Αυτόματη πρόβλεψη χαρακτηριστικών της προσωπικότητας του συγγραφέα μέσω υφομετρικής ανάλυσης κειμένων

Σοφία Γαγιάτσου, Γεώργιος Μαρκόπουλος & Γεώργιος Μικρός*
Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών, *Hamad Bin Khalifa University

ABSTRACT

This paper presents research on the automatic prediction of author personality traits, based on stylometric techniques applied to a corpus of essays written in Modern Greek by high-school students. Participant students have been profiled with the use of two personality questionnaires, based on the typology of Carl Jung and the Model of Five Factors. By referring to current research findings, we examine the effectiveness of several stylometric features in predicting the personality of students. The feature set employed was a combination of word and sentence length, most frequent part-of-speech tags, most frequent character/word bigrams and trigrams, most/least frequent words, function words, as well as hapax/dis legomena. Features were automatically extracted from the corpus of essays by using tools and natural languages processing resources. A number of different machine learning algorithms have been exploited in order to find the best approach in terms of model performance. The results of our research show a competitive approach to the personality prediction problem and validate the use of stylometric features sets for tackling this kind of research questions. New combinations of stylometric features have emerged, along with corresponding computational techniques, giving satisfying solutions to the problem of author personality automatic recognition for Greek, while the value of using stylometric linguistic features was demonstrated.

ΛΕΞΕΙΣ-ΚΛΕΙΔΙΑ: αναγνώριση συγγραφέα, μηχανική μάθηση, πρόβλεψη προσωπικότητας, υπολογιστική γλωσσολογία, υφομετρικά χαρακτηριστικά

1. Εισαγωγή

Ιδιαίτερο ενδιαφέρον παρουσιάζει η μελέτη των τρεχουσών εξελίξεων στην υπολογιστική γλωσσολογία που επιτρέπουν ανεπτυγμένου τύπου αναλύσεις στα κείμενα. Σε αυτό έχει συμβάλει και η ωρίμανση εργαλείων όπως οι λημματοποιητές και οι μορφολογικοί και συντακτικοί αναλυτές, που έχουν διευρύνει το πεδίο εφαρμογών της ανάκτησης πληροφορίας και της εξόρυξης γνώσης. Ένας νέος και σύγχρονος τύπος υπολογιστικής ανάλυσης κειμένων που συναντάται στη βιβλιογραφία και εντάσσεται στον παραπάνω τομέα εφαρμογών θέτει ως στόχο την αναγνώριση του συγγραφέα (authorship identification).

Το ύφος με το οποίο είναι γραμμένο ένα κείμενο έχει αποδειχτεί ότι παρέχει σημαντικές πληροφορίες για την αναγνώριση του συγγραφέα. Ο συγκεκριμένος ερευνητικός τομέας διακρίνεται σε τρεις επιμέρους κλάδους: στην απόδοση ενός κειμένου σε συγκεκριμένο συγγραφέα από ένα πεπερασμένο σύνολο συγγραφέων (authorship attribution), στην απόδοση κειμένου σε συγγραφέα που δεν ανήκει σε κλειστό σύνολο (authorship verification) και στον καθορισμό δημογραφικών και ψυχολογικών χαρακτηριστικών του συγγραφέα (authorship profiling), που περιλαμβάνει την κατηγοριοποίηση του συγγραφέα με κριτήριο το φύλο, την ηλικία, τη μητρική γλώσσα, τη μόρφωση, την προσωπικότητα κ.λπ. (Gagiatsou et al. 2021a).

Το παρόν άρθρο παρουσιάζει τα αποτελέσματα της έρευνας που διεξήχθη στο πλαίσιο της διδακτορικής διατριβής της πρώτης συγγραφέως, η οποία εκπονήθηκε στον τομέα Γλωσσολογίας του Εθνικού και Καποδιστριακού Πανεπιστημίου Αθηνών, για την πρόβλεψη στοιχείων της προσωπικότητας του συγγραφέα μέσω τεχνικών επεξεργασίας φυσικής γλώσσας. Για τον σκοπό αυτό δημιουργήθηκε σώμα κειμένων από εκθέσεις μαθητών Λυκείου, φυσικών ομιλητών της Ελληνικής γλώσσας. Η προσωπικότητα των συμμετεχόντων μαθητών καθορίστηκε με τη χορήγηση και συμπλήρωση ερωτηματολογίων προσωπικότητας. Το ύφος των κειμένων υπέστη ποσοτική επεξεργασία και στη συνέχεια προσδιορίστηκαν τα υφομετρικά χαρακτηριστικά που εξήχθησαν από το σώμα κειμένων, τα οποία χρησιμοποιήθηκαν αποτελεσματικά για την

