nowson & Oberlander (2006) αναφέρεται στη σύγκριση γλωσσολογικών χαρακτηριστικών από μηνύματα ηλεκτρονικού ταχυδρομείου και από κείμενα ιστολογίων, δύο διαφορετικών κειμενικών ειδών, που χρησιμοποιούνται για την πρόβλεψη προσωπικότητας. Το πρώτο σώμα κειμένων αποτελείται από 210 κείμενα ηλεκτρονικού ταχυδρομείου (δύο από κάθε έναν από τους 105 συμμετέχοντες), ενώ το δεύτερο από 1.854 αναρτήσεις ιστολογίων των 71 συμμετεχόντων. Φάνηκε ότι τα κείμενα ηλεκτρονικού ταχυδρομείου έχουν κοινά χαρακτηριστικά τόσο με τα κείμενα των ιστολογίων όσο και με τα γραπτά κείμενα μη ηλεκτρονικής επικοινωνίας (π.χ. εκθέσεις), σε αντίθεση με τις αναρτήσεις των ιστολογίων που διαφοροποιούνται από τα κείμενα μη ηλεκτρονικής επικοινωνίας.

4.3.5 Αυτόματη αναγνώριση προσωπικότητας από κείμενα των μέσων κοινωνικής δικτύωσης

Η ανάγκη για μεγάλο όγκο δεδομένων προς επεξεργασία και εξαγωγή αποτελεσμάτων στον τομέα της αυτόματης πρόβλεψης της προσωπικότητας του συγγραφέα οδήγησε τους ερευνητές στα κοινωνικά μέσα δικτύωσης. Εκεί, εκατομμύρια χρήστες αποκαλύπτουν πολλά στοιχεία για τον εαυτό τους και μέσω των αναρτήσεών τους αλλά και με τον τρόπο με τον οποίο τις δομούν. Γι’ αυτό και η πρόβλεψη της προσωπικότητας από δεδομένα των κοινωνικών δικτύων αποτελεί πεδίο που προσελκύει πολλούς ερευνητές.

Είναι προφανής, άλλωστε, η ύπαρξη στενής σχέσης ανάμεσα στην προσωπικότητα των χρηστών και στον τρόπο με τον οποίο συμπεριφέρονται στα κοινωνικά δίκτυα. Οι χρήστες των ιστοσελίδων κοινωνικής δικτύωσης μπορούν να

ακολουθήσουν κάποιον, να τον κάνουν φίλο, μπορούν επιπλέον να κοινοποιήσουν ένα μήνυμα που έλαβαν σχολιάζοντάς το ή όχι. Από τις διάφορες συμπεριφορές τέτοιου είδους είναι δυνατή η αυτόματη πρόβλεψη στοιχείων του χαρακτήρα τους. Έχει βρεθεί, για παράδειγμα, ότι η Εξωστρέφεια και η Ευσυνειδησία συνδέονται με την ευκολία στη χρήση ιστοσελίδων κοινωνικής δικτύωσης. Επιπλέον, η Εξωστρέφεια συνδέεται με το μέγεθος του κοινωνικού δικτύου του χρήστη (Golbeck et al., 2011). Γενικότερα τα μέσα κοινωνικής δικτύωσης συνδέονται με την Εξωστρέφεια, αφού προσφέρουν τη δυνατότητα κοινωνικοποίησης και αλληλεπίδρασης (Seidman, 2013).

Η πρόβλεψη της προσωπικότητας του χρήστη των κοινωνικών δικτύων μπορεί να αξιοποιηθεί στο διαδικτυακό εμπόριο και στις εφαρμογές σε αυτό με στόχο την δημιουργία διαφημιστικών μηνυμάτων και διεπαφών προσαρμοσμένων στον χαρακτήρα και τις προτιμήσεις των χρηστών. Έτσι, προσφέρονται προϊόντα και υπηρεσίες που ανταποκρίνονται στις ανάγκες τους. Με αυτό τον τρόπο οι χρήστες έχει προκύψει από έρευνες ότι είναι πιο δεκτικοί και εμπιστεύονται περισσότερο πληροφορίες που παρουσιάζονται έτσι όπως ταιριάζει στην ιδιοσυγκρασία τους. Ωστόσο, η ταχύτατη κειμενική παραγωγή στο διαδίκτυο από τα μέσα κοινωνικής δικτύωσης δημιουργεί και προβληματισμούς σχετικά με τον συγγραφέα των κειμένων που αναρτώνται. Η ανωνυμία όλο και συχνότερα καλύπτει παραβιάσεις όπως τα πνευματικά δικαιώματα αλλά και ποινικά αδικήματα. Για αυτόν τον λόγο οι υφομετρικές μελέτες αυξάνονται για τον προσδιορισμό του συγγραφέα και της προσωπικότητάς του.

Στις ενότητες που ακολουθούν γίνεται επισκόπηση των ερευνών που έχουν πραγματοποιηθεί για την πρόβλεψη του συγγραφικού προφίλ των χρηστών των μέσων κοινωνικής δικτύωσης και συγκεκριμένα από προσωπικά ιστολόγια, βίντεο-ιστολόγια, Facebook και Twitter. Πρέπει να επισημανθεί ότι αυτά τα μέσα παρουσιάζουν κάποια χαρακτηριστικά που επηρεάζουν τα αποτελέσματα της έρευνας. Υπάρχουν, δηλαδή, διαφοροποιήσεις ανά χρήστη στην έκταση και τον αριθμό των κειμενικών δεδομένων, στην σύνταξη, στις συντομογραφίες, στο θέμα, στα ορθογραφικά ή γραμματικά λάθη (Chin & Wright, 2014). Ακόμα και το κάθε είδος από τα παραπάνω μέσα παρουσιάζει ιδιαίτερα χαρακτηριστικά που πρέπει να λαμβάνονται υπόψη κατά την δημιουργία των συστημάτων πρόβλεψης.

4.3.5.1 Αυτόματη αναγνώριση προσωπικότητας από ιστολόγια

Η ελευθερία που προσφέρει το διαδίκτυο στην έκφραση έχει ως συνέπεια τα προσωπικά ιστολόγια (weblogs/blogs), ιστοσελίδες δηλαδή με καταχωρήσεις από την πιο πρόσφατη στην παλαιότερη που συνήθως μοιάζουν με προσωπικά ημερολόγια και εκφράζουν τη γνώμη και τα ενδιαφέροντα του συγγραφέα, να παρουσιάζουν μεγάλη διαφοροποίηση και παρόλο που πλέον αποτελούν ένα νέο κειμενικό είδος να ποικίλουν τόσο πολύ, σε σημείο που έχουν γίνει αντικείμενο έρευνας με ιδιαίτερο γλωσσολογικό ενδιαφέρον.

Το συγκεκριμένο κειμενικό είδος παρουσιάζει ομοιότητες τόσο με τον προφορικό όσο και με τον γραπτό λόγο. Είναι δυνατόν να έχει χαρακτηριστικά και μονολόγου και διαλόγου, δεδομένου ότι μπορεί να υπάρξουν απαντήσεις στις αναρτήσεις του χρήστη. Σημαντική επίδραση στη γλώσσα των ιστολογίων έχει η ηλικία των χρηστών, καθώς οι μισοί περίπου είναι μεταξύ 18 και 34.[7]

Οι έρευνες που ακολουθούν έχουν ως στόχο την εξέταση προσωπικών ιστολογίων για την αυτόματη αναγνώριση της προσωπικότητας του συγγραφέα τους μέσω της μελέτης της γλώσσας που χρησιμοποιείται. Ένας άλλος στόχος είναι να ελεγχθεί η ισχύς των γλωσσικών χαρακτηριστικών που έχουν εξαχθεί, δηλαδή κατά πόσο τα συγκεκριμένα χαρακτηριστικά μπορούν να συμβάλλουν αποτελεσματικά στην πρόβλεψη της προσωπικότητας του συγγραφέα.

Οι Nowson & Oberlander (2006) μελέτησαν 1.854 αναρτήσεις από τα 71 προσωπικά ιστολόγια των συμμετεχόντων των οποίων γνώριζαν τον τύπο προσωπικότητας μέσω ερωτηματολογίου (Nowson, 2006). Για κάθε τύπο του μοντέλου των Πέντε Παραγόντων δημιούργησαν γλωσσολογικά χαρακτηριστικά από το LIWC, το MRC, F-measure των μερών του λόγου και κυρίως διγράμματα και τριγράμματα λέξεων (28 από τα 34 χαρακτηριστικά). Με τη μέθοδο της παλινδρόμησης κατέληξαν σε συνδυασμό χαρακτηριστικών. Όσον αφορά τα αποτελέσματα, παραθέτουμε την περίπτωση του Νευρωτισμού κατά την οποία υπολογίζοντας τη σχετική συχνότητα δέκα χαρακτηριστικών σε ένα κείμενο πέτυχαν R^2 67%.

Οι ίδιοι ερευνητές (Oberlander & Nowson, 2006) παρουσίασαν τα αποτελέσματα της εργασίας τους στην αυτόματη κατηγοριοποίηση της προσωπικότητας του συγγραφέα χρησιμοποιώντας το ίδιο σώμα κειμένων από

[7] The Social Media Report: Q3 2011, MN Incite, Nielsen

ιστολόγια. Η κατηγοριοποίηση περιλάμβανε μόνο τους τέσσερις από τους πέντε τύπους, αποκλείστηκε δηλαδή η Δεκτικότητα στην εμπειρία. Έγιναν δοκιμές δίτιμης και πολλαπλής ταξινόμησης με τη χρήση διγραμμάτων και τριγραμμάτων λέξεων και με ταξινομητές βασισμένους στο θεώρημα Bayes και σε Μηχανές Διανυσμάτων Υποστήριξης με αρκετά καλά αποτελέσματα (ακρίβεια πρόβλεψης μεταξύ 75% και 84%) παρά τον μικρό αριθμό των χαρακτηριστικών που χρησιμοποιήθηκαν.

Με νέα τους ανακοίνωση οι Nowson & Oberlander (2007) συμπλήρωσαν τα αποτελέσματα της προηγούμενης έρευνας. Χρησιμοποίησαν δύο σώματα κειμένων, ένα που είχε ήδη χρησιμοποιηθεί και στις δύο παραπάνω έρευνες και ένα νέο, το οποίο περιλάμβανε κείμενα από ιστολόγια 1.672 ατόμων. Χρησιμοποιήθηκαν και πάλι διγράμματα και τριγράμματα λέξεων αλλά με τις απαραίτητες γενικεύσεις. Ο ταξινομητής βασίστηκε στο θεώρημα Bayes και ομαδοποίησε τους συγγραφείς των ιστολογίων με ακρίβεια αρκετά χαμηλότερη από την προηγούμενη έρευνα (περίπου 55%). Σε μικρά δηλαδή σώματα κειμένων τα αποτελέσματα είναι σαφώς καλύτερα.

Οι Gill, Nowson & Oberlander (2009) μελέτησαν τις αναρτήσεις 2.393 συμμετεχόντων των οποίων μέσω ερωτηματολογίων γνώριζαν την προσωπικότητα. Αφού έγινε η κατάλληλη επεξεργασία, το σώμα κειμένων αναλύθηκε με το λογισμικό LIWC και έγινε εξαγωγή χαρακτηριστικών για να διαπιστωθεί το περιεχόμενο των ιστολογίων, η προσωπικότητα του συγγραφέα και το κίνητρό του να γράψει. Επιπλέον, χρησιμοποιήθηκαν τα ακόλουθα χαρακτηριστικά: πρώτο ενικό πρόσωπο αντωνυμιών, τρίτο πρόσωπο αντωνυμιών, λέξεις θετικών και αρνητικών συναισθημάτων, παρελθοντικοί, παροντικοί και μελλοντικοί χρόνοι ρημάτων. Κάθε χαρακτηριστικό του μοντέλου των Πέντε Παραγόντων εξετάστηκε ξεχωριστά. Φάνηκε ότι οι Νευρωτικοί συντάκτες των ιστολογίων γράφουν για αυτοθεραπευτικούς λόγους. Χωρίς να το αναμένουν οι ερευνητές είδαν ότι οι Εξωστρεφείς συμμετέχοντες χρησιμοποιούν συχνά εκφράσεις και θετικά και αρνητικά φορτισμένες. Γενικά, αν και στα ιστολόγια υπάρχει ελευθερία έκφρασης, οι συγγραφείς γράφουν με τρόπο που θα έγραφαν και σε άλλα κειμενικά είδη και η προσωπικότητά τους προβλέπεται από τα γλωσσικά τους χαρακτηριστικά.

Μια άλλη έρευνα (Yarkoni, 2010) θέτει το πρόβλημα της ισχύος των γλωσσικών χαρακτηριστικών και μελετά τη σχέση τους με τον χαρακτήρα των συγγραφέων των προσωπικών ιστολογίων. Τα ιστολόγια που μελετήθηκαν ήταν 694 με περίπου 115.423 λέξεις το καθένα. Με την χρήση του LIWC αναλύθηκαν 66 κατηγορίες και επιπλέον μελετήθηκε η συσχέτιση με συγκεκριμένες λέξεις. Όντως

προέκυψε συσχέτιση ανάμεσα σε αυτές και στα χαρακτηριστικά του μοντέλου των Πέντε Παραγόντων αλλά και σε 30 ακόμα διαφορετικές όψεις της προσωπικότητας.

Με στόχο την σύγκριση των χαρακτηριστικών κατηγοριοποίησης των ιστολογίων βάσει προσωπικότητας αλλά και την αναγνώριση συγκεκριμένων γλωσσικών χαρακτηριστικών που συνδέονται με την προσωπικότητα, εργάστηκαν οι Iacobelli et al. (2011). Με τη χρήση Μηχανών Διανυσμάτων Υποστήριξης πέτυχαν ακρίβεια πρόβλεψης προσωπικότητας από 70,51% για τον Νευρωτισμό μέχρι 84,36% για την Δεκτικότητα στην εμπειρία σε ένα σώμα κειμένων αποτελούμενο από αναρτήσεις 3.000 περίπου συγγραφέων. Ο καλύτερος συνδυασμός χαρακτηριστικών που έφτασε στα παραπάνω αποτελέσματα ήταν τα θεματικά δίλεκτα (stemmed bigrams), συμπεριλαμβανομένων των λειτουργικών λέξεων και οι τύποι δυαδικής λογικής, για την απουσία ή παρουσία των παραπάνω χαρακτηριστικών. Αξίζει να σημειωθεί ότι η χρήση του λογισμικού LIWC δε συνέβαλε στην αποτελεσματικότητα του μοντέλου.

Η έρευνα των Li & Chignell (2010) που αφορά σε ιστολόγια εξετάζει κατά πόσο η συμβατότητα των προσωπικοτήτων των συγγραφέων και των αναγνωστών επηρεάζει την διαδικτυακή αλληλεπίδραση. Πραγματοποιήθηκαν δύο πειράματα: στο πρώτο οκτώ άτομα έγραψαν ιστολόγια μορφής ημερολογίου και σχολιασμού, στο δεύτερο δώδεκα διαφορετικά άτομα έκριναν την προσωπικότητα των συγγραφέων του πρώτου πειράματος μόνο από τα κείμενά τους. Και τα είκοσι άτομα, από το πανεπιστήμιο του Τορόντο, συμπλήρωσαν ερωτηματολόγια προσωπικότητας. Προέκυψε ότι οι συγγραφείς των ιστολογίων προσέλκυσαν περισσότερο με τα κείμενα τους, τα οποία αναλύθηκαν με το LIWC, αναγνώστες που είχαν παρόμοια χαρακτηριστικά προσωπικότητας.

Μια άλλη έρευνα (Qiu et al., 2016) εξετάζει τη σχέση ανάμεσα στην προσωπικότητα και την χρήση της Κινεζικής γλώσσας στα microblogs. Σο πρώτο πείραμα έλαβαν μέρος 470 άτομα από την Κίνα και στο δεύτερο 90 φοιτητές του πανεπιστημίου της Σιγκαπούρης. Συμπλήρωσαν ερωτηματολόγιο προσωπικότητας και έδωσαν την άδεια για τη χρήση των στοιχείων τους από την εφαρμογή στο διαδίκτυο. Οι ερευνητές χρησιμοποίησαν την κινεζική έκδοση του LIWC για να βρουν τη συσχέτιση των λεξικών συχνοτήτων με τα χαρακτηριστικά της προσωπικότητας του μοντέλου των Πέντε Παραγόντων και να κάνουν την σύγκριση με τα αποτελέσματα στην Αγγλική γλώσσα. Κατέληξαν στο συμπέρασμα ότι όσον αφορά τις γλωσσικές

εκφράσεις προσωπικότητας υπάρχουν και καθολικά αλλά και χαρακτηριστικά που συνδέονται με τη συγκεκριμένη γλώσσα.

Επίσης κείμενα της Κινεζικής γλώσσας εξετάζονται ως προς τα χαρακτηριστικά του μοντέλου των Πέντε Παραγόντων (Xue et al., 2017). Πρόκειται για αναρτήσεις σε κινεζικό microblog 994 χρηστών στις οποίες εφαρμόστηκε μια νέα τεχνική μηχανικής μάθησης, εκμάθηση με κατανομή ετικετών (label distribution learning, LDL) για πρώτη φορά στον τομέα της αναγνώρισης προσωπικότητας. Έγινε χρήση 113 χαρακτηριστικών περιεχομένου και 8 LDL αλγορίθμων για την εκπαίδευση του μοντέλου, δύο εκ των οποίων με αποτελέσματα καλύτερα από αυτά των παραδοσιακών τεχνικών παλινδρόμησης.

4.3.5.2 Αυτόματη αναγνώριση προσωπικότητας από βίντεο-ιστολόγια

Ο τομέας της αυτόματης αναγνώρισης της προσωπικότητας κάποιου ατόμου από το βίντεο-ιστολόγιό του (video blog/vlog), στο οποίο ο ομιλητής είναι ένας και το βίντεό του βρίσκεται στο διαδίκτυο, παρουσιάζει μια διαφοροποίηση σε σχέση με τις άλλες κατηγορίες, καθώς δεν αναγνωρίζεται ο πραγματικός του χαρακτήρας αλλά αυτός που εκλαμβάνεται από τους θεατές του βίντεο. Στις έρευνες που ήδη αναφερθήκαμε οι μελετητές γνώριζαν τα πραγματικά στοιχεία της προσωπικότητας των συμμετεχόντων, αφού τους ζητούσαν να συμπληρώσουν ένα ερωτηματολόγιο προσωπικότητας. Στην περίπτωση, όμως, του βίντεο-ιστολογίου που είναι ήδη αναρτημένο στο διαδίκτυο κάτι τέτοιο δεν είναι εφικτό.

Παρόλο που πρόκειται για ένα βίντεο που δεν περιλαμβάνει αλληλεπίδραση, οι δημιουργοί αυτών των ιστολογίων συμπεριφέρονται σαν να συνομιλούν με τον θεατή μέσω της κάμερας. Έτσι, υπάρχουν πολλά παραγλωσσικά και εξωγλωσσικά στοιχεία. Η πρόβλεψη, λοιπόν, της προσωπικότητας από ένα βίντεο-ιστολόγιο με δεδομένες τις παραπάνω ιδιαιτερότητες απαιτεί τόσο ανάλυση του μεταγεγραμμένου από το βίντεο κειμένου όσο και μη λεκτικά στοιχεία, όπως οπτικοακουστικά χαρακτηριστικά. Η πιο χαρακτηριστική έρευνα που μελετά την εντύπωση που δίνει η προσωπικότητα του ατόμου που μιλάει στο βίντεο μέσω μη λεκτικών στοιχείων έγινε από τους Biel &

Gatica-Perez (2013) και παρουσίασε σχετικά καλά αποτελέσματα μόνο για το χαρακτηριστικό της Εξωστρέφειας (R^2 = 36%).

Στο τέλος της ίδιας χρονιάς οι Biel et al. (2013) ανακοίνωσαν τα αποτελέσματα της έρευνας τους στον ίδιο τομέα, αλλά αυτή τη φορά αξιοποιώντας τα λεκτικά στοιχεία από τα βίντεο. Τα δεδομένα τους αποτελούνται από 442 βίντεο-ιστολόγια του YouTube και ένα σύνολο από εντυπώσεις για την προσωπικότητα των ατόμων που μιλούν στα βίντεο που δημιουργήθηκαν από επισημειωτές μέσω του Amazon's Mechanical Turk.

Όσον αφορά στη μεθοδολογία, έγινε χειρωνακτική μεταγραφή του ήχου σε κείμενο. Επιπλέον έγινε και αυτόματη μεταγραφή με τεχνική αυτόματης αναγνώρισης φωνής και στοίχιση με το κείμενο που δημιουργήθηκε χειρωνακτικά. Ακολούθησε η αυτόματη ανάλυση του κειμένου με δύο μεθόδους, με το λογισμικό LIWC και με ν-γράμματα. Με Μηχανές Διανυσμάτων Υποστήριξης με γραμμική, πολυωνυμική και RBF συνάρτηση και με Random Forests (RFs) αλγόριθμους ελέγχθηκε η χρήση των μηχανισμών πρόβλεψης.

Από τα αποτελέσματα που προέκυψαν από το σώμα κειμένων που μεταγράφηκε χειρωνακτικά αξίζει να σημειωθεί ότι το χαρακτηριστικό που παρουσιάζει την υψηλότερη απόδοση (R^2 = 31%) είναι η Προσήνεια. Η χρήση του αυτόματα μεταγεγραμμένου υλικού έδειξε ότι η επιτυχία της πρόβλεψης μειώνεται λόγω των λαθών του συστήματος αναγνώρισης φωνής. Για παράδειγμα η απόδοση της Προσήνειας έφτασε μόνο το 10%. Τέλος, οι ερευνητές σύγκριναν τα παραπάνω αποτελέσματα με αυτά του συνδυασμού ανάλυσης λεκτικών και μη λεκτικών στοιχείων και είδαν ότι η πρόβλεψη λειτουργεί καλύτερα.

Στο Θεματικό εργαστήριο (Workshop) στην Υπολογιστική Αναγνώριση της Προσωπικότητας του 2014 (WCPR14) υπήρχαν κάποιες συμμετοχές που αναφέρονται σε βίντεο-ιστολόγια. Η προσέγγιση των Farnadi et al. (2014b) είναι πολυμεσική, αφού χρησιμοποίησαν και οπτικοακουστικά χαρακτηριστικά αλλά και κειμενικά. Τα δεδομένα αποτελούνται από 404 βίντεο-ιστολόγια του YouTube. Οι εντυπώσεις για την προσωπικότητα των ατόμων που μιλούν στα βίντεο δημιουργήθηκαν από επισημειωτές μέσω του Amazon's Mechanical Turk[8] και του Ten-Item Personality Inventory[9]. Το ακουστικό αρχείο μεταγράφηκε και τα οπτικοακουστικά χαρακτηριστικά (συνολικά 25) εξήχθησαν αυτόματα. Επιπλέον, έγινε εξαγωγή χαρακτηριστικών από διάφορες

[8] https://www.mturk.com/
[9] https://gosling.psy.utexas.edu/scales-weve-developed/ten-item-personality-measure-tipi/

πηγές: το λογισμικό LIWC, το λεξικό συναισθημάτων NRC, την ψυχογλωσσολογική βάση MRC, τον αυτόματο αναλυτή συναισθήματος SentiStrength και το γλωσσολογικό εργαλείο ανάλυσης ενδείξεων SPLICE. Χρησιμοποιήθηκαν 6 αλγόριθμοι πολυπαραγοντικής παλινδρόμησης για να προβλέψουν συνολικά τα πέντε χαρακτηριστικά του μοντέλου των Πέντε Παραγόντων. Η Προσήνεια παρουσίασε την μεγαλύτερη απόδοση (R^2 = 37%) και ακολουθεί η Εξωστρέφεια.

Άλλη μια συμμετοχή στο ίδιο Θεματικό εργαστήριο είναι των Verhoeven, Soler & Daelemans (2014). Οι ερευνητές επιδίωξαν να ελέγξουν εάν τα χαρακτηριστικά που είχαν προτείνει οι Soler & Wanner (2014) για την πρόβλεψη του φύλου του συγγραφέα θα μπορούσαν να έχουν επιτυχή αποτελέσματα και στην πρόβλεψη της προσωπικότητας αποκλειστικά από κείμενο. Τα δεδομένα ήταν τα ίδια για όλους τους συμμετέχοντες, όπως τα παρουσιάσαμε παραπάνω. Όλα τους τα πειράματα έγιναν με Μηχανές Διανυσμάτων Υποστήριξης (αλγόριθμος Scikit-learn's). Χρησιμοποιήθηκαν μονογράμματα λεξικών μονάδων, τριγράμματα χαρακτήρων, χαρακτηριστικά από το LIWC και συνδυασμός των παραπάνω. Αυτά τα χαρακτηριστικά συγκρίθηκαν με τα χαρακτηριστικά των Soler & Wanner (2014) που ανήκουν σε πέντε κατηγορίες: χαρακτήρες (π.χ. συχνότητα σημείων στίξης), λέξεις (π.χ. λεξιλογικός πλούτος), προτάσεις (πλήθος προτάσεων και λέξεων ανά πρόταση), λεξικά (π.χ. θετικά ή αρνητικά συναισθηματικά φορτισμένες λέξεις) και συντακτικά (dependency parser). Τελικά, παρόλο που για κάποια από τα χαρακτηριστικά του μοντέλου ο αλγόριθμος δουλεύει καλά, φαίνεται ότι χρειάζεται περαιτέρω έρευνα, καθώς το σώμα κειμένων ήταν περιορισμένο. Με τη χρήση του φύλου, του λογισμικού LIWC και λεξικά χαρακτηριστικά πέτυχαν f1-score 0.54, κατά μέσο όρο στα πέντε χαρακτηριστικά του μοντέλου.

Οι Alam & Riccardi (2014) στο WCPR14 μελέτησαν γλωσσολογικά, ψυχογλωσσολογικά, συναισθηματικά χαρακτηριστικά εκτός από τα οπτικοακουστικά για να προσεγγίσουν την αυτόματη αναγνώριση της προσωπικότητας των δημιουργών των 404 βίντεο-ιστολογίων. Ενδεικτικά, στα οπτικοακουστικά χαρακτηριστικά εντάσσονται ο χρόνος ομιλίας, ο επιτονισμός, η εγγύτητα στην κάμερα, η κίνηση του σώματος. Όσον αφορά τα λεξικά χαρακτηριστικά, χρησιμοποιήθηκαν τριγράμματα λεξικών μονάδων. Επιπλέον, με τον Μορφολογικό Αναλυτή Stanford έγινε εξαγωγή των μερών του λόγου. Για τα ψυχογλωσσολογικά χαρακτηριστικά χρησιμοποιήθηκε το λογισμικό LIWC και για τα συναισθηματικά χαρακτηριστικά επισημείωσαν λίστες λέξεων με κατηγορίες που εκφράζουν συναίσθημα και υπολόγισαν τις συχνότητές

τους. Τέλος, έλεγξαν εάν είναι αποτελεσματικότερο να προβλέψουν ένα χαρακτηριστικό του μοντέλου των Πέντε Παραγόντων χρησιμοποιώντας τα υπόλοιπα ως χαρακτηριστικά. Κατά μέσο όρο έφτασαν στο 67,3% επιτυχίας με διάφορους συνδυασμούς χαρακτηριστικών.

Με τα ίδια δεδομένα δούλεψαν και οι Sarkar et al. (2014) στο πλαίσιο του WCPR14. Πειραματίστηκαν με πέντε κατηγορίες χαρακτηριστικών: οπτικοακουστικά, κειμενικά (unigram Bag of Words model), στατιστικά λέξεων (πλήθος λέξεων ανά κείμενο, μέσος όρος προτάσεων ανά κείμενο και μέσος όρος λέξεων ανά πρόταση), συναισθηματικά (με το λογισμικό ανάλυσης συναισθήματος SentiStrength κατασκεύασαν πέντε χαρακτηριστικά) και ένα δημογραφικό χαρακτηριστικό, το φύλο του ατόμου που μιλούσε στο βίντεο. Συνολικά, η ομάδα δημιούργησε 1.079 χαρακτηριστικά και χρησιμοποίησε τη λογιστική παλινδρόμηση. Από την παρουσίαση των αποτελεσμάτων αξίζει να σημειωθεί ότι τα οπτικοακουστικά χαρακτηριστικά επηρεάζουν αρνητικά την πρόβλεψη του χαρακτηριστικού της Δεκτικότητας στην εμπειρία, ενώ τα κειμενικά χαρακτηριστικά την ενισχύουν και ο συνδυασμός τους έχει καλά αποτελέσματα στην πρόβλεψη της Προσήνειας.

4.3.5.3 Αυτόματη αναγνώριση προσωπικότητας από το Facebook

Στο υποκεφάλαιο αυτό θα δούμε πώς μπορούν να προβλεφθούν τα χαρακτηριστικά της προσωπικότητας των χρηστών του Facebook, ιστοσελίδας κοινωνικής δικτύωσης με τα περισσότερα μέλη από όλο τον κόσμο (2,7 δισεκατομμύρια ενεργούς χρήστες ανά μήνα)[10], μέσω των πληροφοριών που δημοσιεύουν. Ιδιαίτερα κατατοπιστική είναι η εκτεταμένη ψυχολογική έρευνα των Kosinski et al. (2013) για τα χαρακτηριστικά που επιλέγονται με στόχο την αυτόματη πρόβλεψη, ανάμεσα στα οποία και το πλήθος των επαφών, των φωτογραφιών και των γεγονότων που παρακολουθεί ο χρήστης του Facebook. Η Εξωστρέφεια φάνηκε ότι είναι το χαρακτηριστικό που εκφράζεται περισσότερο από το Facebook.

Σε μια από τις πρώτες έρευνες (Golbeck, Robles & Turner, 2011) που αφορούν στο θέμα συγκεντρώθηκαν μέσω εφαρμογής στοιχεία από 167 χρήστες του Facebook (λίστες φίλων, προσωπικά στοιχεία, προτιμήσεις και δραστηριότητες, στατιστικά και

[10] https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/

γλωσσικά στοιχεία) και οι απαντήσεις ερωτηματολογίου βασισμένου στο μοντέλο των Πέντε Παραγόντων. Αφού έγινε η συσχέτιση των χαρακτηριστικών με την προσωπικότητα, κατέληξαν σε 74 χαρακτηριστικά ανά χρήστη και χρησιμοποίησαν αλγόριθμους παλινδρόμησης. Τελικά κατόρθωσαν να προβλέψουν το κάθε χαρακτηριστικό του μοντέλου των Πέντε Παραγόντων με απόκλιση 11% από την πραγματική τιμή. Ενδιαφέρουσες είναι οι συσχετίσεις του χαρακτηριστικού της Ευσυνειδησίας, που σχετίζεται περισσότερο με τα γλωσσικά χαρακτηριστικά και ιδιαίτερα με κοινωνικές διεργασίες. Η συχνότητα των υβριστικών λέξεων και οι λέξεις αντίληψης-αίσθησης σχετίζονται αρνητικά με την Ευσυνειδησία. Θετικά σχετίζεται με λέξεις κοινωνικότητας όπως και με λέξεις που περιγράφουν ανθρώπους.

Ο Celli (2012) επεξεργάστηκε 5.200 αναρτήσεις 1.100 Ιταλών χρηστών του Facebook με το σύστημα αναγνώρισης προσωπικότητας που ανέπτυξε και πέτυχε κατά μέσο όρο συνολικά για τα πέντε χαρακτηριστικά f= .628. Προέκυψε από την έρευνα ότι οι χρήστες που είναι Δεκτικοί στην εμπειρία έχουν τον υψηλότερο αριθμό αλληλεπιδράσεων και είναι αυτοί που επηρεάζουν τις συζητήσεις.

Στην εργασία των Farnadi et al. (2014a) εξετάστηκε η σχέση ανάμεσα στο συναίσθημα των χρηστών του Facebook και στην ηλικία, το φύλο και την προσωπικότητά τους. Για την εξαγωγή συναισθημάτων από τις δημοσιεύσεις χρησιμοποιήθηκε το NRC λεξικό συναισθημάτων. Αξίζει να σημειωθεί ότι όσοι χρήστες είχαν υψηλό ποσοστό στο χαρακτηριστικό της Δεκτικότητας στην εμπειρία προέκυψε ότι εκφράζουν συναισθήματα συχνότερα.

Οι μελέτες που ακολουθούν πραγματοποιήθηκαν στο πλαίσιο του Θεματικού εργαστηρίου στην Υπολογιστική Αναγνώριση Προσωπικότητας 2013 (Celli et al., 2013). Τα δεδομένα, κοινά για όλους τους συμμετέχοντες αποτελούνταν από α) κείμενα-εκθέσεις περίπου 2.400 φοιτητών (essays) και β) δεδομένα από το Facebook 250 χρηστών (myPersonality).

Οι Verhoeven, Daelemans & De Smedt (2013) πρότειναν για την αναγνώριση της προσωπικότητας την εφαρμογή της μετα-μάθησης, μιας άλλης προσέγγισης της Μηχανικής Μάθησης, η οποία αυξάνει την ακρίβεια των συστημάτων. Εργάστηκαν με δεδομένα από το Facebook, αλλά οι ταξινομητές εκπαιδεύτηκαν στο άλλο είδος, στα κειμενικά δεδομένα. Χρησιμοποίησαν Μηχανές Διανυσμάτων Υποστήριξης με τα 2.000 συχνότερα τριγράμματα χαρακτήρων. Πραγματοποίησαν πέντε δοκιμές μετα-μάθησης, μία για κάθε χαρακτηριστικό του μοντέλου των Πέντε Παραγόντων. Σε κάθε δοκιμή έγινε 10-πτυχη διασταυρούμενη επικύρωση (ten-fold cross-validation). Το

καλύτερο αποτέλεσμα της τιμής F ήταν 0,86 για την πρόβλεψη της Δεκτικότητας στην εμπειρία.

Συγκεκριμένα για την πρόβλεψη του χαρακτηριστικού της Ευσυνειδησίας έγινε μια μελέτη (Tomlinson, Hinote & Bracewell, 2013) με δεδομένα από το Facebook. Η έρευνά τους βασίστηκε σε σημασιολογική ανάλυση και μάλιστα ρημάτων που χρησιμοποιούσαν τα άτομα στις περιγραφές τους και θεματικών ρόλων που τα άτομα χρησιμοποιούσαν για να εκφράσουν δική τους συμπεριφορά ή άλλων. Τα εργαλεία που αξιοποιήθηκαν ήταν το WordNet και το Senti-WordNet. Η ακρίβεια πρόβλεψης της ευσυνειδησίας ενός ατόμου που βρίσκεται πάνω ή κάτω από τον μέσο όρο ήταν 58,13% με 10-πτυχη διασταυρούμενη επικύρωση.

Στην εργασία των Alam, Stepanov & Riccardi (2013) χρησιμοποιήθηκαν τρεις διαφορετικές μέθοδοι ταξινόμησης, Sequential Minimal Optimization for Support Vector Machine, Bayesian Logistic Regression, Multinomial Naive Bayes sparse model, από τις οποίες η τελευταία είχε τα καλύτερα αποτελέσματα (μέσος όρος ακρίβειας 61,79%) με τη χρήση μονογραμμάτων ως χαρακτηριστικών. Η μέθοδος εξαγωγής χαρακτηριστικών που αξιοποιήθηκε ήταν η Bag of Words, δηλαδή η αναπαράσταση των κειμένων ως σειρών και των λέξεων ως στηλών στο πίνακα των δεδομένων. Με αυτό το μοντέλο κειμενικής αναπαράστασης ελέγχεται η διασπορά ή η συχνότητα λέξεων στο κείμενο.

Οι Farnadi et al. (2013) στο ίδιο Θεματικό εργαστήριο αξιοποίησαν τέσσερις κατηγορίες χαρακτηριστικών. Από το λεξικό LIWC 81 χαρακτηριστικά, 7 με στοιχεία κοινωνικής δικτύωσης, 6 σχετικά με τον χρόνο και 6 με ποικίλο περιεχόμενο. Εξέτασαν τρεις διαφορετικούς αλγόριθμους: Μηχανές Διανυσμάτων Υποστήριξης, Nearest Neighbor and Naive Bayes.

Οι Markovikj et al. (2013) προέβλεψαν με επιτυχία τα πέντε στοιχεία του μοντέλου των Πέντε Παραγόντων αξιοποιώντας 725 χαρακτηριστικά χωρισμένα σε πέντε ομάδες: κοινωνικά και δημογραφικά στοιχεία, γλωσσολογικά, μέρη του λόγου, τιμές για λέξεις που δηλώνουν συναίσθημα, κλίμακα έντασης των λέξεων. Για την επικύρωση των μοντέλων χρησιμοποίησαν Μηχανές Διανυσμάτων Υποστήριξης, Simple Minimal Optimization (SMO) and Boost algorithms. Η έρευνά τους τόνισε τη σημασία της επιλογής των κατάλληλων χαρακτηριστικών για την πρόβλεψη της προσωπικότητας.

Και στην Κινεζική γλώσσα έχει μελετηθεί ο τομέας της πρόβλεψης προσωπικότητας (Peng et al., 2015) και συγκεκριμένα σε κείμενα 222 χρηστών του

Facebook. Για την επίλυση του προβλήματος της ταξινόμησης επιλέχθηκαν οι Μηχανές Διανυσμάτων Υποστήριξης. Φάνηκε ότι οι εξωστρεφείς έγραφαν περισσότερες προτάσεις και χρησιμοποιούσαν πιο κοινές λέξεις σε σχέση με τους εσωστρεφείς, αποτέλεσμα το οποίο αποκαλύπτει την επιθυμία τους να μοιράζονται τη ζωή τους με άλλους.

Με εφαρμογή σε κείμενα από το Facebook μελετητές από πανεπιστήμιο της Ινδονησίας (Tandera et al., 2017) πέτυχαν την πρόβλεψη των χαρακτηριστικών του μοντέλου των Πέντε Παραγόντων με μέσο όρο ακρίβειας 74,17%. Στις παραδοσιακές τεχνικές μηχανικής μάθησης χρησιμοποιήθηκαν γλωσσικά χαρακτηριστικά με προσέγγιση κλειστού λεξιλογίου (closed-vocabulary approach) (LIWC, SPLICE), ενώ στις τεχνικές μάθησης σε βάθος χαρακτηριστικά ανοικτού λεξιλογίου (open vocabulary approach).

Η προσέγγιση της βαθιάς μάθησης (Deep Learning) ακολουθήθηκε σε έρευνα (Yu & Markov, 2017) εξετάζοντας διάφορες αρχιτεκτονικές νευρωνικών δικτύων όπως fully-connected (FC) networks, convolutional networks (CNN) και recurrent networks (RNN). Τα κειμενικά δεδομένα αντλήθηκαν από αναρτήσεις 250 χρηστών του Facebook, σώμα κειμένων (myPersonality) που χρησιμοποιήθηκε και σε προηγούμενες έρευνες όπως αναφέρθηκε παραπάνω (Celli et al., 2013). Πιο αποτελεσματικά αποδείχθηκαν τα CNN και μάλιστα με μεγαλύτερο ποσοστό ακόμα και από κάποιους συμμετέχοντες στο Θεματικό εργαστήριο στην Υπολογιστική Αναγνώριση Προσωπικότητας 2013 που χρησιμοποίησαν μηχανική μάθηση.

Το ίδιο σώμα κειμένων με την παραπάνω έρευνα, το myPersonality που περιλαμβάνει αναρτήσεις στο Facebook, επεξεργάστηκαν και οι Tadesse et al. (2018). Εξήχθησαν δύο ομάδες χαρακτηριστικών· τα κειμενικά, 85 ψυχογλωσσολογικά χαρακτηριστικά με τη χρήση του λογισμικού LIWC και 74 με το SPLICE και η συμπεριφορά κοινωνικής αλληλεπίδρασης των χρηστών στο Facebook. Όσον αφορά τα κειμενικά χαρακτηριστικά, ο αλγόριθμος XGBoost πρόβλεψε τη Δεκτικότητα σε νέες εμπειρίες του Big Five με το μεγαλύτερο ποσοστό ακρίβειας (73,3%).

Για την Πορτογαλική γλώσσα πραγματοποιήθηκε έρευνα με δεδομένα από το Facebook της Βραζιλίας με τη χρήση τεχνικών μηχανικής μάθησης (Ramos dos Santos & Paraboni, 2018). Οι πέντε κατηγορίες του μοντέλου των Πέντε Παραγόντων εξετάστηκαν ως δυάδες. Επισημαίνεται το πρόβλημα των περιορισμένων δεδομένων που αντιστοιχούν σε κάποιους ψυχολογικούς τύπους.

Πρόσφατα, πραγματοποιήθηκε έρευνα (Prajwal et al., 2020), κατά την οποία κείμενα εκατό φοιτητών από τις δημοσιεύσεις τους στον προσωπικό τους λογαριασμό στο Facebook υποβλήθηκαν σε επεξεργασία με τεχνικές μηχανικής μάθησης στοχεύοντας στην αναγνώριση της προσωπικότητάς τους. Ως προς τα αποτελέσματα, έγινε σύγκρισή τους με τις τιμές που προέκυψαν από τα ερωτηματολόγια του μοντέλου των Πέντε Παραγόντων και η ποσοστιαία απόκλισή τους κυμάνθηκε μεταξύ 8,71% και 17,7%, που, σύμφωνα με τους ερευνητές, είναι ικανοποιητική.

Βιβλιογραφικά, το θέμα ερευνητικό πεδίο της πρόβλεψης χαρακτηριστικών της προσωπικότητας με δεδομένα από το Facebook προσεγγίζεται από τους Marengo & Settanni (2019), οι οποίοι συμπεραίνουν ότι η ακρίβεια της πρόβλεψης είναι παρόμοια με την ακρίβεια που επιτυγχάνεται από δεδομένα αντλημένα από άλλα μέσα κοινωνικής δικτύωσης. Συνολικά, στο συγκεκριμένο ερευνητικό πεδίο το Facebook είναι το μέσο κοινωνικής δικτύωσης του οποίου το περιεχόμενο έχει μελετηθεί περισσότερο από κάθε άλλο και μάλιστα με πολύ μεγαλύτερη συχνότητα (Bowden-Green, Hinds & Joinson, 2020).

4.3.5.4 Αυτόματη αναγνώριση προσωπικότητας από το Twitter

Από το 2006, οπότε το Twitter άρχισε να χρησιμοποιείται, έγινε παγκοσμίως ένας από τους πιο δημοφιλείς ιστοχώρους κοινωνικής δικτύωσης, που επιτρέπει στους χρήστες του να στέλνουν και να διαβάζουν μικρά και σύντομα μηνύματα (tweets). Υπολογίζεται ότι κατά μέσο όρο κάθε δευτερόλεπτο στέλνονται περίπου 6.000 tweets[11]. Το Twitter αποτελεί το μεγαλύτερο κοινωνικό δίκτυο στον τομέα της ενημέρωσης και βρίσκεται στη δεύτερη θέση παγκοσμίως με κριτήριο την επισκεψιμότητα[12]. Θα ήταν χρήσιμο σε αυτό το σημείο να τονίσουμε μια βασική διαφορά ανάμεσα στο Facebook και στο Twitter. Το πρώτο συνήθως συνδέει άτομα που γνωρίζονται μεταξύ τους, ενώ το δεύτερο είναι ένα μέσο κοινωνικής δικτύωσης στο οποίο οι χρήστες βλέπουν ό,τι έχει αναρτηθεί· σπάνια υπάρχει προστασία των δεδομένων. Έτσι, τα δεδομένα είναι περισσότερα και ελεύθερα παρόλο που υπάρχει ανώτερο όριο χαρακτήρων στις δημοσιεύσεις.

[11] https://www.internetlivestats.com/twitter-statistics/
[12] https://www.alexa.com/topsites

Μια ακόμη μελέτη που επιβεβαιώνει την αυξανόμενη χρήση των κοινωνικών δικτύων στη ζωή του σύγχρονου ανθρώπου και μάλιστα την κοινοποίηση ακόμα και πολύ προσωπικών θεμάτων πραγματοποιήθηκε από επιστήμονες της πληροφορικής και της ψυχολογίας (Preotiuc-Pietro et al., 2015). Χρησιμοποιήθηκαν δεδομένα από το Twitter ατόμων που ανέφεραν στις αναρτήσεις τους ότι είχαν διαγνωστεί με κατάθλιψη και μετατραυματική αγχώδη διαταραχή. Από την ανάλυση της γλώσσας αυτών των ατόμων ήταν δυνατό να προβλεφθεί και η προσωπικότητά τους και μάλιστα συγκεκριμένα στοιχεία της προσωπικότητας προβλέπουν νοητικές ασθένειες. Για παράδειγμα, οι καταθλιπτικοί συμμετέχοντες είχαν το μικρότερο ποσοστό στο χαρακτηριστικό της Δεκτικότητας στην εμπειρία, ενώ τα άτομα με μετατραυματική αγχώδη διαταραχή το υψηλότερο.

Οι Quercia et al. (2011) στον τίτλο του άρθρου τους προβάλλουν την ιδέα ότι το προφίλ μας στο Twitter είναι ο εαυτός μας. Όντως, με δεδομένα από 335 χρήστες του Twitter προέβλεψαν τα στοιχεία του μοντέλου των Πέντε Παραγόντων. Προτείνουν, λοιπόν, εμπορικές εφαρμογές στον χώρο του Marketing (επιλογή κατάλληλων προς τον χρήστη διαφημίσεων), στο σχεδιασμό της διεπαφής της ιστοσελίδας κοινωνικής δικτύωσης και στα συστήματα συστάσεων μουσικής (Recommender Systems).

Οι Golbeck et al. (2011) για να αναλύσουν την προσωπικότητα χρηστών του Twitter, εκπαίδευσαν δύο αλγόριθμους μηχανικής μάθησης (ZeroR and Gaussian Processes) και προέβλεψαν μεταξύ 11% και 18% της πραγματικής τιμής των χαρακτηριστικών του μοντέλου των Πέντε Παραγόντων μέσω των δεδομένων που συγκέντρωσαν από το Twitter. Η Δεκτικότητα στην εμπειρία ήταν ευκολότερα προβλέψιμο χαρακτηριστικό, ενώ δυσκολότερη ήταν η πρόβλεψη του Νευρωτισμού.

Στην έρευνα των Qiu et al. (2012) συμμετείχαν 142 χρήστες του Twitter, η πλειοψηφία των οποίων από Η.Π.Α. και Σιγκαπούρη. Συμπλήρωσαν διαδικτυακά δυο ερωτηματολόγια, ένα με τα στοιχεία τους (δημογραφικά και στοιχεία χρήσης του Twitter) και ένα βασισμένο στο μοντέλο των Πέντε Παραγόντων. Ακολούθησε η ανάκτηση των αναρτήσεων ενός μηνός, η επεξεργασία των δεδομένων και η εξαγωγή χαρακτηριστικών με τη χρήση του LIWC. Το σημαντικό είναι ότι τα χαρακτηριστικά του Νευρωτισμού και της Προσήνειας κατέστη δυνατό να προβλεφθούν μέσω των γλωσσικών δεδομένων από οκτώ παρατηρητές-κριτές. Όσον αφορά την αυτόματη πρόβλεψη, προέκυψαν νέες συσχετίσεις μεταξύ προσωπικότητας και γλωσσικών στοιχείων. Η Εξωστρέφεια βρέθηκε να έχει αρνητική συσχέτιση με λειτουργικές λέξεις

και θετική με λέξεις δηλωτικές συμφωνίας (assent words), π.χ. ναι, συμφωνώ. Η Δεκτικότητα στην εμπειρία συσχετίστηκε αρνητικά με την χρήση επιρρημάτων και θετικά με την χρήση προθέσεων.

Σε σώμα κειμένων αποτελούμενο από 200.000 αναρτήσεις του Twitter και επισημειωμένο με την υβριδική μέθοδο του Celli (2012) φάνηκε ότι οι Νευρωτικοί χρήστες κάνουν περισσότερες αναρτήσεις και δημιουργούν μεγαλύτερες αλυσίδες χρηστών που αλληλεπιδρούν σε σχέση με τους ήρεμους χρήστες.

Οι Brinks & White (2012) αξιοποίησαν τα αποτελέσματα του Ερωτηματολογίου Τύπων Προσωπικότητας Myers-Briggs Type Indicator (MBTI) που χορήγησαν σε χρήστες του Twitter για να προβλέψουν στοιχεία της προσωπικότητάς τους από 960.715 αναρτήσεις τους συνολικά. Ο Naive Bayes ταξινομητής δεν ήταν τόσο αποτελεσματικός, επειδή σύμφωνα με τους ερευνητές χρειάζονται προηγμένες τεχνικές επεξεργασίας φυσικής γλώσσας.

Στον 3ο διεθνή διαγωνισμό ιχνογράφησης χαρακτηριστικών του συγγραφέα (Author Profiling) PAN 2015 δόθηκε στους διαγωνιζόμενους σώμα κειμένων αποτελούμενο από κείμενα του Twitter στα Αγγλικά, Ισπανικά, Ιταλικά και Ολλανδικά. Είκοσι δύο ήταν οι ερευνητικές ομάδες που υπέβαλλαν συμμετοχές χρησιμοποιώντας διάφορες μεθόδους επεξεργασίας του σώματος κειμένων, επιλέγοντας διαφορετικά χαρακτηριστικά και μεθόδους ταξινόμησης. Όσον αφορά στην πρόβλεψη προσωπικότητας πρέπει να σημειωθεί ότι τα λιγότερα λάθη έγιναν στα Ολλανδικά και Ιταλικά με τιμές κάτω του 5% για τα περισσότερα χαρακτηριστικά, λόγω του μικρότερου αριθμού συγγραφέων για τις γλώσσες αυτές. Δεν αναφερόμαστε στον διαγωνισμό του 2016, διότι αφορούσε πρόβλεψη μόνο ηλικίας και φύλου και όχι προσωπικότητας. Παρακάτω θα αναφερθούμε σε κάποιες από τις καλύτερες συμμετοχές του διαγωνισμού. Για πλήρη ανάλυση των αποτελεσμάτων παραπέμπουμε στο Rangel et al. (2015).

Η νικήτρια ομάδα (Álvarez-Carmona et al., 2015) συνδύασε δυο ειδών χαρακτηριστικά, θεματικά και υφομετρικά. Πρόκειται για τεχνική λανθάνουσας σημασιολογικής ανάλυσης (Latent Semantic Analysis) που εξάγει και αναπαριστά τη σημασία των λέξεων και των κειμένων και για τεχνική που αναπαριστά το κείμενο ως διάνυσμα που σχετίζεται με κάθε προφίλ-στόχο (Second Order Attributes). Για την ταξινόμηση χρησιμοποιήθηκαν Μηχανές Διανυσμάτων Υποστήριξης και η ακρίβεια πρόβλεψης ήταν πολύ υψηλή.

Από την δεύτερη καλύτερη συμμετοχή (González-Gallardo et al., 2015) αξιοποιήθηκαν ως υφομετρικά χαρακτηριστικά για την ταξινόμηση των tweets τα ν-γράμματα χαρακτήρων και ν-γράμματα μερών του λόγου. Τα αποτελέσματά τους ήταν πολύ κοντά στην προηγούμενη ομάδα, όπως και αυτά των Grivas, Krithara & Giannakopoulos (2015), που ακολουθούν στην κατάταξη με τη χρήση του ταξινομητή Support Vector Machine Regression with a linear kernel. Η ομάδα προσέγγισε το πρόβλημα με τη χρήση υφομετρικών (όπως TF-IDF of trigrams, Bag of Words, Bag of trigrams) και δομικών χαρακτηριστικών των κειμένων (π.χ. πλήθος των hashtags).

Ιδιαίτερη αναφορά γίνεται και στην ακόλουθη έρευνα που εντάσσεται επίσης στον προαναφερθέντα διαγωνισμό λόγω της χρήσης αποκλειστικά υφομετρικών χαρακτηριστικών, εκ των οποίων τα δεκατέσσερα είναι ανεξάρτητα γλώσσας, ενώ τα υπόλοιπα δεκαπέντε χρησιμοποιούνται μόνο για τα Αγγλικά (Pervaz et al., 2015). Η ερευνητική ομάδα αντιμετώπισε το πρόβλημα της αναγνώρισης του συγγραφικού προφίλ ως θέμα μηχανικής εποπτευόμενης μάθησης. Καλύτερα αποτελέσματα επιτεύχθηκαν για την Ολλανδική γλώσσα.