πρόβλεψη προσωπικότητας. Επιπλέον, αξιολογήθηκαν οι αλγόριθμοι μηχανικής μάθησης ως προς την ικανότητά τους να προβλέπουν με σχετική ακρίβεια συγκεκριμένα χαρακτηριστικά προσωπικότητας ενός/μιας συγγραφέα-μαθητή/μαθήτριας βάσει των γραπτών εκθέσεών του/της.

2. Προσωπικότητα και γλώσσα

Ο τρόπος με τον οποίο χρησιμοποιεί ένα άτομο τη γλώσσα ως κώδικα επικοινωνίας αποκαλύπτει πολλές πληροφορίες για την προσωπικότητά του. 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 κάθε άνθρωπος έχει ένα χαρακτηριστικό τρόπο χρήσης της γλώσσας, ένα είδος συγγραφικού αποτυπώματος (Juola 2008).

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 et al. (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 επίσης την ψυχολογική κατάσταση ενός ατόμου. Σύγχρονες έρευνες, μάλιστα, δείχνουν ότι η σχέση γλώσσας και προσωπικότητας είναι δυνατό να προσδιοριστεί υπολογιστικά.

3. Οι πρώτες μελέτες αυτόματης πρόβλεψης της προσωπικότητας από κείμενο

Μια από τις πρώτες προσπάθειες για την αυτόματη πρόβλεψη της προσωπικότητας των συγγραφέων με τεχνικές μηχανικής μάθησης αφορούσε σε κείμενα με ύφος οικείο και καθημερινό (Argamon et al. 2005, 2007). Η έρευνα κινείται στο πλαίσιο της ψυχολογίας της γλώσσας, αλλά και της υφομετρίας. Αντικείμενο επεξεργασίας της έρευνας υπήρξαν 1.157 εκθέσεις αυθόρμητες και εκθέσεις στις οποίες ανέλυαν τον χαρακτήρα τους 1.106 φοιτητές Ψυχολογίας του πανεπιστημίου του Τέξας στοAustin, που γράφτηκαν μεταξύ των ετών 1997 και 2003. Από το ερωτηματολόγιο προσωπικότητας που συμπλήρωσαν οι φοιτητές, το οποίο ήταν βασισμένο στο μοντέλο των Πέντε Παραγόντων, αξιοποιήθηκαν τα αποτελέσματα για την εξωστρέφεια (extraversion) και τον νευρωτισμό (neuroticism).

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, 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 τη δημιουργία του μοντέλου αναγνώρισης χρησιμοποιήθηκαν Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines) με τις οποίες 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, βλ. Mairesse et al. 2007) έχει καθοδική (top-down) προσέγγιση. Σε αντίθεση με άλλες έρευνες που αναφέρουν μόνο αποτελέσματα ταξινόμησης, σε αυτήν εφαρμόστηκαν μοντέλα ταξινόμησης (classification), παλινδρόμησης (regression) και ιεράρχησης (ranking) για κάθε διάσταση του μοντέλου των Πέντε Παραγόντων. Πραγματοποιήθηκαν διάφορα πειράματα σε ένα

σώμα κειμένων 2.479 εκθέσεων, οι οποίες συντάχθηκαν από φοιτητές ψυχολογίας μέσα σε είκοσι λεπτά στα οποία έγραφαν ό,τι σκεφτόντουσαν, για να διαπιστωθεί 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 στην επόμενη ενότητα με θέμα την αναγνώριση προσωπικότητας από προφορικό λόγο.