Η έρευνα των Plank & Hovy (2015) στηρίζεται στο Ερωτηματολόγιο Τύπων Προσωπικότητας Myers-Briggs Type Indicator (MBTI) που χορηγήθηκε στους συμμετέχοντες και όχι σε αυτό των Πέντε Παραγόντων. Αξιοποιήθηκαν κειμενικά δεδομένα από το Twitter (1,2 εκατομμύρια αναρτήσεις στην Αγγλική γλώσσα από 1.500 χρήστες) και χρησιμοποιήθηκε ταξινομητής λογιστικής παλινδρόμησης. Κατέληξαν στο ότι τα δεδομένα παρέχουν αρκετά γλωσσολογικά στοιχεία για την πρόβλεψη των τεσσάρων διαστάσεων εσωστρέφειας – εξωστρέφειας (72,5% ακρίβεια), σκέψης – συναισθήματος (77,4% ακρίβεια), αλλά όχι και για τις άλλες δύο διαστάσεις για τις οποίες τα ποσοστά ακρίβειας ήταν 61,2% και 55,4%.

Η ερευνητική ομάδα που ανέπτυξε το πολύγλωσσο σώμα κειμένων TwiSty που αναφέρθηκε στην ενότητα για την πρόβλεψη του φύλου του συγγραφέα, μελέτησε και το θέμα της προσωπικότητας (Verhoeven, Daelemans & Plank 2016a, 2016b). Και αυτοί οι ερευνητές χορήγησαν το Ερωτηματολόγιο Τύπων Προσωπικότητας Myers-Briggs Type Indicator (MBTI) έναντι του μοντέλου των Πέντε Παραγόντων που επικρατεί στη βιβλιογραφία. Χρησιμοποίησαν τα ακόλουθα χαρακτηριστικά και για τις έξι γλώσσες (Ολλανδικά, Γερμανικά, Γαλλικά, Ιταλικά, Πορτογαλικά και Ισπανικά): μονογράμματα και διγράμματα λέξεων καθώς και τριγράμματα και τετραγράμματα χαρακτήρων. Το μοντέλο επιβεβαίωσε προηγούμενα αποτελέσματα ερευνών δείχνοντας ότι η πρόβλεψη εσωστρέφειας – εξωστρέφειας και σκέψης –

συναισθήματος επιτυγχάνεται, ενώ δεν ισχύει το ίδιο για τη νόηση - διαίσθηση και κρίση-αντίληψη μόνο με τη χρήση γλωσσολογικών χαρακτηριστικών.

Στην Ινδονησία επιχειρήθηκε με κείμενα από το Twitter η αυτόματη ανάκτηση στοιχείων της προσωπικότητας 97 χρηστών (Lukito et al., 2016), οι οποίοι είχαν απαντήσει διαδικτυακά στις ερωτήσεις του μεταφρασμένου Ερωτηματολογίου Τύπων Προσωπικότητας Myers-Briggs Type Indicator (MBTI). Οι τρεις διαφορετικές προσεγγίσεις που δοκιμάστηκαν ήταν με αλγόριθμο μηχανικής μάθησης Naive Bayes, λεξιλογική προσέγγιση και γραμματικοί κανόνες με μορφολογική επισημείωση. Κατά μέσο όρο ακρίβειας ο Naive Bayes ταξινομητής απέδωσε αποτελεσματικότερα με ποσοστό 80% για Εσωστρέφεια-Εξωστρέφεια και 60% για τις άλλες τρεις δυάδες των χαρακτηριστικών.

Και στην ακόλουθη έρευνα επισημειώθηκαν, με βάση τα δεδομένα από το Ερωτηματολόγιο Τύπων Προσωπικότητας Myers-Briggs Type Indicator (MBTI) για τους συγγραφείς, κείμενα που ανακτήθηκαν από τη σελίδα κοινωνικής δικτύωσης Reddit (Gjurkovic & Šnajder, 2018). Χρησιμοποιήθηκαν γλωσσολογικά χαρακτηριστικά, υφομετρικά και ψυχογλωσσολογικά (μέσω των εργαλείων LIWC και MRC αλλά και μεταδεδομένα από τη δραστηριότητα και τις αναρτήσεις των χρηστών του Reddit. Το πρόβλημα της πρόβλεψης αντιμετωπίστηκε ως δυαδική ταξινόμηση των τεσσάρων διαστάσεων. Με τη χρήση τριών ταξινομητών ελέγχθηκε η ικανότητα πρόβλεψης των προεπιλεγμένων χαρακτηριστικών. Τα αποτελέσματα των γλωσσολογικών χαρακτηριστικών κυμάνθηκαν από 67% για τη δυάδα Σκέψη-Συναίσθημα και 82% για Νόηση-Διαίσθηση.

Αγγλικά κείμενα ανακτημένα από το Twitter αποτέλεσαν το σώμα κειμένων για τους Farnadi et al. (2016) με στόχο την αυτόματη ανάκτηση της προσωπικότητας 44 χρηστών. Στα παραπάνω κείμενα για την εξαγωγή των γλωσσολογικών χαρακτηριστικών βάσει λεξικών χρησιμοποιήθηκαν το λογισμικό LIWC, η ψυχογλωσσολογική βάση MRC, το πρόγραμμα ανάλυσης συναισθήματος SentiStrength, που ήταν πιο αποτελεσματικό για την πρόβλεψη των Ευσυνείδητων και το εργαλείο εξαγωγής γλωσσολογικών ενδείξεων SPLICE, το οποίο αποδείχθηκε ότι μείωσε το ποσοστό λάθους της πρόβλεψης για τους Εξωστρεφείς.

Με σώμα κειμένων αντλημένο από το αγγλικό, ισπανικό και ιταλικό Twitter εργάστηκαν οι Liu, Perez & Nowson (2017) με στόχο την πρόβλεψη των πέντε στοιχείων της προσωπικότητας μέσω νέου συστήματος αρχιτεκτονικής νευρωνικών δικτύων. Πρόκειται για ένα συνθετικό σύστημα που βασίζεται στην βαθιά μάθηση, το

οποίο αναπαριστά ιεραρχικά τους χαρακτήρες, τις λέξεις και τις προτάσεις των κειμένων και τα συνδέει με τα χαρακτηριστικά της προσωπικότητας. Να τονιστεί ότι κανένα στοιχείο του μοντέλου δεν εξαρτάται από συγκεκριμένη γλώσσα. Η αξιολόγηση έγινε τόσο σε σύγκριση με μοντέλα που παρουσιάζουν την υψηλότερη απόδοση ως σήμερα και χρησιμοποιούν γλωσσολογικά χαρακτηριστικά όσο και με αντίστοιχα μοντέλα που είναι ανεξάρτητα τέτοιων χαρακτηριστικών. Τα αποτελέσματα κατατάσσουν την έρευνα των Liu, Perez & Nowson στην πρώτη θέση για τα Αγγλικά και τα Ισπανικά και σε ισότιμη για τα Ιταλικά.

Με το αγγλικό σώμα κειμένων από 152 χρήστες του Twitter που παρείχε το PAN CLEF 2015 (Rangel et al., 2015) εργάστηκαν οι Moreno et al. (2019) για την αυτόματη πρόβλεψη προσωπικότητας. Από τα κειμενικά δεδομένα εξήχθησαν λανθάνοντα χαρακτηριστικά (latent features) και με τρία μοντέλα έγιναν πειράματα ταξινόμησης: linear support vector classifier, logistic regression and random forest. Τα καλύτερα αποτελέσματα ως προς την ακρίβεια (70%) πέτυχαν με το πρώτο από τα προαναφερθέντα μοντέλα.

Σε κείμενα της Αραβικής γλώσσας για πρώτη φορά, ερευνήθηκε η πρόβλεψη της προσωπικότητας των συγγραφέων. Συγκεκριμένα με τη χρήση δεδομένων από 92 Αιγύπτιους χρήστες του Twitter συγκροτήθηκε ένα σύνολο χαρακτηριστικών προσωπικότητας που ονομάστηκε AraPersonality (Salem, Ismail & Aref, 2019).

Σύμφωνα με έρευνα πανεπιστημίου της Ιαπωνίας (Yamada, Sasano & Takeda, 2019) τα κειμενικά δεδομένα αναρτήσεων του Twitter είναι πιο αποτελεσματικά για την πρόβλεψη της προσωπικότητας των χρηστών σε σχέση με τις πληροφορίες που παρέχονται από τις μη λεκτικές τους αντιδράσεις (likes, retweets). Το ερωτηματολόγιο που κλήθηκαν να συμπληρώσουν οι συμμετέχοντες ήταν το ψυχολογικό τεστ που βασίζεται στην τυπολογία Myers-Briggs. Όλες οι αναρτήσεις των 20.364 χρηστών ήταν γραμμένες στην Ιαπωνική γλώσσα και η πρόβλεψη, που αντιμετωπίστηκε και εδώ ως πρόβλημα δίτιμης ταξινόμησης, έγινε με Μηχανές Διανυσμάτων Υποστήριξης. Για την κειμενική πληροφορία χρησιμοποιήθηκαν ως χαρακτηριστικά οι πιο συχνές λέξεις και σε άλλο μοντέλο λέξεις που εμφανίζονταν δέκα ή περισσότερες φορές.

Σε πρόσφατη μελέτη (Sierra et al., 2020), που αφορά στην Ισπανική γλώσσα, έτσι όπως καταγράφηκε από 98 χρήστες του Twitter, η προσωπικότητα των οποίων χαρακτηρίστηκε με το μοντέλο των Πέντε Παραγόντων, διαπιστώθηκε πως η Εξωστρέφεια εκφράζεται κυρίως με τις προσωπικές αντωνυμίες α΄ πληθυντικού και β΄ προσώπου. Τα άρθρα και οι προθέσεις συμβάλλουν στον εντοπισμό των δεκτικών σε

νέες εμπειρίες χρηστών και το χαρακτηριστικό που έχει τα καλύτερα αποτελέσματα πρόβλεψης είναι ο Νευρωτισμός, καθώς οι χρήστες εκφράζουν περισσότερο τα συναισθήματά τους στα κοινωνικά δίκτυα λόγω της ελευθερίας που νιώθουν. Το λογισμικό με το οποίο επισημείωσαν τα κείμενα ήταν το LIWC και η πρόταση των ερευνητών είναι ο εμπλουτισμός των λεξικών του με την αργκό των μέσων κοινωνικής δικτύωσης, ώστε να βελτιωθεί η ανάλυση των κειμένων με στόχο την εξαγωγή χαρακτηριστικών της προσωπικότητας.

Για πρόσφατη βιβλιογραφική επισκόπηση του ερευνητικού πεδίου του αυτόματου προσδιορισμού της προσωπικότητας των χρηστών του Twitter από τα κείμενά τους παραπέμπουμε σε έκδοση (Dandannavar, Mangalwede & Kulkarni, 2018) του Institute of Electrical and Electronics Engineers (IEEE), η οποία περιλαμβάνει σχετικές μελέτες, ψυχομετρικά ερωτηματολόγια που αξιοποιούνται, μεθοδολογία και εφαρμογές.

4.4 Σύνοψη

Το θεωρητικό υπόβαθρο και η βιβλιογραφική επισκόπηση που παρουσιάστηκαν στα κεφάλαια που προηγούνται αποτελούν προϋπόθεση για την παρακολούθηση του τρόπου με τον οποίο πραγματοποιούνται οι έρευνες αυτόματου προσδιορισμού της προσωπικότητας. Ειδικότερα, στο δεύτερο κεφάλαιο επιχειρήθηκε η περιγραφή των θεωριών της προσωπικότητας και καταλήξαμε στην επιλογή δύο με βάση τις οποίες αποδόθηκαν χαρακτηριστικά της προσωπικότητας στους μαθητές-συγγραφείς μέσω των αντίστοιχων ψυχομετρικών τεστ. Στο τρίτο κεφάλαιο έγινε αναφορά στις προσπάθειες ποσοτικοποίησης του ύφους ενός κειμένου με στόχο την αυτόματη αναγνώριση του συγγραφέα και των χαρακτηριστικών του, ενώ στο παρόν κεφάλαιο εστιάσαμε στις μελέτες αναγνώρισης στοιχείων της προσωπικότητας του συγγραφέα διαφόρων κειμενικών ειδών.

Για κάθε μία περίπτωση αναφέραμε το κειμενικό είδος, τη γλώσσα, το πλήθος των συμμετεχόντων, το ψυχομετρικό τεστ που χρησιμοποιήθηκε, την τεχνική που εφαρμόστηκε και τα αποτελέσματα της έρευνας. Έτσι όπως προέκυψε από την επισκόπηση της διεθνούς βιβλιογραφίας, οι έρευνες είναι αρκετές, τα δεδομένα είναι ανομοιογενή μεταξύ τους και εφαρμόζεται ποικιλία μεθόδων και τεχνικών.

Συγκεκριμένα, η γλώσσα που κυριαρχεί στα δεδομένα είναι η Αγγλική και η πλειοψηφία των ερευνών αντλεί κείμενα από λογαριασμούς χρηστών στο Twitter. Ως προς τα ερωτηματολόγια προσωπικότητας, με τα αποτελέσματα των οποίων γίνεται η επαλήθευση της ακρίβειας των αλγορίθμων, οι περισσότερες ερευνητικές ομάδες επιλέγουν αυτό που βασίζεται στο μοντέλο των Πέντε Παραγόντων. Οι μέθοδοι επεξεργασίας του σώματος κειμένων επίσης διαφέρουν, καθώς κάθε ομάδα χρησιμοποιεί διαφορετικά εργαλεία. Πάντως, οι περισσότερες επισημειώνουν τα κείμενα με τη χρήση λεξικών. Διαφοροποίηση εντοπίζεται και στα χαρακτηριστικά που εξάγονται από τα κείμενα, αλλά επικρατούν τα ψυχογλωσσολογικά. Τα πειράματα ταξινόμησης υλοποιούνται με αλγόριθμους μηχανικής μάθησης.

Διαπιστώνουμε, τέλος, πως σε καμία έρευνα δεν έχουν χρησιμοποιηθεί ως κειμενικά δεδομένα εκθέσεις μαθητών και όσον αφορά την Ελληνική γλώσσα η παρούσα διατριβή αποτελεί την πρώτη εκτεταμένη μελέτη. Στους συμμετέχοντες στην έρευνα χορηγήσαμε και τα δύο τεστ προσωπικότητας, που απαντούν στη βιβλιογραφία. Ως προς τη μεθοδολογία γίνεται πλήρης παρουσίαση στο έκτο κεφάλαιο που ακολουθεί, ωστόσο να σημειώσουμε εδώ πως εφαρμόσαμε τεχνικές μηχανικής μάθησης, αλλά χωρίς χρήση λεξικών, αφού όλα τα χαρακτηριστικά που εξήχθησαν από τις μαθητικές εκθέσεις είναι υφομετρικά.

Κεφάλαιο 5

Μηχανική Μάθηση και Προσωπικότητα

5.1 Μηχανική Μάθηση

Ανεξάρτητα από την επιλογή της θεωρίας προσωπικότητας (ψυχαναλυτική προσέγγιση ή μοντέλο των Πέντε Παραγόντων), το είδος των δεδομένων (π.χ. γραπτό ή προφορικό κείμενο) και τα χαρακτηριστικά που θα επιλεγούν για την αναγνώριση της προσωπικότητας (π.χ. γλωσσολογικά, υφομετρικά), οι έρευνες συμπίπτουν στη χρήση αλγορίθμων μηχανικής μάθησης για την ταξινόμηση των τύπων της προσωπικότητας, όπως διαπιστώθηκε στο προηγούμενο κεφάλαιο με τη βιβλιογραφική επισκόπηση. Για το λόγο αυτό ακολουθεί μια εισαγωγή στη μηχανική μάθηση. Στο πλαίσιο της παρούσας διατριβής στόχος είναι η ανάδειξη του βέλτιστου μοντέλου ταξινόμησης για την πρόβλεψη του τύπου προσωπικότητας των συγγραφέων εκθέσεων.

Η Μηχανική Μάθηση (Machine Learning) ως κλάδος της Τεχνητής Νοημοσύνης (Artificial Intelligence) στόχο έχει την κατασκευή και μελέτη συστημάτων που μπορούν να μαθαίνουν από δεδομένα και να κατασκευάζουν μοντέλα για να κάνουν προβλέψεις αυτόματα πάνω σε νέα δεδομένα βελτιώνοντας την απόδοσή τους βάσει προηγούμενης εμπειρίας και σώματος εκπαίδευσης. Ο ορισμός που δίνεται από τον Mitchell (1997) για τη μηχανική μάθηση είναι: «ένα πρόγραμμα υπολογιστή λέγεται ότι μαθαίνει από μια εμπειρία E σε σχέση με ένα σύνολο εργασιών T και μέτρο απόδοσης P, όταν η απόδοσή του σε εργασίες από το T, όπως μετριέται από το P, βελτιώνεται μέσω της εμπειρίας E».

Στο ερευνητικό πρόβλημα της αυτόματης απόδοσης ψυχολογικών χαρακτηριστικών σε συγγραφέα ο ορισμός θα μπορούσε να προσδιοριστεί ως εξής: Έργο T είναι η αναγνώριση της τιμής των χαρακτηριστικών της προσωπικότητας ενός συγγραφέα, απόδοση P είναι το ποσοστό των χαρακτηριστικών που αναγνωρίστηκαν σωστά και Εμπειρία E είναι το σύνολο των συγγραφέων των οποίων γνωρίζουμε τα χαρακτηριστικά μέσω των ερωτηματολογίων προσωπικότητας.

Οι αλγόριθμοι μηχανικής μάθησης κατηγοριοποιούνται σε δύο είδη μάθησης: μη εποπτευόμενη Μάθηση και εποπτευόμενη Μάθηση, τα οποία θα παρουσιαστούν σύντομα. Εκτός από τις δύο αυτές βασικές κατηγορίες μηχανικής μάθησης έχουν αναπτυχθεί και υβριδικές, όπως η ημι-εποπτευόμενη και η ενισχυμένη μάθηση.

Στη μη εποπτευόμενη Μάθηση (Unsupervised Learning) το σύστημα δεν γνωρίζει τη μορφή εξόδου των δεδομένων, πρέπει δηλαδή μόνο του να μάθει τις συσχετίσεις ή ομάδες σε ένα σύνολο δεδομένων δημιουργώντας πρότυπα. Σε αυτή την κατηγορία μάθησης εντάσσεται η διαδικασία της Συσταδοποίησης (Clustering) και της Εκτίμησης Παραμέτρων (Parameter Estimation). Η αξία της μη εποπτευόμενης Μάθησης έγκειται στη διάκριση προτύπων και μοτίβων που δεν έχουν παρατηρηθεί από τον άνθρωπο (Mitchell, 1997).

Στην εποπτευόμενη Μάθηση (Supervised Learning) το σύστημα πρέπει να μάθει επαγωγικά (από ένα σύνολο παραδειγμάτων) μια συνάρτηση η οποία αποτελεί την περιγραφή ενός μοντέλου. Ο επιβλέπων παρέχει τη σωστή τιμή εξόδου της συνάρτησης για τα υπό εξέταση δεδομένα και στόχος είναι η δημιουργία ενός μοντέλου το οποίο θα προβλέπει χαρακτηριστικά σε άγνωστα δεδομένα χρησιμοποιώντας στα δεδομένα εκπαίδευσης παραδείγματα που έχουν αυτά τα χαρακτηριστικά γνωστά. Τα χαρακτηριστικά είναι ανεξάρτητες παρατηρούμενες μεταβλητές, ενώ η κατηγορία είναι μία εξαρτημένη μεταβλητή με τιμή που καθορίζεται από τις τιμές των ανεξάρτητων μεταβλητών (Mitchell, 1997).

Οι αλγόριθμοι εποπτευόμενης Μάθησης διακρίνονται σε δύο κατηγορίες, Παλινδρόμησης (Regression) και Ταξινόμησης (Classification), με τους οποίους ασχολούμαστε στη διατριβή. Η Παλινδρόμηση αναφέρεται στη δημιουργία μοντέλων πρόβλεψης αριθμητικών τιμών (ποσοτικές μεταβλητές). Η έξοδος δεν είναι διακριτή, αλλά συνεχής. Έτσι, ο αλγόριθμος μπορεί να προβλέψει μεταξύ ενός εύρους τιμών.

Η ταξινόμηση, κατά την οποία δημιουργούνται μοντέλα πρόβλεψης διακριτών κατηγοριών/κλάσεων (ποιοτικές μεταβλητές), είναι μια τεχνική εξόρυξης δεδομένων (data mining) κατά την οποία δημιουργείται ένα μοντέλο που κατηγοριοποιεί δεδομένα. Οι κατηγορίες είναι καθορισμένες πριν την ανάλυση των δεδομένων και αυτό είναι το βασικότερο χαρακτηριστικό της ταξινόμησης. Για την επίλυση του προβλήματος της ταξινόμησης απαιτούνται παραδείγματα αντιπροσωπευτικά για κάθε κατηγορία, τα οποία έχουν ήδη ταξινομηθεί. Το σύνολο δεδομένων εισόδου χωρίζεται σε ένα σώμα δεδομένων για να εκπαιδεύσει τον αλγόριθμο, ώστε να προβλέπει (training set) και ένα σώμα δεδομένων ελέγχου (test set) για να αξιολογεί την ακρίβεια

της πρόβλεψης. Από το σώμα δεδομένων εκπαίδευσης ένας αλγόριθμος μηχανικής μάθησης παρατηρώντας τα χαρακτηριστικά των δεδομένων παράγει ένα μοντέλο ικανό να αντιστοιχίζει άγνωστα παραδείγματα στις καθορισμένες κατηγορίες. Το σώμα δεδομένων ελέγχου χρησιμοποιείται για την επικύρωση του μοντέλου, δηλαδή τον υπολογισμό της ακρίβειάς του.

Οι δημοφιλέστερες τεχνικές ταξινόμησης είναι οι αλγόριθμοι κατηγοριοποίησης ή αλλιώς ταξινομητές, όπως τα Naive Bayes μοντέλα, τα Δένδρα Αποφάσεων (Decision Trees), οι Πλησιέστεροι Γείτονες (Nearest Neighbors), τα Νευρωνικά Δίκτυα (Neural networks), οι Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines, SVM).

5.2 Αλγόριθμοι Εποπτευόμενης Μάθησης

Στο προηγούμενο υποκεφάλαιο παρουσιάσαμε δύο είδη αλγορίθμων εποπτευόμενης μάθησης, αυτούς που χρησιμοποιούνται για την παλινδρόμηση και αυτούς που χρησιμοποιούνται για την ταξινόμηση. Επειδή το θέμα της αναγνώρισης της προσωπικότητας του συγγραφέα που αφορά την παρούσα διατριβή, θα αντιμετωπισθεί ως πρόβλημα ταξινόμησης, στις δύο ενότητες που ακολουθούν θα δούμε εν συντομία δύο συγκεκριμένους αλγόριθμους ταξινόμησης, χωρίς τεχνικές λεπτομέρειες, αλλά με όσα στοιχεία χρειάζονται για την κατανόησή τους, καθώς θα αναφερθούμε σε αυτούς στα αποτελέσματα της έρευνάς μας. Πρόκειται για τον Naive Bayes και τον Generalized Linear Model.

5.2.1 Naive Bayes

Ο αλγόριθμος Naive Bayes (Langley, Iba & Thompson, 1992) ανήκει στους αλγόριθμους εποπτευόμενης μάθησης. Πρόκειται για πιθανοτικό ταξινομητή, που βασίζεται στη θεωρία αποφάσεων του θεωρήματος του Bayes με την υπόθεση της ανεξαρτησίας ανάμεσα σε όλα τα χαρακτηριστικά που εξετάζονται. Αυτή η υπόθεση βοηθάει στην απλοποίηση της κατηγοριοποίησης. Για να προβλέψει την κατηγορία μιας άγνωστης παρατήρησης, ο Naive Bayes αλγόριθμος υπολογίζει τις πιθανότητες

για την κάθε διαθέσιμη κατηγορία και κατατάσσει την παρατήρηση στην κατηγορία με τη μεγαλύτερη πιθανότητα.

Παρά την απλοποιημένη προσέγγιση ο Naive Bayes αλγόριθμος είναι εξαιρετικά διαδεδομένος και χρησιμοποιείται και για ταξινόμηση κειμένων με ικανοποιητική απόδοση. Απαιτεί μικρό αριθμό δειγμάτων εκπαίδευσης, άρα σύντομο χρόνο εκπαίδευσης και επομένως η διαδικασία είναι πολύ γρήγορη σύγκριση με άλλους πιο πολύπλοκους αλγόριθμους.

5.2.2 Generalized Linear Model

Στην ίδια κατηγορία αλγορίθμων μηχανικής μάθησης με εκπαίδευση ανήκει και ο Generalized Linear Model (Γενικευμένα Γραμμικά Μοντέλα) (Montgomery, Peck & Vining, 2012). Πρόκειται για επέκταση των παραδοσιακών γραμμικών μοντέλων, μία γενίκευση των μοντέλων γραμμικής παλινδρόμησης. Ο συγκεκριμένος αλγόριθμος ταιριάζει τα γενικευμένα γραμμικά μοντέλα στα δεδομένα μεγιστοποιώντας τον λογάριθμο της συνάρτησης πιθανοφάνειας (log-likelihood).

Η τεχνική του elastic net penalty μπορεί να χρησιμοποιηθεί για την κανονικοποίηση των παραμέτρων. Ο υπολογισμός της προσαρμογής του μοντέλου είναι παράλληλος, εξαιρετικά γρήγορος και κλιμακώνεται πολύ καλά για μοντέλα με περιορισμένο αριθμό δεικτών πρόβλεψης με μη μηδενικούς συντελεστές.

5.3 Αξιολόγηση αλγορίθμων εποπτευόμενης μάθησης

Για να προσδιοριστεί η απόδοση ενός αλγόριθμου χρησιμοποιούνται διάφορες μετρικές. Αυτές που ακολουθούν προέρχονται από τον κλάδο της Ανάκτησης Πληροφορίας και μας παρέχουν τη δυνατότητα να ελέγξουμε την ποιότητα του αλγόριθμου να μάθει ικανοποιητικά να ταξινομεί κείμενα με βάση το ψυχολογικό προφίλ του συγγραφέα τους.

Κάθε μοντέλο παράγει έναν πίνακα ταξινόμησης, όπου αναφέρονται πόσα κείμενα ταξινομήθηκαν ορθά και πόσα ταξινομήθηκαν λανθασμένα. Ο πίνακας στην περίπτωση μιας δίτιμης ταξινόμησης, όταν ελέγχουμε για παράδειγμα την

αποτελεσματικότητα της πρόβλεψης σε κείμενα εσωστρεφών και εξωστρεφών συγγραφέων έχει την ακόλουθη μορφή:

Πίνακας 2: Ταξινόμηση κειμένων Εξωστρεφών-Εσωστρεφών.

Κείμενα Εξωστρεφών συγγραφέων	Κείμενα Εσωστρεφών συγγραφέων	Ταξινομήθηκαν ως
Ψευδώς Αρνητικά (ΨΑ)	Αληθώς Θετικά (ΑΘ)	Κείμενα Εσωστρεφών συγγραφέων
Αληθώς Αρνητικά (ΑΑ)	Ψευδώς Θετικά (ΨΘ)	Κείμενα Εξωστρεφών συγγραφέων

Το πλήθος των κειμένων τα οποία ανήκουν σε εσωστρεφείς συγγραφείς και τα οποία ο αλγόριθμος ταξινόμησε ορθά στους εσωστρεφείς συγγραφείς αποτελούν τα Αληθώς Θετικά (ΑΘ) αποτελέσματα της ταξινόμησης. Αντίθετα, το πλήθος των κειμένων τα οποία ανήκουν σε εσωστρεφείς συγγραφείς αλλά λόγω λάθους του αλγορίθμου ταξινομήθηκαν σε εξωστρεφείς συγγραφείς αποτελούν τα Ψευδώς Θετικά (ΨΘ) αποτελέσματα της ταξινόμησης. Αντίστοιχα, το πλήθος των κειμένων των εξωστρεφών συγγραφέων που ταξινομήθηκαν από τον αλγόριθμο ορθά στους εξωστρεφείς συγγραφείς αποτελούν τα Αληθώς Αρνητικά αποτελέσματα. Τέλος το πλήθος των κειμένων που ανήκουν στους εξωστρεφείς συγγραφείς, αλλά λόγω λάθους του αλγορίθμου ταξινομήθηκαν στους εσωστρεφείς συγγραφείς αποτελεί τα Ψευδώς Αρνητικά (ΨΑ) αποτελέσματα της ταξινόμησης.

Με αυτές τις κατηγορίες ορίζουμε τα εξής μέτρα αξιολόγησης της ταξινόμησης (Μικρός, 2015a:132), τα οποία καθορίζουν την αποτελεσματικότητα των αλγόριθμων και θα χρησιμοποιηθούν ως μετρικές των μοντέλων στην παρούσα διατριβή:

- Ακρίβεια (Accuracy): Η ακρίβεια είναι ο συνολικός αριθμός των ορθών ταξινομήσεων που έκανε ο αλγόριθμος ως προς το σύνολο των κειμένων που είχε

στην διάθεσή του. Η ακρίβεια μιας ταξινόμησης ορίζεται από την παρακάτω εξίσωση: $Ακρίβεια = \frac{ΑΘ+ΑΑ}{ΑΘ+ΨΘ+ΑΑ+ΨΑ}$.

- Ορθότητα (Precision): Η ορθότητα είναι ο αριθμός των ταξινομήσεων που έκανε ο αλγόριθμος (ΑΘ για τους εσωστρεφείς συγγραφείς και ΑΑ για τους εξωστρεφείς συγγραφείς) ως προς το σύνολο των κειμένων που απέδωσε ο αλγόριθμος στον συγκεκριμένο τύπο συγγραφέα. Απαντάται το ερώτημα: Πόσοι από τους συγγραφείς που έχουν προβλεφθεί ως εξωστρεφείς είναι πράγματι εξωστρεφείς; Η ορθότητα της ταξινόμησης για τους εσωστρεφείς και εξωστρεφείς συγγραφείς ορίζεται ως εξής:

$$\text{Ορθότητα (Εσωστρεφών)} = \frac{ΑΘ}{ΑΘ+ΨΘ}$$

$$\text{Ορθότητα (Εξωστρεφών)} = \frac{ΑΑ}{ΑΑ+ΨΑ}$$

- Ανάκληση (Recall): Η ανάκληση των σωστών ταξινομήσεων που έκανε ο αλγόριθμος (ΑΘ για τους εσωστρεφείς συγγραφείς, ΑΑ για τους εξωστρεφείς συγγραφείς) ως προς το σύνολο των κειμένων που ανήκουν στον συγκεκριμένο τύπο συγγραφέα (ΑΘ+ΨΑ για τους εσωστρεφείς συγγραφείς και ΑΑ+ΨΘ για τους εξωστρεφείς συγγραφείς). Απαντάται το ερώτημα: Πόσοι από τους συγγραφείς που είναι πραγματικά εξωστρεφείς προβλέφθηκαν σωστά; Η ανάκληση μιας ταξινόμησης ορίζεται από την παρακάτω εξίσωση:

$$\text{Ανάκληση (Εσωστρεφών)} = \frac{ΑΘ}{ΑΘ+ΨΑ}$$

$$\text{Ανάκληση (Εξωστρεφών)} = \frac{ΑΑ}{ΑΑ+ΨΘ}$$

Οι παραπάνω μετρικές αξιολογούν τη συμπεριφορά του αλγόριθμου ταξινόμησης σε δεδομένα ελέγχου που αποτελούν ένα τυχαίο υποσύνολο των δεδομένων εισόδου. Υπάρχουν, λοιπόν, διάφορες μέθοδοι για τον τρόπο που γίνεται η αξιολόγηση ενός ταξινομητή εποπτευόμενης μάθησης σε περισσότερα τυχαία δείγματα δεδομένων ελέγχου. Μεταξύ των πιο διαδεδομένων είναι η διασταυρούμενη επικύρωση (k-fold cross-validation) (Μικρός, 2015a: 133) κατά την οποία επιλέγονται k τυχαία δείγματα από τα αρχικά δεδομένα και σχηματίζονται αντίστοιχα k δεδομένα

ελέγχου. Η εκπαίδευση του αλγόριθμου επαναλαμβάνεται k φορές και η απόδοσή του αξιολογείται k φορές. Το τελικό μέτρο αξιολόγησης είναι ο μέσος όρος των k επιμέρους αξιολογήσεων που πραγματοποιήθηκαν. Το πλεονέκτημα της μεθόδου είναι ότι προσφέρει αντικειμενικά αποτελέσματα, αποφεύγεται η υπερπροσαρμογή στα δεδομένα και συμπεραίνουμε εύκολα πόσο καλά μπορεί να αποδώσει ο αλγόριθμος σε νέα δεδομένα.

Στην παρούσα διατριβή για την αναγνώριση της προσωπικότητας έχει χρησιμοποιηθεί η πλατφόρμα ανάλυσης δεδομένων RapidMiner, στην οποία θα αναφερθούμε στο επόμενο κεφάλαιο, όπου έχει ενσωματωθεί η μέθοδος αξιολόγησης των προγνωστικών μοντέλων σε εσωτερικούς τελεστές που χωρίζουν το σώμα κειμένων σε δεδομένα εκπαίδευσης και δεδομένα ελέγχου.

Κεφάλαιο 6

Μεθοδολογία της έρευνας

Στο κεφάλαιο αυτό περιγράφεται η μεθοδολογία της παρούσας έρευνας. Δίνονται, δηλαδή, στοιχεία για την αναζήτηση του δείγματος, τη διαδικασία συλλογής των δεδομένων και δημιουργίας του ηλεκτρονικού σώματος κειμένων και περιγράφεται το σώμα κειμένων. Σχολιάζονται τα ερωτηματολόγια προσωπικότητας που χρησιμοποιήθηκαν για την πραγματοποίηση των δύο μελετών της παρούσας διατριβής καθώς και τα αποτελέσματά τους, γίνεται αναφορά στο λογισμικό που χρησιμοποιήθηκε για την υλοποίηση των πειραμάτων και αναλύονται τα υφομετρικά χαρακτηριστικά που εξήχθησαν από τα κείμενα.

6.1 Δημιουργία Ηλεκτρονικού Σώματος Κειμένων

Για να ελεγχθεί η ερευνητική υπόθεση της διατριβής, εάν δηλαδή από γραπτά κείμενα Ελληνικής γλώσσας είναι δυνατό να ανιχνευτούν στοιχεία της προσωπικότητας των συγγραφέων τους, απαραίτητη είναι η ύπαρξη ηλεκτρονικού σώματος κειμένων της Ελληνικής και συγχρόνως η σύνδεση του κάθε συγγραφέα των κειμένων αυτών με ένα ψυχολογικό προφίλ. Λόγω έλλειψης τέτοιου υλικού, το πρώτο βήμα ήταν η συλλογή πρωτογενών κειμενικών δεδομένων από φυσικούς ομιλητές της Ελληνικής γλώσσας με στόχο τη δημιουργία ενός Ηλεκτρονικού Σώματος Κειμένων.

Συγκεκριμένα, ζητήθηκε από μαθητές Λυκείου (κυρίως της Δευτέρας και Τρίτης τάξης) στο πλαίσιο του μαθήματος της Νέας Ελληνικής Γλώσσας να γράψουν εκθέσεις όσο το δυνατό πιο αυθόρμητες στο βαθμό που το επιτρέπει η διδασκαλία της συγγραφής εκθέσεων βάσει της ισχύουσας νομοθεσίας. Η συμμετοχή στην έρευνα ήταν εθελοντική και η διαδικασία της συγγραφής έγινε μέσα στη σχολική αίθουσα και χωρίς να έχει δοθεί το θέμα εκ των προτέρων, ώστε υπό αυτές τις συνθήκες ο λόγος των μαθητών να μην είναι επιτηδευμένος.

Οι μαθητές ενημερώθηκαν αναλυτικά για την έρευνα στην οποία θα συμμετείχαν και για τον τρόπο επεξεργασίας τόσο των κειμένων τους όσο και των ερωτηματολογίων στα οποία απάντησαν ανώνυμα και μας παραχώρησαν την άδεια για τη χρήση τους εκφράζοντας το ενδιαφέρον τους για το αποτέλεσμα.

Για να συγκεντρωθεί ικανοποιητικό για την έρευνά μας πλήθος λέξεων από κάθε μαθητή και δεδομένου ότι η συγκεκριμένη δραστηριότητα παραγωγής λόγου στο Λύκειο με βάση τα τότε ισχύοντα αναλυτικά προγράμματα σπουδών έπρεπε να κυμαίνεται σε έκταση από τετρακόσιες έως εξακόσιες λέξεις ανά έκθεση, δώσαμε τρία θέματα προς συγγραφή κοινά για όλους τους μαθητές, τα οποία είναι τα ακόλουθα:

1. «Σε μια καλλιτεχνική εκδήλωση που γίνεται στο χώρο του Σχολείου μιλώντας ως εκπρόσωπος των συμμαθητών σου:
α) να παρουσιάσεις τη μεγάλη σημασία της τέχνης για το σύγχρονο άνθρωπο και
β) να αναφέρεις τους τρόπους με τους οποίους το σχολείο θα μπορούσε να συμβάλλει στην αισθητική καλλιέργεια των νέων».

2. «Με αφορμή την καθιέρωση της 12ης Ιουνίου ως Παγκόσμιας ημέρας κατά της παιδικής εργασίας να γράψετε ένα σχετικό άρθρο στην εφημερίδα του σχολείου σας. Αφού παρουσιάσετε το πρόβλημα της εκμετάλλευσης της παιδικής εργασίας, να αναφερθείτε στις επιπτώσεις που έχει στον ψυχισμό των παιδιών αυτών, καθώς και στη μετέπειτα κοινωνική και επαγγελματική τους εξέλιξη».

3. «Μεγάλο μέρος των σύγχρονων ανθρώπων δεν χαρακτηρίζονται για την περιβαλλοντική τους ευθύνη. Τα περιθώρια όμως, όσον αφορά την προστασία του περιβάλλοντος, στενεύουν και η ανάγκη να εισαχθεί στην εκπαίδευση η περιβαλλοντική αγωγή κρίνεται επιτακτική.
α) Ποια είναι η αναγκαιότητα-χρησιμότητα της εισαγωγής της περιβαλλοντικής αγωγής στο σχολείο;
β) Ποιο πρέπει να είναι το περιεχόμενό της;
Να αναπτύξετε τις απόψεις σας σε ένα άρθρο που θα δημοσιευτεί στην εφημερίδα του σχολείου σας».

Η επιλογή των τριών αυτών θεμάτων είχε ως κριτήρια την ποικιλία στο περιεχόμενο (τέχνη, εκμετάλλευση παιδικής εργασίας, περιβάλλον), στο κειμενικό

είδος (μία εισήγηση και δύο άρθρα) και το επικοινωνιακό πλαίσιο, καθώς και τη δυνατότητα των μαθητών να προσεγγίσουν το ζητούμενο βάσει των ευρύτερων γνώσεών τους αλλά και της διδακτέας τους ύλης. Σκόπιμα αποφεύχθηκε θέμα σχετικό με προσωπικότητα και ψυχολογία, για να γράψουν ανεπηρέαστοι, χωρίς δηλαδή να σκεφτούν πως με τα κείμενά τους προβάλλεται ο χαρακτήρας τους. Βεβαίως, προϋπόθεση ήταν η αποφυγή προβλημάτων στη ροή των μαθημάτων, γι' αυτό οι τρεις εκθέσεις γράφτηκαν σταδιακά στη διάρκεια ενός ή και δύο διδακτικών ετών ακολουθώντας την ύλη του διδακτικού αντικειμένου.

Η συλλογή των δεδομένων ξεκίνησε το δεύτερο τετράμηνο του διδακτικού έτους 2011-2012 και ολοκληρώθηκε το 2014. Οι εθελοντές μαθητές ήταν περισσότεροι, αλλά για την έρευνα αξιοποιήσαμε το υλικό από 198 και το σώμα κειμένων ανήλθε σε σχεδόν 250.000 λέξεις, αρκετά μεγάλο γνωρίζοντας ότι σήμερα οι μαθητές της δευτεροβάθμιας εκπαίδευσης αντιμετωπίζουν σημαντικές δυσκολίες στην παραγωγή λόγου. Δεν είναι υπερβολή πως ακόμη και το ανώτατο όριο των 600 λέξεων για πολλούς μοιάζει ανέφικτο. Το σώμα κειμένων περιλαμβάνει εκθέσεις με διακύμανση στο κειμενικό μέγεθος (από 125 έως και έκθεση 1.151 λέξεων). Συνολικά, το μικρότερο άθροισμα λέξεων ανά μαθητή είναι 759 και το μεγαλύτερο 2.791 λέξεις.

Ήταν εξαιρετικά δύσκολο να συγκεντρωθεί υλικό από ένα μόνο σχολείο, διότι όπως αναφέραμε η συμμετοχή ήταν εθελοντική και επιπλέον κάποιοι μαθητές δεν ολοκλήρωναν τη διαδικασία (για παράδειγμα έγραφαν μόνο μια έκθεση) ή έγραψαν τρεις πολύ σύντομες εκθέσεις των οποίων ο συνολικός αριθμός λέξεων ήταν κάτω από το όριο του συνόλου των λέξεων ανά συγγραφέα που είχαμε θέσει. Ο μέσος όρος των λέξεων των εκθέσεων ανά μαθητή ανήλθε στις 1.255 λέξεις. Από την άλλη, πρέπει να σημειώσουμε ότι στην έρευνα συμπεριλήφθηκαν και μαθητές οι οποίοι έγραψαν μόνο δύο εκθέσεις, ωστόσο ήταν εκτενείς και επομένως ο συνολικός αριθμός των λέξεων τις καθιστούσε κατάλληλες για επεξεργασία.

Η περιγραφή του ηλεκτρονικού σώματος κειμένων ανά θέμα έκθεσης παρουσιάζεται στον παρακάτω πίνακα.

Πίνακας 3: Περιγραφικά στατιστικά του Ηλεκτρονικού σώματος κειμένων.

Θέμα έκθεσης	Πλήθος εκθέσεων	Σύνολο λέξεων	Μέσος όρος λέξεων ανά έκθεση	Μέγιστο μέγεθος έκθεσης	Ελάχιστο μέγεθος έκθεσης
Τέχνη	189	82.009	433,9	919	163
Παιδική εργασία	197	82.313	417,8	755	232
Περιβάλλον	184	85.545	464,9	1.151	125
Σύνολο	570	249.867	438,9	1.151	125

Η συνεργασία με τους μαθητές πραγματοποιήθηκε σε τέσσερις συναντήσεις, τρεις για τη συγγραφή των τριών εκθέσεων και μία για τη συμπλήρωση των ερωτηματολογίων. Κι επειδή όλα τα κείμενα που αποτελούν το σώμα κειμένων ήταν ανώνυμα, έπρεπε να γίνει η σύνδεση των ερωτηματολογίων με τις εκθέσεις -η συλλογή υλικού ήταν σταδιακή- ώστε να ξέρουμε στο τέλος ότι ο μαθητής που έχει γράψει τις συγκεκριμένες εκθέσεις έχει συμπληρώσει το αντίστοιχο ερωτηματολόγιο. Γι' αυτό, χρησιμοποιήσαμε αύξοντες αριθμούς ανά σχολείο τους οποίους στη συνέχεια ενοποιήσαμε.

Ο χρόνος συγγραφής της κάθε έκθεσης ήταν δύο διδακτικές ώρες, όσες προβλέπονται από το αναλυτικό πρόγραμμα σπουδών για τη διαδικασία της παραγωγής λόγου στο σχολείο. Με τις συνθήκες που δημιουργήθηκαν όχι μόνο αποφεύχθηκε η πρόκληση κούρασης στους μαθητές και η πιθανή μείωση του ενδιαφέροντός τους, αφού όπως αναφέραμε η συγγραφή των εκθέσεων έγινε σταδιακά, αλλά επιπλέον εξέφρασαν την ευχαρίστησή τους για τη συμμετοχή στην έρευνα, γιατί όπως τόνισαν η παραγωγή γραπτού λόγου ενίσχυσε την προσπάθειά τους να βελτιωθούν στο μάθημα της Νέας Ελληνικής Γλώσσας στο οποίο εξετάζονται πανελλαδικά. Επιπρόσθετα, η συμπλήρωση των ερωτηματολογίων τους προκαλούσε την αναμονή για τα αποτελέσματα, τα οποία και τους δόθηκαν.

Σε αυτό το σημείο πρέπει να τονίσουμε ότι οι εκθέσεις των μαθητών ήταν όλες χειρόγραφες. Δεν ακολουθήσαμε τη μέθοδο όλων των άλλων σχετικών ερευνών που καλούν τους συμμετέχοντες να παράγουν κείμενα ηλεκτρονικά και να τα υποβάλλουν

διαδικτυακά. Για τα ελληνικά δεδομένα και μάλιστα τα σχολικά κάτι τέτοιο είναι αδύνατο. Οι μαθητές των ελληνικών σχολείων δεν είναι εξοικειωμένοι με τη συγγραφή εκθέσεων απευθείας σε ηλεκτρονικό υπολογιστή και επιπλέον στις αίθουσες δεν υπάρχουν ηλεκτρονικοί υπολογιστές. Απ' την άλλη πλευρά, αν αναθέταμε στα παιδιά να πληκτρολογήσουν τις εκθέσεις εκτός σχολείου κρίνουμε ότι θα χάναμε πολύτιμα στοιχεία του αυθόρμητου λόγου που προέκυψαν στα χειρόγραφά τους και εκτός αυτού θα διόρθωναν τουλάχιστον τα ορθογραφικά λάθη, τα οποία πιθανόν να είναι ενδεικτικά της ψυχολογίας του συγγραφέα.

Το επόμενο, λοιπόν, χρονοβόρο στάδιο ήταν η ψηφιοποίηση των χειρόγραφων εκθέσεων. Κατά την ψηφιοποίηση, η οποία έγινε εξ ολοκλήρου με πληκτρολόγηση, διατηρήθηκε η πρωτότυπη μορφή των κειμένων σε κάθε επίπεδο, ώστε το ηλεκτρονικό σώμα κειμένων να αποδίδει και την ελάχιστη λεπτομέρεια του χειρόγραφου, ακόμα και την απουσία τόνων ή την παράλειψη κάποιου γράμματος.

6.2 Δημογραφικά στοιχεία των συμμετεχόντων μαθητών

Για τους 198 μαθητές, οι οποίοι είναι οι συγγραφείς των εκθέσεων, που αποτελούν το ηλεκτρονικό σώμα κειμένων της έρευνάς μας έχουμε συγκεντρώσει κάποια προσωπικά στοιχεία. Το ερωτηματολόγιο δημογραφικών στοιχείων που συμπλήρωσαν αποτελείται από έξι ερωτήσεις κλειστού και ανοικτού τύπου και σχεδιάστηκε με στόχο τη συλλογή των βασικών δημογραφικών πληροφοριών των μαθητών που συμμετείχαν στην παρούσα έρευνα. Για όλους τους μαθητές, λοιπόν, έχουμε καταγράψει, το φύλο, την ηλικία, τη μητρική γλώσσα, το σχολείο, την τάξη και την κατεύθυνση ή τον τομέα σπουδών.

Συγκεκριμένα, το ηλεκτρονικό σώμα κειμένων που αναπτύξαμε είναι ισορροπημένο τόσο ως προς τον αριθμό των λέξεων ανά μαθητή όσο και ως προς το φύλο των μαθητών και την ηλικία. Περιλαμβάνει 570 αρχεία-εκθέσεις από 198 μαθητές, ηλικίας μεταξύ 16 και 18 ετών. Από αυτούς οι 88 (44,4%) είναι αγόρια και οι 110 (55,5%) κορίτσια.

Ως προς τη μητρική γλώσσα όλοι οι μαθητές είναι φυσικοί ομιλητές της Ελληνικής γλώσσας, κάποιοι λίγοι μαθητές έχουν γλώσσα καταγωγής την Αλβανική, ωστόσο τους συμπεριλάβαμε στο δείγμα, διότι έχουν γεννηθεί στην Ελλάδα και

φοιτούν σε ελληνικά σχολεία με αποτέλεσμα να είναι φυσικοί ομιλητές και κάποιοι άλλοι είναι δίγλωσσοι με επιπλέον μητρική γλώσσα την Αγγλική, Αρμενική, Γερμανική και Ιταλική.

Οι κατευθύνσεις σπουδών που ακολουθούν οι συμμετέχοντες είναι η θεωρητική, η θετική ή τεχνολογική για το Γενικό Λύκειο και ο τομέας πληροφορικής, μηχανολογίας, ηλεκτρολογίας ή οικονομικών και διοικητικών υπηρεσιών για το Επαγγελματικό Λύκειο. Γενικά, οι 198 μαθητές της έρευνας εντάσσονται σε όλα τα επίπεδα της κλίμακας βαθμολόγησης, συμμετείχαν, δηλαδή, μαθητές πολύ αδύναμοι στα γλωσσικά μαθήματα έως και άριστοι. Οι μαθητές που συμμετείχαν φοιτούσαν στα ακόλουθα σχολεία: Γενικό Λύκειο Αντιμάχειας Κω (94 μαθητές), Γενικό Λύκειο Ν. Ηρακλείου Αττικής (85), Επαγγελματικό Λύκειο Σίφνου (16) και Γενικό Λύκειο Σίφνου (3). Ως προς τις τάξεις φοίτησης, οι 9 μαθητές ήταν της Πρώτης τάξης, οι 142 της Δευτέρας και οι 47 της Τρίτης.

6.3 Ερωτηματολόγια προσωπικότητας

Επιπλέον των δημογραφικών πληροφοριών, για κάθε μαθητή έχουμε τα αποτελέσματα των δύο σταθμισμένων τεστ προσωπικότητας που τους χορηγήθηκαν. Πρόκειται για τα ακόλουθα ερευνητικά εργαλεία: το Ερωτηματολόγιο Τύπων Προσωπικότητας Myers-Briggs Type Indicator (MBTI) (Myers-Briggs, 1962) βασισμένο στο τυπολογικό μοντέλο προσωπικότητας του Carl Jung και το Big Five τεστ προσωπικότητας, που βασίζεται στο μοντέλο των Πέντε Παραγόντων (Costa & McCrae, 1993).

Το Ερωτηματολόγιο Myers-Briggs Type Indicator (MBTI) κατηγοριοποιεί τις ψυχολογικές διαφορές στην προσωπικότητα των ατόμων σε τέσσερις διχοτομήσεις που ως αποτέλεσμα προκύπτουν 16 τύποι προσωπικότητας από την σύντμηση των τεσσάρων αρχικών γραμμάτων των λέξεων Εξωστρέφεια (Extraversion-E), Εσωστρέφεια (Introversion-I), Νόηση (Sensing-S), Διαίσθηση (Intuition-I), Σκέψη (Thinking-T), Συναίσθημα (Feeling-F), Αντίληψη (Perception-P), Κρίση (Judgment-J).

Πίνακας 4: Οι δεκαέξι τύποι προσωπικότητας των Myers-Briggs. (πηγή: http://www.humanmetrics.com/cgi-win/JTypes2.asp).

The 16 personality types			
ESTJ	ISTJ	ENTJ	INTJ
ESTP	ISTP	ENTP	INTP
ESFJ	ISFJ	ENFJ	INFJ
ESFP	ISFP	ENFP	INFP

Το ερωτηματολόγιο προσωπικότητας των Πέντε Παραγόντων είναι το πιο δημοφιλές τεστ προσωπικότητας και στον τομέα της πρόβλεψης του ψυχολογικού προφίλ του συγγραφέα, όπως φάνηκε από τη βιβλιογραφική επισκόπηση στο τέταρτο κεφάλαιο της διατριβής. Περιγράφει την προσωπικότητα μέσω πέντε παραγόντων: Δεκτικότητα στην εμπειρία (Openness to Experience-O), Ευσυνειδησία (Conscientiousness-C), Εξωστρέφεια (Extraversion-E), Προσήνεια (Agreeableness-A) και Νευρωτισμός (Neuroticism-N).

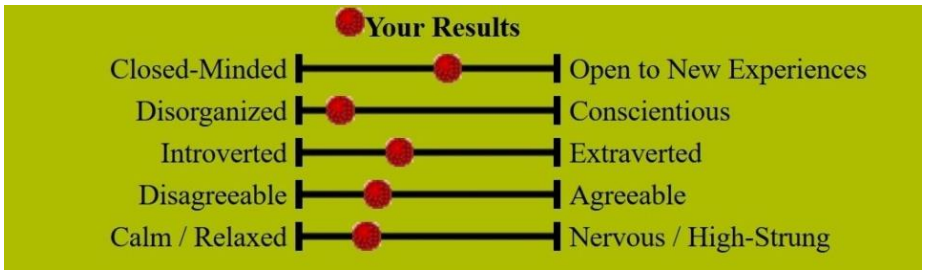
Και τα δύο τεστ προσωπικότητας δόθηκαν στους μαθητές μεταφρασμένα στα Ελληνικά με τρόπο που διασφάλισε την αξιοπιστία και εγκυρότητά τους. Η διαδικασία της μετάφρασης περιελάμβανε αρχικά μια πρώτη μετάφραση από την ερευνήτρια. Στη συνέχεια χορηγήθηκε ατομικά σε πέντε μαθητές και ζητήθηκε να εκφράσουν φωναχτά τι είχαν κατανοήσει από το κάθε ερώτημα, ώστε να καταγραφεί ο τρόπος με τον οποίο γινόταν αντιληπτό το κάθε ένα. Τέλος οι σημειώσεις αυτές αξιοποιήθηκαν, ώστε να γίνουν κάποιες τροποποιήσεις. Όταν τα ερωτηματολόγια δόθηκαν στους μαθητές, ο καθένας τους κλήθηκε να επιλέξει τη δήλωση εκείνη που ταίριαζε ή αντιστοιχούσε καλύτερα στις απόψεις του. (Παράρτημα 1 και 2).

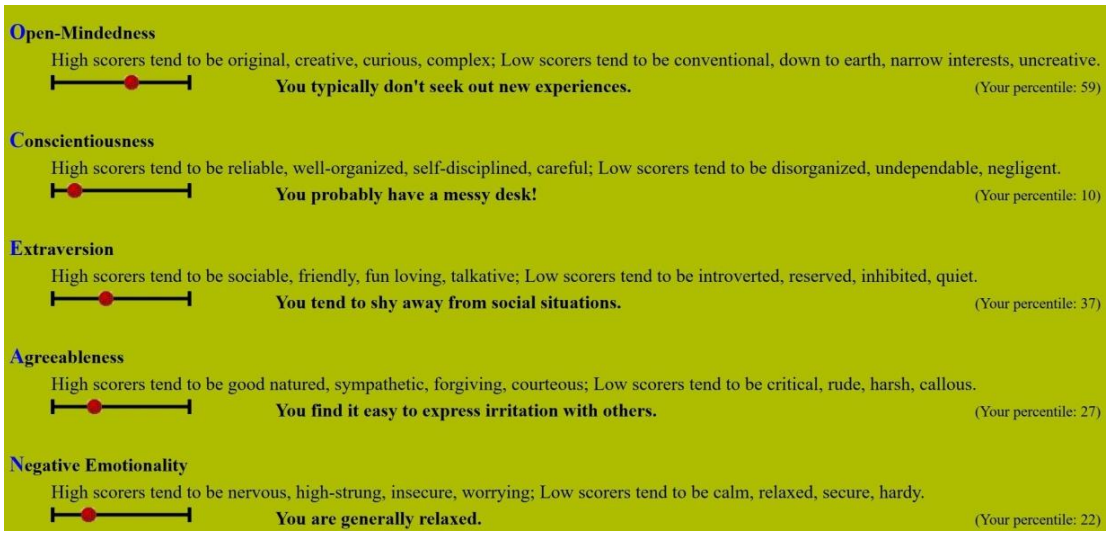
Πρέπει να αναφέρουμε ότι και τα δύο ερωτηματολόγια δόθηκαν στους μαθητές σε έντυπη μορφή και με τον ίδιο τρόπο απάντησαν οι ερωτώμενοι. Δεν ακολουθήσαμε τη μέθοδο των άλλων σχετικών ερευνών που ζητούν από τους συμμετέχοντες να υποβάλλουν τις απαντήσεις στα ερωτηματολόγια διαδικτυακά. Ακολουθήσαμε αυτή τη μέθοδο, διότι θεωρήσαμε πιο αξιόπιστες τις απαντήσεις των μαθητών ως χειρόγραφες, διότι τους δόθηκαν εντός της σχολικής τάξης όπου απαντήθηκαν απορίες και δόθηκαν διευκρινίσεις, κάτι που δε θα γινόταν αν ο κάθε μαθητής συμπλήρωνε τα τεστ διαδικτυακά.

Ακολούθησε το στάδιο της επεξεργασίας των απαντήσεων των μαθητών στις 72 και 45 ερωτήσεις του Ερωτηματολογίου Myers-Briggs Type Indicator και του ερωτηματολογίου προσωπικότητας των Πέντε Παραγόντων αντίστοιχα. Αυτό έγινε με την χειρωνακτική εισαγωγή των απαντήσεών τους στις αντίστοιχες ιστοσελίδες για την έκδοση αποτελεσμάτων (http://www.humanmetrics.com/cgi-win/JTypes2.asp, http://www.outofservice.com/bigfive/).

Στις επόμενες εικόνες παραθέτουμε ένα παράδειγμα αποτελέσματος για κάθε ερωτηματολόγιο προσωπικότητας.

Εικόνα 1: Αποτέλεσμα Ερωτηματολογίου Myers-Briggs Type Indicator (MBTI).





Εικόνα 2: Αποτέλεσμα ερωτηματολογίου προσωπικότητας των Πέντε Παραγόντων.

Οι τιμές των τύπων και των χαρακτηριστικών της προσωπικότητας που προκύπτουν ως αποτελέσματα από τα δύο ερωτηματολόγια είναι αριθμητικές και δείχνουν κατά πόσο κάθε γνώρισμα περιγράφει ένα άτομο. Για τα πειράματα ταξινόμησης μετασχηματίστηκαν οι αριθμητικές τιμές σε δυαδικές, όπως θα εξηγήσουμε στο υποκεφάλαιο 6.6. Στους Πίνακες 5 και 6 φαίνεται το πλήθος των μαθητών που συμμετείχαν στα πειράματα ανά τύπο και χαρακτηριστικό προσωπικότητας βάσει και των δύο ερωτηματολογίων.

Πίνακας 5: Πλήθος μαθητών ανά τύπο προσωπικότητας.

Τύπος προσωπικότητας	**Πλήθος μαθητών**
Extraverted	148
Introverted	35
Sensing	49
Intuitive	115
Thinking	64
Feeling	112
Perceiving	52
Judging	126

Πίνακας 6: Πλήθος μαθητών ανά χαρακτηριστικό προσωπικότητας.

Χαρακτηριστικό Προσωπικότητας	**Πλήθος μαθητών**
Open to New Experiences	29
Closed-Minded	126
Conscientious	71
Disorganised	73
Exrtaverted	55

Χαρακτηριστικό Προσωπικότητας	Πλήθος μαθητών
Intoverted	74
Agreeable	56
Disagreeable	95
Calm	78
Nervous	75

6.4 Ανάλυση δεδομένων με το RAPIDMINER

Αξιοποιώντας τα αποτελέσματα των δύο ερωτηματολογίων προσωπικότητας, που αναφέρθηκαν στο προηγούμενο υποκεφάλαιο, δημιουργήθηκαν υποσώματα κειμένων με στόχο την επεξεργασία τους για την παραγωγή προβλέψεων του ψυχολογικού τύπου των μαθητών. Πρόκειται για διαδικασία που είναι δυνατό να γίνει και αυτόματα με τη χρήση εξειδικευμένων λογισμικών. Στην παρούσα διατριβή το περιβάλλον που επιλέχθηκε για την αυτόματη αναγνώριση της προσωπικότητας των μαθητών είναι το RapidMiner.

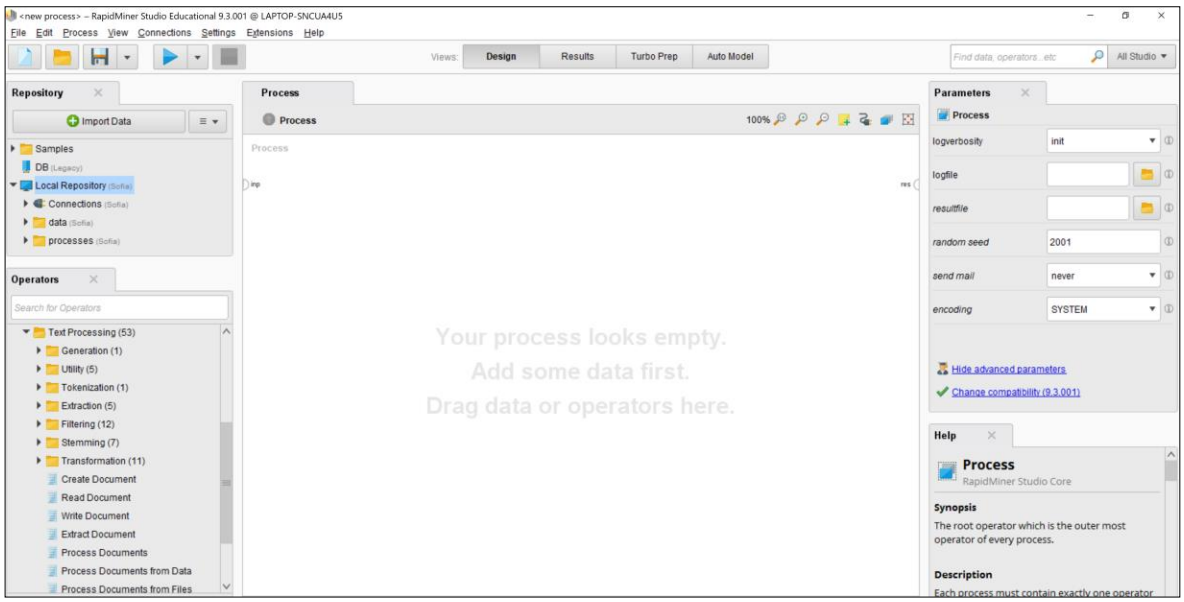
Το RapidMiner είναι ένα λογισμικό ανοικτού κώδικα για την εξόρυξη δεδομένων, τη μηχανική μάθηση και την πρόβλεψη αναλύσεων. Αρχικά, η εφαρμογή παρουσιάστηκε ως YALE (Yet Another Learning Environment), το οποίο αναπτύχθηκε από τους Ingo Mierswa, Ralf Klinkenberg και Simon Fischer στο τμήμα Τεχνητής Νοημοσύνης του Πανεπιστημίου του Dortmund το 2001. Λόγω της κατακόρυφης αύξησης του ενδιαφέροντος για το λογισμικό, οι Ingo Mierswa και Ralf Klinkenberg ίδρυσαν το 2006 την εταιρεία Rapid-I και μετονόμασαν το YALE σε RapidMiner δηλώνοντας έτσι το χαρακτηριστικό της ταχύτητας στην εξόρυξη δεδομένων.

Σήμερα χρησιμοποιείται ευρέως και για επιστημονικές και για εμπορικές εφαρμογές και είναι αναγνωρισμένο σε παγκόσμια κλίμακα. Χαρακτηριστικά αναφέρουμε ότι σε μια έρευνα που διεξήχθη από το ηλεκτρονικό newsletter KDnuggets για εξόρυξη δεδομένων, το RapidMiner το 2010 κατατάχθηκε πρώτο ανάμεσα στα εργαλεία εξόρυξης δεδομένων/ανάλυσης που χρησιμοποιήθηκαν σε πραγματικές εργασίες. Αυτή τη στιγμή το RapidMiner είναι το δεύτερο πιο δημοφιλές εργαλείο

παγκοσμίως[13]. Σε δημοσίευση του ίδιου περιοδικού το 2014 το RapidMiner ανακηρύχθηκε ως το καλύτερο εργαλείο εξόρυξης δεδομένων. Καλύπτει όλα τα στάδια της διαδικασίας εξόρυξης δεδομένων, από την μεταφορά και μετατροπή δεδομένων, ως την περιγραφική και προβλεπτική μοντελοποίηση, ανάπτυξη και αξιολόγηση του μοντέλου.

Η υλοποίηση του RapidMiner γίνεται στη γλώσσα προγραμματισμού Java, ωστόσο δεν απαιτεί τη γραφή κώδικα από τον χρήστη. Οι διαδικασίες εκτελούνται από ένα σύνολο τελεστών (operators) που δέχονται αντικείμενα εισόδου και παράγουν αντικείμενα εξόδου, τα οποία για παράδειγμα μπορεί να είναι αρχεία δεδομένων, μοντέλα, κριτήρια απόδοσης. Το λογισμικό μπορεί να επεκταθεί με πρόσθετα (plugins) βελτιώνοντας έτσι την εφαρμογή του σε κάθε πτυχή της εξόρυξης δεδομένων.

Με το άνοιγμα της κύριας περιοχής σχεδίασης του RapidMiner εμφανίζεται το ακόλουθο παράθυρο:



Εικόνα 3: Η αρχική οθόνη του RapidMiner.

Κάτω αριστερά της οθόνης παρουσιάζονται κατηγοριοποιημένοι σε ομάδες οι διαθέσιμοι τελεστές (Operators) του λογισμικού. Η επεξεργασία των δεδομένων γίνεται μέσω αυτών των τελεστών τοποθετώντας τους κατάλληλους στη σωστή ακολουθία. Συγκεκριμένα για εφαρμογές κατηγοριοποίησης κειμένου χρειάζεται να

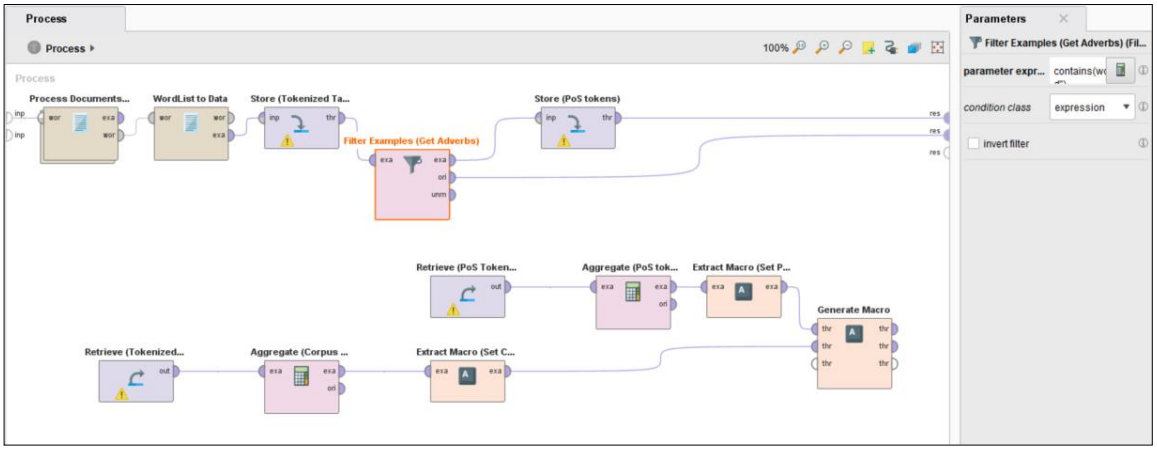
[13] https://www.kdnuggets.com/2020/06/data-science-tools-popularity-animated.html

εγκατασταθούν επιπλέον τελεστές επεξεργασίας κειμένου, ώστε να καταστεί δυνατή η επεξεργασία των κειμένων εισόδου.

Με το RapidMiner σχεδιάσαμε και εκτελέσαμε πολλές δεκάδες διεργασιών (processes) (ένα παράδειγμα παρατίθεται στην Εικόνα 4) για να εξάγουμε υφομετρικά χαρακτηριστικά από τα δύο σώματα κειμένων που δημιουργήσαμε τόσο για την πρόβλεψη των Τύπων προσωπικότητας του Jung όσο και για των χαρακτηριστικών προσωπικότητας του Big Five. Κάθε μία έκθεση των 198 μαθητών αποθηκεύτηκε ως ξεχωριστό αρχείο απλού κειμένου χωρίς μορφοποίηση (τύπου txt) και έτσι προέκυψαν τα 570 αρχεία, τα οποία ομαδοποιήθηκαν σε υπο-σώματα κειμένων, ανάλογα με τα αποτελέσματα των ψυχομετρικών τεστ στα οποία υποβλήθηκαν οι συγγραφείς τους.

Για την εισαγωγή των δεδομένων στο RapidMiner απαιτήθηκε η επεξεργασία τους με εργαλεία αυτόματης κειμενικής ανάλυσης διαφόρων βαθμίδων γλωσσικής τεχνολογίας, τα οποία δεν είναι διαθέσιμα από το λογισμικό για την Ελληνική γλώσσα: Χωρισμός σε λέξεις (Tokenization), Λημματοποίηση (Lemmatization), Μορφολογική Επισημείωση (Part-of-Speech Tagging). Έχει προβλεφθεί, ωστόσο, η εισαγωγή στο RapidMiner των προεπεξεργασμένων δεδομένων ανεξάρτητα από γλώσσα.

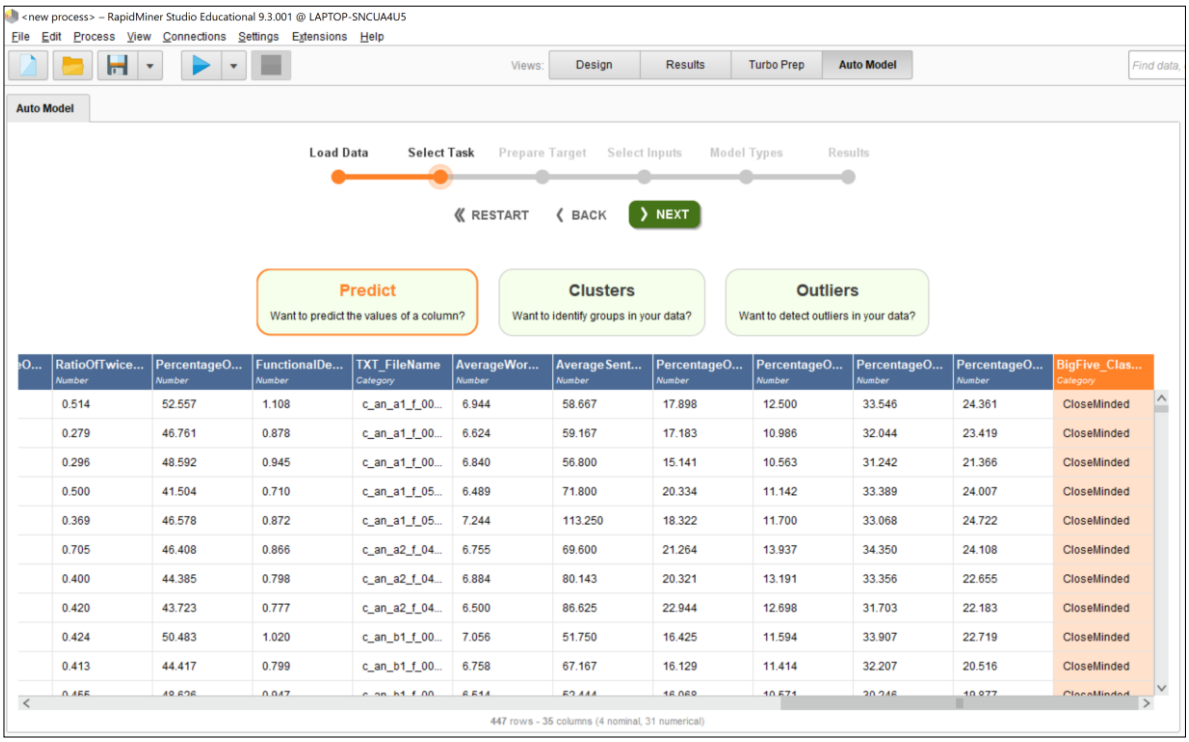
Οι διεργασίες αρχικά εφαρμόστηκαν στα πρωτότυπα κείμενα των μαθητών, αλλά διαπιστώθηκε ότι οι διαφοροποιήσεις από τον κανόνα, σε όποια γλωσσικά επίπεδα και αν παρατηρήθηκαν, επηρέαζαν αρνητικά το αποτέλεσμα της πρόβλεψης. Ως εκ τούτου, οι εκθέσεις που τροφοδότησαν την είσοδο του λογισμικού, διορθώθηκαν χειρωνακτικά, ώστε να αποδίδουν το περιεχόμενο που είχαν αρχικά, χωρίς όμως να χάνουν πληροφορίες λόγω απόκλισης στην ορθογραφική απόδοση, σε μορφολογικό και συντακτικό επίπεδο.



Εικόνα 4: Παράδειγμα Διεργασίας (σχετική συχνότητα εμφάνισης επιρρημάτων).

Μετά την εξαγωγή όλων των υφομετρικών χαρακτηριστικών (εκτενής παρουσίαση γίνεται στο υποκεφάλαιο που ακολουθεί), στην επιλογή των οποίων καταλήξαμε ύστερα από μελέτη της βιβλιογραφίας αλλά και με κριτήριο τη γλωσσική διαίσθηση που μας κατευθύνει στα γλωσσολογικά εκείνα στοιχεία που κάνουν τον γραπτό λόγο κάποιου που είναι εξωστρεφής, για παράδειγμα, να ξεχωρίζει από τον λόγο του εσωστρεφούς, περάσαμε στη δεύτερη φάση της υλοποίησης μέσω του RapidMiner.

Το λογισμικό διαθέτει μια επέκταση, το Auto Model, το οποίο διευκολύνει τη διαδικασία του σχεδιασμού και της επικύρωσης των μοντέλων. Όσον αφορά τη διατριβή, παρέχεται η δυνατότητα δημιουργίας μοντέλων πρόβλεψης (Εικόνα 5) για προβλήματα ταξινόμησης με τη χρήση αλγορίθμων μηχανικής μάθησης. Πριν, βέβαια, από την εκπαίδευση των αλγορίθμων και την εξαγωγή αποτελεσμάτων, απαιτείται η εισαγωγή, με συγκεκριμένη δομή, των υφομετρικών χαρακτηριστικών που επιλέξαμε από το προηγούμενο στάδιο για κάθε τύπο προσωπικότητας των μαθητών που θέλουμε να προβλέψουμε από τις εκθέσεις τους.



Εικόνα 5: Επιλογή διεργασίας πρόβλεψης.

Για τα προβλήματα ταξινόμησης το Auto Model δίνει τη δυνατότητα επιλογής μεταξύ των παρακάτω εννέα αλγορίθμων μηχανικής μάθησης: Naive Bayes,

Generalized Linear Model, Logistic Regression, Fast Large Margin, Deep Learning, Decision Tree, Random Forest, Gradient Boosted Trees, Support Vector Machine. Εάν δεν υπάρχει περιορισμός χρόνου, η καλύτερη λύση είναι να τρέξουν όλοι οι αλγόριθμοι και στη συνέχεια να συγκριθούν οι αποδόσεις τους τόσο σε χρόνο όσο και σε ακρίβεια.

Πίνακας 7: Οι αλγόριθμοι του Auto Model.

1. Naive Bayes
2. Generalized Linear Model
3. Logistic Regression
4. Fast Large Margin
5. Deep Learning
6. Decision Tree
7. Random Forest
8. Gradient Boosted Trees
9. Support Vector Machine

6.5 Εξαγωγή υφομετρικών χαρακτηριστικών

Τα γλωσσικά χαρακτηριστικά τα οποία έχουν χρησιμοποιηθεί ως υφομετρικοί δείκτες σε προηγούμενες έρευνες είναι πολυάριθμα, συνεχώς αυξάνονται και ανήκουν σε όλο το φάσμα των γλωσσικών επιπέδων. Η πολυεπίπεδη συσχέτιση των υφομετρικών χαρακτηριστικών με διαφορετικές κειμενικές λειτουργίες, όπως το συγγραφικό ύφος αλλά και μετακειμενικά χαρακτηριστικά, όπως είναι το θέμα, συντείνει στην αξιοποίησή τους και για τον αυτόματο προσδιορισμό του ψυχολογικού τύπου του συγγραφέα. Στα υποκεφάλαια 6.5.1, 6.5.2 και 6.5.3, που ακολουθούν, γίνεται αναλυτική παρουσίαση των υφομετρικών χαρακτηριστικών που έχουν επιλεχθεί για την έρευνά μας. Γίνεται η σύνδεσή τους με τη βιβλιογραφία, καθώς παρατίθενται οι έρευνες στις οποίες αξιοποιήθηκαν και η συμβολή τους στην αναγνώριση των χαρακτηριστικών της προσωπικότητας.

Μέσω των υφολογικών δεικτών θα επιδιώξουμε την ποσοτικοποίηση της γλώσσας των εκθέσεων των μαθητών που συμμετείχαν στην έρευνα προκειμένου να εξετασθεί η αρχική μας υπόθεση ότι, δηλαδή, η προσωπικότητα αυτών των μαθητών μπορεί αυτόματα να αναγνωριστεί από τα κείμενά τους. Στο υποκεφάλαιο αυτό παρουσιάζονται αναλυτικά τα υφομετρικά γλωσσικά χαρακτηριστικά που τελικά επιλέχθηκαν προς μέτρηση στις εκθέσεις των μαθητών. Η μέτρηση των χαρακτηριστικών έγινε με το λογισμικό στο οποίο αναφερθήκαμε στο αμέσως προηγούμενο υποκεφάλαιο, το RapidMiner.

Πρέπει να τονίσουμε πως το Ηλεκτρονικό σώμα κειμένων αναλύθηκε πριν την επεξεργασία του από το RapidMiner από Λεκτικό Αναλυτή (Tokenizer), Μορφολογικό Αναλυτή (Part of Speech Tagger) και Λημματοποιητή (Lemmatizer) (Prokopidis, Georgantopoulos & Papageorgiou, 2011; Giouli et al., 2006; Papageorgiou et al., 2000).

Για κάθε ομάδα χαρακτηριστικών χρησιμοποιήσαμε διαφορετικά επεξεργασμένο σώμα κειμένων, συγκεκριμένα για το μέσο μήκος λέξης και πρότασης και για τα διγράμματα και τριγράμματα λέξεων και χαρακτήρων (6 χαρακτηριστικά) χρησιμοποιήθηκε το αρχικό σώμα κειμένων χωρίς επεξεργασία (plain text corpus), για την εξαγωγή των σχετικών συχνοτήτων των Μερών του Λόγου (14 χαρακτηριστικά) χρησιμοποιήθηκε το μορφολογικά επισημειωμένο σώμα κειμένων και για τα υπόλοιπα υφομετρικά χαρακτηριστικά (11) το λημματοποιημένο σώμα κειμένων (Πίνακας 8).

Πειράματα επίσης πραγματοποιήθηκαν με μια ακόμα κατηγορία υφομετρικών χαρακτηριστικών. Πρόκειται για τα εξής τέσσερα: οι 100 πιο συχνές λέξεις, τα 100 πιο συχνά δίλεκτα, τα 100 πιο συχνά διγραμμάτων χαρακτήρων και τα 100 πιο συχνά τριγραμμάτων χαρακτήρων, των οποίων υπολογίστηκε για κάθε κείμενο η κανονικοποιημένη συχνότητα.

Παρουσιάζονται ξεχωριστά από τα προαναφερθέντα, διότι εξήχθησαν από το σύνολο του σώματος κειμένων και όχι από τα υποσώματα με βάση την κατηγορία προσωπικότητας στα οποία έγιναν οι μετρήσεις για όλα τα άλλα χαρακτηριστικά.

Πίνακας 8: Όλα τα υφομετρικά χαρακτηριστικά που χρησιμοποιήθηκαν ανά είδος σώματος κειμένων.

Απλό κείμενο (Plain Text Corpus)	***Μορφολογικά επισημειωμένο***	***Λημματοποιημένο (Lemmatized Corpus)***

	(Tagged Corpus)	
Μέσο μήκος λέξης σε χαρακτήρες όλων των λέξεων ενός κειμένου.	Σχετική συχνότητα εμφάνισης ρηματικών τύπων σε ένα κείμενο.	Σχετική συχνότητα εμφάνισης των λειτουργικών λέξεων σε ένα κείμενο.
Μέσο μήκος πρότασης σε λέξεις όλων των προτάσεων ενός κειμένου.	Σχετική συχνότητα εμφάνισης των ρημάτων ενεργητικής φωνής σε ένα κείμενο.	Σχετική συχνότητα εμφάνισης των πιο συχνών λέξεων μέσα σε ένα κείμενο.
Σχετική συχνότητα εμφάνισης των πιο συχνών δίλεκτων μέσα σε ένα κείμενο.	Σχετική συχνότητα εμφάνισης των ρημάτων παθητικής φωνής σε ένα κείμενο.	Σχετική συχνότητα εμφάνισης των πιο συχνών μη λειτουργικών λέξεων μέσα σε ένα κείμενο.
Σχετική συχνότητα εμφάνισης των πιο συχνών τρίλεκτων μέσα σε ένα κείμενο.	Σχετική συχνότητα εμφάνισης των ουσιαστικών σε ένα κείμενο.	Σχετική συχνότητα εμφάνισης των πιο σπάνιων λέξεων μέσα σε ένα κείμενο.
Σχετική συχνότητα εμφάνισης των πιο συχνών διγραμμάτων χαρακτήρων μέσα σε ένα κείμενο.	Σχετική συχνότητα εμφάνισης των επιθέτων σε ένα κείμενο.	Σχετική συχνότητα εμφάνισης των πιο σπάνιων μη λειτουργικών λέξεων μέσα σε ένα κείμενο.
Σχετική συχνότητα εμφάνισης των πιο συχνών τριγραμμάτων χαρακτήρων μέσα σε ένα κείμενο.	Σχετική συχνότητα εμφάνισης των άρθρων σε ένα κείμενο.	Σχετική συχνότητα μη λειτουργικών λέξεων μέσα σε ένα κείμενο.
Σχετική συχνότητα εμφάνισης των 100 πιο συχνών λέξεων.	Σχετική συχνότητα εμφάνισης των αντωνυμιών σε ένα κείμενο.	Σχετική συχνότητα των άπαξ εμφανιζομένων λέξεων μέσα σε ένα κείμενο.

Σχετική συχνότητα εμφάνισης των 100 πιο συχνών δίλεκτων.	Σχετική συχνότητα εμφάνισης των προσωπικών αντωνυμιών σε ένα κείμενο.	Σχετική συχνότητα των δις εμφανιζομένων λέξεων μέσα σε ένα κείμενο.
Σχετική συχνότητα εμφάνισης των 100 πιο συχνών διγραμμάτων χαρακτήρων.	Σχετική συχνότητα εμφάνισης των προσωπικών και κτητικών αντωνυμιών σε ένα κείμενο.	Λόγος των δις προς άπαξ λεγόμενα.
Σχετική συχνότητα εμφάνισης των 100 πιο συχνών τριγραμμάτων χαρακτήρων.	Σχετική συχνότητα εμφάνισης των επιρρημάτων σε ένα κείμενο.	Σχετική συχνότητα εμφάνισης των μη λειτουργικών λέξεων μέσα σε ένα κείμενο.
	Σχετική συχνότητα εμφάνισης των προθέσεων σε ένα κείμενο.	Λόγος των λέξεων περιεχομένου προς τις λειτουργικές.
	Σχετική συχνότητα εμφάνισης των συνδέσμων σε ένα κείμενο.	
	Σχετική συχνότητα εμφάνισης των παρατακτικών συνδέσμων σε ένα κείμενο.	
	Σχετική συχνότητα εμφάνισης των υποτακτικών συνδέσμων σε ένα κείμενο.	

Αναλυτικά παρουσιάζονται στο υποκεφάλαιο που ακολουθεί, όπου παρατίθενται το όνομα (και στα Αγγλικά η σύντμηση με την οποία δηλώθηκε στο RapidMiner, διότι με αυτό το λεκτικό θα φανεί στα διαγράμματα των αποτελεσμάτων), η περιγραφή και ο τύπος υπολογισμού όλων των υφομετρικών χαρακτηριστικών που χρησιμοποιήθηκαν στην έρευνά μας.

6.5.1 Πρώτη κατηγορία υφομετρικών χαρακτηριστικών: από το σώμα κειμένων στην αρχική του μορφή (Plain Text Corpus)

Σε αυτή την κατηγορία χαρακτηριστικών έχουμε εντάξει το μέσο μήκος λέξης σε χαρακτήρες και πρότασης σε λέξεις, καθώς και τα διγράμματα και τριγράμματα χαρακτήρων και λέξεων. Ανήκουν σε μια κατηγορία, διότι για την εξαγωγή τους δεν χρειάζεται προεπεξεργασία του σώματος κειμένων, όπως στις δύο προηγούμενες.

Το πλήθος χαρακτήρων ανά λέξη συγκαταλέγονται στα ισχυρά υφομετρικά χαρακτηριστικά σύμφωνα με τους Iqbal et al. (2010).

Το μήκος λέξης ανήκει στους φωνολογικούς δείκτες, ενώ της πρότασης στους συντακτικούς, καθώς και από αυτό καθορίζεται η συνθετότητα της προτασιακής δομής. Είναι από τα παλαιότερα υφομετρικά χαρακτηριστικά που χρησιμοποιήθηκαν. Έχουν μελετηθεί αρκετά και η σύγχρονη υφομετρική έρευνα έχει δείξει ότι το λεξικό και το προτασιακό μήκος, είναι πολυσυσχετιζόμενες μεταβλητές και πολλές φορές σχετίζονται με τον συγγραφέα, το θέμα, το γένος κ.ά. (Μικρός, 2015a: 102).

Τα N-γράμματα είναι μία συνεχόμενη ακολουθία ν τεμαχίων από μία συγκεκριμένη ακολουθία κειμένου. Έτσι, έχουμε ν-γράμματα με βάση τον χαρακτήρα, τη συλλαβή, τη λέξη ή άλλη γλωσσική δομή. Οι τιμές του ν δημιουργούν ν-γράμματα διαφορετικού μήκους. Για παράδειγμα, δημιουργούνται διγράμματα χαρακτήρων, δηλαδή ν-γράμματα χαρακτήρων με ν=2, που αποτελούν συνεχόμενες ακολουθίες δύο χαρακτήρων.

Και σε αυτή την κατηγορία χαρακτηριστικών υπάρχουν πολλές αναφορές στη βιβλιογραφία, από τις οποίες φαίνεται ότι πρόκειται για δείκτες που αποκαλύπτουν διαφορές στους τύπους προσωπικότητας των συγγραφέων: Gill & Oberlander, 2002; Oberlander & Gill, 2004a; Nowson & Oberlander, 2006; Celli, 2012; Wright, 2012; Biel et al., 2013; González-Gallardo et al., 2015; Wright & Chin, 2016.

Τα χαρακτηριστικά που χρησιμοποιήθηκαν στην παρούσα διατριβή και ανήκουν σε αυτή την κατηγορία είναι:

1. **Μέσο μήκος λέξης σε χαρακτήρες όλων των λέξεων ενός κειμένου. (AverageWordLength)**
   Εξετάζονται οι λέξεις που από 1 έως απεριόριστο πλήθος χαρακτήρων.

$$Μέσο\ Μήκος\ σε\ χαρακτήρες\ Όλων\ των\ Λέξεων\ ενός\ Κειμ. = Average(μήκος\ λέξης\ σε\ χαρακτήρες, όλες\ οι\ Λέξεις\ του\ Κειμένου$$

2. **Μέσο μήκος πρότασης σε λέξεις όλων των προτάσεων ενός κειμένου. (AverageSentenceLengthInWords)**
   Εξετάζεται κάθε πρόταση που περιλαμβάνει από 2 λέξεις και πάνω.

$$Μέσο\ Μήκος\ Πρότασης\ σε\ λέξεις\ Όλων\ των\ Προτάσ.\ ενός\ Κειμ. = Average(μήκος\ πρότ.\ σε\ λέξ.,\ όλες\ οι\ Προτ.\ του\ Κειμ.$$

3. **Σχετική συχνότητα εμφάνισης των πιο συχνών δίλεκτων μέσα σε ένα κείμενο.**
   (PercentageOfTopMostFreqWordBiGramsCoverageInFile)
   Πρώτα υπολογίζεται η συχνότητα εμφάνισης του κάθε δίλεκτου του υπό επεξεργασία κειμένου. Οι συχνότητες αυτές διατάσσονται κατά φθίνουσα σειρά. Το 10% των κορυφαίων τιμών της διατεταγμένης αυτής σειράς συχνοτήτων αθροίζεται, το δε αποτέλεσμα δίνει το πλήθος των εμφανίσεων των συχνότερων δίλεκτων του κειμένου. Το πλήθος αυτό διαιρείται με το σύνολο των εμφανίσεων όλων των λέξεων του κειμένου και το αποτέλεσμα πολλαπλασιάζεται επί 100 για να δώσει την σημειούμενη συχνότητα.

$$Σχετ.\ Συχν.\ των\ Συχνότερων\ Δίλεκτων\ σε\ Κείμενο = 100 \cdot \frac{Πλήθος\ Εμφανίσεων\ Συχνότερων\ Δίλεκτων\ στο\ Κείμενο}{Πλήθος\ Εμφανίσεων\ Όλων\ των\ Λέξεων\ του\ Κειμένου}$$

4. **Σχετική συχνότητα εμφάνισης των πιο συχνών τρίλεκτων μέσα σε ένα κείμενο.**
   (PercentageOfTopMostFreqWordTriGramsCoverageInFile)
   Πρώτα υπολογίζεται η συχνότητα εμφάνισης του κάθε τρίλεκτου του υπό επεξεργασία κειμένου. Οι συχνότητες αυτές διατάσσονται κατά φθίνουσα σειρά. Το 10% των κορυφαίων τιμών της διατεταγμένης αυτής σειράς συχνοτήτων αθροίζεται, το δε αποτέλεσμα δίνει το πλήθος των εμφανίσεων των συχνότερων τρίλεκτων του κειμένου. Το πλήθος αυτό διαιρείται με το σύνολο των εμφανίσεων όλων των λέξεων του κειμένου και το αποτέλεσμα πολλαπλασιάζεται επί 100 για να δώσει την σημειούμενη συχνότητα.

$$Σχετ.Συχν.των\ Συχνότερων\ Τρίλεκτων\ σε\ Κείμενο = 100 \cdot \frac{Πλήθος\ Εμφανίσεων\ Συχνότερων\ Τρίλεκτων\ στο\ Κείμενο}{Πλήθος\ Εμφανίσεων\ Όλων\ των\ Λέξεων\ του\ Κειμένου}$$

5. **Σχετική συχνότητα εμφάνισης των πιο συχνών διγραμμάτων χαρακτήρων μέσα σε ένα κείμενο.**
   (PercentageOfTopMostFreqCharBiGramsCoverageInFile)
   Πρώτα υπολογίζεται η συχνότητα εμφάνισης τού κάθε διγράμματος χαρακτήρων του υπό επεξεργασία κειμένου. Οι συχνότητες αυτές διατάσσονται κατά φθίνουσα σειρά. Το 10% των κορυφαίων τιμών της διατεταγμένης αυτής σειράς συχνοτήτων αθροίζεται, το δε αποτέλεσμα δίνει το πλήθος των εμφανίσεων των συχνότερων διγραμμάτων χαρακτήρων του κειμένου. Το πλήθος αυτό διαιρείται με το σύνολο των εμφανίσεων όλων των διγραμμάτων χαρακτήρων του κειμένου και το αποτέλεσμα πολλαπλασιάζεται επί 100 για να δώσει την σημειούμενη συχνότητα.

$$Σχετ.Συχν.των\ Συχνότερων\ Διγραμμάτων\ χαρακτήρων\ σε\ Κείμενο = 100 \cdot \frac{Πλήθος\ Εμφανίσεων\ Συχνότερων\ Διγραμμάτων\ χαρ.στο\ Κείμενο}{Πλήθος\ Εμφανίσεων\ Όλων\ των\ Διγραμμάτων\ χαρ.του\ Κειμένου}$$

6. **Σχετική συχνότητα εμφάνισης των πιο συχνών τριγραμμάτων χαρακτήρων μέσα σε ένα κείμενο.**
   (PercentageOfTopMostFreqCharTriGramsCoverageInFile)
   Πρώτα υπολογίζεται η συχνότητα εμφάνισης του κάθε τριγράμματος χαρακτήρων του υπό επεξεργασία κειμένου. Οι συχνότητες αυτές διατάσσονται κατά φθίνουσα σειρά. Το 10% των κορυφαίων τιμών της διατεταγμένης αυτής σειράς συχνοτήτων αθροίζεται, το δε αποτέλεσμα δίνει το πλήθος των εμφανίσεων των συχνότερων τριγραμμάτων χαρακτήρων του κειμένου. Το πλήθος αυτό διαιρείται με το σύνολο των εμφανίσεων όλων των τριγραμμάτων χαρακτήρων του κειμένου και το αποτέλεσμα πολλαπλασιάζεται επί 100 για να δώσει την σημειούμενη συχνότητα.

$$Σχετ.Συχν.των\ Συχνότερων\ Τριγραμμάτων\ χαρακτήρων\ σε\ Κείμενο = 100 \cdot \frac{Πλήθος\ Εμφανίσεων\ Συχνότερων\ Τριγραμμάτων\ χαρ.στο\ Κείμενο}{Πλήθος\ Εμφανίσεων\ Όλων\ των\ Τριγραμμάτων\ χαρ.του\ Κειμένου}$$

7. **Σχετική συχνότητα εμφάνισης των 100 πιο συχνών λέξεων μέσα σε ένα κείμενο. (Wug_1... Wug_100)**
   Πρώτα υπολογίζονται η συχνότητα εμφάνισης της κάθε λέξης του σώματος κειμένων. Οι συχνότητες αυτές διατάσσονται κατά φθίνουσα σειρά. Οι 100 κορυφαίες τιμές της διατεταγμένης αυτής σειράς συχνοτήτων αθροίζεται, το δε αποτέλεσμα δίνει το πλήθος των εμφανίσεων των συχνότερων λέξεων του σώματος κειμένων. Το πλήθος αυτό διαιρείται με το σύνολο των εμφανίσεων όλων των λέξεων του υπό επεξεργασία κειμένου και το αποτέλεσμα πολλαπλασιάζεται επί 100 για να δώσει την σημειούμενη συχνότητα.

8. **Σχετική συχνότητα εμφάνισης των 100 πιο συχνών δίλεκτων μέσα σε ένα κείμενο. (Wbg_1... Wbg_100)**
   Πρώτα υπολογίζεται η συχνότητα εμφάνισης του κάθε δίλεκτου του σώματος κειμένων. Οι συχνότητες αυτές διατάσσονται κατά φθίνουσα σειρά. Οι 100 κορυφαίες τιμές της διατεταγμένης αυτής σειράς συχνοτήτων αθροίζεται, το δε αποτέλεσμα δίνει το πλήθος των εμφανίσεων των συχνότερων δίλεκτων του σώματος κειμένων. Το πλήθος αυτό διαιρείται με το σύνολο των εμφανίσεων

όλων των λέξεων του υπό επεξεργασία κειμένου και το αποτέλεσμα πολλαπλασιάζεται επί 100 για να δώσει την σημειούμενη συχνότητα.

9. **Σχετική συχνότητα εμφάνισης των 100 πιο συχνών διγραμμάτων χαρακτήρων μέσα σε ένα κείμενο. (Cbg_1... Cbg_100)**
   Πρώτα υπολογίζεται η συχνότητα εμφάνισης τού κάθε διγράμματος χαρακτήρων του σώματος κειμένων. Οι συχνότητες αυτές διατάσσονται κατά φθίνουσα σειρά. Οι 100 κορυφαίες τιμές της διατεταγμένης αυτής σειράς συχνοτήτων αθροίζεται, το δε αποτέλεσμα δίνει το πλήθος των εμφανίσεων των συχνότερων διγραμμάτων χαρακτήρων του σώματος κειμένων. Το πλήθος αυτό διαιρείται με το σύνολο των εμφανίσεων όλων των διγραμμάτων χαρακτήρων του υπό επεξεργασία κειμένου και το αποτέλεσμα πολλαπλασιάζεται επί 100 για να δώσει την σημειούμενη συχνότητα.

10. **Σχετική συχνότητα εμφάνισης των 100 πιο συχνών τριγραμμάτων χαρακτήρων μέσα σε ένα κείμενο. (Ctg_1... Ctg_100)**
    Πρώτα υπολογίζεται η συχνότητα εμφάνισης του κάθε τριγράμματος χαρακτήρων του σώματος κειμένων. Οι συχνότητες αυτές διατάσσονται κατά φθίνουσα σειρά. Οι 100 κορυφαίες τιμές της διατεταγμένης αυτής σειράς συχνοτήτων αθροίζεται, το δε αποτέλεσμα δίνει το πλήθος των εμφανίσεων των συχνότερων τριγραμμάτων χαρακτήρων σώματος κειμένων. Το πλήθος αυτό διαιρείται με το σύνολο των εμφανίσεων όλων των τριγραμμάτων χαρακτήρων του υπό επεξεργασία κειμένου και το αποτέλεσμα πολλαπλασιάζεται επί 100 για να δώσει την σημειούμενη συχνότητα.

6.5.2 Δεύτερη Κατηγορία υφομετρικών χαρακτηριστικών: από το μορφολογικά επισημειωμένο σώμα κειμένων (Tagged Corpus)

Η χρήση των Μερών του Λόγου ως υφομετρικών χαρακτηριστικών αυξήθηκε, όταν η πρόοδος στο ερευνητικό πεδίο της Επεξεργασίας Φυσικής Γλώσσας έκανε ευρύτερα διαθέσιμους τους Μορφολογικούς Αναλυτές. Η πιο βασική μορφή

αξιοποίησης των Μερών του Λόγου ως υφομετρικών δεικτών είναι η χρήση της συχνότητας του καθενός (κανονικοποιημένη ως προς το συνολικό μέγεθος του κειμένου) ως χαρακτηριστικό στο διάνυσμα του κειμένου (Mikros, 2006). Για την αναγνώριση της προσωπικότητας του συγγραφέα πολλές είναι οι έρευνες που αξιοποίησαν αυτό το χαρακτηριστικό: Nowson & Oberlander, 2006; Lee et al., 2007; Markovikj et al., 2013; Alam & Riccardi, 2014; Oberlander & Gill, 2004a. Στην τελευταία μάλιστα, τα Μέρη του Λόγου φάνηκε ότι συμβάλλουν σημαντικά στην πρόβλεψη του νευρωτικού συγγραφέα.

Παρακάτω, παραθέτουμε αποτελέσματα ερευνών με τη χρήση συγκεκριμένων μερών του λόγου. Οι Pennebaker & King (1999) διαπίστωσαν ότι οι νευρωτικοί φοιτητές που συμμετείχαν στην έρευνα έγραφαν κυρίως στο πρώτο ενικό γραμματικό πρόσωπο. Όσοι ήταν δεκτικοί σε νέες εμπειρίες, χρησιμοποίησαν περισσότερα άρθρα και λιγότερες λέξεις σε πρώτο ενικό πρόσωπο και χρόνο ενεστώτα. Άτομα με υψηλό σκορ στην προσήνεια βρέθηκε ότι χρησιμοποιούν περισσότερο το πρώτο ενικό πρόσωπο και λιγότερα άρθρα. Οι Mairesse & Walker (2007) στο λόγο των εξωστρεφών βρήκαν λίγα άρθρα, πολλά ρήματα, επιρρήματα και αντωνυμίες. Οι Gill, Nowson & Oberlander (2009) μεταξύ των χαρακτηριστικών περιέλαβαν το πρώτο ενικό πρόσωπο και το τρίτο πρόσωπο αντωνυμιών. Ένα από τα χαρακτηριστικά που χρησιμοποίησε στη μελέτη του ο Wright (2012) ήταν η προσωπική αντωνυμία του πρώτου προσώπου (“I”) σε όλες της τις μορφές. Για τους Qiu et al. (2012) η Δεκτικότητα στην εμπειρία συσχετίστηκε αρνητικά με την χρήση επιρρημάτων και θετικά με την χρήση προθέσεων.

Από το μορφολογικά επισημειωμένο σώμα κειμένων της παρούσας έρευνας εξήχθησαν τα ακόλουθα χαρακτηριστικά:

11. **Σχετική συχνότητα εμφάνισης ρηματικών τύπων σε ένα κείμενο. (VerbsFreq)**

    Προσμετρώνται όλοι οι κλιτικοί τύποι του κάθε ρήματος που απαντά στο κείμενο.

$$Σχετ.Συχν.Ρημάτων\ σε\ Κείμενο = 100 \cdot \frac{Πλήθος\ Εμφανίσεων\ Ρημάτων\ στο\ Κείμενο}{Πλήθος\ Εμφανίσεων\ Όλων\ των\ Λέξεων\ του\ Κειμένου}$$

12. **Σχετική συχνότητα εμφάνισης των ρημάτων ενεργητικής φωνής σε ένα κείμενο.**
    (ActiveVoiceVerbsFreq)
    Προσμετρώνται όλοι οι κλιτικοί τύποι του κάθε ρήματος ενεργητικής φωνής που απαντά στο κείμενο.

$$Σχετ.Συχν.Ρημάτων\ Ενεργητικής\ Φωνής\ σε\ Κείμενο = 100 \cdot \frac{Πλήθος\ Εμφανίσεων\ Ρημάτων\ Ενεργ.Φωνής\ στο\ Κείμενο}{Πλήθος\ Εμφανίσεων\ Όλων\ των\ Λέξεων\ του\ Κειμένου}$$

13. **Σχετική συχνότητα εμφάνισης των ρημάτων παθητικής φωνής σε ένα κείμενο.**
    (PassiveVoiceVerbsFreq)
    Προσμετρώνται όλοι οι κλιτικοί τύποι του κάθε ρήματος παθητικής φωνής που απαντά στο κείμενο.

$$Σχετ.Συχν.Ρημάτων\ Παθητικής\ Φωνής\ σε\ Κείμενο = 100 \cdot \frac{Πλήθος\ Εμφανίσεων\ Ρημάτων\ Παθητ.Φωνής\ στο\ Κείμενο}{Πλήθος\ Εμφανίσεων\ Όλων\ των\ Λέξεων\ του\ Κειμένου}$$

14. **Σχετική συχνότητα εμφάνισης των ουσιαστικών σε ένα κείμενο. (NounsFreq)**
    Προσμετρώνται όλοι οι κλιτικοί τύποι του κάθε ουσιαστικού που απαντά στο κείμενο.

$$Σχετ.Συχν.Ουσιαστικών\ σε\ Κείμενο = 100 \cdot \frac{Πλήθος\ Εμφανίσεων\ Ουσιαστικών\ στο\ Κείμενο}{Πλήθος\ Εμφανίσεων\ Όλων\ των\ Λέξεων\ του\ Κειμένου}$$

15. **Σχετική συχνότητα εμφάνισης των επιθέτων σε ένα κείμενο. (AdjectivesFreq)**
    Προσμετρώνται όλοι οι κλιτικοί τύποι του κάθε επιθέτου που απαντά στο κείμενο.

$$Σχετ.Συχν.Επιθέτων\ σε\ Κείμενο = 100 \cdot \frac{Πλήθος\ Εμφανίσεων\ Επιθέτων\ στο\ Κείμενο}{Πλήθος\ Εμφανίσεων\ Όλων\ των\ Λέξεων\ του\ Κειμένου}$$

16. **Σχετική συχνότητα εμφάνισης των άρθρων σε ένα κείμενο. (ArticlesFreq)**
    Προσμετρώνται όλοι οι κλιτικοί τύποι του κάθε άρθρου που απαντά στο κείμενο.

$$Σχετ.Συχν.Άρθρων\ σε\ Κείμενο = 100 \cdot \frac{Πλήθος\ Εμφανίσεων\ Άρθρων\ στο\ Κείμενο}{Πλήθος\ Εμφανίσεων\ Όλων\ των\ Λέξεων\ του\ Κειμένου}$$

17. **Σχετική συχνότητα εμφάνισης των αντωνυμιών σε ένα κείμενο. (PronounsFreq)**
    Προσμετρώνται όλοι οι κλιτικοί τύποι της κάθε αντωνυμίας που απαντά στο κείμενο. Δεν γίνεται διάκριση μεταξύ κτητικών, αναφορικών κλπ. αντωνυμιών.

$$Σχετ.Συχν.Αντωνυμιών\ σε\ Κείμενο = 100 \cdot \frac{Πλήθος\ Εμφανίσεων\ Αντωνυμιών\ στο\ Κείμενο}{Πλήθος\ Εμφανίσεων\ Όλων\ των\ Λέξεων\ του\ Κειμένου}$$

18. **Σχετική συχνότητα εμφάνισης των προσωπικών αντωνυμιών σε ένα κείμενο.**
    (PersonalPronounsFreq)
    Προσμετρώνται όλοι οι κλιτικοί τύποι της κάθε προσωπικής αντωνυμίας που απαντά στο κείμενο.

$$Σχετ.Συχν.Προσωπικών\ Αντωνυμιών\ σε\ Κείμενο = 100 \cdot \frac{Πλήθος\ Εμφανίσεων\ Προσωπ.Αντωνυμιών\ στο\ Κείμενο}{Πλήθος\ Εμφανίσεων\ Όλων\ των\ Λέξεων\ του\ Κειμένου}$$

19. **Σχετική συχνότητα εμφάνισης των προσωπικών και κτητικών αντωνυμιών σε ένα κείμενο.**
    (PersonalAndPossessivePronounsFreq)
    Προσμετρώνται όλοι οι κλιτικοί τύποι της κάθε προσωπικής και κτητικής αντωνυμίας που απαντά στο κείμενο.

$$Σχετ.Συχν.Προσωπ.\&\ Κτητ.Αντωνυμιών\ σε\ Κείμενο = 100 \cdot \frac{Πλήθος\ Εμφανίσεων\ Προσωπ.\&\ Κτητ.Αντων.στο\ Κείμενο}{Πλήθος\ Εμφανίσεων\ Όλων\ των\ Λέξεων\ του\ Κειμένου}$$

20. **Σχετική συχνότητα εμφάνισης των επιρρημάτων σε ένα κείμενο. (AdverbsFreq)**

$$Σχετ.Συχν.ΕπιρρημάτωνσεΚείμενο = 100 \cdot \frac{Πλήθος\ Εμφανίσεων\ Επιρρημάτων\ στο\ Κείμενο}{Πλήθος\ Εμφανίσεων\ Όλων\ των\ Λέξεων\ του\ Κειμένου}$$

21. **Σχετική συχνότητα εμφάνισης των προθέσεων σε ένα κείμενο. (PrepositionsFreq)**

$$Σχετ.Συχν.Προθέσεων\ σε\ Κείμενο = 100 \cdot \frac{Πλήθος\ Εμφανίσεων\ Προθέσεων\ στο\ Κείμενο}{Πλήθος\ Εμφανίσεων\ Όλων\ των\ Λέξεων\ του\ Κειμένου}$$

22. **Σχετική συχνότητα εμφάνισης των συνδέσμων σε ένα κείμενο. (ConjunctionsFreq)**

$$Σχετ.Συχν.Συνδέσμων\ σε\ Κείμενο = 100 \cdot \frac{Πλήθος\ Εμφανίσεων\ Συνδέσμων\ στο\ Κείμενο}{Πλήθος\ Εμφανίσεων\ Όλων\ των\ Λέξεων\ του\ Κειμένου}$$

23. **Σχετική συχνότητα εμφάνισης των παρατακτικών συνδέσμων σε ένα κείμενο.**
(CoordinativeConjunctionsFreq)

$$Σχετ.Συχν.Παρατακτικών\ \ Συνδέσμων\ σε\ Κείμεν = 100 \cdot \frac{Πλήθος\ Εμφανίσεων\ Συνδυαστ.Συνδέσμων\ στο\ Κείμενο}{Πλήθος\ Εμφανίσεων\ Όλων\ των\ Λέξεων\ του\ Κειμένου}$$

24. **Σχετική συχνότητα εμφάνισης των υποτακτικών συνδέσμων σε ένα κείμενο.**
(SubordinativeConjunctionsFreq)

$$Σχετ.Συχν.Υποτακτικών\ Συνδέσμων\ σε\ Κείμενο = 100 \cdot \frac{Πλήθος\ Εμφανίσεων\ Υποτακτ.Συνδέσμων\ στο\ Κείμενο}{Πλήθος\ Εμφανίσεων\ Όλων\ των\ Λέξεων\ του\ Κειμένου}$$

6.5.3 Τρίτη Κατηγορία υφομετρικών χαρακτηριστικών: από το λημματοποιημένο σώμα κειμένων (Lemmatized Corpus)

Όσον αφορά τις λειτουργικές λέξεις, υπάρχει ασάφεια του ορισμού τους και η επιστημονική κοινότητα δεν συμφωνεί για το ποιες ακριβώς λέξεις χαρακτηρίζονται ως λειτουργικές. Έτσι, διαφορετικές μελέτες χρησιμοποιούν διαφορετικό αριθμό λειτουργικών λέξεων. Πρέπει, λοιπόν, να διευκρινίσουμε ότι στην παρούσα διατριβή χρησιμοποιήσαμε τις 635 λειτουργικές λέξεις που έχουν καταγραφεί για τον Μορφολογικό Αναλυτή του Ινστιτούτου Επεξεργασίας του Λόγου στις οποίες περιλαμβάνονται άρθρα, αντωνυμίες, προθέσεις, επιρρήματα, σύνδεσμοι, βοηθητικά ρήματα, μόρια και επιφωνήματα.