Από 88 κατηγορίες λέξεων που εντοπίστηκαν με το λογισμικό ανάλυσης κειμένου Linguistic Inquiry and Word Count (LIWC, Pennebaker et al. 2015) εξήχθησαν γλωσσικά χαρακτηριστικά που περιλαμβάνουν συντακτική και σημασιολογική πληροφορία. Επιπλέον, με τη χρήση της ψυχογλωσσολογικής βάσης The Medical Research Council Psycholinguistic Database Machine Usable Dictionary MRC (Wilson 1988) προέκυψαν 14 ψυχογλωσσολογικά χαρακτηριστικά. Για την αξιολόγηση των μοντέλων ταξινόμησης έγινε χρήση των εξής 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 σε νέες εμπειρίες (openness to experience), καθώς πέντε στους έξι αλγορίθμους ξεπέρασαν το σημείο αναφοράς (ακρίβεια 62,1%) και τέσσερις παρουσιάζουν την καλύτερη απόδοση γι' αυτή τη διάσταση. Η ευσυνειδησία (conscientiousness) είναι η δυσκολότερη διάσταση για μοντελοποίηση. Τα καλύτερα αποτελέσματα ταξινόμησης επιτεύχθηκαν με τη χρήση του αλγορίθμου Naive Bayes με 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 όλες τις διαστάσεις του μοντέλου των Πέντε Παραγόντων και περισσότερο για την δεκτικότητα σε νέες εμπειρίες.

4. Μεθοδολογία της έρευνας

Ακολούθως περιγράφεται η μεθοδολογία της παρούσας έρευνας. Περιγράφεται το σώμα κειμένων, σχολιάζονται τα ερωτηματολόγια προσωπικότητας που χορηγήθηκαν στους μαθητές, γίνεται αναφορά στο λογισμικό που χρησιμοποιήθηκε για την υλοποίηση των πειραμάτων και παρουσιάζονται τα υφομετρικά χαρακτηριστικά που εξήχθησαν από τα κείμενα.

4.1. Το ηλεκτρονικό σώμα κειμένων

Για να φέρει αποτελέσματα η πρόβλεψη προσωπικότητας είναι απαραίτητο να αντιμετωπιστεί ένα πρωταρχικό πρόβλημα, η δυσκολία συλλογής δεδομένων επισημειωμένων ως προς την προσωπικότητα του/της συγγραφέα τους. Για τη διεξαγωγή, λοιπόν, της έρευνάς μας απαραίτητη ήταν η συλλογή πρωτογενών κειμενικών δεδομένων από φυσικούς ομιλητές/ομιλήτριες της ελληνικής με στόχο τη δημιουργία ενός ηλεκτρονικού σώματος κειμένων.

Ειδικότερα, ζητήθηκε από μαθητές/μαθήτριες Λυκείου στο πλαίσιο του μαθήματος της Νεοελληνικής Γλώσσας να γράψουν εκθέσεις. Η συμμετοχή στην έρευνα ήταν εθελοντική και η διαδικασία 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/μαθητριών να μην είναι επιτηδευμένος. Δόθηκαν τα εξής τρία θέματα προς συγγραφή, η επιλογή των οποίων έγινε με κριτήρια την ποικιλία στο περιεχόμενο, στο κειμενικό είδος και το επικοινωνιακό πλαίσιο:

- «Σε μια καλλιτεχνική εκδήλωση που γίνεται στον 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».

- «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».
- «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».

Για την έρευνα αξιοποιήθηκε το υλικό από 198 μαθητές/μαθήτριες και το σώμα κειμένων ανήλθε σε σχεδόν 250.000 λέξεις. Ο μέσος όρος του αριθμού των λέξεων των εκθέσεων ανά μαθητή/μαθήτρια ανήλθε στις 1.255 λέξεις. Το ηλεκτρονικό σώμα κειμένων που αναπτύθηκε 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/μαθήτριες, ηλικίας μεταξύ 16 και 18 ετών. Από αυτούς οι 88 (44,4%) είναι αγόρια και οι 110 (55,5%) κορίτσια. Οι εκθέσεις ήταν όλες χειρόγραφες και ακολούθησε η διαδικασία ψηφιοποίησής τους με πληκτρολόγηση.