Εξαιτίας της υψηλής συχνότητάς τους στη γλώσσα και του γραμματικοποιημένου τους ρόλου (έχουν περιορισμένη λεξιλογική σημασία) είναι δύσκολο να ελεγχθεί η χρήση των λειτουργικών λέξεων συνειδητά από τον συγγραφέα. Συγχρόνως, επειδή ποικίλει η συχνότητα των διάφορων λειτουργικών λέξεων ανά συγγραφέα ή ανά κειμενικό γένος, θεωρείται ότι υπάρχει συσχέτισή τους με το συγγραφικό ύφος. Γι' αυτό το λόγο χρησιμοποιούνται οι λειτουργικές λέξεις ως υφομετρικό χαρακτηριστικό σε έρευνες αναγνώρισης του ψυχολογικού προφίλ του συγγραφέα (Luyckx & Daelemans, 2008a, 2008b; Iacobelli et al., 2011. Κατά τους Chung & Pennebaker (2007) οι λειτουργικές λέξεις είναι φορείς ψυχολογικής σημασιολογίας και ρυθμίζουν την κοινωνική διάδραση των χρηστών τους. Οι Argamon et al. (2005) και Argamon et al. (2007) χρησιμοποίησαν ως υφομετρικό χαρακτηριστικό 675 λειτουργικές λέξεις, το οποίο λειτούργησε ως το αποτελεσματικότερο για την πρόβλεψη της Εξωστρέφειας. Αντίθετα, σε άλλη έρευνα (Qiu et al., 2012) βρέθηκε ότι οι εξωστρεφείς χρησιμοποιούν λιγότερες λειτουργικές λέξεις.

Σε αντίθεση με τις λειτουργικές λέξεις, οι λέξεις περιεχομένου επιλέγονται συνειδητά από τον συγγραφέα και θα υποθέταμε ότι δεν εντάσσονται στα υφομετρικά χαρακτηριστικά με μεγάλη προβλεπτικότητα για τον χαρακτήρα του. Ωστόσο, οι μελέτες δείχνουν ότι πρόκειται για αξιόπιστο υφομετρικό δείκτη. Κατά τη μελέτη γραπτών εκθέσεων που γράφτηκαν από φοιτητές της Ψυχολογίας (Pennebaker & King, 1999) διαπιστώθηκε η συσχέτιση ανάμεσα σε γλωσσικά χαρακτηριστικά και τα χαρακτηριστικά προσωπικότητας του μοντέλου των Πέντε Παραγόντων. Οι

εξωστρεφείς χρησιμοποίησαν περισσότερες λέξεις κοινωνικού περιεχομένου και ο λόγος όσων είχαν αυξημένο το χαρακτηριστικό της ευσυνειδησίας συσχετίζεται θετικά με τη χρήση θετικών συναισθηματικά λέξεων και λιγότερων αρνήσεων.

Βασικό υφομετρικό χαρακτηριστικό του λεξιλογικού τομέα είναι η συχνότητα των λέξεων σε ένα κείμενο. Μια λίστα λεξιλογικής συχνότητας περιλαμβάνει λέξεις σε φθίνουσα σειρά με πρώτη λέξη την πιο συχνή. Όπως έχει ήδη αναφερθεί στο τρίτο κεφάλαιο, σύμφωνα με τον νόμο του Zipf σε ένα ηλεκτρονικό σώμα κειμένων η συχνότητα οποιουδήποτε λεξικού τύπου φυσικής γλώσσας είναι αντιστρόφως ανάλογη με την κατάταξή της στη λίστα λεξιλογικής συχνότητας. Για την αναγνώριση προσωπικότητας του συγγραφέα χρησιμοποιούνται ως χαρακτηριστικά οι πιο συχνές λέξεις (Coniam, 2004; Ramos dos Santos & Paraboni, 2018; Yamada, Sasano & Takeda, 2019). Στην παρούσα διατριβή, λοιπόν, εξετάζουμε τη συχνότητα εμφάνισης στο σώμα κειμένων που αναπτύξαμε και των λειτουργικών και των λέξεων περιεχομένου

Η έρευνα, βέβαια, στο πλαίσιο των λεξιλογικών χαρακτηριστικών έχει επεκταθεί εκτός της χρήσης των λεξικών συχνοτήτων ως υφομετρικών χαρακτηριστικών και σε άλλους σύνθετους δείκτες, οι οποίοι είναι ευρύτερα γνωστοί στη βιβλιογραφία ως δείκτες λεξιλογικού «πλούτου» ή δείκτες λεξιλογικής διαφοροποίησης. Οι συγκεκριμένοι δείκτες προσπαθούν με μία αριθμητική τιμή να περιγράψουν το πόσο διαφορετικό λεξιλόγιο χρησιμοποιεί ένας συγγραφέας στα έργα του και επομένως να προσεγγίσουν τον βαθμό του λεξιλογικού «πλούτου» που τον χαρακτηρίζει (Μικρός, 2015a: 78). Από αυτούς βάσει βιβλιογραφίας επιλέξαμε να υπολογίσουμε τα άπαξ λεγόμενα, τα δις λεγόμενα, τον λόγο των δις προς άπαξ και τη λειτουργική πυκνότητα, τον λόγο δηλαδή των λέξεων περιεχομένου προς τις λειτουργικές λέξεις (García & Martín, 2007).

25. **Σχετική συχνότητα εμφάνισης των λειτουργικών λέξεων σε ένα κείμενο. (PercentageOfAllStopWordsCoverageInFile)**

    Προσμετρώνται, όπου υπάρχουν, όλοι οι κλιτικοί τύποι της κάθε λειτουργικής λέξης που απαντά στο κείμενο.

$$Σχετ.Συχν.Λειτουργικών\ Λέξεων\ σε\ Κείμενο = 100 \cdot \frac{Πλήθος\ Εμφανίσεων\ Όλων\ των\ Λειτ.Λέξεων\ στο\ Κείμενο}{Πλήθος\ Εμφανίσεων\ Όλων\ των\ Λέξεων\ του\ Κειμένου}$$

26. **Σχετική συχνότητα εμφάνισης των πιο συχνών λέξεων μέσα σε ένα κείμενο. (PercentageOfTopMostFreqTokensCoverageInFile)**

    Πρώτα υπολογίζεται η συχνότητα εμφάνισης τής κάθε λέξης τού υπό επεξεργασία κειμένου. Οι συχνότητες αυτές διατάσσονται κατά φθίνουσα σειρά. Το 10% των κορυφαίων τιμών της διατεταγμένης αυτής σειράς συχνοτήτων αθροίζεται, το δε αποτέλεσμα δίνει το πλήθος των εμφανίσεων των συχνότερων λέξεων του κειμένου. Το πλήθος αυτό διαιρείται με το σύνολο των εμφανίσεων όλων των λέξεων του κειμένου και το αποτέλεσμα πολλαπλασιάζεται επί 100 για να δώσει την σημειούμενη συχνότητα. Προσμετρώνται όλοι οι κλιτικοί τύποι της κάθε λέξης που απαντά στο κείμενο.

$$Σχετ.Συχν.Συχνότερων\ Λέξεων\ σε\ Κείμενο = 100 \cdot \frac{Πλήθος\ Εμφανίσεων\ Συχνότερων\ Λέξεων\ στο\ Κείμενο}{Πλήθος\ Εμφανίσεων\ Όλων\ των\ Λέξεων\ του\ Κειμένου}$$

27. **Σχετική συχνότητα εμφάνισης των πιο συχνών μη λειτουργικών λέξεων μέσα σε ένα κείμενο.**

    (PercentageOfTopMostFreqNonStopWordsCoverageInFile)

    Πρώτα υπολογίζεται η συχνότητα εμφάνισης τής κάθε μη λειτουργικής λέξης του υπό επεξεργασία κειμένου. Οι συχνότητες αυτές διατάσσονται κατά φθίνουσα σειρά. Το 10% των κορυφαίων τιμών τής διατεταγμένης αυτής σειράς συχνοτήτων αθροίζεται, το δε αποτέλεσμα δίνει το πλήθος των εμφανίσεων των συχνότερων μη λειτουργικών λέξεων του κειμένου. Το πλήθος αυτό διαιρείται με το σύνολο των εμφανίσεων όλων των μη λειτουργικών λέξεων του κειμένου και το αποτέλεσμα πολλαπλασιάζεται επί 100 για να δώσει την σημειούμενη συχνότητα. Προσμετρώνται όλοι οι κλιτικοί τύποι της κάθε μη λειτουργικής λέξης που απαντά στο κείμενο.

$$Σχετ.Συχν.Συχνότερων\ Μη\ Λειτ.Λέξεων\ σε\ Κείμενο = 100 \cdot \frac{Πλήθος\ Εμφανίσεων\ Συχν.Μη\ Λειτ.Λέξεων\ στο\ Κείμενο}{Πλήθος\ Εμφανίσεων\ Όλων\ των\ Μη\ Λειτ.Λέξεων\ του\ Κειμ.}$$

28. **Σχετική συχνότητα εμφάνισης των πιο σπάνιων λέξεων μέσα σε ένα κείμενο.**

    (PercentageOfBottomLeastFreqTokensCoverageInFile)

    Πρώτα υπολογίζεται η συχνότητα εμφάνισης τής κάθε λέξης του υπό επεξεργασία κειμένου και δημιουργείται ένας πίνακας που στην μία στήλη του περιλαμβάνονται οι λέξεις του κειμένου (κλιτικοί τύποι του ίδιου λήμματος εκλαμβάνονται ως διαφορετικές λέξεις) και στην άλλη στήλη του περιέχεται η συχνότητα εμφάνισης τής κάθε λέξης. Από τον πίνακα αυτόν επιλέγονται οι συχνότητες εμφάνισης που ισούνται με 1 ή 2 και αυτές αθροίζονται. Το αποτέλεσμα δίνει το πλήθος των εμφανίσεων των σπανιότερων λέξεων του κειμένου. Το πλήθος αυτό διαιρείται με το σύνολο των εμφανίσεων όλων των λέξεων του κειμένου και το αποτέλεσμα πολλαπλασιάζεται επί 100 για να δώσει την σημειούμενη συχνότητα.

$$Σχετ.Συχν.Σπανιότερων\ Λέξεων\ σε\ Κείμενο = 100 \cdot \frac{Πλήθος\ Εμφανίσεων\ Σπανιότερων\ Λέξεων\ στο\ Κείμενο}{Πλήθος\ Εμφανίσεων\ Όλων\ των\ Λέξεων\ του\ Κειμένου}$$

29. **Σχετική συχνότητα εμφάνισης των πιο σπάνιων μη λειτουργικών λέξεων μέσα σε ένα κείμενο.**

    (PercentageOfBottomLeastFreqNonStopWordsCoverageInFile)

    Πρώτα υπολογίζεται η συχνότητα εμφάνισης τής κάθε μη λειτουργικής λέξης του υπό επεξεργασία κειμένου και δημιουργείται ένας πίνακας που στην μία στήλη του περιλαμβάνονται οι μη λειτουργικές λέξεις του κειμένου (κλιτικοί τύποι του ίδιου λήμματος εκλαμβάνονται ως διαφορετικές λέξεις) και στην άλλη στήλη του περιέχεται η συχνότητα εμφάνισης τής κάθε μιας μη λειτουργικής λέξης. Από τον πίνακα αυτόν επιλέγονται οι συχνότητες εμφάνισης που ισούνται με 1 ή 2 και αυτές αθροίζονται. Το αποτέλεσμα δίνει το πλήθος των εμφανίσεων των σπανιότερων μη λειτουργικών λέξεων του κειμένου. Το πλήθος αυτό διαιρείται με το σύνολο των εμφανίσεων όλων των μη λειτουργικών λέξεων του κειμένου και το αποτέλεσμα πολλαπλασιάζεται επί 100 για να δώσει την σημειούμενη συχνότητα.

$$Σχετ.Συχν.Σπανιότερων\ Μη\ Λειτ.Λέξεων\ σε\ Κείμενο = 100 \cdot \frac{Πλήθος\ Εμφανίσεων\ Σπανιότερων\ Μη\ Λειτ.Λέξ.στο\ Κείμ.}{Πλήθος\ Εμφανίσεων\ Όλων\ των\ Μη\ Λειτ.Λέξεων\ του\ Κειμ.}$$

30. **Σχετική συχνότητα μη λειτουργικών λέξεων μέσα σε ένα κείμενο. (NumOfSingleNonStopWordsPerAllWordsOccurrencesInFile)**

    Πρώτα δημιουργείται ένας πίνακας που στην μία στήλη του περιλαμβάνονται οι μη λειτουργικές λέξεις του κειμένου (κλιτικοί τύποι του ίδιου λήμματος εκλαμβάνονται ως διαφορετικές λέξεις) και στην άλλη στήλη του περιέχεται η συχνότητα εμφάνισης τής κάθε μιας μη λειτουργικής λέξης. Το πλήθος των γραμμών αυτού του πίνακα δίνει το πλήθος των μη λειτουργικών λέξεων του κειμένου, δηλ., το πόσες είναι οι μη λειτουργικές λέξεις μέσα στο κείμενο, όπου οι κλιτικοί τύποι μιας μη λειτουργικής λέξης λογίζονται ως διαφορετικές μη λειτουργικές λέξεις. Με άλλα λόγια, το πλήθος των μη λειτουργικών λέξεων υπολογίζεται μετρώντας μία μόνο φορά την κάθε μία μη λειτουργική λέξη, είτε αυτή απαντά ως λήμμα είτε ως κλιτικός τύπος. Το πλήθος αυτό διαιρείται με το σύνολο των εμφανίσεων όλων των λέξεων του κειμένου και το αποτέλεσμα πολλαπλασιάζεται επί 100 για να δώσει την σημειούμενη συχνότητα.

$$Σχετ. Συχν. Μη\ Λειτουργικών\ Λέξεων\ σε\ Κείμενο = 100 \cdot \frac{Πλήθος\ Όλων\ των\ Μη\ Λειτουργικών\ Λέξεων\ στο\ Κείμενο}{Πλήθος\ Εμφανίσεων\ Όλων\ των\ Λέξεων\ του\ Κειμένου}$$

31. **Σχετική συχνότητα των άπαξ εμφανιζομένων λέξεων μέσα σε ένα κείμενο. (PercentageOfTokensAppearingOnceCoverageInFile)**

    Πρώτα υπολογίζεται πόσες είναι οι λέξεις του κειμένου που απαντούν σε αυτό μία μόνο φορά. Οι κλιτικοί τύπου του ίδιου λήμματος λογίζονται ως διαφορετικές λέξεις. Το πλήθος αυτών των άπαξ εμφανιζομένων λέξεων διαιρείται με το σύνολο των εμφανίσεων όλων των λέξεων του κειμένου και το αποτέλεσμα πολλαπλασιάζεται επί 100 για να δώσει την σημειούμενη συχνότητα.

$$Σχετ. Συχν. των\ Άπαξ\ Εμφανιζομένων\ Λέξεων\ σε\ Κείμενο = 100 \cdot \frac{Πλήθος\ Άπαξ\ Εμφανιζομένων\ Λέξεων\ στο\ Κείμενο}{Πλήθος\ Εμφανίσεων\ Όλων\ των\ Λέξεων\ του\ Κειμένου}$$

32. **Σχετική συχνότητα των δις εμφανιζομένων λέξεων μέσα σε ένα κείμενο. (PercentageOfTokensAppearingTwiceCoverageInFile)**

    Πρώτα υπολογίζεται πόσες είναι οι λέξεις του κειμένου που απαντούν σε αυτό δύο μόνο φορές. Οι κλιτικοί τύπου του ίδιου λήμματος λογίζονται ως διαφορετικές λέξεις. Το πλήθος αυτών των δις εμφανιζομένων λέξεων επί δύο διαιρείται με το σύνολο των εμφανίσεων όλων των λέξεων του κειμένου και το αποτέλεσμα πολλαπλασιάζεται επί 100 για να δώσει την σημειούμενη συχνότητα.

$$Σχετ. Συχν. των Δις Εμφανιζομένων Λέξεων σε Κείμενο = 100 \cdot \frac{(Πλήθος\ Δις\ Εμφανιζομένων\ Λέξεων\ στο\ Κείμενο) \cdot 2}{Πλήθος\ Εμφανίσεων\ Όλων\ των\ Λέξεων\ του\ Κειμένου}$$

33. **Λόγος των δις προς άπαξ λεγόμενα. (RatioOfTwiceOverOnceAppearingTokens)**

$$Λόγος\ των\ Δις\ προς\ των\ Άπαξ\ Εμφανιζομ. Λέξ. σε\ Κείμ. = 100 \cdot \frac{(Πλήθος\ Δις\ Εμφανιζομένων\ Λέξεων\ στο\ Κείμ.) \cdot 2}{Πλήθος\ Άπαξ\ Εμφανιζομένων\ Λέξεων\ στο\ Κείμ.}$$

34. **Σχετική συχνότητα εμφάνισης των μη λειτουργικών λέξεων μέσα σε ένα κείμενο. (PercentageOfAllNonStopWordsCoverageInFile)**

    Πρώτα υπολογίζεται η συχνότητα εμφάνισης τής κάθε μη λειτουργικής λέξης τού υπό επεξεργασία κειμένου. Οι συχνότητες αυτές αθροίζονται, το δε αποτέλεσμα δίνει το πλήθος των εμφανίσεων των όλων των μη λειτουργικών λέξεων του κειμένου. Το πλήθος αυτό διαιρείται με το σύνολο των εμφανίσεων όλων των λέξεων του κειμένου και το αποτέλεσμα πολλαπλασιάζεται επί 100 για να δώσει την σημειούμενη συχνότητα. Προσμετρώνται όλοι οι κλιτικοί τύποι τους κάθε μη λειτουργικής λέξης που απαντά στο κείμενο.

$$Σχετ. Συχν. των\ Μη\ Λειτ. Λέξεων σε\ Κείμενο = 100 \cdot \frac{Πλήθος\ Εμφανίσεων\ των\ \ Μη\ Λειτ. Λέξεων\ στο\ Κείμενο}{Πλήθος\ Εμφανίσεων\ Όλων\ των\ Λέξεων\ του\ Κειμένου}$$

35. **Λόγος των λέξεων περιεχομένου προς τις λειτουργικές (λειτουργική πυκνότητα).**
    (FunctionalDensity)

$$Λειτουργική\ Πυκνότητα\ σε\ Κείμενο = 100 \cdot \frac{Πλήθος\ Εμφανίσεων\ Όλων\ των\ \ Μη\ Λειτ.\ Λέξεων\ στο\ Κείμενο}{Πλήθος\ Εμφανίσεων\ Όλων\ των\ Λειτ.\ Λέξεων\ του\ Κειμένου}$$

6.6 Η πρόβλεψη προσωπικότητας ως πρόβλημα δυαδικής ταξινόμησης

Ένα ζήτημα που προκύπτει κατά την υλοποίηση συστήματος μηχανικής αναγνώρισης της προσωπικότητας από κείμενο είναι η θεώρηση ή η αγνόηση της ουδέτερης κλάσης/κατηγορίας. Με τον όρο ουδέτερη κλάση τόσο στο Ερωτηματολόγιο Τύπων Προσωπικότητας Myers-Briggs Type Indicator (MBTI) όσο και στο ερωτηματολόγιο των Πέντε Παραγόντων εννοούμε, για κάθε τύπο ή χαρακτηριστικό αντίστοιχα, την κατηγορία στην οποία εντάσσονται τα κείμενα των μαθητών που δεν ανήκουν ούτε στη θετική ούτε στην αρνητική κλάση. Διευκρινίζουμε, λοιπόν, ότι στην παρούσα διατριβή το πρόβλημα της πρόβλεψης της προσωπικότητας αντιμετωπίστηκε ως δυαδική (binary) ταξινόμηση των τεσσάρων διαστάσεων και των πέντε χαρακτηριστικών προσωπικότητας, διότι είναι η πιο συχνή προσέγγιση βάσει βιβλιογραφίας. Επιπλέον, διότι κρίναμε πως για να είναι η πρόβλεψη έγκυρη τα υφομετρικά χαρακτηριστικά που εξήχθησαν έπρεπε να εφαρμοστούν σε κείμενα των οποίων οι συγγραφείς ανήκουν σαφώς σε μια θετική ή αρνητική κατηγορία. Έτσι, αφήσαμε εκτός έρευνας τα κείμενα των μαθητών, οι οποίοι βάσει των αποτελεσμάτων των δύο ερωτηματολογίων κατατάχθηκαν στην ενδιάμεση κατηγορία (π.χ. Ούτε εξωστρεφής ούτε εσωστρεφής). Ακολουθούν οι πίνακες που παραθέτουν αναλυτικά τις κατηγορίες τους διαμορφώθηκαν με βάση τα δύο ερωτηματολόγια καθώς και το πλήθος των εκθέσεων ανά κατηγορία και τον αριθμό των λέξεών τους.

Πίνακας 9: Κατηγοριοποίηση του σώματος κειμένων με βάση το Ερωτηματολόγιο Τύπων Προσωπικότητας Myers-Briggs Type Indicator (MBTI).

Τύπος Προσωπικότητας	Πλήθος εκθέσεων	Σύνολο Λέξεων
Extraverted	422	185.696
Introverted	103	44.948
Neither Introverted Nor Extraverted	45	19.223

Τύπος Προσωπικότητας	Πλήθος εκθέσεων	Σύνολο Λέξεων
iNtuitive	331	151.396
Sensing	141	59.268
Neither iNtuitive Nor Sensing	98	39.203

Τύπος Προσωπικότητας	Πλήθος εκθέσεων	Σύνολο Λέξεων
Feeling	321	140.224
Thinking	184	80.393
Neither Thinking Nor Feeling	65	29.250

Τύπος Προσωπικότητας	Πλήθος εκθέσεων	Σύνολο Λέξεων
Judging	364	160.827
Perceiving	149	64.032
Neither Perceiving Nor Judging	57	25.008

Πίνακας 10: Κατηγοριοποίηση του σώματος κειμένων με βάση το Ερωτηματολόγιο του μοντέλου των Πέντε Παραγόντων.

Χαρακτηριστικό Προσωπικότητας	Πλήθος εκθέσεων	Σύνολο Λέξεων
Open to New Experiences	83	94.830
Closed-Minded	364	154.278
Not typically open to new experiences	120	54.175

Χαρακτηριστικό Προσωπικότητας	Πλήθος εκθέσεων	Σύνολο Λέξεων
Conscientious	204	91.367
Disorganized	210	87.039
Neither Disorganized or Conscientious	153	70.702

Χαρακτηριστικό Προσωπικότητας	Πλήθος εκθέσεων	Σύνολο Λέξεων
Extraverted	157	68.369
Introverted	212	96.654
Neither Introverted or Extraverted	198	84.085

Χαρακτηριστικό Προσωπικότητας	Πλήθος εκθέσεων	Σύνολο Λέξεων
Agreeable	271	121.166
Disagreeable	164	68.221
Neither Disagreeable or Agreeable	132	59.721

Χαρακτηριστικό Προσωπικότητας	**Πλήθος εκθέσεων**	**Σύνολο Λέξεων**
Nervous	216	97.052
Calm	224	96.490
Neither Calm or Nervous	127	55.566

Κεφάλαιο 7

Αποτελέσματα

Σε αυτό το κεφάλαιο θα εξεταστεί η πειραματική διαδικασία που ακολουθήθηκε με σκοπό την αυτόματη κατηγοριοποίηση των εκθέσεων των μαθητών βάσει της προσωπικότητάς τους έτσι όπως αποτυπώθηκε στα ερωτηματολόγια προσωπικότητας που συμπλήρωσαν. Με άλλα λόγια, αξιολογούμε τους αλγόριθμους μηχανικής μάθησης ως προς την προβλεπτική τους ικανότητα χρησιμοποιώντας ως δεδομένα εκπαίδευσης και εν μέρει ελέγχου τις εκθέσεις των μαθητών, των οποίων η προσωπικότητα έχει ήδη αξιολογηθεί μέσω των ερωτηματολογίων (Myers-Briggs Type Indicator και Big Five).

Παρουσιάζονται τα αποτελέσματα των πειραμάτων ταξινόμησης με εφαρμογή των εννέα αλγορίθμων μηχανικής μάθησης μέσω της επέκτασης AutoModel του λογισμικού RapidMiner. Συγκεκριμένα, παρουσιάζεται τελευταίο στάδιο λειτουργίας του AutoModel, όπου φαίνονται τα μοντέλα που δημιουργήθηκαν και τα αποτελέσματά τους. Ανά πάσα στιγμή, παρόλο που τα μοντέλα τρέχουν στο παρασκήνιο, μπορεί κάποιος να ανοίξει το περιβάλλον της κάθε διεργασίας που παράγει ένα μοντέλο να το ελέγξει και να δει αναλυτικά τα αποτελέσματα.

7.1 Αποτελέσματα Αυτόματης Ταξινόμησης μαθητικών εκθέσεων (Ερωτηματολόγιο Myers-Briggs Type Indicator)

Στα υποκεφάλαια που ακολουθούν παρουσιάζονται τα ερευνητικά αποτελέσματα που αφορούν στο ερωτηματολόγιο Myers-Briggs Type Indicator (MBTI). Για κάθε τύπο προσωπικότητας δυαδικής μορφής παρατίθενται τρεις εικόνες: τα αποτελέσματα πρόβλεψης σε εκατοστιαίο ποσοστό των 219 μοντέλων από τους εννέα αλγόριθμους που εφαρμόστηκαν στα κειμενικά δεδομένα, ο πίνακας με τα μέτρα αξιολόγησης του αποτελεσματικότερου αλγόριθμου και τα βάρη (weights) που

επέδρασαν θετικά στην προβλεπτική του ικανότητα. Για την πρόβλεψη όλων των τύπων του συγκεκριμένου ερωτηματολογίου ο αλγόριθμος που είχε τα καλύτερα αποτελέσματα ήταν ο Naive Bayes. Τα ποσοστά ακρίβειας της πρόβλεψης του Naive Bayes κυμαίνονται από 69% έως 81%, με μέσο όρο 76,5%. Συγκεκριμένα, πέτυχε ακρίβεια 81% για την πρόβλεψη του τύπου της Εξωστρέφειας των μαθητών, 80% για την πρόβλεψη της Διαίσθησης, 69% του Συναισθήματος και 76% της Κρίσης. Ο Πίνακας 11 παρουσιάζει την απόδοση όσον αφορά στις μετρήσεις ακρίβειας, ορθότητας και ανάκλησης αυτού του ταξινομητή.

Πίνακας 11: Η απόδοση του Naive Bayes.

Τύπος προσωπικότητας	**Ακρίβεια**	**Ορθότητα**	**Ανάκληση**
Εξωστρέφεια	80,67%	80,54%	100%
Διαίσθηση	79,85%	81,31%	92,55%
Συναίσθημα	68,75%	67,69%	96,70%
Κρίση	75,68%	76,15%	95,19%

Πριν την παρουσίαση του βέλτιστου αλγόριθμου κρίνουμε απαραίτητη την παράθεση κάποιων από τις πολλές δεκάδες πειράματα που πραγματοποιήθηκαν μέχρι να επιτευχθεί το καλύτερο αποτέλεσμα. Σε πίνακες θα δούμε τα υφομετρικά χαρακτηριστικά που επιλέχθηκαν ανά τύπο προσωπικότητας, το υψηλότερο ποσοστό ακρίβειας και το όνομα του αλγόριθμου ή των αλγόριθμων που το πέτυχαν (Τα πλήρη ονόματα των χαρακτηριστικών στα Ελληνικά αναφέρονται στα υποκεφάλαια 6.5.1 έως 6.5.3). Η τελευταία σε κάθε πίνακα σειρά αναφέρεται στα χαρακτηριστικά που χρησιμοποιήθηκαν με τα οποία ο αλγόριθμος Naive Bayes απέδωσε καλύτερα.

Πίνακας 12: Πειράματα για πρόβλεψη Εξωστρέφειας.

Πρόβλεψη Εξωστρέφειας		
Υφομετρικά χαρακτηριστικά	**Ακρίβεια**	**Αλγόριθμος**

Πρόβλεψη Εξωστρέφειας		
Όλα τα χαρακτηριστικά	80%	Generalized Linear Model, Fast Large Margin, Random Forest, Support Vector Machine
Όλα χαρακτηριστικά του plain text corpus	80%	Generalized Linear Model, Logistic Regression, Fast Large Margin
Όλα τα χαρακτηριστικά του tagged corpus	80%	Generalized Linear Model, Logistic Regression, Fast Large Margin, Random Forest, Support Vector Machine
Όλα τα χαρακτηριστικά του lemmatized corpus	80%	Generalized Linear Model, Logistic Regression, Fast Large Margin, Deep learning, Random Forest, Support Vector Machine
AverageWordLength, AverageSentenceLength, PercentageOfTopMostFreqBiGrams, PercentageOfTopMostFreqTriGrams	80%	Naive Bayes, Generalized Linear Model, Logistic Regression, Fast Large Margin, Deep Learning, Decision Tree, Random Forest, Support Vector Machine
VerbsFreq,ActiveVoiceVerbsFreq, PassiveVoiceVerbsFreq	81%	Deep learning
VerbsFreq, ActiveVoiceVerbsFreq, PassiveVoiceVerbsFreq, PercentageOfBottomLeastFreqTokens	81%	Gradient Boosted Trees
PersonalPronounsFreq, ActiveVoiceVerbsFreq, AverageSentenceLength,	81%	Deep learning

Πρόβλεψη Εξωστρέφειας		
PercentageOfTopMostFreqNonStopWords		
PersonalPronounsFreq, ActiveVoiceVerbsFreq, AverageSentenceLength, PercentageOfTopMostFreqNonStopWords, PercentageOfTokensAppearingTwice	***81%***	***Naive Bayes***

Πίνακας 13: Πειράματα για πρόβλεψη Διαίσθησης.

Πρόβλεψη Διαίσθησης		
Υφομετρικά χαρακτηριστικά	**Ακρίβεια**	**Αλγόριθμος**
Όλα τα χαρακτηριστικά	74%	Fast Large Margin
Όλα χαρακτηριστικά του plain text corpus	75%	Gradient Boosted Trees
Όλα τα χαρακτηριστικά του tagged corpus	73%	Deep learning, Support Vector Machine
Όλα τα χαρακτηριστικά του lemmatized corpus	72%	Generalized Linear Model, Logistic Regression
AverageWordLength, AverageSentenceLength, PercentageOfTopMostFreqBiGrams, PercentageOfTopMostFreqTriGrams	77%	Deep Learning

Πρόβλεψη Διαίσθησης		
Υφομετρικά χαρακτηριστικά	**Ακρίβεια**	**Αλγόριθμος**
Όλα τα χαρακτηριστικά του lemmatized corpus, AverageWordLength, AverageSentenceLength, PercentageOfTopMostFreqBiGrams, PercentageOfTopMostFreqTriGrams	73%	Naive Bayes
Όλα τα χαρακτηριστικά του tagged corpus, PercentageOfTopMostFreqTokens+PercentageOfTopMostFreqNonStopWords+ PercentageOfBottomLeastFreqTokens+PercentageOfBottomLeastFreqNonStopWords+PercentageOfTokensAppearingOnce	74%	Logistic Regression
Όλα τα χαρακτηριστικά του tagged corpus (-VerbsFreq, PassiveVoiceVerbsFreq, PersonalAndPossessivePronounsFreq), PercentageOfBottomLeastFreqTokens+PercentageOfBottomLeastFreqNonStopWords+PercentageOfTokensAppearingOnce	75%	Generalized Linear Model, Logistic Regression
Όλα τα χαρακτηριστικά του tagged corpus (-VerbsFreq, PassiveVoiceVerbsFreq,	76%	Deep learning

Πρόβλεψη Διαίσθησης		
Υφομετρικά χαρακτηριστικά	**Ακρίβεια**	**Αλγόριθμος**
PersonalAndPossessivePronounsFreq), AverageWordLength, AverageSentenceLength, PercentageOfTopMostFreqBiGrams,PercentageOfTopMostFreqTriGrams,PercentageOfBottomLeastFreqTokens+PercentageOfBottomLeastFreqNonStopWords+PercentageOfTokensAppearingOnce		
AverageWordLength, AverageSentenceLength, PercentageOfTopMostFreqBiGrams,PercentageOfTopMostFreqTriGrams, PercentageOfTokensAppearingOnce	76%	Naive Bayes
AverageWordLength, AverageSentenceLength, PercentageOfTopMostFreqBiGrams,PercentageOfTopMostFreqTriGrams, NumOfSingleNonStopWordsPerAllWords	77%	Naive Bayes
AverageWordLength, AverageSentenceLength, PercentageOfTopMostFreqBiGrams,PercentageOfTopMostFreqTriGrams,	78%	Naive Bayes

Πρόβλεψη Διαίσθησης		
Υφομετρικά χαρακτηριστικά	**Ακρίβεια**	**Αλγόριθμος**
NumOfSingleNonStopWordsPerAllWords, PercentageOfBottomLeastFreqTokens		
AverageWordLength, PercentageOfTopMostFreqBiGrams,PercentageOfTopMostFreqTriGrams, PercentageOfTopMostFreqCharTriGrams, NumOfSingleNonStopWordsPerAllWords, PercentageOfTokensAppearingOnce, PercentageOfBottomLeastFreqTokens, PersonalPronounsFreq	79%	Naive Bayes
AverageWordLength, PercentageOfTopMostFreqBiGrams,PercentageOfTopMostFreqTriGrams, PercentageOfTopMostFreqCharTriGrams, PersonalPronounsFreq, NumOfSingleNonStopWordsPerAllWords, PercentageOfTokensAppearingOnce, PercentageOfBottomLeastFreq	***80%***	***Naive Bayes***

Πρόβλεψη Διαίσθησης		
Υφομετρικά χαρακτηριστικά	**Ακρίβεια**	**Αλγόριθμος**
Tokens,PercentageOfAllNonStopWords		

Πίνακας 14: Πειράματα για πρόβλεψη Συναισθήματος.

Πρόβλεψη Συναισθήματος		
Υφομετρικά χαρακτηριστικά	**Ακρίβεια**	**Αλγόριθμος**
Όλα τα χαρακτηριστικά	63%	Generalized Linear Model, Logistic Regression, Decision Tree, Random Forest, Gradient Boosted Trees, Support Vector Machine
Όλα χαρακτηριστικά του plain text corpus	64%	Deep learning
Όλα τα χαρακτηριστικά του tagged corpus	63%	Fast Large Margin, Decision Tree, Random Forest, Support Vector Machine
Όλα τα χαρακτηριστικά του lemmatized corpus	63%	Generalized Linear Model, Logistic Regression, Fast Large Margin, Random Forest, Decision Tree, Support Vector Machine
AverageWordLength, AverageSentenceLength, PercentageOfTopMostFreqBiGrams,	66%	Generalized Linear Model, Logistic Regression

Πρόβλεψη Συναισθήματος		
Υφομετρικά χαρακτηριστικά	**Ακρίβεια**	**Αλγόριθμος**
PercentageOfTopMostFreqTriGrams		
Όλα χαρακτηριστικά του plain text corpus, Όλα τα χαρακτηριστικά του lemmatized corpus	64%	Gradient Boosted Trees
Όλα τα χαρακτηριστικά του lemmatized corpus, AverageWordLength, AverageSentenceLength, PercentageOfTopMostFreqBiGrams, PercentageOfTopMostFreqTriGrams	65%	Gradient Boosted Trees
AverageWordLength, AverageSentenceLength, PercentageOfTopMostFreqBiGrams, PercentageOfTopMostFreqTriGrams, AdjectivesFreq	65%	Naive Bayes, Generalized Linear Model, Logistic Regression
AverageWordLength, AverageSentenceLength, PercentageOfTopMostFreqBiGrams, PercentageOfTopMostFreqTriGrams, RatioOfTwiceOverOnceAppearing	66%	Generalized Linear Model, Logistic Regression

Πρόβλεψη Συναισθήματος		
Υφομετρικά χαρακτηριστικά	**Ακρίβεια**	**Αλγόριθμος**
PercentageOfTopMostFreqNonStopWords, VerbsFreq, NounsFreq, AdjectivesFreq	67%	Naive Bayes
PercentageOfTopMostFreqNonStopWords, VerbsFreq, NounsFreq, AdjectivesFreq, PersonalPronounsFreq	68%	Naive Bayes
PercentageOfTopMostFreqNonStopWords, VerbsFreq, NounsFreq, AdjectivesFreq, PersonalPronounsFreq, PassiveVoiceVerbsFreq	68%	Naive Bayes
VerbsFreq, NounsFreq, AdjectivesFreq, PersonalAndPossessivePronounsFreq, AdverbsFreq, PercentageOfTopMostFreqNonStopWords	***69%***	***Naive Bayes***

Πίνακας 15: Πειράματα για πρόβλεψη Κρίσης.

Πρόβλεψη Κρίσης		
Υφομετρικά χαρακτηριστικά	**Ακρίβεια**	**Αλγόριθμος**
Όλα τα χαρακτηριστικά	71%	Fast Large Margin, Random Forest, Support Vector Machine
Όλα χαρακτηριστικά του plain text corpus	73%	Naive Bayes

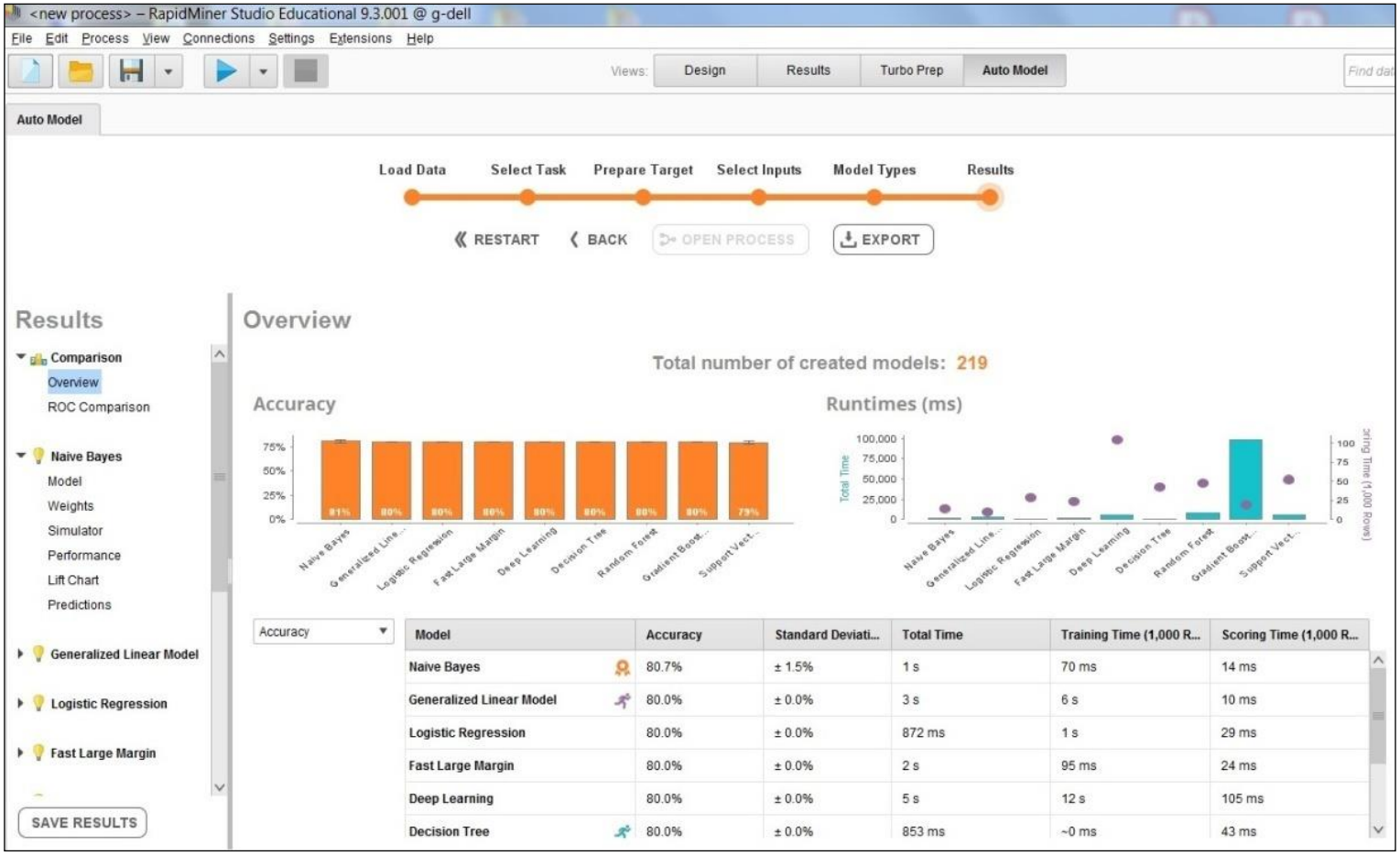
Πρόβλεψη Κρίσης		
Υφομετρικά χαρακτηριστικά	**Ακρίβεια**	**Αλγόριθμος**
Όλα τα χαρακτηριστικά του tagged corpus	71%	Generalized Linear Model, Logistic Regression, Fast Large Margin, Random Forest, Decision Tree, Support Vector Machine
Όλα τα χαρακτηριστικά του lemmatized corpus	71%	Generalized Linear Model, Logistic Regression, Fast Large Margin, Random Forest, Decision Tree, Gradient Boosted Trees, Support Vector Machine
Όλα τα χαρακτηριστικά του tagged corpus (-ActiveVoiceVerbsFreq, PassiveVoiceVerbsFreq, PersonalAndPossessivePronounsFreq), PercentageOfTopMostFreqNonStopWords+PercentageOfBottomLeastFreqNonStopWords+RatioOfTwiceOverOnceAppearingTokens	72%	Generalized Linear Model
Όλα τα χαρακτηριστικά του tagged corpus (-ActiveVoiceVerbsFreq, PassiveVoiceVerbsFreq, PersonalAndPossessivePronounsFreq), PercentageOfTopMostFreqNon	72%	Logistic Regression

Πρόβλεψη Κρίσης		
Υφομετρικά χαρακτηριστικά	**Ακρίβεια**	**Αλγόριθμος**
StopWords, PercentageOfBottomLeastFreqNonStopWords, RatioOfTwiceOverOnceAppearingTokens, PercentageOfBottomLeastFreqTokens		
Όλα τα χαρακτηριστικά του tagged corpus (-VerbsFreq, ActiveVoiceVerbsFreq, PassiveVoiceVerbsFreq, PersonalAndPossessivePronounsFreq), PercentageOfTopMostFreqNonStopWords, PercentageOfBottomLeastFreqNonStopWords, RatioOfTwiceOverOnceAppearingTokens, PercentageOfBottomLeastFreqTokens	72%	Generalized Linear Model, Logistic Regression
AverageWordLength, AverageSentenceLength, PercentageOfTopMostFreqBiGrams, PercentageOfTopMostFreqTriGrams	73%	Naive Bayes
AverageWordLength, PercentageOfTopMostFreqBiGrams,	74%	Naive Bayes

Πρόβλεψη Κρίσης		
Υφομετρικά χαρακτηριστικά	**Ακρίβεια**	**Αλγόριθμος**
PercentageOfTopMostFreqTriGrams, PronounsFreq		
Όλα χαρακτηριστικά του plain text corpus, ArticlesFreq	74%	Naive Bayes
AverageWordLength, PercentageOfTopMostFreqBiGrams, PercentageOfTopMostFreqTriGrams, PronounsFreq, PrepositionsFreq	75%	Naive Bayes
Όλα χαρακτηριστικά του plain text corpus, ArticlesFreq, PassiveVoiceVerbsFreq	75%	Naive Bayes
Όλα χαρακτηριστικά του plain text corpus, ArticlesFreq, PersonalAndPossessivePronounsFreq	***76%***	***Naive Bayes***

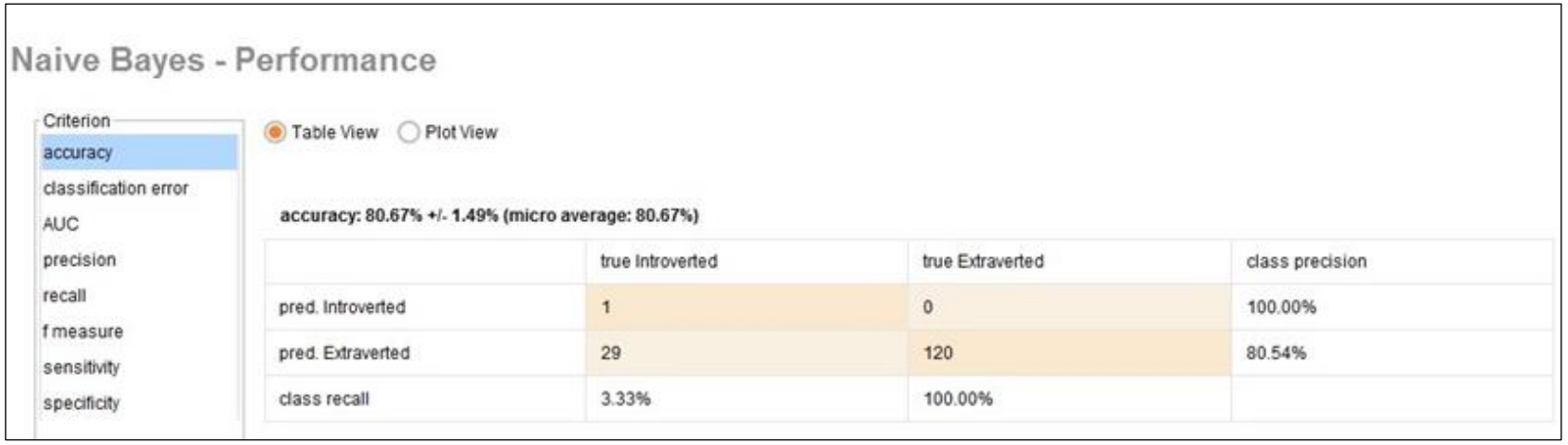
7.1.1 Εσωστρέφεια - Εξωστρέφεια

Στη γενική εικόνα αποτελεσμάτων του AutoModel αποτυπώνεται η ακρίβεια όλων των μοντέλων και ο χρόνος που χρειάστηκε το καθένα να εφαρμοστεί. Ο τύπος προσωπικότητας που επιλέχθηκε για πρόβλεψη είναι η Εξωστρέφεια, διότι οι περισσότεροι μαθητές του δείγματος είναι εξωστρεφείς και το λογισμικό προτείνει την επιλογή της κατηγορίας με τον μεγαλύτερο αριθμό δεδομένων. Παρατηρούμε ότι ο Naive Bayes αλγόριθμος πέτυχε το καλύτερο αποτέλεσμα, 81% σε ελάχιστο χρόνο (ένα δευτερόλεπτο). Οι υπόλοιποι αλγόριθμοι είχαν ακρίβεια 80% εκτός από τις Μηχανές Διανυσμάτων Υποστήριξης με 79%.



Εικόνα 6: Αποτελέσματα για Εξωστρέφεια.

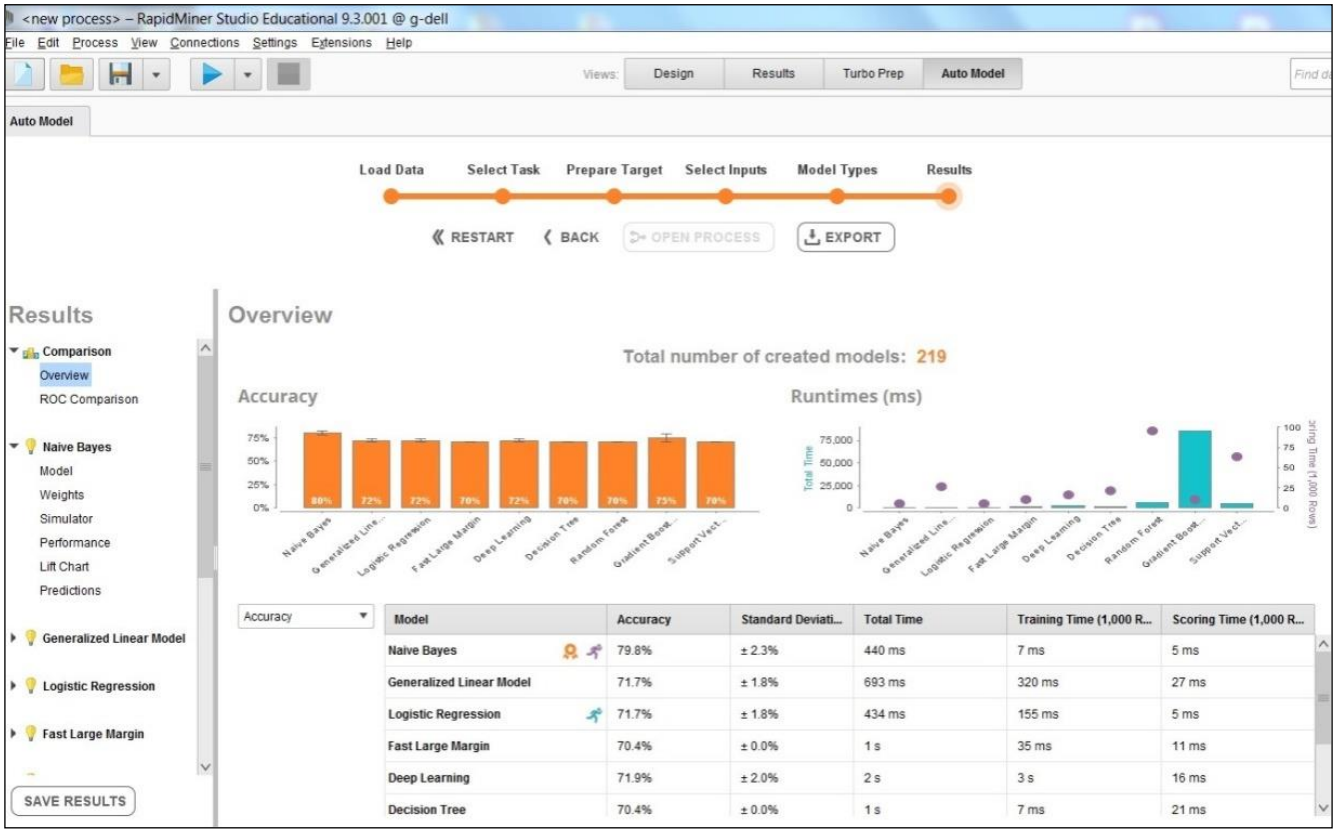
Ακολουθεί ο πίνακας με τις μετρικές αξιολόγησης του μοντέλου. Η ακρίβεια πρόβλεψης για τον τύπο της Εξωστρέφειας είναι 80,67%, η ορθότητα 80,54% και η ανάκληση 100%. Παρατηρούμε στον πίνακα πως η ικανότητα ανάκτησης κειμένων μόνο Εξωστρεφών μαθητών είναι 80,54%, αφού ο αλγόριθμος επέστρεψε και 29 κείμενα Εσωστρεφών μαθητών. Αντίθετα η ανάκληση, δηλαδή, η ικανότητα ανάκτησης όλων των κειμένων των Εξωστρεφών είναι 100%.



Εικόνα 7: Απόδοση για Εξωστρέφεια.

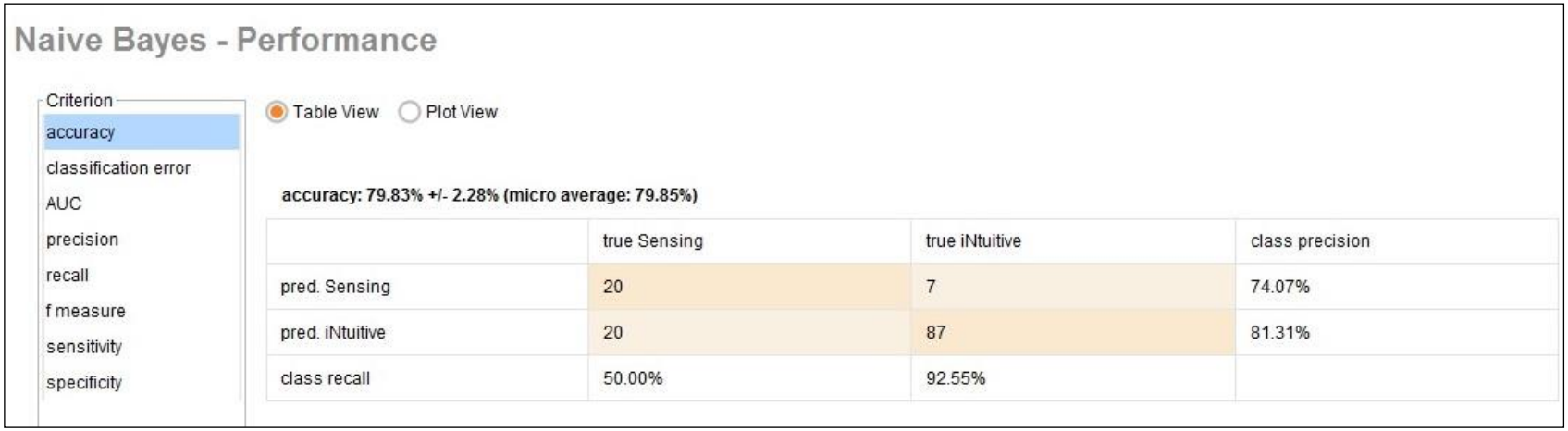
7.1.2 Διαίσθηση - Νόηση

Η Διαίσθηση, που επικράτησε έναντι της Νόησης στο συγκεκριμένο δείγμα μαθητών που είχαμε, προβλέφθηκε από τον Naive Bayes αλγόριθμο με ακρίβεια 80%. Το αμέσως χαμηλότερο ποσοστό ήταν το 75% από τον Gradient Boosted Trees, ενώ ακολουθούν τα υπόλοιπα μοντέλα με ακρίβεια από 72% έως 70% το ελάχιστο.



Εικόνα 8: Αποτελέσματα για Διαίσθηση.

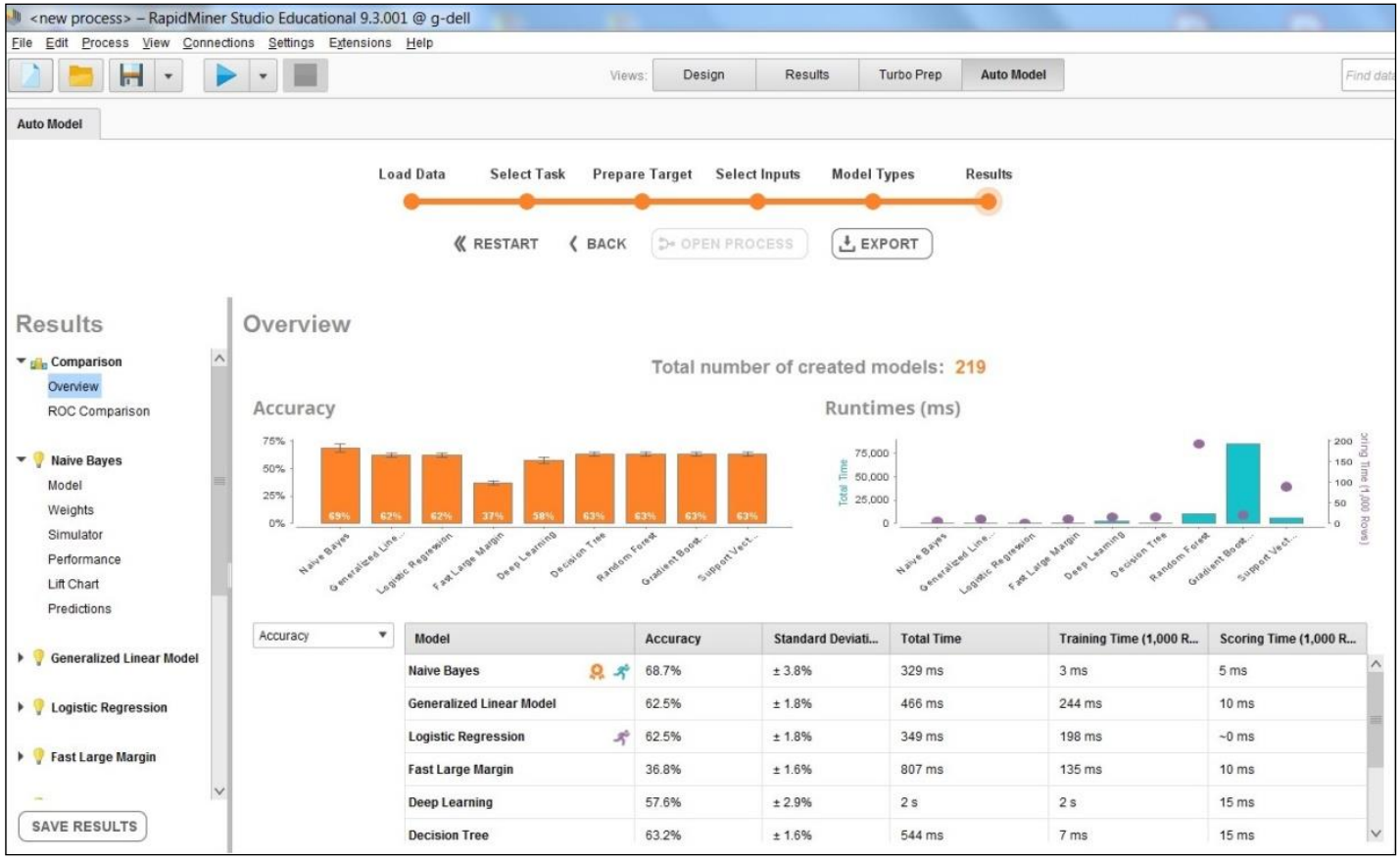
Όσον αφορά στις μετρικές αξιολόγησης του μοντέλου, η ακρίβεια πρόβλεψης για τον τύπο της Διαίσθησης είναι 79,83 %, η ορθότητα 81,31% και η ανάκληση 92,55%.



Εικόνα 9: Απόδοση για Διαίσθηση.

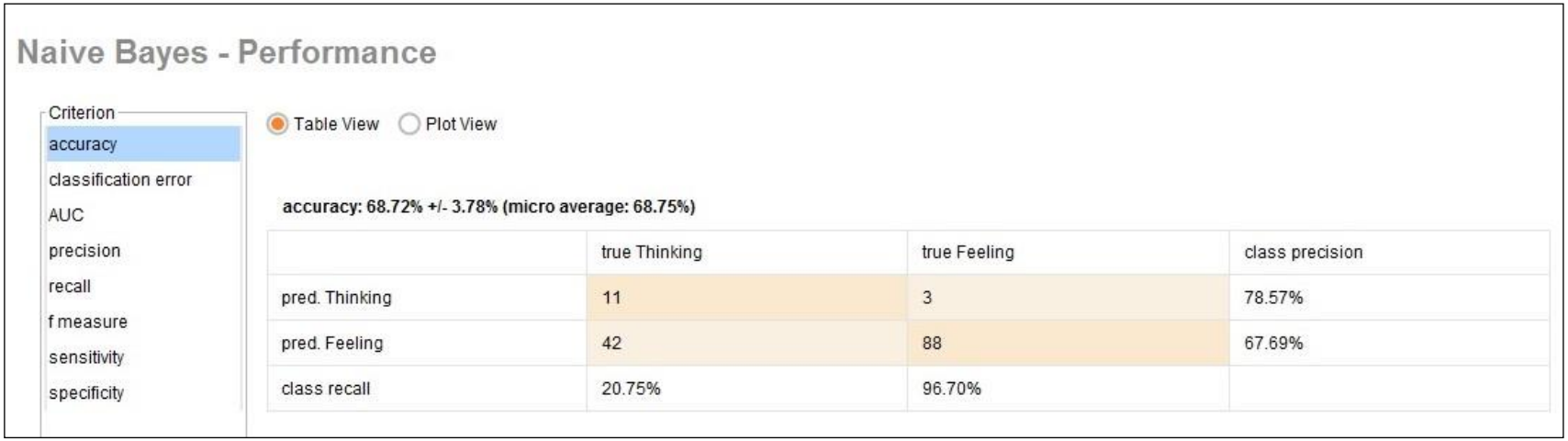
7.1.3 Σκέψη - Συναίσθημα

Ο Naive Bayes αλγόριθμος πέτυχε την πρόβλεψη του Συναισθήματος με ακρίβεια 69%, ποσοστό υψηλότερο σε σχέση με όσους ακολουθούν από 63% έως 37%.



Εικόνα 10: Αποτελέσματα για Συναίσθημα.

Ο αλγόριθμος αξιολογείται με 68,72% ακρίβεια, 67,69% ορθότητα και 96,70% ανάκληση.

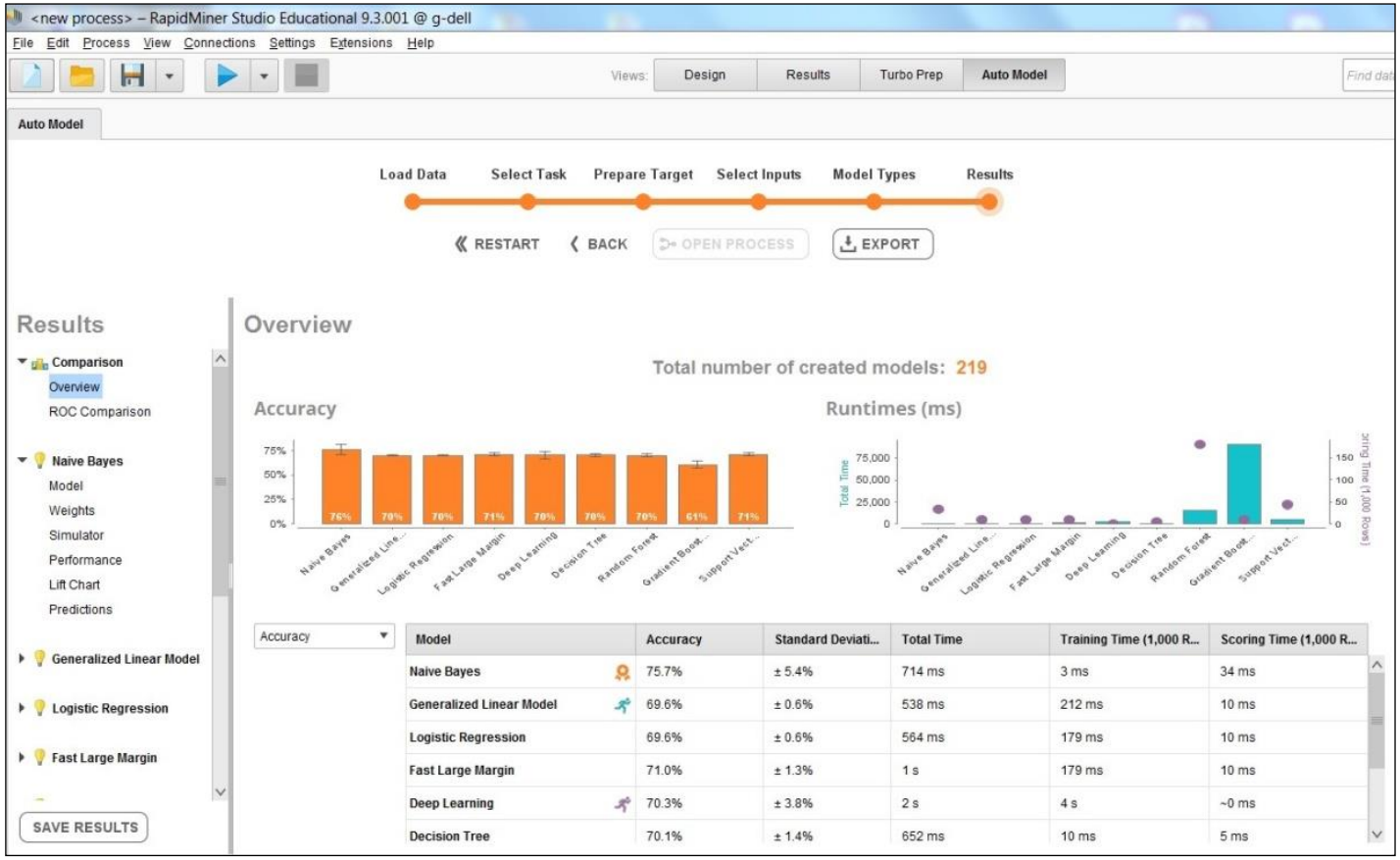


	true Thinking	true Feeling	class precision
pred. Thinking	11	3	78.57%
pred. Feeling	42	88	67.69%
class recall	20.75%	96.70%	

Εικόνα 11: Απόδοση για Συναίσθημα.

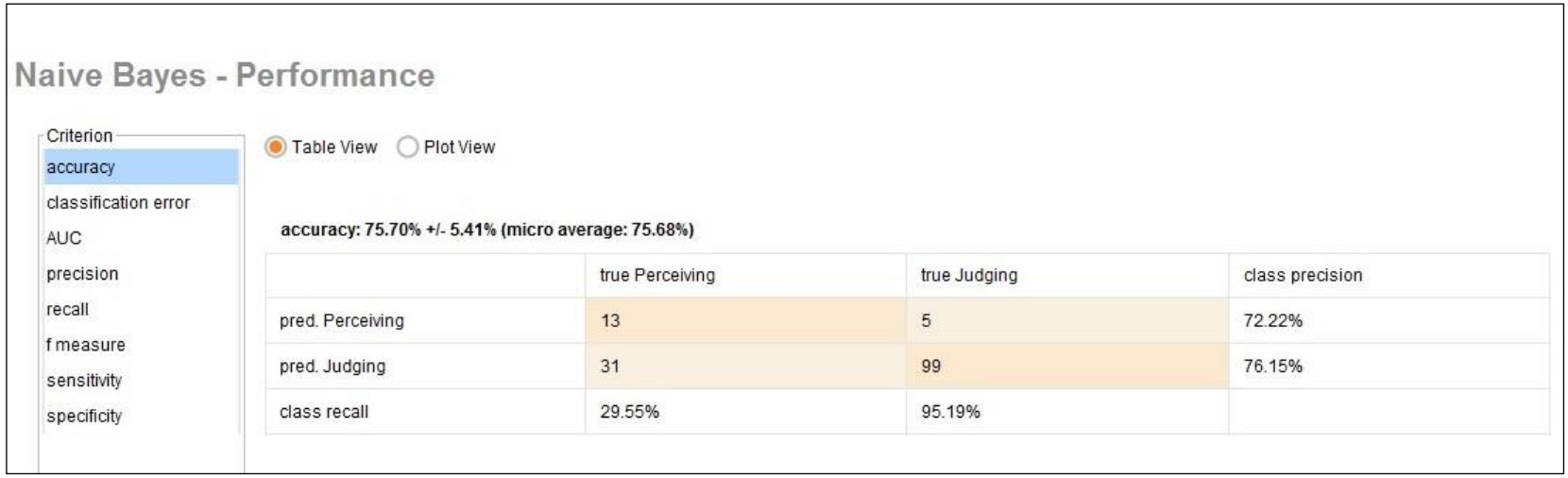
7.1.4 Αντίληψη - Κρίση

Η απόδοση του ταξινομητή Naive Bayes κρίνεται ικανοποιητική για την Κρίση, αφού το ποσοστό ακρίβειάς του είναι 76% ,σε αντίθεση με τους υπόλοιπους οκτώ που ξεκινούν από 61% και δεν ξεπερνούν το 71%.



Εικόνα 12: Αποτελέσματα για Κρίση.

Πιο συγκεκριμένα, με ορθότητα 76,15% και ανάκληση 95,19% η ακρίβεια του αλγόριθμου ήταν 75,70%.



	true Perceiving	true Judging	class precision
pred. Perceiving	13	5	72.22%
pred. Judging	31	99	76.15%
class recall	29.55%	95.19%	

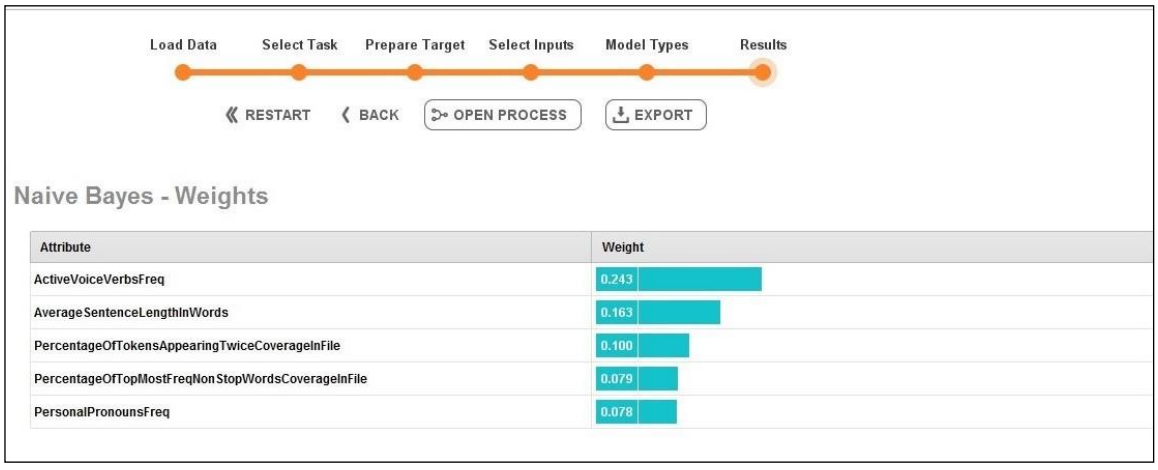
Εικόνα 13: Απόδοση για Κρίση.

7.1.5 Ανάλυση αποτελεσμάτων

Στο παρόν υποκεφάλαιο επιχειρείται η ανάλυση των αποτελεσμάτων πρόβλεψης της προσωπικότητας. Στόχος της μελέτης ήταν η ταξινόμηση των εκθέσεων των μαθητών σε τύπους προσωπικότητας με τη χρήση υφομετρικών δεικτών. Πρέπει λοιπόν να ελεγχθεί εάν και ποια από αυτά τα χαρακτηριστικά είναι σημαντικά τόσο για την εκπαίδευση των αλγορίθμων μηχανικής μάθησης όσο και για την προβλεπτική τους ικανότητα. Η ακρίβεια της πρόβλεψης επιβεβαιώνεται από τα αποτελέσματα του ερωτηματολογίου προσωπικότητας Myers-Briggs Type Indicator (MBTI).

Καταρχάς, ο αλγόριθμος Naive Bayes κρίνεται ως ο αποτελεσματικότερος με μέσο όρο ακρίβειας στην πρόβλεψη για όλους τους τύπους προσωπικότητας 76,5%. Ο τύπος με το μεγαλύτερο ποσοστό ακρίβειας είναι η Εξωστρέφεια με 81% έναντι των άλλων (80%, 76%, 69%). Οι παρακάτω εικόνες με τα βάρη δείχνουν ποια υφομετρικά χαρακτηριστικά έχουν τη μεγαλύτερη επίδραση στην πρόβλεψη του Naive Bayes αλγορίθμου.

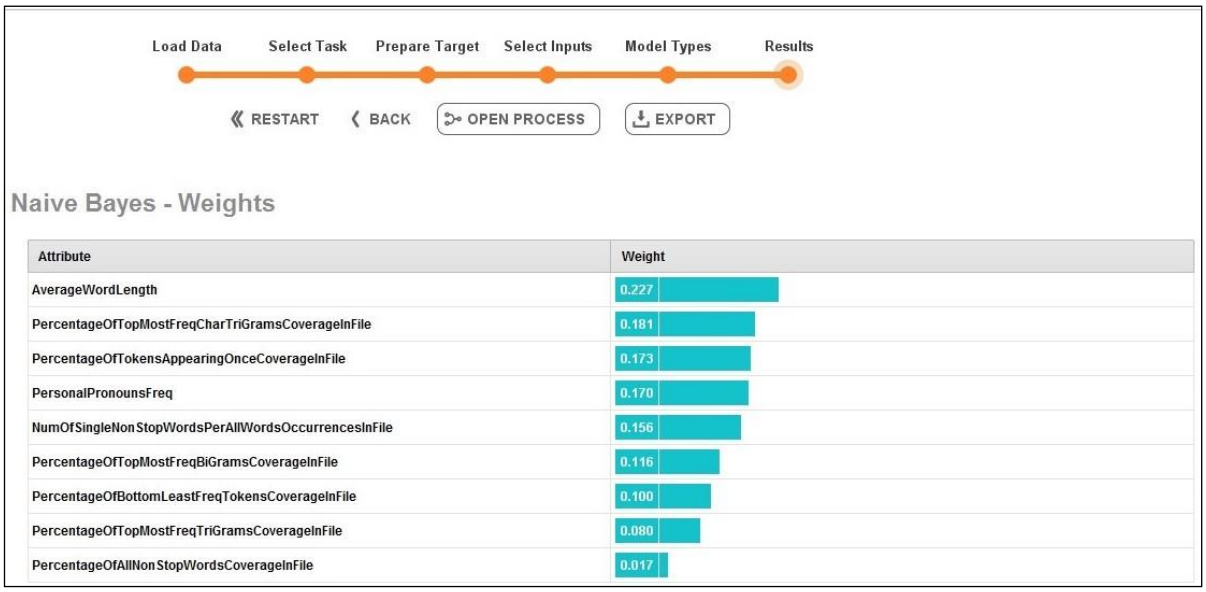
Για την Εξωστρέφεια επέδρασε σημαντικά η χρήση ρηματικών τύπων ενεργητικής φωνής (0,243). Ακολουθούν το μέσο μήκος πρότασης σε λέξεις όλων των προτάσεων (0,163), οι λέξεις που απαντούν δύο μόνο φορές σε ένα κείμενο (0,100), οι πιο συχνές λέξεις περιεχομένου (0,079) και τέλος οι προσωπικές αντωνυμίες (0,078).



Εικόνα 14: Βάρη για Εξωστρέφεια.

Στην εικόνα με τα βάρη που ακολουθεί αποτυπώνεται η προβλεπτική ικανότητα των υφομετρικών χαρακτηριστικών για την Διαίσθηση στον συγκεκριμένο αλγόριθμο. Τη σημαντικότερη επίδραση είχε το μέσο μήκος λέξης σε χαρακτήρες (0,227). Έπονται

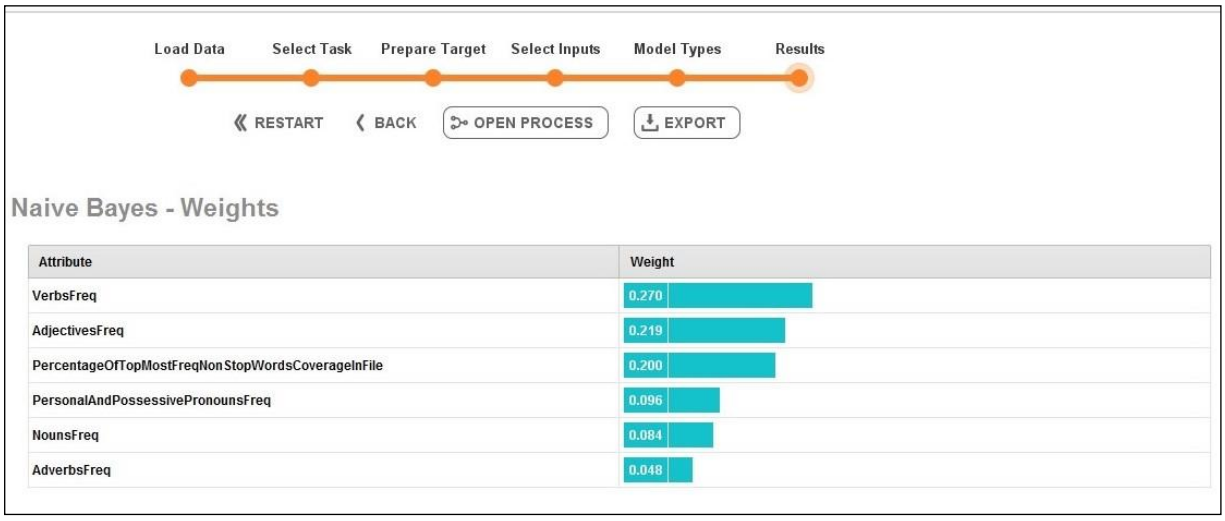
τα συχνότερα τριγράμματα χαρακτήρων (0,181), οι άπαξ εμφανιζόμενες λέξεις (0,173), οι προσωπικές αντωνυμίες (0,170), οι λέξεις περιεχομένου (0,156), τα συχνότερα δίλεκτα (0,116), οι πιο σπάνιες λέξεις (0,100), τα συχνότερα τρίλεκτα (0,080) και όλες οι λέξεις περιεχομένου (0,017).



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

Εικόνα 15: Βάρη για Διαίσθηση.

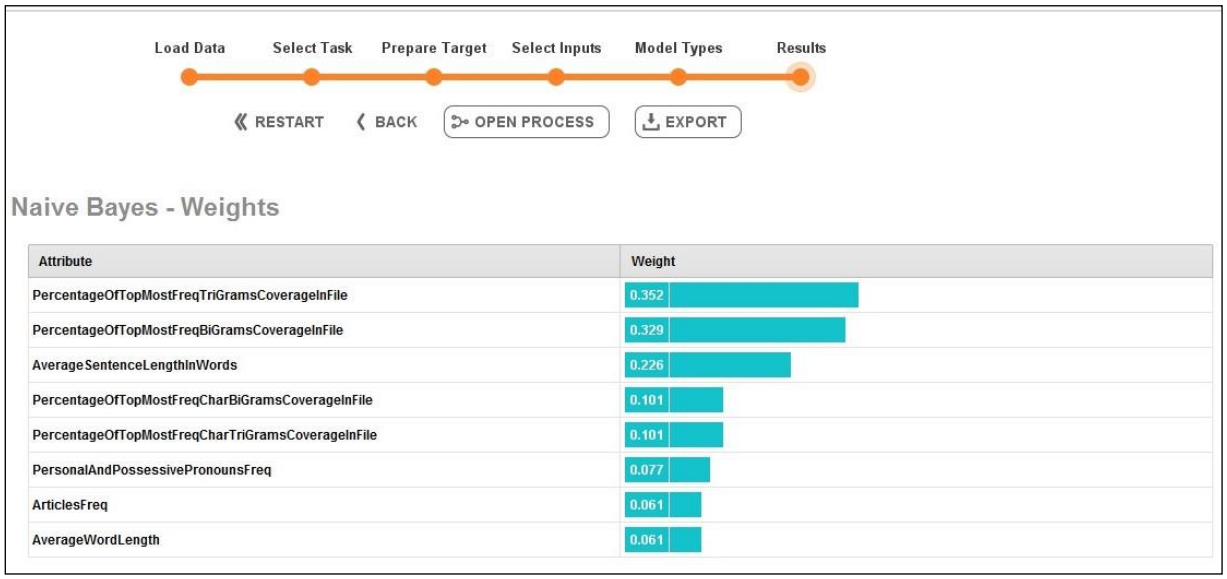
Τα υφομετρικά χαρακτηριστικά που επηρέασαν το αποτέλεσμα της ταξινόμησης των εκθέσεων ως προς το Συναίσθημα είναι τα ρήματα (0,270), τα επίθετα (0,219), οι συχνότερες μη λειτουργικές λέξεις (0,200), οι προσωπικές και κτητικές αντωνυμίες (0,096), τα ουσιαστικά (0,084) και τα επιρρήματα (0,048).



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

Εικόνα 16: Βάρη για Συναίσθημα.

Τέλος, τα οκτώ υφομετρικά χαρακτηριστικά που συνέβαλαν στην πρόβλεψη της Κρίσης ήταν κατά σειρά φθίνουσα: τα συχνότερα τρίλεκτα (0,352), τα συχνότερα δίλεκτα (0,329), το μέσο μήκος πρότασης σε λέξεις (0,226), τα συχνότερα διγράμματα χαρακτήρων και τα συχνότερα τριγράμματα χαρακτήρων με ίδιο ποσοστό (0,101), οι προσωπικές και κτητικές αντωνυμίες (0,077), τα άρθρα (0,061) και το μέσο μήκος λέξης σε χαρακτήρες (0,061).



Εικόνα 17: Βάρη για Κρίση.

7.2 Αποτελέσματα Αυτόματης Ταξινόμησης μαθητικών εκθέσεων (Ερωτηματολόγιο προσωπικότητας των Πέντε Παραγόντων)

Στα υποκεφάλαια που ακολουθούν παρουσιάζονται τα ερευνητικά αποτελέσματα που αφορούν στο ερωτηματολόγιο προσωπικότητας των Πέντε Παραγόντων. Για κάθε χαρακτηριστικό προσωπικότητας δυαδικής μορφής παρατίθενται τρεις εικόνες: τα αποτελέσματα πρόβλεψης σε εκατοστιαίο ποσοστό των 219 μοντέλων από τους εννέα αλγόριθμους που εφαρμόστηκαν στα κειμενικά δεδομένα, ο πίνακας με τα μέτρα αξιολόγησης του αποτελεσματικότερου αλγόριθμου και τα βάρη (weights) που επέδρασαν θετικά στην προβλεπτική του ικανότητα.

Για την πρόβλεψη όλων των χαρακτηριστικών του συγκεκριμένου ερωτηματολογίου ο αλγόριθμος που είχε τα καλύτερα αποτελέσματα ήταν ο Generalized Linear Model. Τα ποσοστά ακρίβειας της πρόβλεψης του Generalized Linear Model κυμαίνονται από 66% έως 86%, με μέσο όρο 72,2%. Συγκεκριμένα, πέτυχε ακρίβεια 86% για την πρόβλεψη Δεκτικότητας των μαθητών στην εμπειρία, 71% για την πρόβλεψη της Ευσυνειδησίας, 68% της Εξωστρέφειας, 70% της Προσήνειας και 66% του Νευρωτισμού. Ο Πίνακας 16 παρουσιάζει την απόδοση όσον αφορά στις μετρήσεις ακρίβειας, ορθότητας και ανάκλησης αυτού του ταξινομητή.

Πίνακας 16: Η απόδοση του Generalized Linear Model.

Χαρακτηριστικό προσωπικότητας	**Ακρίβεια**	**Ορθότητα**	**Ανάκληση**
ClosedMinded	85,94%	85,37%	100%
Disorganised	71,19%	68,57%	80%
Introverted	67,62%	66,67%	86,67%
Agreeable	70,16%	67,86%	98,70%
Calm	65,60%	64,79%	71,88%

Όπως και για την Τυπολογία του Jung, πριν την παρουσίαση του βέλτιστου αλγόριθμου για την αναγνώριση των χαρακτηριστικών προσωπικότητας του μοντέλου των Πέντε Παραγόντων, κρίνουμε απαραίτητη την παράθεση κάποιων από τις πολλές δεκάδες πειράματα που πραγματοποιήθηκαν μέχρι να επιτευχθεί το καλύτερο αποτέλεσμα. Σε πίνακες θα δούμε τα υφομετρικά χαρακτηριστικά που επιλέχθηκαν ανά χαρακτηριστικό προσωπικότητας, το υψηλότερο ποσοστό ακρίβειας και το όνομα του αλγόριθμου ή των αλγόριθμων που το πέτυχαν. Η τελευταία σε κάθε πίνακα σειρά αναφέρεται στα χαρακτηριστικά που χρησιμοποιήθηκαν με τα οποία ο αλγόριθμος Generalized Linear Model απέδωσε καλύτερα.

Πίνακας 17: Πειράματα για πρόβλεψη Δεκτικότητας στην εμπειρία.

Πρόβλεψη Δεκτικότητας στην εμπειρία (Closed Minded)		
Υφομετρικά χαρακτηριστικά	**Ακρίβεια**	**Αλγόριθμος**
Όλα τα χαρακτηριστικά	82%	Generalized Linear Model
Όλα χαρακτηριστικά του plain text corpus	82%	Random Forest
Όλα τα χαρακτηριστικά του tagged corpus	81%	Generalized Linear Model, Logistic Regression, Fast Large Margin, Decision Tree, Random Forest, Support Vector Machine
Όλα τα χαρακτηριστικά του lemmatized corpus	81%	Generalized Linear Model, Logistic Regression, Fast Large Margin, Decision tree, Random Forest, Support Vector Machine
AverageWordLength, AverageSentenceLength, PercentageOfTopMostFreqBiGrams, PercentageOfTopMostFreqTriGrams	81%	Naive Bayes, Generalized Linear Model, Logistic Regression, Fast Large Margin, Decision Tree, Random Forest, Support Vector Machine
Όλα χαρακτηριστικά του plain text corpus, Όλα τα χαρακτηριστικά του tagged corpus	83%	Deep Learning
Όλα τα χαρακτηριστικά του lemmatized corpus, AverageWordLength, AverageSentenceLength, PercentageOfTopMostFreqBiGrams,	82%	Deep Learning

Πρόβλεψη Δεκτικότητας στην εμπειρία (Closed Minded)		
Υφομετρικά χαρακτηριστικά	**Ακρίβεια**	**Αλγόριθμος**
PercentageOfTopMostFreqTriGrams		
Όλα τα χαρακτηριστικά του tagged corpus- (ActiveVoiceVerbsFreq, PassiveVoiceVerbsFreq, PersonalAndPossessivePronounsFreq)	82%	Generalized Linear Model, Logistic Regression
AverageWordLength, AverageSentenceLength, PercentageOfTopMostFreqBiGrams, PercentageOfTopMostFreqTriGrams, PercentageOfBottomLeastFreqTokens	82%	Gradient boosted trees
AverageWordLength, AverageSentenceLength, PercentageOfTopMostFreqBiGrams, PercentageOfTopMostFreqTriGrams, ActiveVoiceVerbsFreq, PassiveVoiceVerbsFreq	83%	Gradient boosted trees
AverageWordLength, AverageSentenceLength, PercentageOfTopMostFreqBiGrams, PercentageOfTopMostFreqTriGrams,	81%	Naive Bayes, Generalized Linear Model, Logistic Regression, Fast Large Margin, Decision Tree, Random Forest, Support Vector Machine

Πρόβλεψη Δεκτικότητας στην εμπειρία (Closed Minded)		
Υφομετρικά χαρακτηριστικά	**Ακρίβεια**	**Αλγόριθμος**
Όλα τα χαρακτηριστικά του tagged corpus-(ActiveVoiceVerbsFreq, PassiveVoiceVerbsFreq, PersonalAndPossessivePronounsFreq), PercentageOfTopMostFreqCharTriGrams	84%	Generalized Linear Model, Logistic Regression
Όλα τα χαρακτηριστικά του tagged corpus-(ActiveVoiceVerbsFreq, PassiveVoiceVerbsFreq, PersonalAndPossessivePronounsFreq),PercentageOfTokensAppearingTwice	***86%***	***Generalized Linear Model***

Πίνακας 18: Πειράματα για πρόβλεψη Ευσυνειδησίας.

Πρόβλεψη Ευσυνειδησίας (Disorganised)		
Υφομετρικά χαρακτηριστικά	**Ακρίβεια**	**Αλγόριθμος**
Όλα τα χαρακτηριστικά	59%	Logistic Regression
Όλα χαρακτηριστικά του plain text corpus	58%	Naive Bayes, Generalized Linear Model, Logistic Regression
Όλα τα χαρακτηριστικά του tagged corpus	57%	Naive Bayes
Όλα τα χαρακτηριστικά του lemmatized corpus	65%	Generalized Linear Model

Πρόβλεψη Ευσυνειδησίας (Disorganised)		
Υφομετρικά χαρακτηριστικά	**Ακρίβεια**	**Αλγόριθμος**
AverageWordLength, AverageSentenceLength, PercentageOfTopMostFreqBiGrams, PercentageOfTopMostFreqTriGrams	62%	Naive Bayes, Generalized Linear Model, Logistic Regression
Όλα τα χαρακτηριστικά του tagged corpus- (ActiveVoiceVerbsFreq, PassiveVoiceVerbsFreq, PersonalAndPossessivePronounsFreq)	62%	Naive Bayes
Όλα τα χαρακτηριστικά του lemmatized corpus, PercentageOfTopMostFreqBiGrams	60%	Naive Bayes, Generalized Linear Model
Όλα τα χαρακτηριστικά του lemmatized corpus- (PercentageOfAllStopWords), PercentageOfTopMostFreqBiGrams	62%	Naive Bayes
Όλα τα χαρακτηριστικά του lemmatized corpus- (PercentageOfAllStopWords), PercentageOfTopMostFreqBiGrams, NounsFreq	63%	Naive Bayes
Όλα τα χαρακτηριστικά του lemmatized corpus- (PercentageOfAllStopWords),	64%	Naive Bayes, Generalized Linear Model, Logistic Regression

Πρόβλεψη Ευσυνειδησίας (Disorganised)		
Υφομετρικά χαρακτηριστικά	**Ακρίβεια**	**Αλγόριθμος**
PercentageOfTopMostFreqBiGrams, NounsFreq, ArticlesFreq		
Όλα τα χαρακτηριστικά του lemmatized corpus, NounsFreq, ArticlesFreq, AverageWordLength	67%	Generalized Linear Model
Όλα τα χαρακτηριστικά του lemmatized corpus, SubordinativeConjunctionsFreq	67%	Generalized Linear Model
Όλα τα χαρακτηριστικά του lemmatized corpus, PercentageOfTopMostFreqBiGrams, SubordinativeConjunctionsFreq	70%	Generalized Linear Model
Όλα τα χαρακτηριστικά του lemmatized corpus, SubordinativeConjunctionsFreq, AverageWordLength, PercentageOfTopMostFreqBiGrams	***71%***	***Generalized Linear Model***

Πίνακας 19: Πειράματα για πρόβλεψη Εξωστρέφειας.

Πρόβλεψη Εξωστρέφειας (Intoverted)		
Υφομετρικά χαρακτηριστικά	**Ακρίβεια**	**Αλγόριθμος**
Όλα τα χαρακτηριστικά	57%	Support Vector Machine

Πρόβλεψη Εξωστρέφειας (Intoverted)		
Υφομετρικά χαρακτηριστικά	**Ακρίβεια**	**Αλγόριθμος**
Όλα χαρακτηριστικά του plain text corpus	59%	Decision Tree
Όλα τα χαρακτηριστικά του tagged corpus	63%	Deep Learning
Όλα τα χαρακτηριστικά του lemmatized corpus	57%	Random Forest
AverageWordLength, AverageSentenceLength, PercentageOfTopMostFreqBiGrams, PercentageOfTopMostFreqTriGrams	61%	Support Vector Machine
Όλα τα χαρακτηριστικά του tagged corpus-(ActiveVoiceVerbsFreq, PassiveVoiceVerbsFreq, PersonalAndPossessivePronounsFreq)	65%	Deep Learning
Όλα τα χαρακτηριστικά του lemmatized corpus, AverageWordLength, AverageSentenceLength, PercentageOfTopMostFreqBiGrams, PercentageOfTopMostFreqTriGrams	58%	Random Forest
Όλα τα χαρακτηριστικά του tagged corpus-(ArticlesFreq,ActiveVoiceVerb	67%	Deep Learning

Πρόβλεψη Εξωστρέφειας (Intoverted)		
Υφομετρικά χαρακτηριστικά	**Ακρίβεια**	**Αλγόριθμος**
sFreq, PassiveVoiceVerbsFreq, PersonalAndPossessivePronounsFreq), PercentageOfBottomLeastFreqNonStopWords		
AverageWordLength, AverageSentenceLength, PercentageOfTopMostFreqBiGrams, PercentageOfTopMostFreqTriGrams, RatioOfTwiceOverOnceAppearingTokens	62%	Generalized Linear Model
AverageWordLength, AverageSentenceLength, PercentageOfTopMostFreqBiGrams, RatioOfTwiceOverOnceAppearingTokens, PersonalAndPossessivePronounsFreq, PrepositionsFreq	63%	Generalized Linear Model
AverageSentenceLength, PercentageOfTopMostFreqBiGrams, RatioOfTwiceOverOnceAppearingTokens, PersonalAndPossessivePronounsFreq, PrepositionsFreq	64%	Generalized Linear Model

Πρόβλεψη Εξωστρέφειας (Intoverted)		
Υφομετρικά χαρακτηριστικά	**Ακρίβεια**	**Αλγόριθμος**
AverageSentenceLength, PercentageOfTopMostFreqBiGrams, RatioOfTwiceOverOnceAppearingTokens, PersonalAndPossessivePronounsFreq, PrepositionsFreq, AdverbsFreq	65%	Generalized Linear Model
AverageSentenceLength, PercentageOfTopMostFreqBiGrams, RatioOfTwiceOverOnceAppearingTokens, PersonalAndPossessivePronounsFreq, PrepositionsFreq, AdverbsFreq, PercentageOfBottomLeastFreq NonStopWords	66%	Generalized Linear Model
AverageSentenceLength, PercentageOfTopMostFreqBiGrams, RatioOfTwiceOverOnceAppearingTokens, PersonalAndPossessivePronounsFreq, PrepositionsFreq, AdverbsFreq, ConjunctionsFreq	67%	Generalized Linear Model
AverageSentenceLength, PercentageOfTopMostFreqBi Grams,	***68%***	***Generalized Linear Model***

Πρόβλεψη Εξωστρέφειας (Intoverted)		
Υφομετρικά χαρακτηριστικά	**Ακρίβεια**	**Αλγόριθμος**
PersonalAndPossessivePronounsFreq, PrepositionsFreq, AdverbsFreq,ConjunctionsFreq,RatioOfTwiceOverOnceAppearingTokens,PercentageOfTokensAppearingTwice,PercentageOfBottomLeastFreqNonStop Words		

Πίνακας 20: Πειράματα για πρόβλεψη Προσήνειας.

Πρόβλεψη Προσήνειας (Agreeable)		
Υφομετρικά χαρακτηριστικά	**Ακρίβεια**	**Αλγόριθμος**
Όλα τα χαρακτηριστικά	65%	Deep Learning
Όλα χαρακτηριστικά του plain text corpus	66%	Naive Bayes
Όλα τα χαρακτηριστικά του tagged corpus	63%	Random Forest
Όλα τα χαρακτηριστικά του lemmatized corpus	66%	Logistic Regression
AverageWordLength, AverageSentenceLength, PercentageOfTopMostFreqBiGrams, PercentageOfTopMostFreqTriGrams	65%	Generalized Linear Model, Logistic Regression

Πρόβλεψη Προσήνειας (Agreeable)		
Υφομετρικά χαρακτηριστικά	**Ακρίβεια**	**Αλγόριθμος**
Όλα τα χαρακτηριστικά του tagged corpus- (ActiveVoiceVerbsFreq, PassiveVoiceVerbsFreq, PersonalAndPossessivePronounsFreq)	63%	Fast large margin, Random Forest
AverageWordLength, AverageSentenceLength, PercentageOfTopMostFreqBiGrams, PercentageOfTopMostFreqTriGrams, PercentageOfTokensAppearingTwice	68%	Naive Bayes
AverageWordLength, AverageSentenceLength, PercentageOfTopMostFreqBiGrams, PercentageOfTopMostFreqTriGrams, PrepositionsFreq	69%	Naive Bayes
AverageWordLength, AverageSentenceLength, PercentageOfTopMostFreqBiGrams, PercentageOfTopMostFreqTriGrams, PrepositionsFreq, VerbsFreq	70%	Naive Bayes

Πρόβλεψη Προσήνειας (Agreeable)		
Υφομετρικά χαρακτηριστικά	**Ακρίβεια**	**Αλγόριθμος**
Όλα τα χαρακτηριστικά του lemmatized corpus-(FunctionalDensity)	66%	Generalized Linear Model, Logistic Regression
Όλα τα χαρακτηριστικά του lemmatized corpus-FunctionalDensity, PercentageOfAllStopWords), AverageWordLength, AverageSentenceLength	69%	Fast Large Margin
AverageWordLength, PercentageOfTopMostFreqBiGrams, RatioOfTwiceOverOnceAppearingTokens, PercentageOfTopMostFreqTokens, PersonalAndPossessivePronounsFreq, VerbsFreq, PrepositionsFreq	67%	Generalized Linear Model
AverageWordLength, PercentageOfTopMostFreqBiGrams, RatioOfTwiceOverOnceAppearingTokens, PercentageOfTopMostFreqTokens, PersonalAndPossessivePronounsFreq, VerbsFreq, PrepositionsFreq,	69%	Generalized Linear Model

Πρόβλεψη Προσήνειας (Agreeable)		
Υφομετρικά χαρακτηριστικά	**Ακρίβεια**	**Αλγόριθμος**
PercentageOfTopMostFreqTriGrams, ActiveVoiceVerbsFreq		
AverageWordLength, AverageSentenceLength, PercentageOfTopMostFreqBiGrams, PercentageOfTopMostFreqCharTriGrams, PersonalAndPossessivePronounsFreq, VerbsFreq, PrepositionsFreq, ActiveVoiceVerbsFreq, RatioOfTwiceOverOnceAppearingTokens, PercentageOfTopMostFreqTokens	***70%***	***Generalized Linear Model***

Πίνακας 21: Πειράματα για πρόβλεψη Νευρωτισμού.

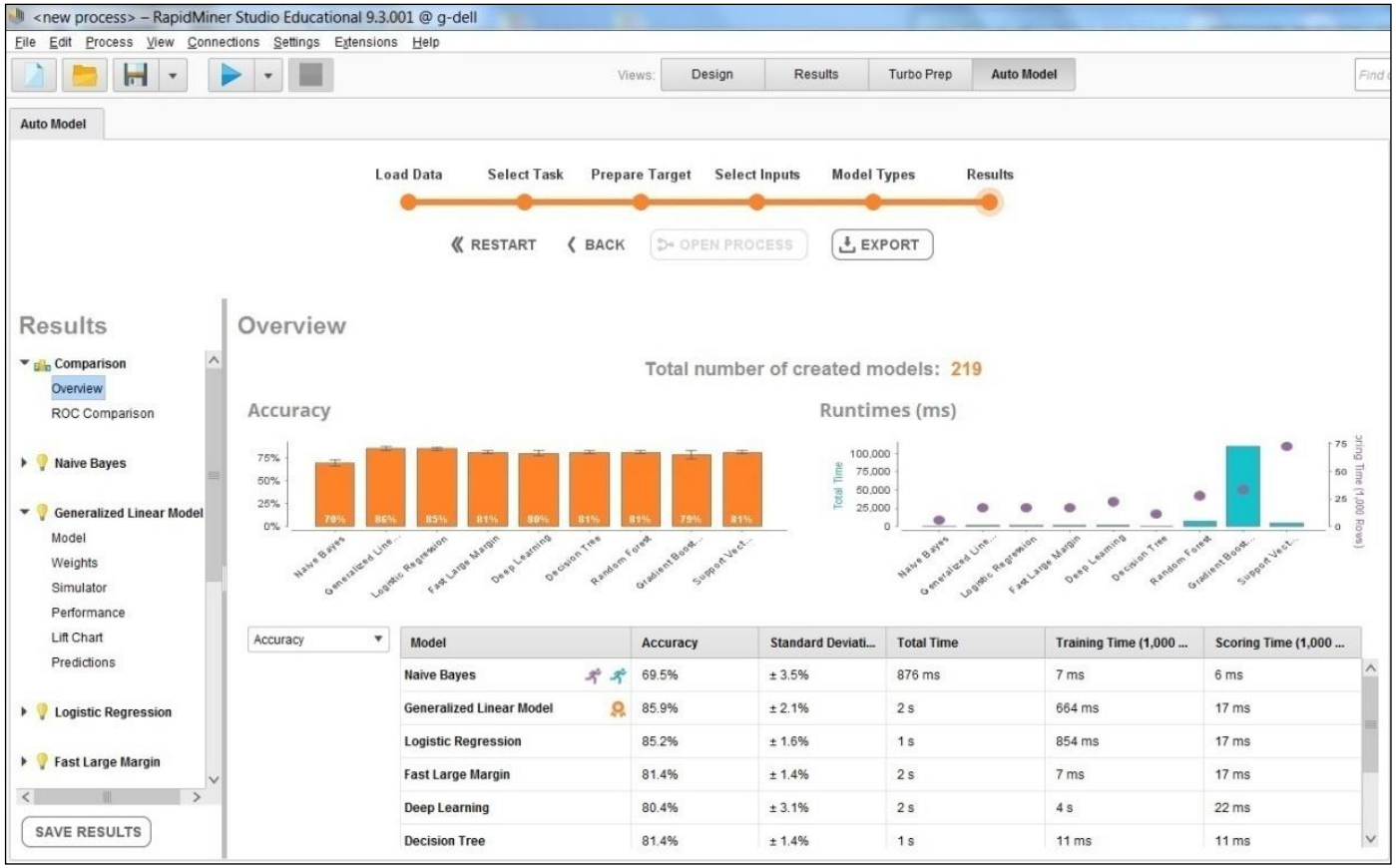
Πρόβλεψη Νευρωτισμού (Calm)		
Υφομετρικά χαρακτηριστικά	**Ακρίβεια**	**Αλγόριθμος**
Όλα τα χαρακτηριστικά	61%	Gradient boosted trees
Όλα χαρακτηριστικά του plain text corpus	53%	Generalized Linear Model, Logistic Regression
Όλα τα χαρακτηριστικά του tagged corpus	59%	Generalized Linear Model

Πρόβλεψη Νευρωτισμού (Calm)		
Υφομετρικά χαρακτηριστικά	**Ακρίβεια**	**Αλγόριθμος**
Όλα τα χαρακτηριστικά του lemmatized corpus	55%	Random Forest
AverageWordLength, AverageSentenceLength, PercentageOfTopMostFreqBiGrams, PercentageOfTopMostFreqTriGrams	53%	Fast Large Margin
Όλα τα χαρακτηριστικά του tagged corpus-(ActiveVoiceVerbsFreq, PassiveVoiceVerbsFreq, PersonalAndPossessivePronounsFreq)	62%	Deep Learning
Όλα τα χαρακτηριστικά του tagged corpus-(ActiveVoiceVerbsFreq, PassiveVoiceVerbsFreq, PronounsFreq, SubordinativeConjunctionsFreq,), AverageWordLength	62%	Generalized Linear Model
Most Frequent ch-trigrams, όλα τα χαρακτηριστικά του tagged corpus-(ActiveVoiceVerbsFreq, PassiveVoiceVerbsFreq, PersonalAndPossessivePronounsFreq), PercentageOfAllStopWords	63%	Generalized Linear Model

Πρόβλεψη Νευρωτισμού (Calm)		
Υφομετρικά χαρακτηριστικά	**Ακρίβεια**	**Αλγόριθμος**
Most Frequent ch-trigrams, όλα τα χαρακτηριστικά του tagged corpus-(ActiveVoiceVerbsFreq, PassiveVoiceVerbsFreq, PersonalAndPossessivePronounsFreq, ArticlesFreq), PercentageOfAllStopWords	64%	Generalized Linear Model
Most Frequent ch-trigrams, όλα τα χαρακτηριστικά του tagged corpus-(ActiveVoiceVerbsFreq, PassiveVoiceVerbsFreq, PersonalAndPossessivePronounsFreq, VerbsFreq, ArticlesFreq),PercentageOfTopMostFreqNonStopWords	65%	Generalized Linear Model
Most Frequent ch-trigrams, PercentageOfTopMostFreqTriGrams, NounsFreq, AdverbsFreq, SubordinativeConjunctionsFreq, PassiveVoiceVerbsFreq	***66%***	***Generalized Linear Model***

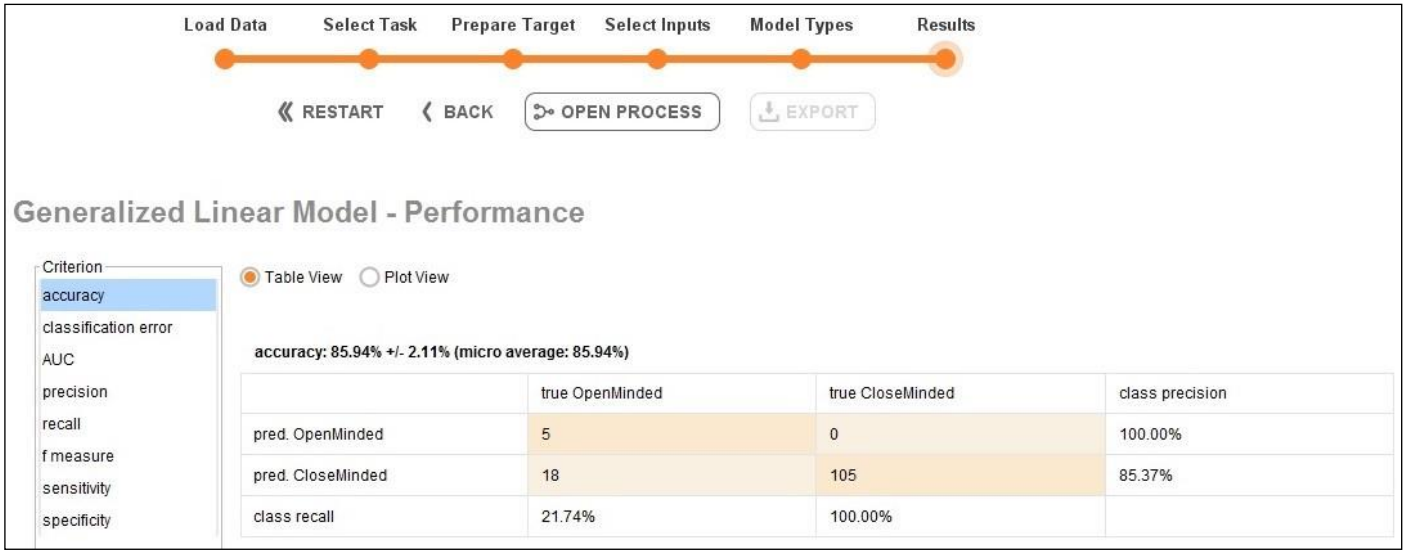
7.2.1 Δεκτικότητα στην εμπειρία

Οι μαθητές που δεν είναι δεκτικοί σε νέες εμπειρίες αναγνωρίστηκαν από τις εκθέσεις τους με ακρίβεια 86%. Οι υπόλοιποι αλγόριθμοι πέτυχαν από 85% έως 70%.



Εικόνα 18: Αποτελέσματα για Δεκτικότητα στην εμπειρία

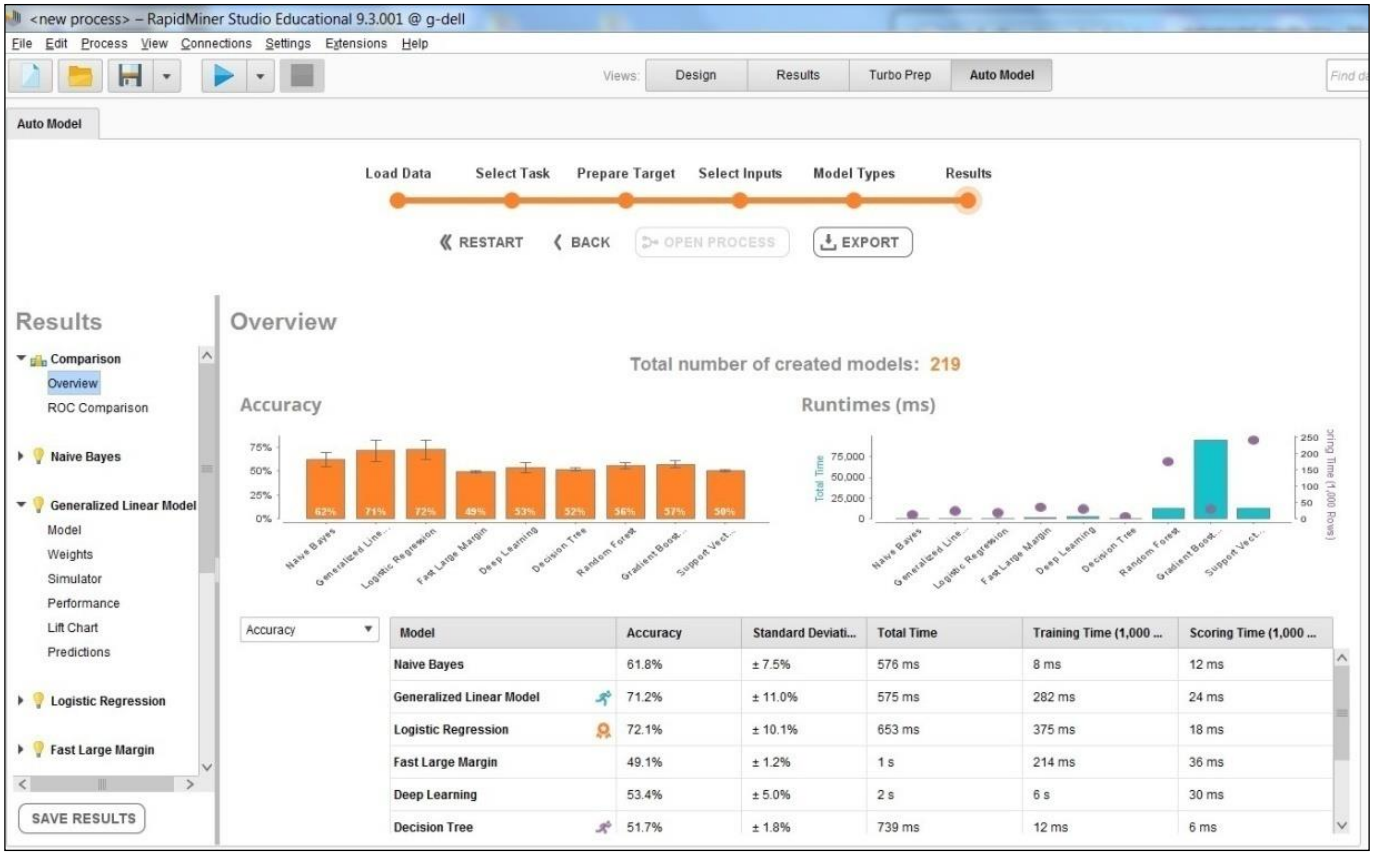
Ο Generalized Linear Model αξιολογήθηκε με ορθότητα 85,37% και ανάκληση 100%, επομένως ακρίβεια 85, 94%.



Εικόνα 19: Απόδοση για Δεκτικότητα στην εμπειρία

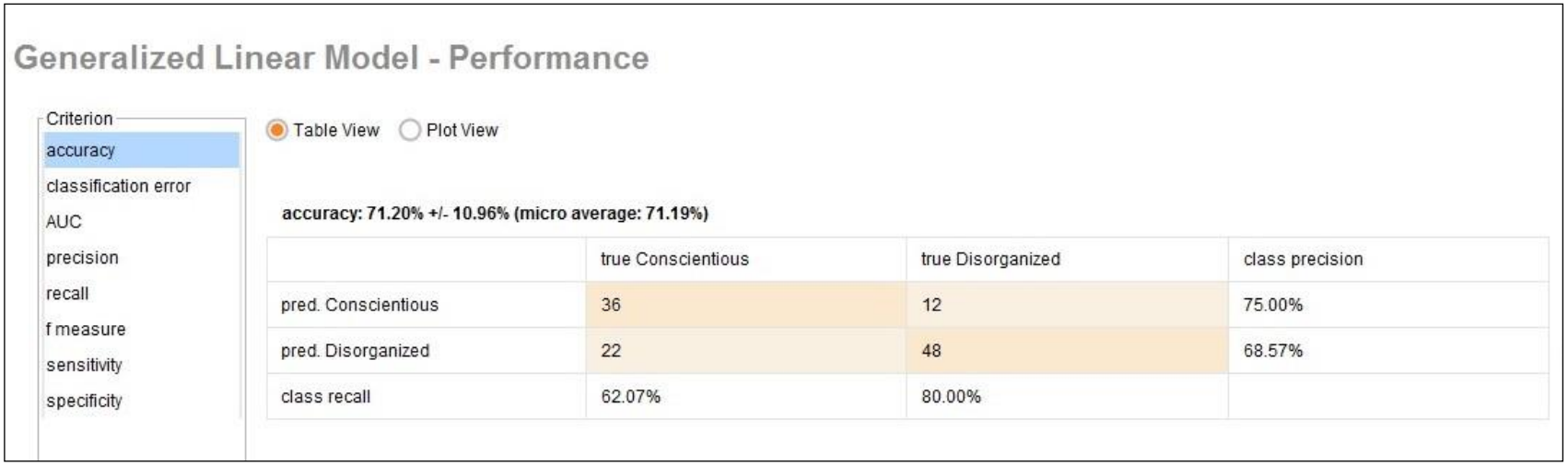
7.2.2 Ευσυνειδησία

Με ακρίβεια 71% ο Generalized Linear Model αναγνώρισε το χαρακτηριστικό της Ευσυνειδησίας στις εκθέσεις των μαθητών. Παρατηρούμε, βέβαια, ότι δεν είναι αυτός ο αλγόριθμος με το υψηλότερο ποσοστό, αλλά ο Logistic Regression (72%). Για λόγους, όμως, σύγκρισης με τα υπόλοιπα χαρακτηριστικά του ερωτηματολογίου επιλέγουμε τον πρώτο αλγόριθμο.



Εικόνα 20: Αποτελέσματα για Ευσυνειδησία

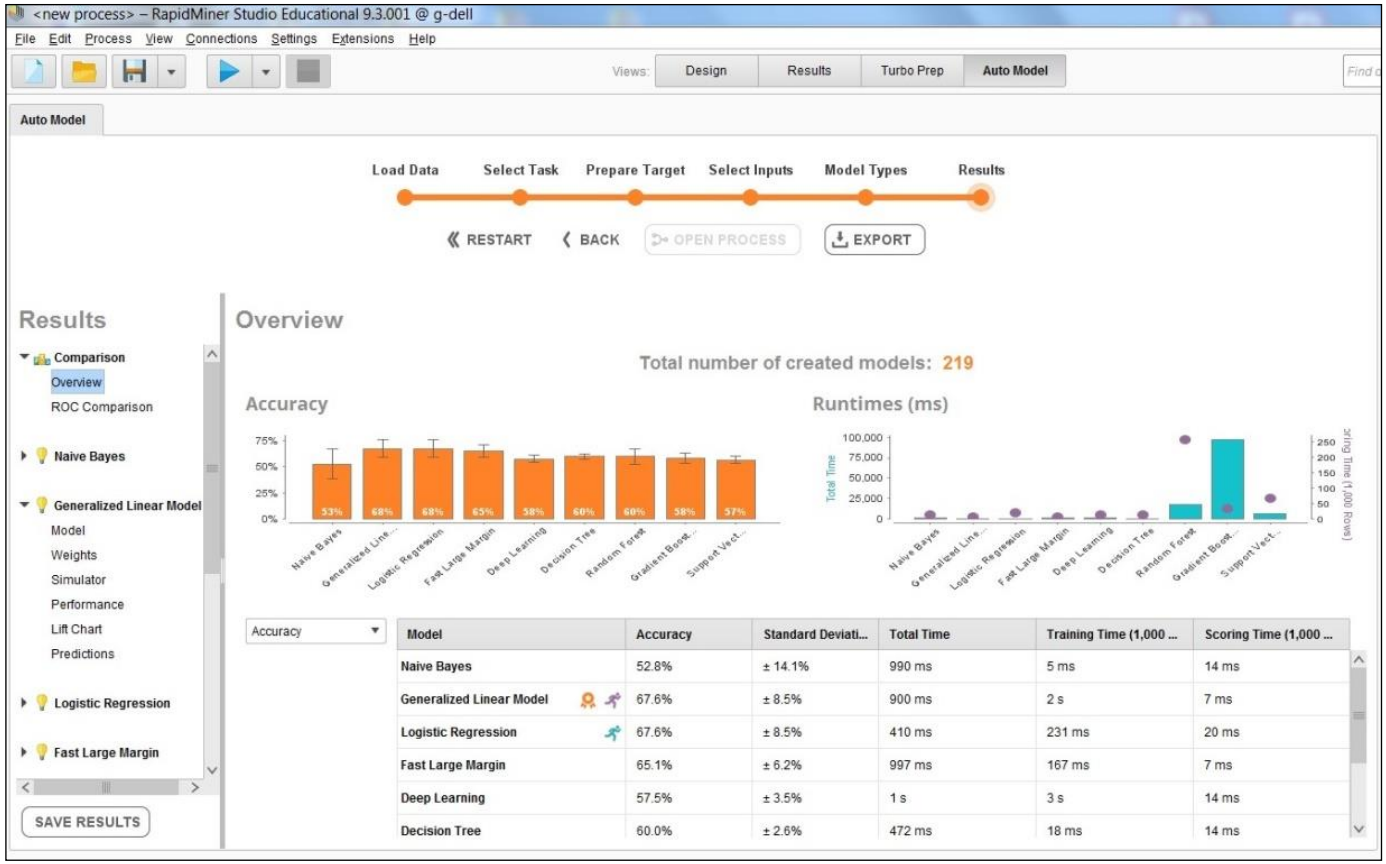
Όσον αφορά στις μετρικές αξιολόγησης του μοντέλου, η ακρίβεια πρόβλεψης για την Ευσυνειδησία είναι 71,20%, η ορθότητα 68,57% και η ανάκληση 80%.



Εικόνα 21: Απόδοση για Ευσυνειδησία

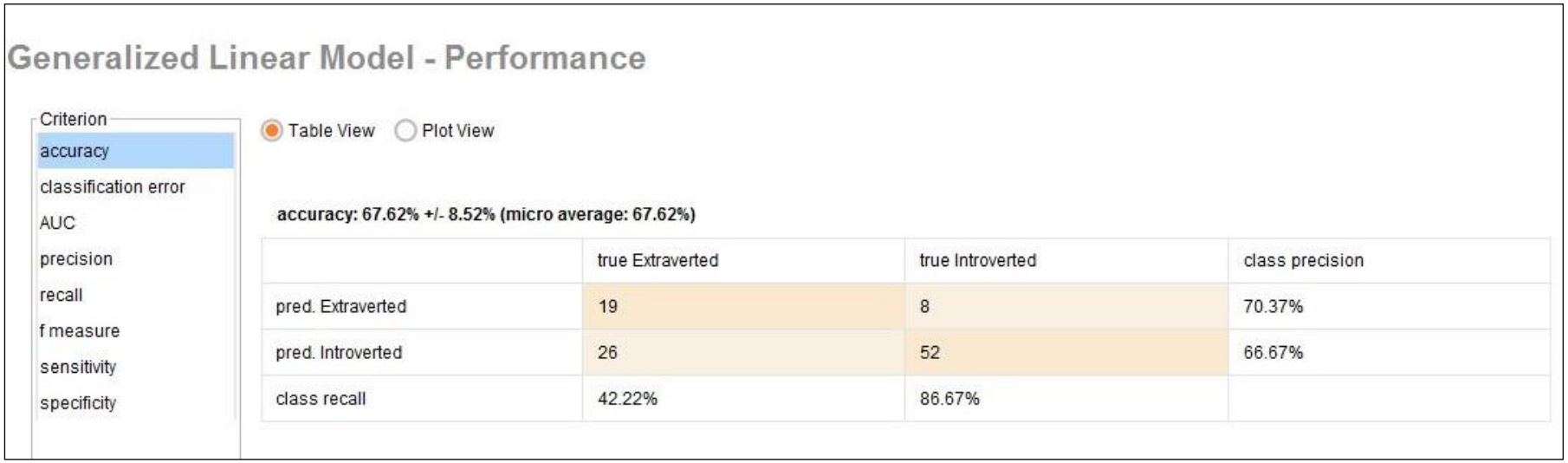
7.2.3 Εξωστρέφεια

Η Εσωστρέφεια, που επικράτησε έναντι της Εξωστρέφειας στο συγκεκριμένο δείγμα μαθητών που είχαμε, προβλέφθηκε από τον αλγόριθμο με ακρίβεια 68%. Το αμέσως χαμηλότερο ποσοστό ήταν 65%, ενώ ακολουθούν τα υπόλοιπα μοντέλα με ακρίβεια από 60% έως 53% το ελάχιστο.



Εικόνα 22: Αποτελέσματα για Εξωστρέφεια

Ο Generalized Linear Model αξιολογήθηκε με ορθότητα 66,67% και ανάκληση 86,67%, επομένως ακρίβεια 67,62%.

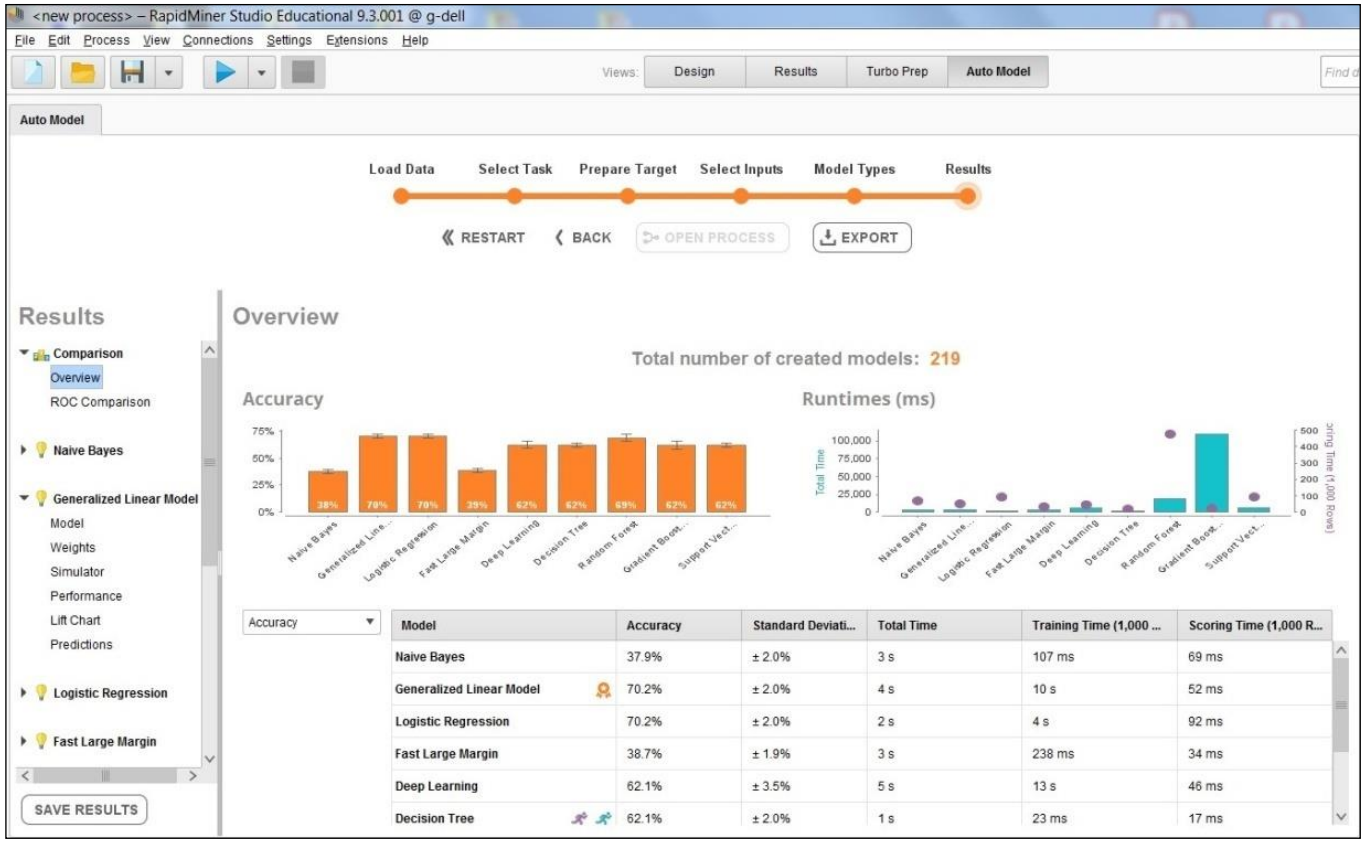


	true Extraverted	true Introverted	class precision
pred. Extraverted	19	8	70.37%
pred. Introverted	26	52	66.67%
class recall	42.22%	86.67%	

Εικόνα 23: Απόδοση για Εξωστρέφεια

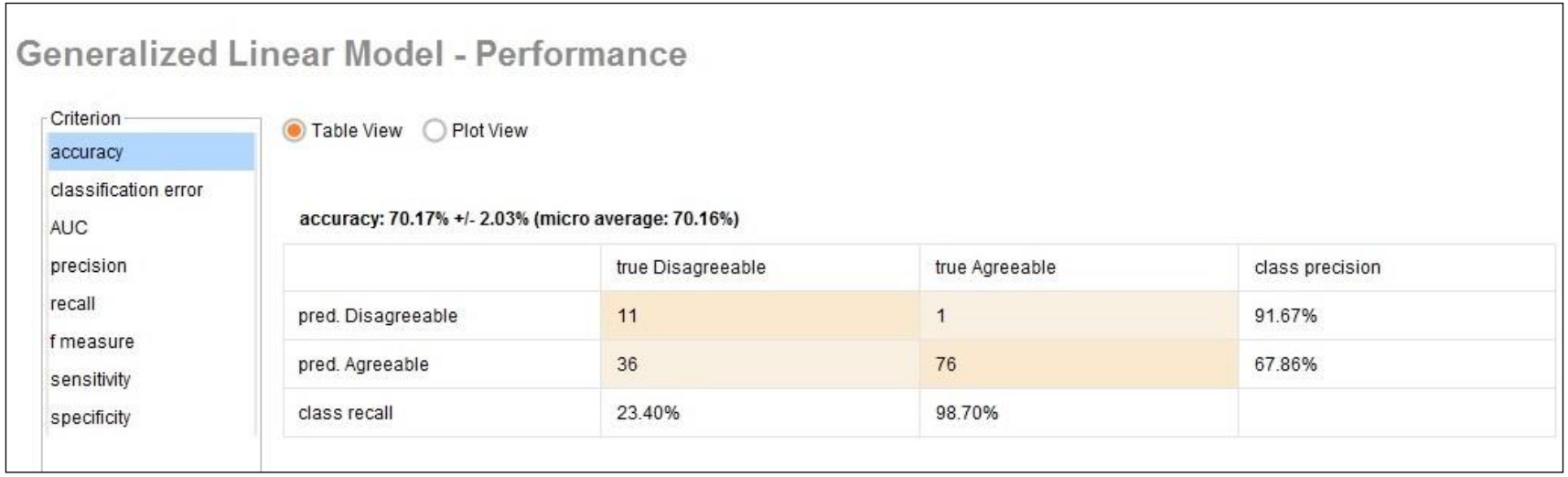
7.2.4 Προσήνεια

Με ακρίβεια 70% αναγνωρίστηκε το συγκεκριμένο χαρακτηριστικό. Το ίδιο ποσοστό πέτυχε και ο Logistic Regression και έπονται με 69% ο Random Forest και οι υπόλοιποι έξι αλγόριθμοι από 62% έως 38%.



Εικόνα 24: Αποτελέσματα για Προσήνεια

Ο αλγόριθμος αξιολογείται, όπως είδαμε και στη γενική εικόνα, με ακρίβεια 70,17%, ορθότητα 67,86% και ανάκληση 98,70%.

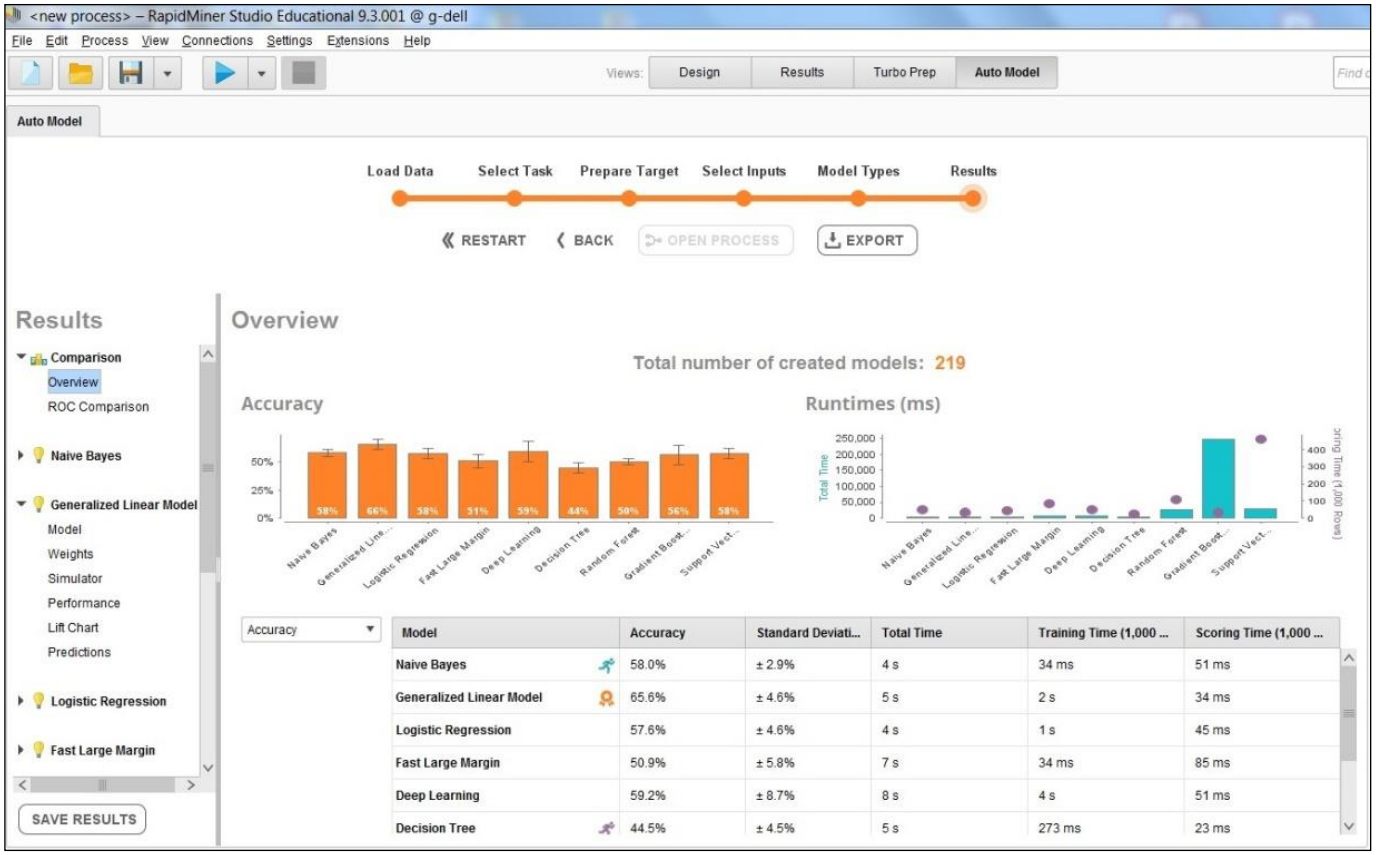


	true Disagreeable	true Agreeable	class precision
pred. Disagreeable	11	1	91.67%
pred. Agreeable	36	76	67.86%
class recall	23.40%	98.70%	

Εικόνα 25: Απόδοση για Προσήνεια

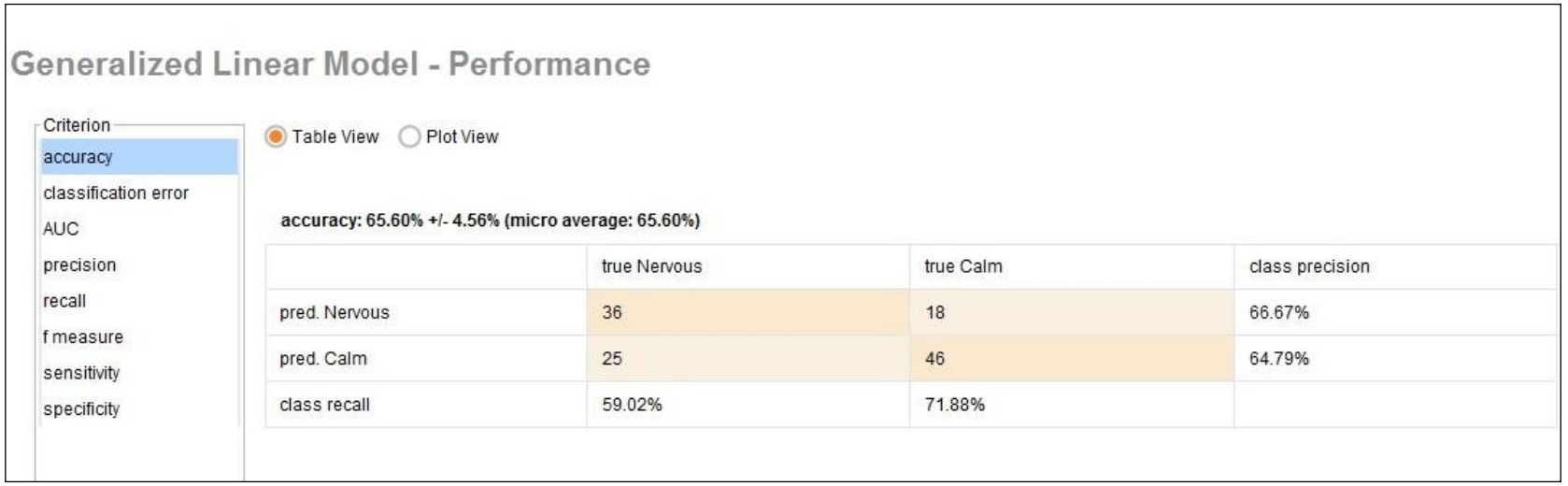
7.2.5 Νευρωτισμός

Η αναγνώριση του χαρακτηριστικού του Νευρωτισμού παρουσίασε το χαμηλότερο ποσοστό ακρίβειας σε σχέση με τα υπόλοιπα, 66%. Οι άλλοι οκτώ αλγόριθμοι του RapidMiner έχουν ακόμα χαμηλότερη απόδοση (44%-59%).



Εικόνα 26: Αποτελέσματα για Νευρωτισμό

Και από τον πίνακα της αξιολόγησης προκύπτει ορθότητα 64,79%, ανάκληση 71,88% και τελικά ακρίβεια 65,60%.



	true Nervous	true Calm	class precision
pred. Nervous	36	18	66.67%
pred. Calm	25	46	64.79%
class recall	59.02%	71.88%	

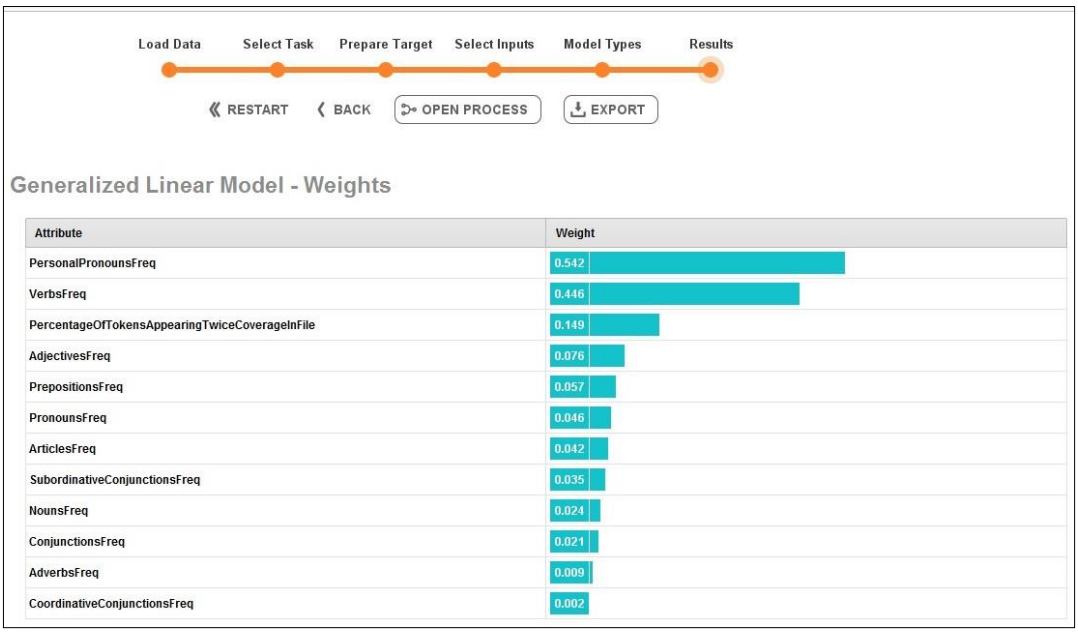
Εικόνα 27: Απόδοση για Νευρωτισμό

7.2.6 Ανάλυση αποτελεσμάτων

Στο παρόν υποκεφάλαιο επιχειρείται η ανάλυση των αποτελεσμάτων πρόβλεψης της προσωπικότητας. Στόχος της μελέτης ήταν η ταξινόμηση των εκθέσεων των μαθητών σε χαρακτηριστικά προσωπικότητας με τη χρήση υφομετρικών δεικτών. Πρέπει λοιπόν να ελεγχθεί εάν και ποια από αυτά τα χαρακτηριστικά είναι σημαντικά τόσο για την εκπαίδευση των αλγορίθμων μηχανικής μάθησης όσο και για την προβλεπτική τους ικανότητα. Η ακρίβεια της πρόβλεψης επιβεβαιώνεται από τα αποτελέσματα του ερωτηματολογίου προσωπικότητας των Πέντε Παραγόντων.

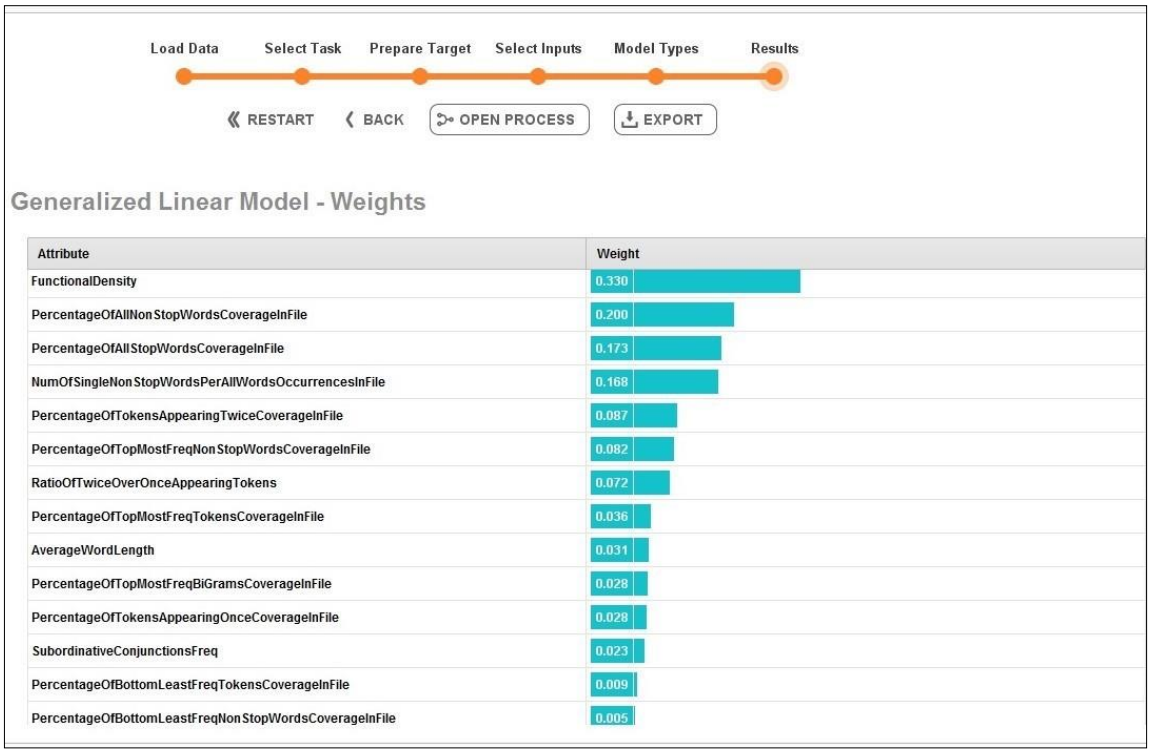
Καταρχάς, ο αλγόριθμος Generalized Linear Model κρίνεται ως ο αποτελεσματικότερος με μέσο όρο ακρίβειας στην πρόβλεψη για όλα τα χαρακτηριστικά προσωπικότητας 72,2%. Το χαρακτηριστικό με το μεγαλύτερο ποσοστό ακρίβειας είναι η Δεκτικότητα στην εμπειρία με 86% έναντι των άλλων (71%, 70%, 68%, 66%).

Τα υφομετρικά χαρακτηριστικά που επιλέξαμε και διαπιστώσαμε ότι ο συνδυασμός τους είχε το καλύτερο αποτέλεσμα αξιολογήθηκαν από το RapidMiner ως προς την επίδραση που έχουν στο μοντέλο για την πρόβλεψη της Δεκτικότητας στην εμπειρία. Έτσι, το πιο σημαντικό κρίθηκε η χρήση προσωπικών αντωνυμιών (0,542). Ακολουθούν τα ρήματα (0,446) και οι λέξεις που εμφανίζονται δύο φορές σε κάθε κείμενο (0,149). Πιο χαμηλά ποσοστά έχουν τα επίθετα (0,076), οι προθέσεις (0,057), οι αντωνυμίες (0,046), τα άρθρα (0,042), οι υποτακτικοί σύνδεσμοι (0,035), τα ουσιαστικά (0,024), οι σύνδεσμοι (0,021), τα επιρρήματα (0,009) και οι παρατακτικοί σύνδεσμοι (0,002).



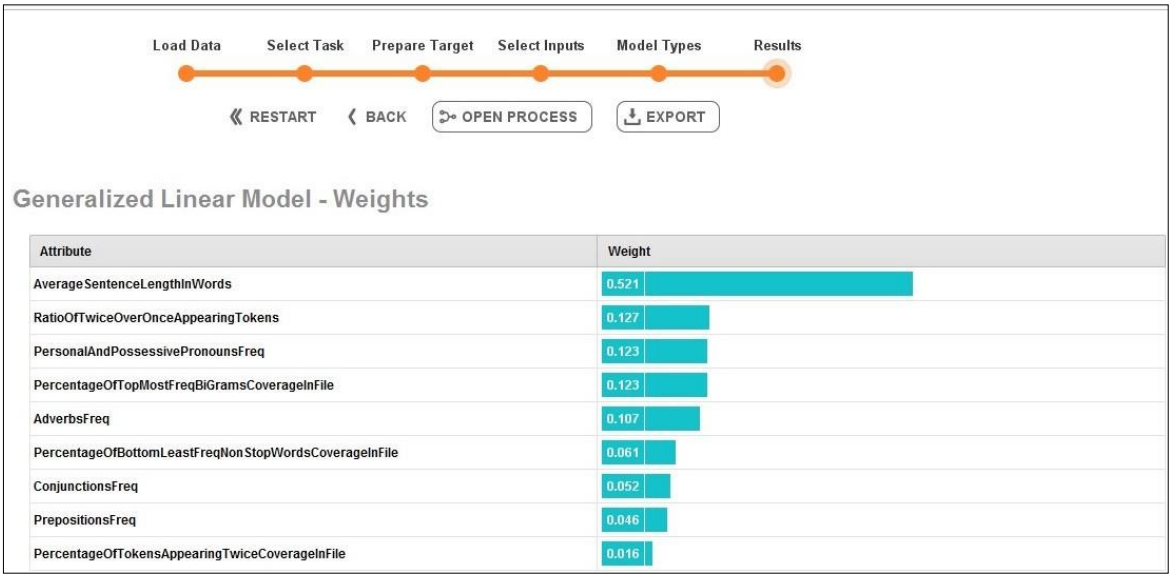
Εικόνα 28: Βάρη για Δεκτικότητα στην εμπειρία

Το πιο σημαντικό υφομετρικό χαρακτηριστικό για την πρόβλεψη της Ευσυνειδησίας είναι η λειτουργική πυκνότητα (0,330), οι μη λειτουργικές λέξεις (0,200), οι λειτουργικές λέξεις (0,173), οι εμφανίσεις των μη λειτουργικών λέξεων (0,168), οι λέξεις που απαντούν δύο φορές (0,087), οι πιο συχνές μη λειτουργικές λέξεις (0,082), ο λόγος των δις προς άπαξ λεγόμενα (0,072), οι πιο συχνές λέξεις (0,036), το μέσο μήκος λέξης σε χαρακτήρες (0,031), τα συχνότερα δίλεκτα (0,028), οι λέξεις που απαντούν μία φορά (0,028), οι υποτακτικοί σύνδεσμοι (0,023), οι πιο σπάνιες λέξεις (0,009) και οι πιο σπάνιες μη λειτουργικές λέξεις (0,005).



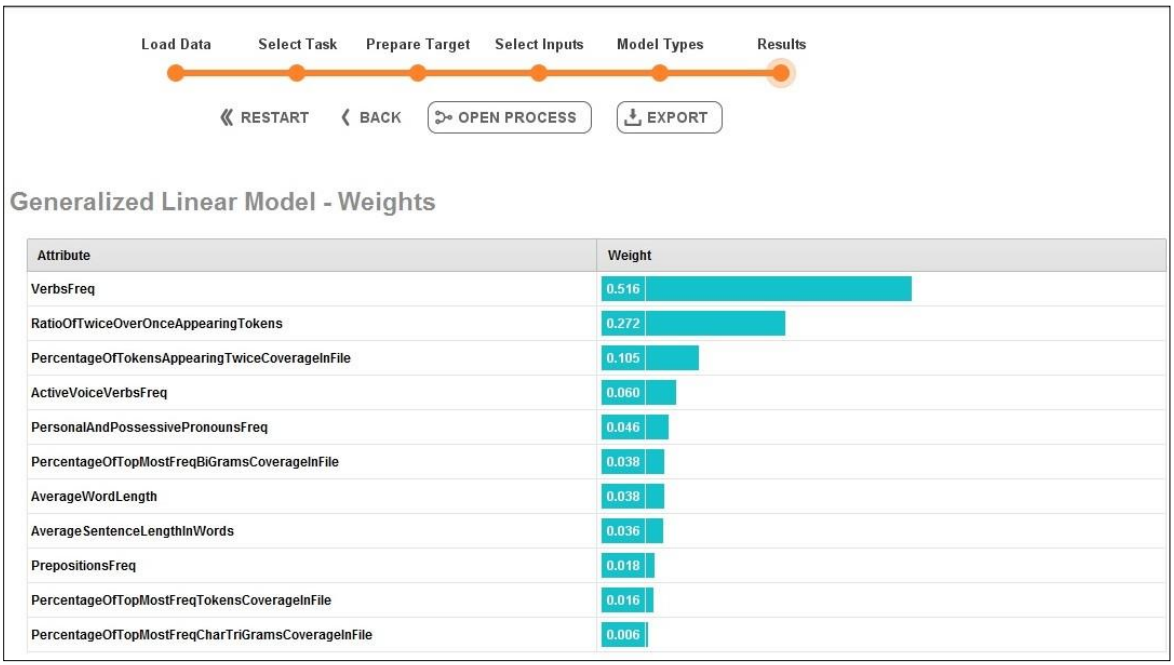
Εικόνα 29: Βάρη για Ευσυνειδησία

Η παρακάτω εικόνα με τα βάρη δείχνει ποια υφομετρικά χαρακτηριστικά έχουν τη μεγαλύτερη επίδραση στην πρόβλεψη για το συγκεκριμένο μοντέλο. Για την Εξωστρέφεια επέδρασαν σημαντικά κατά φθίνουσα σειρά: το μέσο μήκος πρότασης σε λέξεις (0,521), ο λόγος των δις προς άπαξ λεγόμενα (0,127), οι προσωπικές και κτητικές αντωνυμίες 0,123), τα συχνότερα δίλεκτα (0,123), τα επιρρήματα (0,107), οι πιο σπάνιες μη λειτουργικές λέξεις (0,061), οι σύνδεσμοι (0,052), οι προθέσεις (0,046) και οι λέξεις που απαντούν δύο μόνο φορές (0,016).



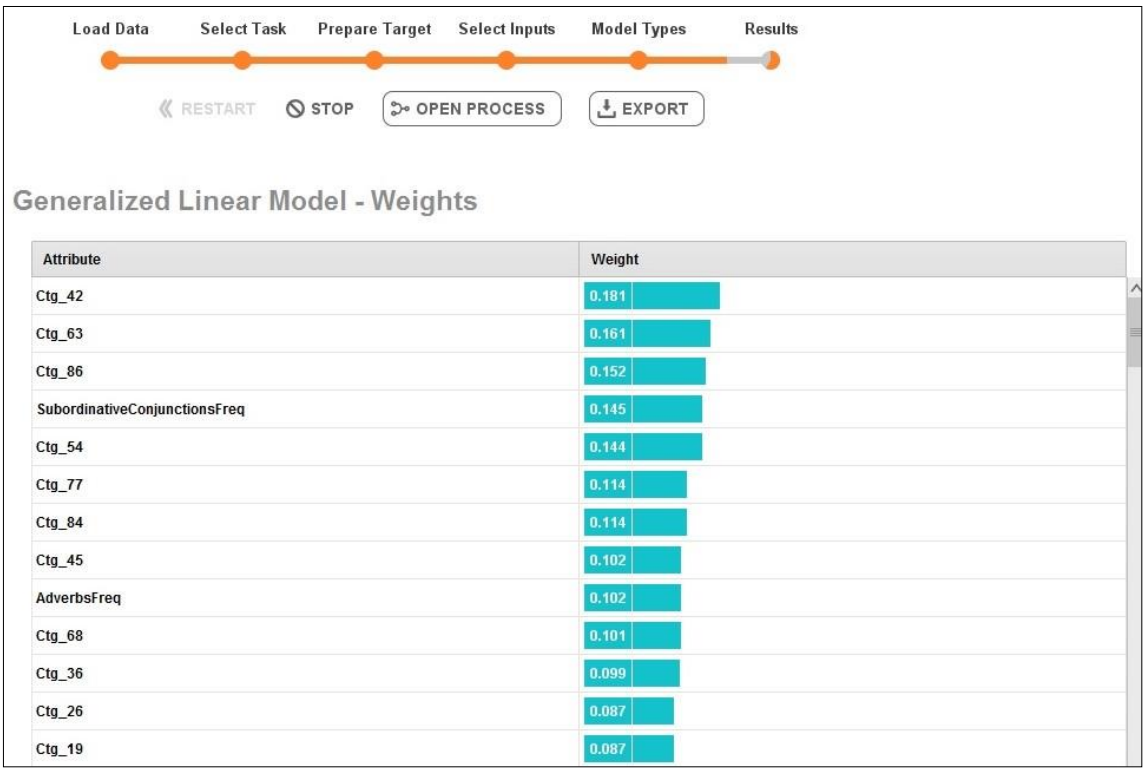
Εικόνα 30: Βάρη για Εξωστρέφεια

Τα χαρακτηριστικά που συνέβαλαν στην απόδοση του αλγόριθμου για την πρόβλεψη της Προσήνειας, έτσι όπως αποτυπώνονται στο RapidMiner, είναι: τα ρήματα (0,516) ο λόγος των δις προς άπαξ λεγόμενα (0,272), οι λέξεις που απαντούν δύο μόνο φορές (0,105), οι ρηματικοί τύποι ενεργητικής φωνής (0,060), οι προσωπικές και κτητικές αντωνυμίες (0,046), τα πιο συχνά δίλεκτα (0,038), το μέσο μήκος λέξης σε χαρακτήρες (0,038), το μέσο μήκος πρότασης σε λέξεις (0,036), οι προθέσεις (0,018), οι πιο συχνές λέξεις (0,016) και τα πιο συχνά τριγράμματα χαρακτήρων (0,006).



Εικόνα 31: Βάρη για Προσήνεια

Για την πρόβλεψη του Νευρωτισμού καταλήξαμε σε ένα άλλο είδος υφομετρικού χαρακτηριστικού. Πρόκειται για τα 100 συχνότερα τριγράμματα χαρακτήρων που μετρήθηκαν σε όλο το ηλεκτρονικό σώμα κειμένων και όχι στα επιμέρους δεδομένα που χωρίστηκαν ανά χαρακτηριστικό και εκεί μετρήθηκαν όλα τα υπόλοιπα 31 υφομετρικά χαρακτηριστικά. Έτσι, εκτός από τα πιο συχνά τριγράμματα χαρακτήρων όλου του σώματος κειμένων, επηρέασαν και οι υποτακτικοί σύνδεσμοι (0,145), τα επιρρήματα (0,102), τα ουσιαστικά (0,081), τα πιο συχνά τρίλεκτα (0,045) και οι ρηματικοί τύποι παθητικής φωνής (0,014).



Εικόνα 32: Βάρη για Νευρωτισμό

Κεφάλαιο 8

Συμπεράσματα και Προοπτικές

Στο τελευταίο αυτό κεφάλαιο συζητούνται τα συμπεράσματα που εξήχθησαν σχετικά με την προβλεπτική ικανότητα των αλγορίθμων μηχανικής μάθησης για την ταξινόμηση των μαθητών με βάση τα χαρακτηριστικά της προσωπικότητάς τους, όπως αυτά εξάγονται από τις εκθέσεις τους. Παρουσιάζονται, επίσης, οι προτάσεις για περαιτέρω έρευνα μετά την ολοκλήρωση των κύριων ερευνητικών σταδίων εκπόνησης της παρούσας διατριβής και κάτω από το πρίσμα των σχετικά πρόσφατα δημοσιευμένων αποτελεσμάτων έρευνας στο εξεταζόμενο πεδίο. Το κεφάλαιο κλείνει με την επισήμανση των σπουδαίων προοπτικών για ανάπτυξη καινοτόμων εφαρμογών που δίνει η αξιοποίηση των αποτελεσμάτων τόσο της παρούσας διατριβής όσο και άλλων συναφών ερευνητικών αποτελεσμάτων.

8.1 Σύνοψη - Συμπεράσματα

Στο πλαίσιο της παρούσας διδακτορικής διατριβής πραγματοποιήθηκε εκτενής επισκόπηση της διεθνούς βιβλιογραφίας στο ερευνητικό πεδίο του καθορισμού δημογραφικών και ψυχολογικών χαρακτηριστικών του συγγραφέα (Authorship Profiling), με ιδιαίτερη έμφαση στην προσωπικότητά του από γραπτό λόγο. Στη βιβλιογραφία επικρατούν δύο προσεγγίσεις για την αυτόματη αναγνώριση της προσωπικότητας του συγγραφέα. Στην ανοδική προσέγγιση, οι ερευνητές ξεκινώντας από τα δεδομένα αναζητούν γλωσσολογικά στοιχεία, τα οποία σχετίζονται με τα χαρακτηριστικά προσωπικότητας και εξάγονται από τα σώματα κειμένων. Αντίθετα, στην καθοδική προσέγγιση αξιοποιούνται λεξικά, για να ελεγχθεί πιθανή συσχέτιση με τα χαρακτηριστικά προσωπικότητας. Επιλέγονται εργαλεία που εξαρτώνται από τη γλώσσα του κειμένου και μέσω αυτών επιλέγονται τα χαρακτηριστικά που θα εξαχθούν από τα σώματα κειμένων και θα τροφοδοτήσουν το μοντέλο πρόβλεψης προσωπικότητας. Και οι δυο προσεγγίσεις έχουν πλεονεκτήματα αλλά και

περιορισμούς (βλ. εν. 4.2), λόγω των οποίων η επιστημονική έρευνα προσανατολίζεται σε υβριδικές μεθόδους που συνδυάζουν τη χρήση λεξικού με μηχανική μάθηση, ώστε να επωφεληθούν από τα πλεονεκτήματα και των δύο προσεγγίσεων, δηλαδή της ταχύτητας και της ακρίβειας αντίστοιχα.

Το ερευνητικό μέρος της εργασίας είχε ως στόχο τον προσδιορισμό της προσωπικότητας μαθητών Λυκείου από ένα ηλεκτρονικό σώμα κειμένων που συντέθηκε με τις εκθέσεις που έγραψαν στην Ελληνική γλώσσα, αφού προηγουμένως το ψυχολογικό προφίλ του κάθε μαθητή, που συμμετείχε στο πείραμα, είχε αποτυπωθεί μέσω δύο προτυποποιημένων ερωτηματολογίων προσωπικότητας, που ευρέως χρησιμοποιούνται διεθνώς σε παρόμοιες ερευνητικές προσπάθειες. Η βασική μεθοδολογία, που ακολουθήθηκε για την επίτευξη του προαναφερθέντος στόχου της διατριβής, στηρίχθηκε στην τεχνική της ταξινόμησης κειμένων, την εκπαίδευση, δηλαδή, αλγορίθμων εποπτευόμενης μάθησης, οι οποίοι υλοποιήθηκαν με το λογισμικό RapidMiner. Τα χαρακτηριστικά που εξήχθησαν από τις εκθέσεις των μαθητών και συνέβαλαν στην ταξινόμηση ήταν αποκλειστικά υφομετρικά. Τελικός στόχος ήταν να προσδιοριστεί ο βαθμός συμφωνίας του αλγόριθμου ταξινόμησης των μαθητών με τα αποτελέσματα των ψυχομετρικών δοκιμών στις ίδιες κατηγορίες που έχουν προκύψει από τις δοκιμές σε ένα σύνολο άγνωστων ως προς τον αλγόριθμο εκθέσεων.

Ο επικρατέστερος αλγόριθμος για την αναγνώριση των τύπων προσωπικότητας όπως αποδίδονται από το Ερωτηματολόγιο Τύπων Προσωπικότητας Myers-Briggs Type Indicator (MBTI) είναι ο Naive Bayes με ποσοστά ακρίβειας 81% για την Εξωστρέφεια (Exrtaverted), 80% για τη Διαίσθηση (iNtuitive), 69% για το Συναίσθημα (Feeling) και 76% για την Κρίση (Judging). Για την πρόβλεψη των χαρακτηριστικών προσωπικότητας βάσει του μοντέλου των Πέντε Παραγόντων επικράτησε ο Generalized Linear Model αλγόριθμος με ακρίβεια 86% για τους μαθητές που δεν είναι δεκτικοί σε νέες εμπειρίες (ClosedMinded), 71% για τους αποδιοργανωμένους (Disorganised), 68% για τους εσωστρεφείς (Intoverted), 70% για τους προσηνείς (Agreeable) και 66% για τους ήρεμους (Calm).

Καθίσταται σαφές (βλ. Κεφ. 7) πως τα ποσοστά ακρίβειας των υπολογιστικών μοντέλων που κατασκευάσαμε για την αναγνώριση προσωπικότητας από κείμενα είναι υψηλά σε σχέση με την υπάρχουσα βιβλιογραφία. Είναι πολύ σημαντικό, βέβαια, να τονίσουμε ότι στο πεδίο έρευνας στο οποίο εντάσσεται η παρούσα διατριβή δεν υπάρχουν δεδομένα αναφοράς με τα οποία να μετρώνται και να συγκρίνονται αντικειμενικά οι επιδόσεις των διαφόρων μεθόδων ανίχνευσης χαρακτηριστικών της

προσωπικότητας. Καμία από τις υπάρχουσες έρευνες δεν χρησιμοποιεί συγκρίσιμες μεθόδους που να έχουν εφαρμοστεί σε ίδια ή συγκρίσιμα σύνολα κειμενικών δεδομένων στην ίδια γλώσσα. Επομένως, τα ποσοστά από την βιβλιογραφία, που ακολουθούν, αφορούν μεν σε έρευνες με κειμενικά δεδομένα εκθέσεις, όχι όμως μαθητών αλλά ενηλίκων, που γράφτηκαν υπό άλλες συνθήκες και σε διαφορετική γλώσσα και βέβαια με άλλα χαρακτηριστικά όχι πάντα υφομετρικά. Έχοντας τα παραπάνω υπόψη, αναφέρουμε ότι για την πρόβλεψη των χαρακτηριστικών προσωπικότητας των μαθητών βάσει του Ερωτηματολογίου Τύπων Προσωπικότητας Myers-Briggs Type Indicator (MBTI) πετύχαμε μέσο όρο ακρίβειας 76,5%, ποσοστό αυξημένο σε σχέση με το 68,62% που καταγράφηκε για την Ολλανδική γλώσσα (Luyckx & Daelemans, 2008a). Αντίστοιχα, για την πρόβλεψη των χαρακτηριστικών προσωπικότητας βάσει του μοντέλου των Πέντε Παραγόντων ο μέσος όρος ακρίβειας στη βιβλιογραφία κυμαίνεται από 57% (Mairesse et al., 2007) έως 60,6% (Mehta et al., 2020), ενώ στην έρευνα που παρουσιάστηκε στην παρούσα διατριβή το ποσοστό είναι 72,2%.

Συνοψίζοντας, τα αποτελέσματα της έρευνάς μας δείχνουν ότι οι υφομετρικές μεταβλητές μπορούν να χρησιμοποιηθούν ως αξιόπιστοι δείκτες πρόβλεψης του ψυχολογικού προφίλ του συγγραφέα. Η πολυεπίπεδη υφομετρική επεξεργασία των μαθητικών εκθέσεων και η ποσοτικοποίηση της γλώσσας τους επιβεβαίωσαν τη διαφορά στο ύφος του γράφοντος ανάλογα με την ιδιοσυγκρασία του, ερευνητική υπόθεση που μελετήθηκε βιβλιογραφικά και τέθηκε από την αρχή της διατριβής και ισχύει επομένως και για την Ελληνική γλώσσα.

Μελέτες στο πεδίο της ταξινόμησης κειμένων με βάση τα χαρακτηριστικά της προσωπικότητας έχουν ικανοποιητικά αποτελέσματα και συνεχώς βελτιώνονται είτε με νέα γλωσσικά χαρακτηριστικά είτε με συλλογή νέων κειμενικών δεδομένων. Οι πρώτες έρευνες εφαρμόστηκαν σε αγγλικά κειμενικά δεδομένα, ωστόσο τα τελευταία χρόνια παρατηρείται έντονη δραστηριότητα και για πολλές άλλες γλώσσες. Επειδή, όμως, η εύρεση κειμενικών πόρων είναι προϋπόθεση για τη διεξαγωγή της έρευνας και ο μεγάλος όγκος ψηφιακά διαθέσιμων γλωσσικών δεδομένων αντιπροσωπεύει μια πλούσια πηγή για την ανάλυση της χρήσης φυσικής γλώσσας, αρκεί βέβαια να συνοδεύεται από τα μεταδεδομένα για τον συγγραφέα, η έρευνα έχει στραφεί στη δημιουργία σωμάτων κειμένων από αναρτήσεις στα μέσα κοινωνικής δικτύωσης.

8.2 Συνεισφορά της διατριβής

Η συνεισφορά της πραγματοποιηθείσας έρευνας στο πλαίσιο της παρούσας διατριβής έγκειται στα ακόλουθα:

- Η διατριβή αποτελεί συμβολή στο διεθνώς ενεργό και ανοικτό ερευνητικό πεδίο της αυτοματοποιημένης αναγνώρισης στοιχείων της προσωπικότητας ενός ανθρώπου, ιδιαίτερα στο υποπεδίο αναγνώρισης προσωπικότητας με βάση την γλώσσα, και, ακόμη πιο ειδικά, στην περιοχή της αναγνώρισης με βάση τον γραπτό λόγο και τα υφολογικά χαρακτηριστικά του συγγραφέα. Στην τελευταία αυτή ερευνητική περιοχή η παρούσα διατριβή αποτελεί για τα ελληνικά δεδομένα την πρώτη πρωτότυπη και ολοκληρωμένη εργασία, η οποία ανοίγει έναν νέο άξονα ερευνητικής δράσης στον ελληνικό χώρο της έρευνας.

- Πρόκειται για την πρώτη μελέτη ανίχνευσης χαρακτηριστικών της προσωπικότητας σε γραπτά κείμενα μαθητών Λυκείου τόσο στα Ελληνικά όσο και σε άλλες γλώσσες.

- Η διατριβή στηρίζεται σε “state of the art” υπολογιστικούς αλγορίθμους και σε σύγχρονα εργαλεία πληροφορικής, πράγμα που είναι συμβατό με την τρέχουσα και μελλοντική τάση της οριζόντιας αξιοποίησης της πληροφορικής στο σύνολο σχεδόν των ερευνητικών πεδίων.

- Παρουσιάστηκε λεπτομερής επισκόπηση της διεθνούς βιβλιογραφίας στον τομέα του καθορισμού των δημογραφικών και κυρίως των ψυχολογικών χαρακτηριστικών του συγγραφέα από τις πρώτες ερευνητικές προσπάθειες μέχρι τις πιο πρόσφατες εξελίξεις. Έγινε αναφορά ανά μελέτη στα κειμενικά δεδομένα, τη μεθοδολογία, τα χαρακτηριστικά, τους αλγόριθμους που χρησιμοποιήθηκαν και τελικά στα αποτελέσματα.

- Αναπτύχθηκε ένα νέο ηλεκτρονικό σώμα κειμένων στην Ελληνική γλώσσα για τις ανάγκες της έρευνας της προσωπικότητας του συγγραφέα. Περιλαμβάνει σχεδόν 250.000 λέξεις σε μορφή εκθέσεων γραμμένες από 198 μαθητές

Λυκείου. Πρόκειται για συλλογή πρωτογενών κειμενικών δεδομένων από φυσικούς ομιλητές της Ελληνικής γλώσσας, που πραγματοποιήθηκε στη διάρκεια τριών διδακτικών ετών, ψηφιοποιήθηκε με πληκτρολόγηση και συνοδεύεται από δημογραφικά και ψυχολογικά χαρακτηριστικά των συγγραφέων, όπως αυτά προκύπτουν από τα δύο ερωτηματολόγια προσωπικότητας που χορηγήθηκαν στους μαθητές, το Ερωτηματολόγιο Τύπων Προσωπικότητας Myers-Briggs Type Indicator (MBTI) και το Ερωτηματολόγιο Προσωπικότητας του μοντέλου των Πέντε Παραγόντων.

- Προτάθηκαν υφομετρικά γλωσσολογικά χαρακτηριστικά για την αυτόματη πρόβλεψη της προσωπικότητας του συγγραφέα, τα οποία διαφοροποιούνται ανάλογα με το είδος της επισημείωσης του σώματος κειμένων και με τη χρήση τους επιτυγχάνονται αρκετά υψηλά ποσοστά προβλεπτικής ικανότητας των αλγορίθμων. Είναι σημαντικό πως υπάρχει πλέον για την Ελληνική γλώσσα ένα σύνολο υφομετρικών δεικτών που έχουν δοκιμαστεί με εννέα αλγόριθμους και έχουν προκύψει συγκεκριμένα αποτελέσματα.

- Απαντήθηκε το ερώτημα ποιοι είναι οι επικρατέστεροι αλγόριθμοι μηχανικής μάθησης με την εκπαίδευση των οποίων επιτεύχθηκε το βέλτιστο αποτέλεσμα.

- Δημιουργήθηκε η υποδομή σε εργαλεία, σε δεδομένα και σε αλγόριθμους επεξεργασίας των δεδομένων για εξέλιξη της έρευνας στον τομέα της αναγνώρισης της προσωπικότητας του συγγραφέα, αλλά και της Επεξεργασίας Φυσικής Γλώσσας γενικότερα.

- Τίθενται οι βάσεις και παρέχονται τα υπολογιστικά εργαλεία για συγκρότηση και επεξεργασία μεγαλύτερων σωμάτων κειμένων, καθώς και για ανάπτυξη αυτοματοποιημένης διαδικασίας για την ψυχολογική αξιολόγηση εφήβων μαθητών και την αποτελεσματική διαχείρισή τους σε μια από τις πιο ευαίσθητες περιόδους της ζωής τους.

- Γίνεται διαθέσιμη στους εκπαιδευτικούς μία νέα μέθοδος κατανόησης και αποκωδικοποίησης της προσωπικότητας των μαθητών τους μέσω του γραπτού

τους λόγου και επομένως μπορούν να έχουν στη διάθεσή τους ένα εργαλείο για προσαρμοσμένη ακόμα και εξατομικευμένη μάθηση.

8.3 Προοπτικές και τομείς συνέχισης της έρευνας

Τα αποτελέσματα που μόλις προαναφέρθηκαν είναι δυνατόν να ενισχυθούν και να διευρυνθούν σε μελλοντική έρευνα και να αξιοποιηθούν για ανάπτυξη σχετικών υπολογιστικών εργαλείων ερευνητικών ή/και εμπορικών. Βασική στόχευση για την συνέχιση της ερευνητικής εργασίας στον κλάδο της αυτόματης αναγνώρισης της προσωπικότητας του συγγραφέα είναι η διερεύνηση νέων χαρακτηριστικών αλλά και η δοκιμή επιπλέον υφομετρικών χαρακτηριστικών. Στην παρούσα διατριβή αξιοποιήσαμε μόνο γλωσσολογικά υφομετρικά χαρακτηριστικά. Σε μελλοντική έρευνα θα δοκιμαστούν και άλλα νέα χαρακτηριστικά ή χαρακτηριστικά που ήδη απαντούν στη βιβλιογραφία, όπως χαρακτηριστικά περιεχομένου, ψυχογλωσσολογικά, συντακτικά, τα οποία, βέβαια, απαιτούν, μεταξύ άλλων, δημιουργία λεξικού για την επισημείωση του σώματος κειμένων. Επιπλέον, χρειάζεται να γίνει επιλογή χαρακτηριστικών ανάλογα με το σώμα κειμένων, καθώς, για παράδειγμα, διαφορετικά γράφει κάποιος σε μια σχολική έκθεση ή ένα δοκίμιο και διαφορετικά στα μέσα κοινωνικής δικτύωσης.

Τα αποτελέσματα των ποσοτικών αναλύσεων σχεδιάζεται να αξιοποιηθούν για να διερευνηθεί σε βάθος η σχέση των υφομετρικών δεικτών που επιλέχθηκαν με συγκεκριμένα ψυχολογικά χαρακτηριστικά. Περαιτέρω έρευνα και διερεύνηση απαιτείται, δηλαδή, για την ερμηνεία της χρήσης των υφομετρικών χαρακτηριστικών που επελέγησαν, ώστε να εξηγηθεί πώς η χρήση συγκεκριμένων μερών του λόγου, για παράδειγμα, συνδέεται με κάποιο ψυχολογικό χαρακτηριστικό.

Είναι διαπιστωμένη η ανάγκη για ανάπτυξη μεγάλου όγκου αντιπροσωπευτικών δεδομένων και αποτελεί προαπαιτούμενο για οποιαδήποτε σχετική έρευνα. Χρειάζεται η δημιουργία ηλεκτρονικών σωμάτων ελληνικών κειμένων εκπαίδευσης των χρησιμοποιούμενων αλγοριθμικών μεθόδων, αλλά και αντίστοιχων σωμάτων αναφοράς για τον έλεγχο της απόδοσης των μεθόδων αυτών. Όσον αφορά στην παρούσα ερευνητική εργασία σχεδιάζεται να αυξηθεί το μέγεθος του σώματος κειμένων που χρησιμοποιήθηκε με επιπλέον εκθέσεις μαθητών. Για την εξαγωγή πιο

αξιόπιστων συμπερασμάτων η επαύξηση του σώματος κειμένων πρέπει να διασφαλίζει την ισορροπία στα κειμενικά είδη και στα διαφορετικά προφίλ προσωπικότητας. Για να γίνει, δηλαδή, σύγκριση αποτελεσμάτων ως προς την αυτόματη αναγνώριση προσωπικότητας θα μπορούσε να εμπλουτισθεί το σώμα κειμένων με εκθέσεις διαφορετικής θεματολογίας και κειμενικού γένους, καθώς και με επαρκή δεδομένα για κάθε ψυχολογικό τύπο.

Μία σημαντική προοπτική για μεταδιδακτορική έρευνα ή νέα διδακτορικά είναι η ανάπτυξη ενός υπολογιστικού εργαλείου που θα βοηθάει τους εκπαιδευτικούς στην στάθμιση του χαρακτήρα των μαθητών και θα συμβάλλει θετικά στην εν γένει εκπαιδευτική διαδικασία, στον επαγγελματικό προσανατολισμό, στις κοινωνικές σχέσεις.

Επιπλέον, η ανάπτυξη της γλωσσικής τεχνολογίας για τα Αγγλικά, που προβλέπεται να συνεχίσει να συμβαίνει με αυξανόμενους ρυθμούς (λόγω ραγδαίας εξέλιξης της τεχνητής νοημοσύνης, της έντονα αναδυόμενης δυναμικής για εξομοίωση της επικοινωνίας ανθρώπου-μηχανής με την επικοινωνία ανθρώπου με άνθρωπο κλπ.), στον βαθμό που θα συμπαρασύρει και την ανάπτυξη της γλωσσικής τεχνολογίας και για τα Ελληνικά, αναμένεται να δώσει τα επόμενα χρόνια εξελιγμένους πόρους γλωσσικής τεχνολογίας βελτιστοποιημένους για την Ελληνική γλώσσα, οι οποίοι αξιοποιούμενοι, ως έχουν ή σε συνδυασμό με κατάλληλα στατιστικά εργαλεία, προσδοκάται βάσιμα ότι θα αυξήσουν την απόδοση των τεχνικών αυτόματης αναγνώρισης των ψυχολογικών ή άλλων χαρακτηριστικών συγγραφέων.

Οι μέθοδοι υφομετρικής ανάλυσης πρέπει, επίσης, να εξελιχθούν με την ανάπτυξη εξειδικευμένων λογισμικών υφομετρικής ανάλυσης. Το RapidMiner είναι σίγουρα ένα εργαλείο που βοηθάει τους ερευνητές του χώρου που δεν έχουν ιδιαίτερες γνώσεις μηχανικού υπολογιστών, ωστόσο, επειδή έχει εφαρμογή σε πολύ μεγάλο εύρος διαφορετικών μεταξύ τους γνωστικών πεδίων, δεν μπορεί παρά να υστερεί σε σύγκριση με ένα υπολογιστικό περιβάλλον εστιασμένο στην υφομετρική ανάλυση.

Σχετικά με τα αποτελέσματα της έρευνάς μας, δεν μπορεί να γίνει άμεση και αντικειμενική σύγκριση με τα αντίστοιχα της διεθνούς βιβλιογραφίας, γιατί οι μελέτες αφορούν σε άλλες γλώσσες. Από τη βιβλιογραφική επισκόπηση διαπιστώθηκε ότι οι περισσότερες μελέτες επιχειρούν την πρόβλεψη της προσωπικότητας από κείμενα της Αγγλικής και επομένως τα πορίσματα που αφορούν στα εργαλεία και στα χαρακτηριστικά αναφέρονται σε αυτή τη γλώσσα. Σε μελλοντική έρευνα χρειάζεται να αντιμετωπισθεί το ζήτημα συγκριτικά, για να ελεγχθεί εάν και για τα Ελληνικά ισχύουν

διαπιστώσεις σχετικά με τα ψυχολογικά χαρακτηριστικά (εάν π.χ. ο εξωστρεφής και στην Ελληνική γλώσσα χρησιμοποιεί στο λόγο του περισσότερες προσωπικές αντωνυμίες όπως και στην Αγγλική).

Όσον αφορά στην ερμηνεία χρήσης των συγκεκριμένων γλωσσολογικών επιλογών από τους συμμετέχοντες στην έρευνά μας μαθητές, τα πρώτα συμπεράσματα που εξαγάγαμε μπορούν να συμπληρωθούν και να τεκμηριωθούν με μια διεπιστημονική μελέτη γλωσσολογίας και ψυχολογίας, η οποία εύλογα μπορεί να ακολουθήσει την εργασία που παρουσιάσθηκε στην παρούσα διατριβή.

8.4 Προοπτικές εφαρμογών και καινοτομίας

Από την ανάλυση που έγινε μέχρι τώρα γίνεται φανερό ότι η δυνατότητα αυτοματοποιημένης διαδικασίας αναγνώρισης του ψυχολογικού προφίλ των χρηστών ανοίγει τον δρόμο για μια πλειάδα εφαρμογών, οι οποίες μπορούν να επηρεάσουν θετικά την καθημερινή ζωή των ανθρώπων, καθώς και την οργάνωση και λειτουργία των κοινωνιών τους. Αναμένεται, λοιπόν, να συνεχισθεί βραχυ-μεσοπρόθεσμα με εντατικότερους ρυθμούς η επένδυση ανθρωποπροσπάθειας στο συγκεκριμένο ερευνητικό πεδίο, προκειμένου αφενός να προωθηθεί η παραγωγή νέας γνώσης και αφετέρου να παραχθούν αντίστοιχες καινοτόμες εφαρμογές.

Μια τέτοια προοπτική έρευνας και καινοτομίας μπορεί να γίνει περισσότερο εύληπτη υπό το πρίσμα των τομέων εφαρμογών, οι οποίοι ενδεικτικά αναφέρονται στην συνέχεια και στους οποίους ως ένα βαθμό έχουν αρχίσει ήδη να παράγονται αποτελέσματα (εφαρμογές, νέα γνώση, εργαλεία κλπ.) ή αναμένεται να εμφανισθούν σε κοντινό μέλλον:

- Η αυτοματοποιημένη ανίχνευση χαρακτηριστικών προσωπικότητας ενός ανθρώπου, με βάση τον γραπτό λόγο και τα υφολογικά χαρακτηριστικά του συγγραφέα, θα προωθήσει την έρευνα επάνω σε νέες πτυχές του δυναμικού φάσματος της ανθρώπινης ψυχής και θα συμβάλει έτσι στην ανακάλυψη περισσότερο σύνθετων και λεπτοφυών σχέσεων μεταξύ της συμπεριφοράς των ανθρώπων και της προσωπικότητάς τους.
- Η διαχείριση ανθρώπινων πόρων σε επαγγελματικό ή μη περιβάλλον είναι ένας ακόμα τομέας όπου μπορεί να εφαρμοστούν τα πορίσματα της έρευνας. Η

αυτοματοποιημένη αναγνώριση του ψυχολογικού προφίλ θα υποβοηθήσει π.χ. στην αποτελεσματικότερη επιλογή του πιο κατάλληλου υποψήφιου για μια θέση εργασίας με ειδικές απαιτήσεις, αλλά και στην καλύτερη διαχείριση του προσωπικού μιας εταιρείας, ενός οργανισμού, ενός εκπαιδευτικού φορέα κλπ.

- Μια άλλη εφαρμογή των αποτελεσμάτων της παρούσας διατριβής, αλλά και γενικότερα των ώριμων αποτελεσμάτων στο ερευνητικό πεδίο, στο οποίο αυτή η διατριβή εστίασε, είναι στον τομέα της δημιουργίας προφίλ χρηστών (user profiling), όπου η αυτόματη αναγνώριση και κατάλληλη αξιοποίηση των χαρακτηριστικών της προσωπικότητας επιτρέπει την ευστοχότερη παροχή π.χ. εξατομικευμένων συμβουλών για θέματα κοινωνικών σχέσεων, ψυχικής και διανοητικής υγείας, καθώς και υπηρεσιών marketing προσαρμοσμένων στον ψυχολογικό τύπο και τις προτιμήσεις του χρήστη (διαδικτυακό εμπόριο, μηχανές αναζήτησης, recommendation systems). Επίσης, επιτρέπει την σχεδίαση πιο προσωποποιημένων διεπαφών χρήστη σε εφαρμογές λογισμικού για το διαδίκτυο, για έξυπνες συσκευές για επιτραπέζιους υπολογιστές κ.ά.
- Επιπρόσθετα, η σημασιολογική αποσαφήνιση λέξεων με βάση τα ανιχνευμένα χαρακτηριστικά προσωπικότητας του συγγραφέα (π.χ. αναγνώριση σαρκαστικού ή μη σαρκαστικού ύφους, αναγνώριση/ανάλυση συναισθήματος) είναι δυνατό να πραγματοποιηθεί με την αξιοποίηση των τεχνικών που παρουσιάστηκαν.
- Τέλος, η αυτόματη αναγνώριση του ψυχολογικού προφίλ του χρήστη με βάση τον προφορικό ή γραπτό λόγο του θα επιτρέψει στις μηχανές να προσαρμόζουν το ύφος των απαντήσεών τους και την εν γένει συνθετική ομιλία τους στα ιδιαίτερα χαρακτηριστικά των ανθρώπων με τους οποίους συνομιλούν και εξυπηρετούν, πράγμα που θα προαγάγει την εξέλιξη των ρομποτικών μηχανών και το επίπεδο επικοινωνίας των ανθρώπων με αυτές.

Βιβλιογραφικές αναφορές

Abbasi A. & Chen H. (2008). Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems (TOIS), 26*(2), 1-29.

Abdalgane M. (2019). Language and gender. *British Journal of English Linguistics 8*(1), 1-8.

Akrami N., Fernquist J., Isbister T., Kaati L. & Pelzer B. (2019). Automatic extraction of personality from text: Challenges and opportunities. *2019 IEEE International Conference on Big Data (Big Data),* pp. 3156-3164, Los Angeles, CA, USA, , doi: 10.1109/BigData47090.2019.9005467

Alam F. & Riccardi G. (2014). Predicting personality traits using multimodal information. *Proceedings of the 2014 ACM Multimedia on Workshop on Computational Personality Recognition*, November 2014, Orlando Florida USA, (pp.15-18).

Alam F. & Riccardi G. (2013). Comparative study of speaker personality traits recognition in conversational and broadcast news speech. *Proceedings of 14th Annual Conference of the International Speech Communication Association*, 25-29 August 2013, Lyon, France, (pp. 2851-2855).

Alam F., Stepanov E. A. & Riccardi G. (2013). Personality traits recognition on social network - Facebook. *Proceedings of Workshop on Computational Personality Recognition,* 11 July 2013, Boston, (pp. 6-9).

Allport W. G. (1961). *Pattern and Growth in Personality*. New York: Holt, Rinehart & Winston.

Álvarez-Carmona M. A., López-Monroy A. P., Montes-y-Gómez M., Villaseñor-Pineda L. & Escalante H. J. (2015). INAOE's participation at PAN'15: Author profiling task notebook for PAN at CLEF 2015. In Linda Cappellato, Nicola Ferro, Gareth Jones, and Eric San Juan, editors, *CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers*.

Argamon S., Dhawle S., Koppel M. & Pennebaker J. W. (2005). Lexical predictors of personality type. *Proceedings of Joint Annual Meeting of the Interface and the Classification Society of North America*.

Argamon S., Koppel M., Fine J. & Shimoni A. R. (2003). Gender, genre, and writing style in formal written texts. *Text-Interdisciplinary Journal for the Study of Discourse 23*(3), 321-346.

Argamon S., Koppel M., Pennebaker J. W. & Schler J. (2009). Automatically profiling the author of an anonymous text. *Communications of the Association for Computing Machinery (CACM) 52(*2), 119-123.

Argamon S., Whitelaw C., Chase P., Dhawle S., Hota S. R., Garg N. & Levitan S. (2007). Stylistic text classification using functional lexical features. *Journal of the American Society of Information Science and Technology, 58*(6), 802-822. doi: 10.1002/asi.v58:6

Baayen H. R., van Halteren H. & Tweedie F. J. (1996). Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing, 11*(3), 121-132. doi: 10.1093/llc/11.3.121

Bachrach Y., Kosinski M., Graepel T., Kohli P. & Stillwell D. (2012). Personality and patterns of Facebook usage. *Proceedings of 4th Annual ACM Web Science Conference*, Evanston Illinois, (pp. 36-45).

Bandura A. (1986). *Social Foundations of Thought and Action: A Social Cognitive Theory*. Englewood Cliffs, NJ: Prentice-Hall.

Biel J.-I. & Gatica-Perez D. (2013). The YouTube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs. *IEEE Transactions on Multimedia, 15*(1), 41-55.

Biel J.-I., Tsiminaki V., Dines J. & Gatica-Perez D. (2013). Hi YouTube! Personality impressions and verbal content in social video. *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, 9-13 December 2013, Sydney, Australia, (pp. 119-126).

Bowden-Green T., Hinds J. & Joinson A. (2020). How is extraversion related to social media use? A literature review. *Personality and Individual Differences*, vol. 164.

Bradac J. J. (1990). Language attitudes and impression formation. In H. Giles & W.P. Robinson (eds.), *Handbook of Language and Social Psychology*, 387-412. Chichester: Wiley.

Brinks D. & White H. (2012). Detection of Myers-Briggs type indicator via text based computer-mediated communication. CS 229 *Machine Learning Projects*, Stanford.

Burger J. D. & Henderson J. C. (2006). An exploration of observable features related to blogger age. *Proceedings of the AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs*, California, USA, (pp.15-20).

Burman D. D., Bitan T. & Booth R. J. (2008). Sex differences in neural processing of language among children. *Neuropsychologia 46*(5), 1349-1362.

Burrows J. F. (1987). *Computation Into Criticism: A Study of Jane Austen's Novels and an Experiment in Method*. Oxford: Clarendon Press.

Cattell R. B. (1957). *Personality and Motivation Structure and Measurement.* New York: Harcourt, Brace, & World.

Cattell R. B., Eber H. W. & Tatsuoka M. M. (1970). *Handbook for the Sixteen Personality Factor Questionnaire*. Champaign, IL: Institute for Personality and Ability Testing.

Celli F. (2012). *Adaptive Personality Recognition from Text*. Ph.D. Thesis, University of Trento: Italy.

Celli F., Pianesi F., Stillwell D. & Kosinski M. (2013). Workshop on computational personality recognition (Shared Task). *Proceedings of WCPR13, in conjunction with International Conference on Weblogs and Social Media-13*.

Celli F. & Poesio M. (2014). PR2: A language independent unsupervised tool for personality recognition from text. In Arxiv.org/abs/1402.2796.

Child I. L. (1968). Personality in culture. In Borgatta E. & Lambert W. W. (Eds.), *Handbook of Personality Theory and Research*, 80-101. Chicago: Rand McNally.

Chin D. N. & Wright W. R. (2014). Social media sources for personality profiling. *Proceedings of the 2nd Workshop Emotions and Personality in Personalized Services*, 7-11 July 2014, Aalborg, Denmark, (pp.79–85).

Chittaranjan G., Blom J. & Gatica-Perez D. (2012). Mining large-scale smartphone data for personality studies. *Personal and Ubiquitous Computing 17*(3). doi: 10.1007/s00779-011-0490-1

Chung C. & Pennebaker J. (2007). The psychological functions of function words. In Fiedler K. (Ed.), *Social Communication*, 343-359, New York: Psychology Press.

Cohen S. A., Minor S. K., Baillie E. L. & Dahir M. A. (2008). Clarifying the linguistic signature: Measuring personality from natural speech. *Journal of Personality Assessment, 90*(6), 559–563.

Coniam D. (2004). Concordancing oneself: Constructing individual textual profiles. *International Journal of Corpus Linguistics 9*(2), 271-298.

Cornett H. E. (2014). Gender differences in syntactic development among English speaking adolescents. *Inquiries Journal/Student Pulse, 6*(3),1–6.

Corney M. W. (2003). *Analysing e-mail Text Authorship for Forensic Purposes.* Master, Queensland University of Technology: Queensland.

Costa Jr. P. T. & McCrae R. R. (1985). *The NEO Personality Inventory Manual*. Odessa, Fla.: Psychological Assessment Resources.

Costa Jr. P. T. & McCrae R. R. (1993). *NEO-PI-R: Professional Manual*. Odessa, Fla.: Psychological Assessment Resources.

Covington M. A., He C., Brown C., Naci L., McClain J. T., Fjordbak B. S., Semple J. & Brown J. (2005). Schizophrenia and the structure of language: The linguist's view. *Schizophrenia Research 77*, 85-98.

Craig H. D. & Kinney A. F. (2009). *Shakespeare, Computers and the Mystery of Authorship*. Cambridge: Cambridge University Press.

Daelemans W. (2013). Explanation in computational stylometry. *Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics*, 24-30 March 2013, Samos, Greece, (pp. 451–462).

Dandannavar P. S., Mangalwede S. R. & Kulkarni P. M. (2018). Social media text-A source for personality prediction. *International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)*, 21 -23 December 2018, Belgaum, India, (pp. 62-65), IEEE.

Dang Duc P., Giang Binh T. & Son Bao P. (2009). Author profiling for vietnamese blogs. *Proceedings of the International Conference on Asian Language Processing, 2009 (IALP '09)*, 7-9 December 2009, Singapore, (pp. 190-194).

Dell'Orletta F. & Nissim M. (2018). Overview of the EVALITA 2018 cross-genre gender prediction (GxG) Task. *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18),* 12-13 December 2018, Turin, Italy.

de Montjoye Yves-A., Quoidbach J., Robic F. & Pentland A. S. (2013). Predicting people personality using novel mobile phone-based metrics. *Social Computing, Behavioral-Cultural Modeling and Prediction, 7812*, 48-55.

de Oliveira R., Karatzoglou A., Armenta A. & Oliver N. (2011). Towards a psychographic user model from mobile phone usage. *Proceedings of the International Conference on Human Factors in Computing Systems, Extended Abstracts Volume*, 7-12 May 2011, Vancouver, BC, Canada.

Digman J.M. & Takemoto-Chock N.K. (1981). Factors in the natural language of personality: Re-analysis, comparison and interpretation of six major studies. *Multivariate Behavioral Research, 16*, 149-170.

Digman J.M. (1988). Classical theories of trait organization and the Big Five factors of personality. Presented at the *Annual Meeting of American Psychological Association, Atlanta.*

Dreikurs R. (1975). *Οι Βασικές Αρχές της Αντλεριανής Ψυχολογίας,* μετάφραση Πανταζή, Αθήνα: Κέδρος.

Eysenck H. J. & Rachman (1965). *The Causes and Cures of Neurosis*. San Diego, Ca: Knapp.

Eysenck H. J. (1952). The effects of psychotherapy: An evaluation. *Journal of Consulting Psychology, 16*, 319-324.

Eysenck H. J. (1990). Genetic and environmental contributions to individual differences: The three major dimensions of personality. *Journal of Personality***,** *58*(1), 245-261.

Eysenck H. J. (1992). Four ways five factors are not basic. *Personality and Individual Differences, 13*, 667-673.

Farnadi G., Sitaraman G., Sushmita S., Celli F., Kosinski M., Stillwell D., Davalos S., Moens M.-F. & De Cock M.. (2016). Computational personality recognition in social media. *User Modeling and User-Adapted Interaction. June 2016, 26*(2), 109-142.

Farnadi G., Sitaraman G., Rohani M., Kosinski M., Stillwell D., Moens M.-F., Davalos S. & De Cock M.. (2014a). How are you doing? Emotions and personality in Facebook. *Proceedings of 2nd Workshop on Emotions and Personality in Personalized Services, Workshop at 22nd Conference on User Modelling, Adaptation and Personalization*, (pp. 45-56).

Farnadi G., Sushmita S., Sitaraman G., Ton N., De Cock M. & Davalos S.. (2014b). A multivariate regression approach to personality impression recognition of vloggers. *Proceedings of the 2nd International Workshop in Computational Personality Recognition, Workshop at 22nd Conference on Multi Media,* (pp. 1-6).

Farnadi G., Zoghbi S., Moens M.-F. & De Cock M.. (2013). Recognising personality traits using facebook status updates. *Proceedings of the Workshop on Computational Personality Recognition at the 7th International AAAI Conference on Weblogs and Social Media (ICWSM13),* (pp.14-18).

Forstall C. & Scheirer W.(2010). Features from frequency: Authorship and stylistic analysis using repetitive sound. *Journal of the Chicago Colloquium on Digital Humanities and Computer Science*, *1*(2), 1-23.

Freud S.. (1923). *The Ego and the Id*. Trans. J. Riviere and rev. and ed. J. Strachey. New York: Norton, 1960.

Galton F.. (1884). *The Measurement of Character.* Dennis, Wayne (Ed). (1949). Readings in general psychology, pp. 435-444. New York, NY, US: Prentice-Hall.

García A. M. & Martín J. C. (2007). Function words in authorship attribution studies. *Literary and Linguistic Computing*, *22*(1), 49-66.

Gjurkovic M. & Šnajder J. (2018). Reddit: A gold mine for personality prediction. *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media,* 6 June 2018, New Orleans, Louisiana, USA, (pp. 87-97).

Gill A. J. (2003). *Personality and Language: The Projection and Perception of Personality in Computer-mediated Communication*. Ph.D. Thesis, University of Edinburgh: Scotland.

Gill A. J. & French R. M. (2007). Level of representation and semantic distance: Rating author personality from texts. *Proceedings of the 2nd European Cognitive Science Conference*, 23-27 May 2007, Delphi, Greece, (pp. 682-687)

Gill A. J. & Oberlander J. (2002). Taking care of the linguistic features of extraversion. *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, 7-10 August 2002, Fairfax, Virginia, USA, (pp. 363-368).

Gill A. J., Nowson S. & Oberlander J. (2009). What are they blogging about? Personality, topic and motivation in blogs. *Proceedings of the Third*

International Conference on Weblogs and Social Media, 17–20 May 2009, San Jose, California, USA.

Gill A. J., Nowson S. & Oberlander J. (2006). Language and personality in computer-mediated communication: A cross-genre comparison. *Journal of Computer-Mediated Communication*.

Giouli V., Konstandinidis A., Desipri E. & Papageorgiou H. (2006). Multi-domain multi-lingual named entity recognition: Revisiting & grounding the resources issue. *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2006),* Genoa, Italy, (pp. 59-64).

Golbeck J., Robles C., Edmondson M. & Turner K. (2011). Predicting personality from Twitter. *Privacy, Security, Risk and Trust (PASSAT), 2011 IEEE Third International Conference on Social Computing (SocialCom),* 9-11 October 2011, Boston, MA, (pp149–156).

Golbeck J., Robles C. & Turner K.. (2011). Predicting personality with social media. *Proceedings of the 2011 Annual Conference Extended Abstracts on Human Factors in Computing Systems*, (pp. 253-262).

Goldberg L. R. (1990). An alternative description of personality. The Big-Five factor structure. *Journal of Personality and Social Psychology, 59*, 1216-1229.

Gollub T., Potthast M., Beyer A., Busse M., Rangel F., Rosso P., Stamatatos E. & Benno S. (2013). Recent trends in digital text forensics and its evaluation plagiarism detection, author identification, and author profiling. *Proceedings of the CLEF Intiative: Information Access Evaluation. Multilinguality, Multimodality, and Visualization.* Springer LNCS, 8138, (pp. 282-301).

González-Gallardo C. E., Montes A., Sierra G., Núñez-Juárez J. A., Salinas-López A. J. & Ek J. (2015). Tweets classification using corpus dependent tags, character and POS N-grams. *Working Notes of CLEF 2015 Conference and Labs of the Evaluation forum*, 8-11 September 2015, Toulouse, France.

Goswami S., Sarkar S. & Rustagi M. (2009). Stylometric analysis of bloggers' age and gender. *Proceedings of the Third International Conference on Weblogs and Social Media, ICWSM 2009,* 17-20 May 2009, San Jose, California, USA.

Grivas A., Krithara A. & Giannakopoulos G. (2015). Author profiling using stylometric and structural feature groupings. *Working Notes of CLEF 2015 Conference and Labs of the Evaluation forum*, 8-11 September 2015, Toulouse, France.

Hall K., Borba R. & Hiramoto M. (2020). Language and gender. *The International Encyclopedia of Linguistic Anthropology*, https://doi.org/10.1002/9781118786093.iela0143

van Halteren H. (2004). Linguistic profiling for author recognition and verification. *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. ACL '04, Association for Computational Linguistics,* Stroudsburg, PA, USA.

van Halteren H., Baayen H., Tweedie F., Haverkort M. & Neijt A. (2005). New machine learning methods demonstrate the existence of a human stylome. *Journal of Quantitative Linguistics, 12*(1), 65-77.

Herrmann J. B., van Dalen-Oskam K. & Schöch C. (2015). Revisiting style, a key concept in literary studies. *Journal of Literary Theory 9*(1), 25-52.

Hilgard E. (1962). *Introduction to Psychology* (3rd ed.), London: Methuen.

Hinds J. & Joinson A. (2019). Human and computer personality prediction from digital footprints. *Current Directions in Psychological Science 2*(2), 204-211.

Hirschberg N. (1978). A correct treatment of traits. In H. London (Ed.), *Personality: A new look at metatheories*, 45-48. New York: Macmillan.

Hjelle L. A. & Ziegler D. J. (1992). *Personality Theories, Basic Assumptions, Research and Applications* (3rd ed.), McGraw - Hill, Inc.

Holmes D. I. & Forsyth R. S. (1995). The Federalist revisited: New directions in authorship attribution. *Literary and Linguistic Computing*, *10*(2), 111-127. doi: 10.1093/llc/10.2.111

Holtgraves T. (2011). Text messaging, personality, and the social context. *Journal of Research in Personality*, *45*(1), 92-99.

Hota S. R., Argamon S., Koppel M. & Zigdon I. (2006). Performing gender: Automatic stylistic analysis of Shakespeare's characters. *Proceedings of Digital Humanities 2006,* Paris, France.

Hsieh F., Dias R.& Paraboni I. (2018). Author profiling from Facebook corpora. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 7-12 May 2018, Miyazaki, Japan.

Iacobelli F., Gill A. J., Nowson S. & Oberlander J. (2011). Large scale personality classification of bloggers. *Proceedings of the 4th International Conference on Affective Computing and Intelligent Interaction*, Berlin, Heidelberg: Springer-Verlag.

Iqbal F., Khan L.A., Fung B. C. M. & Debbabi M. (2010). E-mail authorship verification for forensic investigation. *Proceedings of the 2010 ACM Symposium on Applied Computing*, Sierre Switzerland, (pp. 1591–1598).

Ivanov V. A., Riccardi G., Sporka J. A. & Franc J. (2011). Recognition of personality traits from human spoken conversations. *Proceedings of Interspeech 2011, International Conference*, 27-31 August 2011, Florence, Italy, (pp.1549-1552).

John O.P., Hampson S. E. & Goldberg L. R. (1991). The basic level in personality-trait hierarchies: Studies of trait use and accessibility in different contexts. *Journal of Personality and Social Psychology, 60*, 348-361.

John O. P. & Srivastana S. (1999). The Big-Five Trait Taxonomy: History, Measurement, and Theoretical Perspectives, In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (2nd ed.), pp.102–138. New York: Guilford Press.

Jung C. (1954). *On the Nature of the Psyche*. 1988 ed. London: Ark Paperbacks. (contained in Collected Works Vol. 8)

Juola P. (2008). Authorship attribution. *Foundations and Trends in Information Retrieval, 1*(3), 233-334.

Kelly A. G. (1955). *The Psychology of Personal Constructs*. Volumes 1 and 2, New York: Norton.

Kermanidis K. (2012). Mining authors' personality traits from modern Greek spontaneous text. *Proceedings of Workshop on Corpora for Research on Emotion Sentiment & Social Signals, in conjunction with LREC2012*, 26 May 2012, Istanbul, Turkey, (pp. 90-93).

Kestemont M., Stamatatos E., Manjavacas E., Daelemans W., Potthast M. & Stein B. (2019). Overview of the cross-domain authorship attribution task at PAN 2019. In Linda Cappellato, Nicola Ferro, David E. Losada, and Henning Müller, editors, *CLEF 2019 Labs and Workshops, Notebook Papers*, September 2019.

Komianos V., Moustaka E., Andreou M., Banou E., Fanarioti S. & Kermanidis K. L. (2012). Predicting personality traits from spontaneous modern Greek text: Overcoming the barriers. *International Conference on Artificial Intelligence Applications and Innovations Advances in Information and Communication Technology*, L. Iliadis et al. (Eds.): AIAI 2012 Workshops, IFIP AICT 382, (pp. 530-539).

Koppel M., Argamon S. & Shimoni A. R. (2002). Automatically categorizing written texts by author gender. *Literary and Linguistic Computing, 17*(4), 401-412.

Koppel M., Schler J. & Argamon S. (2009). Computational methods in authorship attribution. *Journal of the American Society of Information Science and Technology, 60* (1), 9-26.

Koppel M., Schler J. & Zigdon K. (2005). Determining an author's native language by mining a text for errors. *Proceedings of the eleventh International Conference on Knowledge Discovery in Data Mining,* 21-24 August 2005, Chicago Illinois, USA, (pp. 624-628).

Kosinski M., Bachrach Y., Kohli P., Stillwell D. & Graepel T. (2013). Manifestations of user personality in website choice and behaviour on online social networks. *Machine Learning 95*(3), 1–24.

Langley P., Iba W. & Thompson K. (1992). An analysis of bayesian classifiers. *Proceedings of the Tenth National Conference on Artificial Intelligence*, 12-16 July, San Jose, California, pp. 223-228).

Lee H. C., Kim K., Seo Seok Y. & Chung K. C.. (2007). The relations between personality and language Use. *The Journal of General Psychology*, *134*(4), 405–413.

Li J. & Chignell M. (2010). Birds of a feather: How personality influences blog writing and reading. *International Journal of Human-Computer Studies, 68*(9), 589-602.

Liu F., Perez J. & Nowson S. (2017). A Language-independent and compositional model for personality trait recognition from short texts. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL),* Valencia, Spain.

Lukito L. C., Erwin A., Purnama J. & Danoekoesoemo W. (2016). Social media user personality classification using computational linguistic. *Proceedings of 8th International Conference on Information Technology and Electrical Engineering,* 5-6 October 2016, Yogyakarta, Indonesia, (pp. 1-6).

Luyckx K. (2010). *Scalability Issues in Authorship Attribution*. Ph.D. Thesis. Brussels: Uitgeverij UPA University Press Antwerp.

Luyckx K. & Daelemans W. (2008a). Personae: A corpus for author and personality prediction from text. *Proceedings of the 6th International Language Resources*

and Evaluation Conference (LREC 2008), 28-30 May 2008, Marrakech, Morocco, (pp. 2981-2987).

Luyckx K. & Daelemans W. (2008b). Using syntactic features to predict author personality from text. *Proceedings of Digital Humanities 2008,* Oulu, Finland, (pp. 146-149).

MacKinnon I. & Warren R. H. (2006). Age and geographic inferences of the livejournal social network. In: Airoldi E., Blei D.M., Fienberg S.E., Goldenberg A., Xing E.P., Zheng A.X. (eds) Statistical Network Analysis: Models, Issues, and New Directions. ICML 2006. *Lecture Notes in Computer Science*, *4503*, Berlin, Heidelberg: Springer.

Mairesse F.s & Walker M. (2007). PERSONAGE: Personality generation for dialogue. *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL),* (pp. 496-503).

Mairesse F.& Walker M. A. (2006). Words mark the nerds: Computational models of personality recognition through language. *Proceedings of the 28th Annual Conference of the Cognitive Science Society,* 26-29 July 2006, Vancouver, Canada, (pp. 543-548).

Mairesse F., Walker M. A., Mehl M. R. & Moore R. K. (2007). Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research, 30*, 457-500.

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

Marengo D. & Settanni M. (2019). Mining Facebook data for Personality prediction: An overview. In: Baumeister H., Montag C. (eds) *Digital Phenotyping and Mobile Sensing. Studies in Neuroscience, Psychology and Behavioral Economics:* Springer.

Markovikj D., Gievska S., Kosinski M. & Stillwell D. (2013). Mining Facebook data for predictive personality modeling. *Proceedings of Workshop on Computational Personality Recognition at 7th International Conference on Weblogs and Social Media,* Boston, Mass, USA, (pp. 23-26).

Marquardt J., Farnadi G., Vasudevan G., Moens M.-F., Davalos S., Teredesai A. & De Cock M. (2014). Age and gender identification in social media. *Proceedings of*

CLEF 2014, Evaluation Labs, 15-18 September 2014, Sheffield, UK, (pp. 1129-1136).

Mascol C. (1888). Curves of Pauline and Pseudo-Pauline Style I. *Unitarian Review, 30*, 452-460.

McCrae R. R. & Costa Jr. P. T. (1996). Toward a new generation of personality theories: Theoretical contexts for the five-factor model. In J. S.Wiggins, editor, *The five-factor model of personality: Theoretical perspectives*, 51–87, Guilford, New York.

Mehl R. M., Gosling D. S. & Pennebaker W. J. (2006). Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. *Journal of Personality and Social Psychology 90*(5), 862–877.

Mehta Y., Fatehi S., Kazameini A., Stachl C., Cambria E. & Eetemadi S. (2020). Bottom-up and top-down: Predicting personality with psycholinguistic and language model features. *2020 IEEE International Conference on Data Mining (ICDM)*.

Mendenhall T. C. (1887). The characteristic curves of composition. *Science, 11*, 237-249.

Mikros G. K. (2013a). 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)*, 16-19 April 2013, Belgrade, Serbia. Belgrade: Academic Mind, pp. 21-32.

Mikros G. K. (2013b). Systematic stylometric differences in men and women authors: A corpus-based study. In R. Köhler & G. Altmann (Eds.), *Issues in Quantitative Linguistics 3. Dedicated to Karl-Heinz Best on the occasion of his 70th birthday*, 206-223. Lüdenscheid: RAM - Verlag.

Mikros G. K. (2007). Stylometric experiments in Modern Greek: Investigating authorship in homogeneous newswire texts. In R. Köhler, G. Altmann & P. Grzybek (Eds.), *Exact Methods in the Study of Language and Text*, 445-456. Berlin / New York: Mouton de Gruyter.

Mikros G. K. (2006). Authorship attribution in Modern Greek newswire corpora. In O. Uzuner, S. Argamon & J. Karlgren (Eds.), *Proceedings of the SIGIR 2006 International Workshop on Directions in Computational Analysis of Stylistics in Text Retrieval, (*pp. 43-47), Seattle, Washington, USA: ACM.

Mikros G. K. (2005). Quantitative linguistics in Greece: An overview. In G. Altmann, R. Köhler & R. G. Piotrowski (Eds.), *Quantitative Linguistics. An International Handbook*, 136-142. Berlin: Walter De Gruyter.

Mikros G. K. & Markopoulos G. (2017). 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.

Mikros G. K. & Perifanos K. A. (2013). Authorship attribution in Greek tweets using author's multilevel N-Gram profiles. In E. Hovy, V. Markman, C. H. Martell & D. Uthus (Eds.), *Papers from the 2013 AAAI Spring Symposium "Analyzing Microtext"*, 25-27 March 2013, Stanford, California (pp. 17-23). Palo Alto, California: AAAI Press.

Mitchell T. (1997). *Machine Learning*. McGraw Hill.

Modaresi P., Liebeck M. & Conrad S.. (2016). Exploring the effects of cross-genre machine learning for author profiling, PAN 2016-Notebook for PAN at CLEF 2016. In Balog K., Cappellato L., Ferro N. & Macdonald C., editors, *CLEF 2016 Evaluation Labs and Workshop-Working Notes Papers*, 5-8 September 2016, Évora, Portugal.

Mohammadi G. & Vinciarelli A. (2012). Automatic personality perception: Prediction of trait attribution based on prosodic features. *IEEE Transactions on Affective Computing 3* (3), 273-284.

Mohammad S. & Kiritchenko S. (2013). Using nuances of emotion to identify personality. *AAAI Technical Report WS-13-01 Computational Personality Recognition* (Shared Task).

Mohammad S. & Turney P. (2013). Crowdsourcing a word-emotion association lexicon. *Computational Intelligence, 29* (3), 436-465.

Montgomery D. C., Peck E. A. & Vining G. G. (2012). *Introduction to Linear Regression Analysis*, 5th Edition. Hoboken: Wiley.

Moreno D. R. J., Gomez J. C., Almanza-Ojeda D.-L. & Ibarra-Manzano M.-A. (2019). Prediction of personality traits in Twitter users with latent features. *International Conference on Electronics, Communications and Computers (CONIELECOMP)*, Cholula, Mexico, (pp. 176-181).

Mosteller F. & Wallace D. L. (1984). *Applied Bayesian and Classical Inference. The case of The Federalist Papers* (2nd ed.). New York: Springer-Verlag.

Murphy Gardner. (1966). *Personality: A Biosocial Approach to Origins and Structure*. NY: Basic Books.

Myers-Briggs I. (1962). *The Myers-Briggs Type Indicator*. Palo Alto, California: Consulting Psychologists Press.

Myers-Briggs I. & Myers P. B. (1980). *Gifts Differing: Understanding Personality Type*. Mountain View, CA: Davies-Black Publishing.

Neuman Y. (2016). *Computational Personality Analysis. Introduction, Practical Applications and Novel Directions*. Switzerland: Springer International Publishing.

Neuman Y., Assaf D., Cohen Y. & Knoll L. J. (2015). Profiling school shooters: Automatic text-based analysis. *Front. Psychiatry, 6*, 86.

Newman M. L., Pennebaker J. W., Berry D. S. & Richards J. M. (2003). Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin, 29*, 665-675.

Nguyen D., Smith N.A. & Rosé C. P. (2011). Author age prediction from text using linear regression. *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, Portland, USA, (pp. 115-123). Association for Computational Linguistics.

Nowson S. (2006). *The Language of Weblogs: A Study of Genre and Individual Differences*. Ph.D. Thesis, University of Edinburgh.

Nowson S. & Gill A. J. (2014). Look! Who's talking? Projection of extraversion across different social contexts. *Proceedings of the 2014 ACM Multi Media on Workshop on Computational Personality Recognition*, 7 November 2014, Orlando, Florida, USA, (pp. 23-26).

Nowson S. & Oberlander J. (2007). Identifying more bloggers. Towards large scale personality classification of personal weblogs. *Proceedings of the International Conference on Weblogs and Social Media 2007,* 26-28 March 2007, Boulder, Colorado, USA.

Nowson S. & Oberlander J. (2006). Differentiating document type and author personality from linguistic features. *Proceedings of the 11th Australasian Document Computing Symposium,* 11 December 2006, Brisbane, Australia.

Oberlander J. & Gill A. J. (2006). Language with character: A stratified corpus comparison of individual differences in e-mail communication. *Discourse Processes, 42,* 239-270.

Oberlander J. & Gill A. J. (2004a). Individual differences and implicit language: Personality, parts-of-speech and pervasiveness. *Proceedings of the 26th Annual Conference of the Cognitive Science Society,* 4-7 August 2004, Chicago, USA, (pp. 1035-1040).

Oberlander J. & Gill A. (2004b). Language generation and personality: Two dimensions, two stages, two hemispheres? *Proceedings from the AAAI Spring Symposium on Architectures for Modeling Emotion: Cross-Disciplinary Foundations,* 22-24 March 2004, (pp. 104-111).

Oberlander J. & Nowson S. (2006). Whose thumb is it anyway? Classifying author personality from weblog text. *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics ACL,* 17-21 July 2006, Sydney, Australia, (pp. 627-634).

Papageorgiou H., Prokopidis P., Giouli V. & Piperidis S. (2000). A Unified POS Tagging Architecture and its Application to Greek. *Proceedings of the 2nd Language Resources and Evaluation Conference*, Athens, (pp. 1455-1462).

Pavlov I. (1927). *Conditioned Reflexes*. London: Oxford University Press.

Peersman C., Daelemans W. & Van Vaerenbergh L. (2011). Predicting age and gender in online social networks. *Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents, SMUC'11,* 28 October 2011, Glasgow, Scotland, UK, (pp. 37-44).

Peng K.-H., Liou L.-H., Chang C.-S. & Lee D.-S. (2015). Predicting personality traits of Chinese users based on Facebook wall posts. *2015 24th Wireless and Optical Communication Conference (WOCC)*, Taipei, (pp. 9-14).

Pennebaker J. W., Boyd R. L., Jordan K. & Blackburn K. (2015). *The Development and Psychometric Properties of LIWC2015.* Austin, TX: University of Texas at Austin.

Pennebaker J. W. & King L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology, 77*, 1296-1312.

Pennebaker J. W. & Lay T. C. (2002). Language use and personality during crises: Analyses of Mayor Rudolph Giuliani's press conferences. *Journal of Research in Personality, 36,* 271-282.

Pennebaker J. W., Mehl M. R. & Niederhoffer K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology, 54*, 547-577.

Pennebaker J. W. & Stone L. D. (2003). Words of wisdom: Language use over the life span. *Journal of Personality and Social Psychology, 85*(2), 291-301.

Pervaz I., Ameer I., Sittar A., Muhammad R. & Nawab A. (2015). Identification of author personality traits using stylistic features-Notebook for PAN at CLEF 2015. In Cappellato L., Ferro N., Jones G. & Juan E., editors, *CLEF 2015 Evaluation Labs and Workshop-Working Notes Papers*, 8-11 September 2015, Toulouse, France.

Pervin L. A. & John O. P. (1997). *Personality: Theory and Research*. John Wiley & Sons.

Pianesi F., Mana N., Cappelletti A., Lepri B. & Zancanaro M. (2008). Multimodal recognition of personality traits in social interactions. *Proceedings of the 10th International Conference on Multimodal Interfaces ICMI'08,* 20-22 October 2008, Chania, Crete, Greece.

Piedmont R. L. (1998). *The Revised NEO Personality Inventory: Clinical and Research Applications.* New York. NY: Plenum Publishing Corporation.

Plank B. & Hovy D. (2015). Personality traits on twitter-or-how to get 1,500 personality tests in a week. *6th Workshop on computational approaches to subjectivity, sentiment and social media analysis: WASSA 2015, Association for Computational Linguistics*, Lisboa, Portugal, (pp. 92-98).

Prajwal S., Shahid A., Patel S. R., Srihari H. G.K. & Aditya C.R. (2020). Personality and traits score prediction from social media for students. *International Journal of Engineering Research and Technology, 9*(7).

Preotiuc-Pietro D., Eichstaedt J., Park G., Sap M., Smith L., Tobolsky V., Schwartz H. A. & Ungar L. (2015). The role of personality, age and gender in tweeting about mental illnesses. *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, NAACL*, 5 June 2015, Denver, Colorado.

Prokopidis P., Georgantopoulos B. & Papageorgiou H. (2011). A suite of NLP tools for Greek. *The 10th International Conference of Greek Linguistics*, 1-4 September 2011, Komotini, Greece.

Qiu L., Lin H., Ramsay J. & Yang F. (2012). You are what you tweet: Personality expression and perception on Twitter. *Journal of Research in Personality, 46*(6), 710–718.

Qiu L., Lu J., Ramsay J., Yang S., Qu W. & Zhu T. (2016). Personality expression in Chinese language use. *International Journal of Psychology*.

Quercia D., Kosinski M., Stillwell D. & Crowcroft J. (2011). Our Twitter profiles, ourselves: Predicting personality with Twitter. *Proceedings of the 2011 IEEE International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing*, (pp. 180-185).

Ramos dos Santos W. & Paraboni I. (2018). Personality facets recognition from text. *arXiv:1810.02980v1* [cs.CL].

Rangel F. & Rosso P. (2019). Overview of the 7th author profiling task at PAN 2019: Bots and gender profiling. In Cappellato L., Ferro N., Losada D. & Müller H., editors, *CLEF 2019 Labs and Workshops, Notebook Papers*.

Rangel F., Rosso P., Potthast M. & Stein B. (2017). Overview of the 5th Author Profiling Task at PAN 2017: Gender and language variety identification in Twitter. *Working Notes Papers of the CLEF, 2017.*

Rangel F., Rosso P., Verhoeven B., Daelemans W., Potthast M. & Stein B. (2016). Overview of the 4th author profiling task at PAN 2016: Cross-genre evaluations. *Working Notes Papers of the CLEF, 2016.*

Rangel F., Celli F., Rosso P., Potthast M., Stein B. & Daelemans W. (2015). Overview of the 3rd author profiling task at PAN 2015. In Cappellato L., Ferro N., Gareth J. & San Juan E. (Eds). *CLEF 2015 Evaluation Labs and Workshop - Working Notes Papers.*

Rangel F., Rosso P., Koppel M., Stamatatos E. & Inches G. (2013). Overview of the author profiling task at PAN 2013. In Forner P., Navigli R. & Tufis D., editors, *CLEF 2013 Evaluation Labs and Workshop - Working Notes Papers.*

Rogers C. R. (1951). *Client-centered Therapy: Its Current Practice, Implications and Theory*. Boston: Houghton Mifflin.

Rosenthal S. & McKeown K. (2011). Age prediction in blogs: A study of style, content, and online behavior in pre- and post-social media generations. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies,1*, (pp.763-772). USA: Association for Computational Linguistics.

Rude S. S., Gortner E.-M. & Pennebaker J. W. (2004). Language use of depressed and depression-vulnerable college students. *Cognition and Emotion, 18*, 1121-1133.

Rudman J. (1997). The state of authorship attribution Studies: Some problems and solutions. *Computers and Humanities, 31*, 351-365.

Rüegger D., Stieger M., Nißen M., Allemand M., Fleisch E. & Kowatsch T. (2020). How are personality states associated with smartphone data?. *European Journal of Personality. 34*, 687-713.

Salem S. M., Ismail S. S. & Aref M. (2019). Personality traits for Egyptian Twitter users dataset. *Proceedings of the 2019 8th International Conference on Software and Information Engineering*, 9-12 April, Cairo, Egypt, (pp. 206-211).

Sanford F. H. (1942). Speech and personality: A comparative case study. *Journal of Personality, 10*, 169-198.

Sarkar C., Bhatia S., Agarwal A. & Li J. (2014). Feature analysis for computational personality recognition using YouTube personality data set. *Proceedings of the 2014 ACM Multi Media on Workshop on Computational Personality Recognition*, November 2014, Orlando Florida USA, (pp. 11-14).

Schler J., Koppel M., Argamon S. & Pennebaker J. (2006). Effects of age and gender on blogging. *Proceedings of the 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*, 27-29 March 2006, Stanford, California, (pp. 199-205).

Schwartz A. H., Eichstaedt J. C., Kern M. L., Dziurzynski L., Ramones S. M., Agrawal M., … Ungar L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLOS ONE 8*(9): e73791. doi:10.1371/journal.pone.0073791

Seidman G. (2013). Self-presentation and belonging on Facebook: How personality influences social media use and motivations. *Personality and Individual Differences 54*(3), 402-407.

Shaywitz B., Shaywitz S., Pugh K., Constable T., Skudlarski P., Fulbright R., Bronen R., Fletcher J., Shankweiler D., Katz L. & Gore J. (1995). Sex differences in the functional organization of the brain for language. *Nature, 373*, 607-609.

Sierra G., Bel-Enguix G., Osornio-Arteaga A., Cabrera-Mora A., Garcia-Nieto L., Bustos A., Romo-Anaya A.-M. & Silva-Cuevas V. (2020). An exploration of personality traits detection in a Spanish Twitter corpus. *Proceedings of the*

Eleventh International Conference on Language Resources and Evaluation (LREC 2020).

Skinner B. F. (1953). *Science and Human Behavior*. New York: The Macmillan Company.

Soler Company J. & Leo W. (2016). A semi-supervised approach for gender identification. *Proceedings of the 10th International Language Resources and Evaluation Conference (LREC 2016),* Portorož Slovenia.

Soler Company J. & Wanner L. (2014). How to use less features and reach better performance in author gender identification. *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014),* Reykjavik, Iceland.

Staiano J., Lepri B., Subramanian R., Sebe N. & Pianesi F. (2011). Automatic modeling of personality states in small group interactions. *Proceedings of the 19th International conference on Multimedia,* Scottsdale, AZ, USA, (pp. 989-992).

Staiano J., Pianesi F., Lepri B., Sebe N., Aharony N. & Pentland A. P. (2012). Friends don't lie: Inferring personality traits from social network structure. *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, September 2012, Pittsburgh, Pennsylvania, (pp. 321-330).

Tadesse M. M., Lin H., Xu B. & Yang L. (2018). Personality predictions based on user behaviour on the Facebook social media platform. *IEEE Access, 6,* 61959-61969.

Tam J. & Martell C. H. (2009). Age detection in chat. *Proceedings of the 3rd IEEE International Conference on Semantic Computing.* Berkeley, USA.

Tandera T., Hendro, Suhartono D., Wongso R. & Prasetio Yen L. (2017). Personality prediction system from Facebook users. *2nd International Conference on Computer Science and Computational Intelligence 2017, ICCSCI 2017*, 13-14 October 2017, Bali, Indonesia.

Tausczik Yla R. & Pennebaker J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology 29*(1), 24-54.

Thelwall M., Buckley K., Paltoglou G., Cai D. & Kappas A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology, 61*(12), 2544-2558.

Thetford W. & Walsh R. (1985). Theories of personality and psychopathology: Schools derived from psychology and philosophy. In: H. Kaplan & B. Sadock (eds) *Comprehensive Textbook of Psychiatry*, Baltimore: Williams and Wilkins.

Thomas K. (1995). Dimensions of personality. In I. Roth (Eds.), *Introduction to Psychology, 1*, 373-416, Milton Keynes, The Open University.

Tomlinson M. T., Hinote D. & Bracewell D. B. (2013). Predicting conscientiousness through semantic analysis of Facebook posts. *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, 8-11 July, Massachusetts, USA.

Verhoeven B., Soler Company J. & Daelemans W. (2014). Evaluating content-independent features for personality recognition. *2nd Workshop on Computational Personality Recognition (WCPR14)*, 7 November 2014, Orlando, FL, USA.

Verhoeven B. & Daelemans W. (2014). CLiPS Stylometry investigation (CSI) corpus: A Dutch corpus for the detection of age, gender, personality, sentiment and deception in text. *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014).* Reykjavik, Iceland.

Verhoeven B., Daelemans W. & De Smedt T. (2013). Ensemble methods for personality recognition. *Proceedings of the Workshop on Computational Personality Recognition at 7th International Conference on Weblogs and Social Media*, Boston, Mass, USA, (pp. 35-38).

Verhoeven B., Daelemans W. & Plank B. (2016a). TwiSty: A multilingual Twitter stylometry corpus for gender and personality profiling. *Proceedings of the 10th Annual Conference on Language Resources and Evaluation (LREC 2016)*. Portorož, Slovenia.

Verhoeven B., Daelemans W. & Plank B. (2016b). Creating TwiSty: Corpus development and statistics. *Computational Linguistics and Psycholinguistics Research Center CLiPS Technical Report Series*, University of Antwerp, Belgium, CTRS-006.

Watson J. (1936). *Psychology from the Standpoint of a Behaviorist*. Philadelphia: Lippincott.

Wiegmann M., Stein B. & Potthast M. (2019). Overview of the celebrity profiling task at PAN 2019. In Linda Cappellato, Nicola Ferro, David E. Losada, and Henning Müller, editors, *CLEF 2019 Labs and Workshops, Notebook Papers.*

Wilson M. (1988). MRC Psycholinguistic database: Machine usable dictionary, Version 2.00. *Behavioural Research Methods, Instruments and Computers, 20,* 6-11.

Wong Sze-Meng J. & Dras M. (2009). Contrastive analysis and native language identification. *Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association, ALTA 2019,* 4-6 December 2009, Sydney, Australia.

Wright W. R. (2012). *Literature Review*. Retrieved from: http://www2.hawaii.edu/~wrightwr/WilliamWright_literature_review.pdf

Wright W. R. & Chin D. N. (2016). Personality profiling from text: Language features tied to personality across corpora. *Proceedings of 24th ACM Conference on User Modeling, Adaptation and Personalisation (UMAP 2016)* Halifax, Canada (Extended Proceedings).

Xue D., Hong Z., Guo S., Gao L., Wu L., Zheng J. & Zhao N. (2017). Personality recognition on social media with label distribution learning. *IEEE Access, 5*, 13478-13488.

Yamada K., Sasano R. & Takeda K. (2019). Incorporating textual information on user behavior for personality prediction. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, 28 July-2 August 2019, Florence, Italy, (pp. 177-182).

Yarkoni T. (2010). Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality, 44,* 363–373.

Yu B., Mei Q. & Zhai C. (2005). English usage comparison between native and non-native English speakers in academic writing. *Proceedings of Conference of the Association for Computers and the Humanities Association for Literary and Linguistic Computing(ACH/ALLC 2005),* Victoria, BC, Canada.

Yu J. & Markov K. (2017). Deep learning based personality recognition from Facebook status updates. *2017 IEEE 8th International Conference on Awareness Science and Technology (iCAST),* 8-10 November 2017, Taichung, Taiwan.

Yule U. (1939). On sentence-length as a statistical characteristic of style in prose: With application to two cases of disputed authorship. *Biometrika, 30*(3/4), 363-390.

Zangerle E., Tschuggnall M., Specht G., Potthast M. & Stein B. (2019). Overview of the style change detection task at PAN 2019. In Cappellato L., Ferro N., Losada D. & Müller H., editors, *CLEF 2019 Labs and Workshops, Notebook Papers*.

Zhang C. & Zhang P. (2010). Predicting gender from blog posts. *Technical Report.* University of Massachusetts Amherst, USA.

Zipf G. K. (1935). *The Psychobiology of Language*. Cambridge, Mass.: M.I.T. Press.

Γεωργακοπούλου Αλ. & Γούτσος Δ. (2008). *Κείμενο και Επικοινωνία*, Αθήνα: Ελληνικά Γράμματα.

Κακριδή-Φερράρι Μ. (2005). Γλώσσα και κοινωνικό περιβάλλον: Ζητήματα κοινωνιογλωσσολογίας (Α΄μέρος). *Περιοδικό Παρουσία*, Παράρτημα αρ. 64, Αθήνα.

Κιαπόκας Μ. (1996). *Ιπποκράτης ο Κώος και Ιπποκρατικός Όρκος*. Έκδοση του πνευματικού κέντρου του δήμου Κω.

Κουτουβίδης Ν., Μηνογιάννη Α. & Βάρσου ΣΜ. (2004). Η συμβολή της θεωρίας της προσωπικότητας στην κατανόηση της εγκληματικής συμπεριφοράς, *Εγκέφαλος*, *41*(2).

Λεξικό της Κοινής Νεοελληνικής. (1998). Θεσσαλονίκη: Ινστιτούτο Νεοελληνικών Σπουδών [Ίδρυμα Μανόλη Τριανταφυλλίδη].

Μικρός Γ. Κ. (2015a). *Υπολογιστική Υφολογία*. Ελληνικά Ακαδημαϊκά Συγγράμματα και Βοηθήματα. Αποθετήριο Κάλλιπος: https://repository.kallipos.gr/handle/11419/4860

Μικρός Γ.Κ. (2015b). *Εισαγωγή στην ανάλυση γλωσσικών δεδομένων. Η Ελληνική γλώσσα μέσα από αριθμούς.* Έκδοση: 1.0. Αθήνα. Διαθέσιμο από τη δικτυακή διεύθυνση: http://opencourses.uoa.gr/courses/ILL103.

Μικρός Γ. Κ. (2013). Δυσκολία κατανόησης του ξενόγλωσσου κειμένου και υφομετρία. Μια νέα προσέγγιση στην αναγνωσιμότητα κειμένων από Έλληνες που μαθαίνουν την Ιταλική ως ξένη γλώσσα. *Στο Ελλάδα-Ιταλία: Διαπολιτισμικές προσεγγίσεις* (σσ. 287-309). Αθήνα: ΕΚΠΑ.

Μπαμπινιώτης Γ. (2002). *Λεξικό της Νέας Ελληνικής Γλώσσας*. Αθήνα: Κέντρο Λεξικολογίας. (δεύτερη έκδοση).

Παπαστάμου Στ. (1989). *Εγχειρίδιο Κοινωνικής Ψυχολογίας*, Αθήνα: Οδυσσέας.

Παυλόπουλος Β. & Μπεζεβέγκης Ηλ. (1999). Το μοντέλο των Πέντε Παραγόντων της προσωπικότητας για παιδιά: Μια διαπολιτιστική μελέτη. *Ψυχολογία 6*(2), 174-182.

Περήφανος Κ. (2019). *Ανάλυση ελληνικών σωμάτων κειμένων με τη χρήση τεχνικών μηχανικής μάθησης: Υπολογιστική αναπαράσταση της ιδιολέκτου*. Διδακτορική διατριβή, Αθήνα: Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών.

Πολίτου-Μαρμαρινού Ε., Μικρός Γ. Κ. & Δημητρούλια Τ. (2011). Εφαρμογή υφομετρικών τεχνικών στην αναγνώριση πατρότητας κειμένου: Πρωτότυπα έργα και μεταφράσεις του Παπαδιαμάντη. *Γ' Διεθνές Συνέδριο για τον Παπαδιαμάντη, "Ο Παπαδιαμάντης μεταφράζων και μεταφραζόμενος", Β' Κύκλος*, 7-8 Οκτωβρίου 2011, Αθήνα.

Ποταμιάνος Γ. & Παπαστάμου Στ. (2008). Η μελέτη της προσωπικότητας: μια κριτική προσέγγιση. Στο βιβλίο: *Θεωρίες Προσωπικότητας και Κλινική Πρακτική*. Αθήνα: Ελληνικά Γράμματα.

Τσαούσης Ι. (1999). Αναζητώντας τη δομή της προσωπικότητας: Το μοντέλο των Πέντε Παραγόντων. *Ψυχολογία*, *6*, 88-103.

Παράρτημα I: Jung Typology Test

This free personality test is based on Carl Jung's and Isabel Briggs Myers' personality type theory.

When responding to the statements, please choose the response you agree with most (YES or NO):

1. You are almost never late for your appointments.
   YES NO
2. You like to be engaged in an active and fast-paced job.
   YES NO
3. You enjoy having a wide circle of acquaintances.
   YES NO
4. You feel involved when watching TV soaps.
   YES NO
5. You are usually the first to react to a sudden event: the telephone ringing or unexpected question.
   YES NO
6. You are more interested in a general idea than in the details of its realization.
   YES NO
7. You tend to be unbiased even if this might endanger your good relations with people.
   YES NO
8. Strict observance of the established rules is likely to prevent a good outcome.
   YES NO
9. It's difficult to get you excited.
   YES NO
10. It is in your nature to assume responsibility.
    YES NO
11. You often think about humankind and its destiny.
    YES NO

12. You believe the best decision is one that can be easily changed.
    YES NO
13. Objective criticism is always useful in any activity.
    YES NO
14. You prefer to act immediately rather than speculate about various options.
    YES NO
15. You trust reason rather than feelings.
    YES NO
16. You are inclined to rely more on improvisation than on careful planning.
    YES NO
17. You spend your leisure time actively socializing with a group of people, attending parties, shopping, etc.
    YES NO
18. You usually plan your actions in advance.
    YES NO
19. Your actions are frequently influenced by emotions.
    YES NO
20. You are a person somewhat reserved and distant in communication.
    YES NO
21. You know how to put every minute of your time to good purpose.
    YES NO
22. You readily help people while asking nothing in return.
    YES NO
23. You often contemplate about the complexity of life.
    YES NO
24. After prolonged socializing you feel you need to get away and be alone.
    YES NO
25. You often do jobs in a hurry.
    YES NO
26. You easily see the general principle behind specific occurrences.
    YES NO
27. You frequently and easily express your feelings and emotions.
    YES NO

28. You find it difficult to speak loudly.
    YES NO
29. You get bored if you have to read theoretical books.
    YES NO
30. You tend to sympathize with other people.
    YES NO
31. You value justice higher than mercy.
    YES NO
32. You rapidly get involved in social life at a new workplace.
    YES NO
33. The more people with whom you speak, the better you feel.
    YES NO
34. You tend to rely on your experience rather than on theoretical alternatives.
    YES NO
35. You like to keep a check on how things are progressing.
    YES NO
36. You easily empathize with the concerns of other people.
    YES NO
37. Often you prefer to read a book than go to a party.
    YES NO
38. You enjoy being at the center of events in which other people are directly involved.
    YES NO
39. You are more inclined to experiment than to follow familiar approaches.
    YES NO
40. You avoid being bound by obligations.
    YES NO
41. You are strongly touched by the stories about people's troubles.
    YES NO
42. Deadlines seem to you to be of relative, rather than absolute, importance.
    YES NO
43. You prefer to isolate yourself from outside noises.
    YES NO

44. It's essential for you to try things with your own hands.
    YES NO
45. You think that almost everything can be analyzed.
    YES NO
46. You do your best to complete a task on time.
    YES NO
47. You take pleasure in putting things in order.
    YES NO
48. You feel at ease in a crowd.
    YES NO
49. You have good control over your desires and temptations.
    YES NO
50. You easily understand new theoretical principles.
    YES NO
51. The process of searching for a solution is more important to you than the solution itself.
    YES NO
52. You usually place yourself nearer to the side than in the center of the room.
    YES NO
53. When solving a problem you would rather follow a familiar approach than seek a new one.
    YES NO
54. You try to stand firmly by your principles.
    YES NO
55. A thirst for adventure is close to your heart.
    YES NO
56. You prefer meeting in small groups to interaction with lots of people.
    YES NO
57. When considering a situation you pay more attention to the current situation and less to a possible sequence of events.
    YES NO
58. You consider the scientific approach to be the best.
    YES NO

59. You find it difficult to talk about your feelings.
YES NO
60. You often spend time thinking of how things could be improved.
YES NO
61. Your decisions are based more on the feelings of a moment than on the careful planning.
YES NO
62. You prefer to spend your leisure time alone or relaxing in a tranquil family atmosphere.
YES NO
63. You feel more comfortable sticking to conventional ways.
YES NO
64. You are easily affected by strong emotions.
YES NO
65. You are always looking for opportunities.
YES NO
66. Your desk, workbench etc. is usually neat and orderly.
YES NO
67. As a rule, current preoccupations worry you more than your future plans.
YES NO
68. You get pleasure from solitary walks.
YES NO
69. It is easy for you to communicate in social situations.
YES NO
70. You are consistent in your habits.
YES NO
71. You willingly involve yourself in matters which engage your sympathies.
YES NO
72. You easily perceive various ways in which events could develop.
YES NO

Το μεταφρασμένο Jung Typology Ερωτηματολόγιο

Το παρόν δωρεάν ερωτηματολόγιο προσωπικότητας βασίζεται στη θεωρία των Τύπων του Carl Jung και της Isabel Briggs Myers.

Μετά από κάθε δήλωση κύκλωσε την απάντηση (ΝΑΙ ή ΟΧΙ) που σε εκφράζει:

1. Είσαι συνήθως συνεπής στα ραντεβού σου.
ΝΑΙ ΟΧΙ

2. Σου αρέσει να απασχολείσαι σε δουλειά ενεργητική και με γρήγορους ρυθμούς.
ΝΑΙ ΟΧΙ

3. Απολαμβάνεις να έχεις έναν ευρύ κύκλο γνωριμιών.
ΝΑΙ ΟΧΙ

4. Αισθάνεσαι ότι εμπλέκεσαι συναισθηματικά όταν παρακολουθείς τηλεοπτικές σαπουνόπερες.
ΝΑΙ ΟΧΙ

5. Είσαι συνήθως ο πρώτος που αντιδρά σε ένα ξαφνικό γεγονός: το χτύπημα του τηλέφωνου ή μια αναπάντεχη ερώτηση.
ΝΑΙ ΟΧΙ

6. Σε ενδιαφέρει περισσότερο μια γενική ιδέα παρά οι λεπτομέρειες για την πραγματοποίησή της.
ΝΑΙ ΟΧΙ

7. Τείνεις να είσαι αμερόληπτος ακόμα κι αν αυτό διακινδυνεύει τις καλές σου σχέσεις με τους άλλους.
ΝΑΙ ΟΧΙ

8. Η αυστηρή τήρηση των θεσπισμένων κανόνων είναι πιθανό να αποτρέψει μια καλή έκβαση.
ΝΑΙ ΟΧΙ

9. Δύσκολα ενθουσιάζεσαι.
ΝΑΙ ΟΧΙ

10. Είναι στη φύση σου να αναλαμβάνεις ευθύνες.
ΝΑΙ ΟΧΙ

11. Συχνά προβληματίζεσαι για το ανθρώπινο γένος και τον προορισμό του.

ΝΑΙ ΟΧΙ

12. Πιστεύεις ότι η καλύτερη απόφαση είναι αυτή που μπορεί εύκολα να αλλάξει.

ΝΑΙ ΟΧΙ

13. Η αντικειμενική κριτική είναι πάντα χρήσιμη σε κάθε δραστηριότητα.

ΝΑΙ ΟΧΙ

14. Προτιμάς να ενεργείς αμέσως παρά να σκέφτεσαι τις διαφορετικές επιλογές.

ΝΑΙ ΟΧΙ

15. Εμπιστεύεσαι τη λογική παρά τα συναισθήματα.

ΝΑΙ ΟΧΙ

16. Έχεις την τάση να εμπιστεύεσαι περισσότερο τον αυτοσχεδιασμό παρά ένα προσεκτικό σχεδιασμό.

ΝΑΙ ΟΧΙ

17. Περνάς τον ελεύθερο χρόνο σου δραστήρια καθώς κοινωνικοποιείσαι με μια ομάδα ανθρώπων, πηγαίνοντας σε πάρτι, κάνοντας ψώνια κλπ.

ΝΑΙ ΟΧΙ

18. Συνήθως σχεδιάζεις τις πράξεις σου εκ των προτέρων.

ΝΑΙ ΟΧΙ

19. Οι πράξεις σου επηρεάζονται συχνά από το συναίσθημα.

ΝΑΙ ΟΧΙ

20. Είσαι άνθρωπος κάπως επιφυλακτικός και απόμακρος στην επικοινωνία.

ΝΑΙ ΟΧΙ

21. Ξέρεις πώς να αξιοποιείς κάθε λεπτό του χρόνου σου.

ΝΑΙ ΟΧΙ

22. Βοηθάς πρόθυμα τους ανθρώπους χωρίς να ζητάς αντάλλαγμα.

ΝΑΙ ΟΧΙ

23. Συχνά αναλογίζεσαι την πολυπλοκότητα της ζωής.

ΝΑΙ ΟΧΙ

24. Μετά από παρατεταμένη κοινωνικοποίηση αισθάνεσαι ότι χρειάζεσαι να ξεφύγεις και να μείνεις μόνος.

ΝΑΙ ΟΧΙ

25. Συχνά κάνεις δουλειές βιαστικά.

ΝΑΙ ΟΧΙ

26. Με ευκολία βλέπεις τη γενική αρχή πίσω από συγκεκριμένα συμβάντα.

ΝΑΙ ΟΧΙ

27. Εκφράζεις συχνά και εύκολα συναισθήματά σου.

ΝΑΙ ΟΧΙ

28. Το βρίσκεις δύσκολο να μιλάς δυνατά.

ΝΑΙ ΟΧΙ

29. Βαριέσαι όταν πρέπει να διαβάσεις θεωρητικά βιβλία.

ΝΑΙ ΟΧΙ

30. Έχεις την τάση να συμπονάς τους άλλους.

ΝΑΙ ΟΧΙ

31. Εκτιμάς τη δικαιοσύνη περισσότερο από τον οίκτο.

ΝΑΙ ΟΧΙ

32. Αναμειγνύεσαι πολύ γρήγορα στην κοινωνική ζωή ενός νέου χώρου εργασίας.

ΝΑΙ ΟΧΙ

33. Όσο περισσότεροι είναι οι άνθρωποι με τους οποίους μιλάς, τόσο καλύτερα νιώθεις.

ΝΑΙ ΟΧΙ

34. Έχεις την τάση να εμπιστεύεσαι την εμπειρία σου παρά τις θεωρητικές εναλλακτικές.

ΝΑΙ ΟΧΙ

35. Σου αρέσει να ελέγχεις την πρόοδο των πραγμάτων.

ΝΑΙ ΟΧΙ

36. Εύκολα συμμετέχεις συναισθηματικά στις έγνοιες των άλλων.

ΝΑΙ ΟΧΙ

37. Συχνά προτιμάς να διαβάσεις ένα βιβλίο από το να πας σε ένα πάρτι.

ΝΑΙ ΟΧΙ

38. Απολαμβάνεις να βρίσκεσαι στο κέντρο των γεγονότων στα οποία άλλοι είναι άμεσα εμπλεκόμενοι.

ΝΑΙ ΟΧΙ

39. Έχεις περισσότερο την προδιάθεση να πειραματίζεσαι παρά να ακολουθείς οικείες προσεγγίσεις.

ΝΑΙ ΟΧΙ

40. Αποφεύγεις να δεσμεύεσαι με υποχρεώσεις.

ΝΑΙ ΟΧΙ

41. Συγκινείσαι έντονα με ιστορίες για προβλήματα άλλων.

ΝΑΙ ΟΧΙ

42. Οι προθεσμίες μοιάζουν να είναι για σένα σχετικής παρά απόλυτης σημασίας.

ΝΑΙ ΟΧΙ

43. Προτιμάς να απομονώνεσαι από εξωτερικούς θορύβους.

ΝΑΙ ΟΧΙ

44. Είναι ουσιώδες για σένα να επιχειρείς πράγματα με τα ίδια σου τα χέρια.

ΝΑΙ ΟΧΙ

45. Πιστεύεις ότι σχεδόν οτιδήποτε μπορεί να αναλυθεί.

ΝΑΙ ΟΧΙ

46. Κάνεις ό,τι μπορείς για να ολοκληρώσεις μια δουλειά στην ώρα της.

ΝΑΙ ΟΧΙ

47. Ευχαριστιέσαι να τακτοποιείς πράγματα.

ΝΑΙ ΟΧΙ

48. Νιώθεις άνετα ανάμεσα στο πλήθος.

ΝΑΙ ΟΧΙ

49. Έχεις καλό έλεγχο των επιθυμιών σου και των προκλήσεων.

ΝΑΙ ΟΧΙ

50. Κατανοείς εύκολα νέες θεωρητικές αρχές.

ΝΑΙ ΟΧΙ

51. Η διαδικασία της αναζήτησης μιας λύσης είναι πιο σημαντική για σένα από την ίδια τη λύση.

ΝΑΙ ΟΧΙ

52. Συνήθως θέτεις τον εαυτό σου πιο κοντά στην άκρη παρά στο κέντρο ενός δωματίου.

ΝΑΙ ΟΧΙ

53. Όταν λύνεις ένα πρόβλημα, ακολουθείς τη γνώριμη προσέγγιση από το να ψάχνεις για μια καινούργια.

ΝΑΙ ΟΧΙ

54. Προσπαθείς να τηρείς τις αρχές σου.

ΝΑΙ ΟΧΙ

55. Η δίψα για περιπέτεια είναι σημαντική για σένα.

ΝΑΙ ΟΧΙ

56. Προτιμάς τη συναναστροφή με μικρές ομάδες από την αλληλεπίδραση με πολλούς ανθρώπους.

ΝΑΙ ΟΧΙ

57. Όταν σκέφτεσαι μια κατάσταση προσέχεις περισσότερο τη συγκεκριμένη κατάσταση και λιγότερο την πιθανή διαδοχή των γεγονότων.

ΝΑΙ ΟΧΙ

58. Θεωρείς ότι η επιστημονική προσέγγιση είναι η καλύτερη.

ΝΑΙ ΟΧΙ

59. Δυσκολεύεσαι να μιλήσεις για τα συναισθήματά σου.

ΝΑΙ ΟΧΙ

60. Συχνά περνάς χρόνο σκεπτόμενος πώς θα μπορούσαν να βελτιωθούν τα πράγματα.

ΝΑΙ ΟΧΙ

61. Οι αποφάσεις σου βασίζονται περισσότερο στο συναίσθημα της στιγμής παρά σε προσεκτικό σχεδιασμό.

ΝΑΙ ΟΧΙ

62. Προτιμάς να περνάς τον ελεύθερο χρόνο σου μόνος/χαλαρώνοντας σε ένα ήρεμο οικογενειακό περιβάλλον.

ΝΑΙ ΟΧΙ

63. Νιώθεις πιο άνετα όταν επιμένεις σε συνηθισμένες μεθόδους.

ΝΑΙ ΟΧΙ

64. Εύκολα επηρεάζεσαι από δυνατά συναισθήματα.

ΝΑΙ ΟΧΙ

65. Ψάχνεις πάντα για ευκαιρίες.

ΝΑΙ ΟΧΙ

66. Το γραφείο σου, ο χώρος εργασίας σου κλπ είναι συνήθως καθαρά και τακτοποιημένα.

ΝΑΙ ΟΧΙ

67. Κατά κανόνα, οι τρέχουσες έννοιες σε ανησυχούν περισσότερο από τα μελλοντικά σου σχέδια.

ΝΑΙ ΟΧΙ

68. Οι μοναχικοί περίπατοι σε ευχαριστούν.

ΝΑΙ ΟΧΙ

69. Σου είναι εύκολο να επικοινωνείς σε κοινωνικές περιστάσεις.

ΝΑΙ ΟΧΙ

70. Είσαι συνεπής στις συνήθειές σου.

ΝΑΙ ΟΧΙ

71. Πρόθυμα αναμειγνύεσαι σε ζητήματα που σε κάνουν να συμπάσχεις συναισθηματικά.

ΝΑΙ ΟΧΙ

72. Εύκολα αντιλαμβάνεσαι διαφορετικούς τρόπους με τους οποίους μπορούν να εξελιχθούν τα γεγονότα.

ΝΑΙ ΟΧΙ

Παράρτημα ΙΙ: The Big Five Personality Test

Directions: The following statements concern your perception about yourself in a variety of situations. Your task is to indicate the strength of your agreement with each statement, utilizing a scale in which 1 denotes strong disagreement, 5 denotes strong agreement, and 2, 3, and 4 represent intermediate judgments. In the boxes after each statement, click a number from 1 to 5 from the following scale:

1. Strongly disagree
2. Disagree
3. Neither disagree nor agree
4. Agree
5. Strongly agree

There are no "right" or "wrong" answers, so select the number that most closely reflects you on each statement.

I see myself as someone who...

1. ...Is talkative

Strongly disagree 1 2 3 4 5 Strongly agree

2. ...Tends to find fault with others

Strongly disagree 1 2 3 4 5 Strongly agree

3. ...Does a thorough job

Strongly disagree 1 2 3 4 5 Strongly agree

4. ...Is depressed, blue

Strongly disagree 1 2 3 4 5 Strongly agree

5. ...Is original, comes up with new ideas

Strongly disagree 1 2 3 4 5 Strongly agree

6. ...Is reserved

Strongly disagree 1 2 3 4 5 Strongly agree

7. ...Is helpful and unselfish with others

Strongly disagree 1 2 3 4 5 Strongly agree

8. ...Can be somewhat careless

Strongly disagree 1 2 3 4 5 Strongly agree

9. ...Is relaxed, handles stress well

Strongly disagree 1 2 3 4 5 Strongly agree

10. ...Is curious about many different things

Strongly disagree 1 2 3 4 5 Strongly agree

11. ...Is full of energy

Strongly disagree 1 2 3 4 5 Strongly agree

12. ...Starts quarrels with others

Strongly disagree 1 2 3 4 5 Strongly agree

13. ...Is a reliable worker

Strongly disagree 1 2 3 4 5 Strongly agree

14. ...Can be tense

Strongly disagree 1 2 3 4 5 Strongly agree

15. ...Is ingenious, a deep thinker

Strongly disagree 1 2 3 4 5 Strongly agree

16. ...Generates a lot of enthusiasm

Strongly disagree 1 2 3 4 5 Strongly agree

17. ...Has a forgiving nature

Strongly disagree 1 2 3 4 5 Strongly agree

18. ...Tends to be disorganized

Strongly disagree 1 2 3 4 5 Strongly agree

19. ...Worries a lot

Strongly disagree 1 2 3 4 5 Strongly agree

20. ...Has an active imagination

Strongly disagree 1 2 3 4 5 Strongly agree

21. ...Tends to be quiet

Strongly disagree 1 2 3 4 5 Strongly agree

22. ...Is generally trusting

Strongly disagree 1 2 3 4 5 Strongly agree

23. ...Tends to be lazy

Strongly disagree 1 2 3 4 5 Strongly agree

24. ...Is emotionally stable, not easily upset

Strongly disagree 1 2 3 4 5 Strongly agree

25. ...Is inventive

Strongly disagree 1 2 3 4 5 Strongly agree

26. ...Has an assertive personality

Strongly disagree 1 2 3 4 5 Strongly agree

27. ...Can be cold and aloof

Strongly disagree 1 2 3 4 5 Strongly agree

28. ...Perseveres until the task is finished

Strongly disagree 1 2 3 4 5 Strongly agree

29. ...Can be moody

Strongly disagree 1 2 3 4 5 Strongly agree

30. ...Values artistic, aesthetic experiences

Strongly disagree 1 2 3 4 5 Strongly agree

31. ...Is sometimes shy, inhibited

Strongly disagree 1 2 3 4 5 Strongly agree

32. ...Is considerate and kind to almost everyone

Strongly disagree 1 2 3 4 5 Strongly agree

33. ...Does things efficiently

Strongly disagree 1 2 3 4 5 Strongly agree

34. ...Remains calm in tense situations

Strongly disagree 1 2 3 4 5 Strongly agree

35. ...Prefers work that is routine

Strongly disagree 1 2 3 4 5 Strongly agree

36. ...Is outgoing, sociable

Strongly disagree 1 2 3 4 5 Strongly agree

37. ...Is sometimes rude to others

Strongly disagree 1 2 3 4 5 Strongly agree

38. ...Makes plans and follows through with them

Strongly disagree 1 2 3 4 5 Strongly agree

39. ...Gets nervous easily

Strongly disagree 1 2 3 4 5 Strongly agree

40. ...Likes to reflect, play with ideas

Strongly disagree 1 2 3 4 5 Strongly agree

41. ...Has few artistic interests

Strongly disagree 1 2 3 4 5 Strongly agree

42. ...Likes to cooperate with others

Strongly disagree 1 2 3 4 5 Strongly agree

43. ...Is easily distracted

Strongly disagree 1 2 3 4 5 Strongly agree

44. ...Is sophisticated in art, music, or literature

Strongly disagree 1 2 3 4 5 Strongly agree

45. ...Is politically liberal

Strongly disagree 1 2 3 4 5 Strongly agree

Το μεταφρασμένο Big Five Ερωτηματολόγιο

Οδηγίες: Οι παρακάτω δηλώσεις αφορούν την αντίληψη που έχεις για τον εαυτό σου σε διάφορες περιστάσεις. Ο στόχος είναι να δείξεις το βαθμό συμφωνίας σου σε κάθε δήλωση, χρησιμοποιώντας μια κλίμακα στην οποία ο αριθμός 1 δηλώνει απόλυτη διαφωνία ενώ ο αριθμός 5 απόλυτη συμφωνία και οι αριθμοί 2, 3, και 4 εκφράζουν ενδιάμεσες κρίσεις. Μετά από κάθε δήλωση κύκλωσε έναν αριθμό από το 1 ως το 5 από την παρακάτω κλίμακα:

1. Διαφωνώ απόλυτα
2. Διαφωνώ
3. Ούτε διαφωνώ ούτε συμφωνώ
4. Συμφωνώ
5. Συμφωνώ απόλυτα

Δεν υπάρχουν σωστές ή λανθασμένες απαντήσεις, επομένως επίλεξε τον αριθμό που εκφράζει περισσότερο τον εαυτό σου σε κάθε δήλωση.

Βλέπω τον εαυτό μου ως κάποιον που...

1. …είναι ομιλητικός

Διαφωνώ απόλυτα 1 2 3 4 5 Συμφωνώ απόλυτα

2. ...έχει την τάση να επικρίνει τους άλλους

Διαφωνώ απόλυτα 1 2 3 4 5 Συμφωνώ απόλυτα

3. ...είναι προσεχτικός στην εργασία του

Διαφωνώ απόλυτα 1 2 3 4 5 Συμφωνώ απόλυτα

4. ...είναι μελαγχολικός

Διαφωνώ απόλυτα 1 2 3 4 5 Συμφωνώ απόλυτα

5. ...είναι πρωτότυπος, έχει νέες ιδέες

Διαφωνώ απόλυτα 1 2 3 4 5 Συμφωνώ απόλυτα

6. ...είναι επιφυλακτικός

Διαφωνώ απόλυτα 1 2 3 4 5 Συμφωνώ απόλυτα

7. ...είναι εξυπηρετικός και ανιδιοτελής

Διαφωνώ απόλυτα 1 2 3 4 5 Συμφωνώ απόλυτα

8. ...μπορεί να είναι κάπως απρόσεκτος

Διαφωνώ απόλυτα 1 2 3 4 5 Συμφωνώ απόλυτα

9. ...είναι άνετος, διαχειρίζεται καλά το άγχος

Διαφωνώ απόλυτα 1 2 3 4 5 Συμφωνώ απόλυτα

10. ...είναι περίεργος για πολλά διαφορετικά πράγματα

Διαφωνώ απόλυτα 1 2 3 4 5 Συμφωνώ απόλυτα

11. ...είναι γεμάτος ενέργεια

Διαφωνώ απόλυτα 1 2 3 4 5 Συμφωνώ απόλυτα

12. ...ξεκινάει φιλονικίες με άλλους

Διαφωνώ απόλυτα 1 2 3 4 5 Συμφωνώ απόλυτα

13. ... είναι αξιόπιστος εργαζόμενος

Διαφωνώ απόλυτα 1 2 3 4 5 Συμφωνώ απόλυτα

14. ... μπορεί να είναι σε υπερένταση

Διαφωνώ απόλυτα 1 2 3 4 5 Συμφωνώ απόλυτα

15. ...είναι ευφυής, σε βάθος στοχαστής

Διαφωνώ απόλυτα 1 2 3 4 5 Συμφωνώ απόλυτα

16. ...προκαλεί πολύ ενθουσιασμό

Διαφωνώ απόλυτα 1 2 3 4 5 Συμφωνώ απόλυτα

17. ...είναι επιεικής

Διαφωνώ απόλυτα 1 2 3 4 5 Συμφωνώ απόλυτα

18. ...έχει την τάση να αποδιοργανώνεται

Διαφωνώ απόλυτα 1 2 3 4 5 Συμφωνώ απόλυτα

19. ...ανησυχεί πολύ

Διαφωνώ απόλυτα 1 2 3 4 5 Συμφωνώ απόλυτα

20. ...έχει ζωηρή φαντασία

Διαφωνώ απόλυτα 1 2 3 4 5 Συμφωνώ απόλυτα

21. ...έχει την τάση να είναι ήρεμος

Διαφωνώ απόλυτα 1 2 3 4 5 Συμφωνώ απόλυτα

22. ...είναι γενικά έμπιστος

Διαφωνώ απόλυτα 1 2 3 4 5 Συμφωνώ απόλυτα

23. ...τείνει να είναι οκνηρός, να τεμπελιάζει

Διαφωνώ απόλυτα 1 2 3 4 5 Συμφωνώ απόλυτα

24. ...είναι συναισθηματικά σταθερός, δεν ταράζεται εύκολα

Διαφωνώ απόλυτα 1 2 3 4 5 Συμφωνώ απόλυτα

25. ...είναι επινοητικός

Διαφωνώ απόλυτα 1 2 3 4 5 Συμφωνώ απόλυτα

26. ...είναι κατηγορηματικός, έχει ισχυρή προσωπικότητα

Διαφωνώ απόλυτα 1 2 3 4 5 Συμφωνώ απόλυτα

27. ...μπορεί να είναι ψυχρός και ακατάδεκτος

Διαφωνώ απόλυτα 1 2 3 4 5 Συμφωνώ απόλυτα

28. ...επιμένει μέχρι να ολοκληρωθεί μια εργασία

Διαφωνώ απόλυτα 1 2 3 4 5 Συμφωνώ απόλυτα

29. ...μπορεί να είναι κακόκεφος

Διαφωνώ απόλυτα 1 2 3 4 5 Συμφωνώ απόλυτα

30. ...εκτιμά τις καλλιτεχνικές και αισθητικές δραστηριότητες

Διαφωνώ απόλυτα 1 2 3 4 5 Συμφωνώ απόλυτα

31. ...είναι κάποιες φορές ντροπαλός, συνεσταλμένος

Διαφωνώ απόλυτα 1 2 3 4 5 Συμφωνώ απόλυτα

32. ...είναι διακριτικός και ευγενικός σχεδόν με όλους

Διαφωνώ απόλυτα 1 2 3 4 5 Συμφωνώ απόλυτα

33. ...κάνει πράγματα αποτελεσματικά

Διαφωνώ απόλυτα 1 2 3 4 5 Συμφωνώ απόλυτα

34. ...παραμένει ψύχραιμος σε τεταμένες καταστάσεις

Διαφωνώ απόλυτα 1 2 3 4 5 Συμφωνώ απόλυτα

35. ...προτιμά μια συνηθισμένη δουλειά, ρουτίνας

Διαφωνώ απόλυτα 1 2 3 4 5 Συμφωνώ απόλυτα

36. ...είναι κοινωνικός

Διαφωνώ απόλυτα 1 2 3 4 5 Συμφωνώ απόλυτα

37. ...είναι μερικές φορές αγενής

Διαφωνώ απόλυτα 1 2 3 4 5 Συμφωνώ απόλυτα

38. ...καταστρώνει σχέδια και τα ολοκληρώνει

Διαφωνώ απόλυτα 1 2 3 4 5 Συμφωνώ απόλυτα

39. ...νευριάζει εύκολα

Διαφωνώ απόλυτα 1 2 3 4 5 Συμφωνώ απόλυτα

40. ...του αρέσει να στοχάζεται, να παίζει με τις ιδέες

Διαφωνώ απόλυτα 1 2 3 4 5 Συμφωνώ απόλυτα

41. ...έχει λίγα καλλιτεχνικά ενδιαφέροντα

Διαφωνώ απόλυτα 1 2 3 4 5 Συμφωνώ απόλυτα

42. ...του αρέσει να συνεργάζεται με άλλους

Διαφωνώ απόλυτα 1 2 3 4 5 Συμφωνώ απόλυτα

43. ...εύκολα αποσπάται

Διαφωνώ απόλυτα 1 2 3 4 5 Συμφωνώ απόλυτα

44. ...είναι γνώστης της τέχνης, της μουσικής ή της λογοτεχνίας

Διαφωνώ απόλυτα 1 2 3 4 5 Συμφωνώ απόλυτα

45. ...είναι πολιτικά φιλελεύθερος

Διαφωνώ απόλυτα 1 2 3 4 5 Συμφωνώ απόλυτα