4.2. Τα ερωτηματολόγια προσωπικότητας

Για κάθε μαθητή/μαθήτρια είναι διαθέσιμα τα αποτελέσματα των δύο σταθμισμένων τεστ προσωπικότητας που τους χορηγήθηκαν μεταφρασμένα στα ελληνικά και συμπληρώθηκαν εντός της σχολικής αίθουσας μία εβδομάδα μετά την ολοκλήρωση της συγγραφής των εκθέσεων. Πρόκειται για το ερωτηματολόγιο τύπων προσωπικότητας Myers-Briggs Type Indicator (Myers-Briggs 1962), βασισμένο στο τυπολογικό μοντέλο προσωπικότητας του Carl Jung και το τεστ προσωπικότητας Big Five, που βασίζεται στο μοντέλο των Πέντε Παραγόντων (Costa & McCrae 1993).

Το ερωτηματολόγιο τύπων προσωπικότητας κατηγοριοποιεί τις ψυχολογικές διαφορές στην προσωπικότητα των ατόμων σε τέσσερις διχοτομήσεις. Όλοι οι πιθανοί συνδυασμοί αποδίδουν δεκαέξι τύπους προσωπικότητας: εξωστρεφής (extraverted-E)-εσωστρεφής (introverted-I), αντιληπτικός (sensing-S)-διαισθητικός (intuitive-N), νοητικός (thinking-T)-συναισθηματικός (feeling-F), οργανωτικός (judging-J)-προσαρμοστικός (perceiving-P). Κάθε τύπος προσδιορίζεται από ένα ακρωνύμιο τεσσάρων γραμμάτων του αντίστοιχου συνδυασμού. Για παράδειγμα, το ακρωνύμιο ENFJ δηλώνει τον εξωστρεφή, διαισθητικό, συναισθηματικό, οργανωτικό τύπο.

Το ερωτηματολόγιο προσωπικότητας των Πέντε Παραγόντων είναι το πιο δημοφιλές τεστ προσωπικότητας και στον τομέα της πρόβλεψης του ψυχολογικού προφίλ του συγγραφέα, όπως προέκυψε από τη βιβλιογραφική επισκόπηση. Περιγράφει την προσωπικότητα μέσω πέντε διαστάσεων: δεκτικότητα σε νέες εμπειρίες (openness to experience-O), ευσυνειδησία (conscientiousness-C), εξωστρέφεια (extraversion-E), προσήνεια (agreeableness-A), νευρωτισμός (neuroticism-N).

Οι τιμές των τύπων και των διαστάσεων της προσωπικότητας (βλ. Πίνακα 1) που προκύπτουν ως αποτελέσματα από τα δύο ερωτηματολόγια είναι αριθμητικές και δείχνουν κατά πόσο κάθε γνώρισμα περιγράφει ένα άτομο.

Πίνακας 1. Τύποι και διαστάσεις της προσωπικότητας βάσει των ερωτηματολογίων

Τύπος προσωπικότητας	*Διάσταση προσωπικότητας*
Εξωστρεφής (Extraverted)	Δεκτικότητα σε νέες εμπειρίες (Openness to Experiences)
Εσωστρεφής (Introverted)	Μη δεκτικότητα σε νέες εμπειρίες (Closed- Mindedness)
Αντιληπτικός (Sensing)	Ευσυνειδησία (Conscientiousness)
Διαισθητικός (iNtuitive)	Έλλειψη οργάνωσης (Disorganisation)
Νοητικός (Thinking)	Εξωστρέφεια (Extraversion)
Συναισθηματικός (Feeling)	Εσωστρέφεια (Introversion)
Οργανωτικός (Judging)	Προσήνεια (Agreeableness)
Προσαρμοστικός (Perceiving)	Απουσία προσήνειας (Disagreeableness)
	Νευρωτισμός (Neuroticism)
	Συναισθηματική σταθερότητα (Calmness)

4.3. Ανάλυση δεδομένων

Το λογισμικό που επιλέχθηκε για την αυτόματη αναγνώριση της προσωπικότητας των μαθητών είναι το RapidMiner (Mierswa & Klinkenberg 2018), καθώς είναι αναγνωρισμένο σε παγκόσμια κλίμακα και από τα δημοφιλέστερα εργαλεία εξόρυξης δεδομένων παγκοσμίως. Η τροφοδότηση του λογισμικού απαιτούσε προηγουμένως την επεξεργασία των κειμενικών δεδομένων με εργαλεία αυτόματης κειμενικής ανάλυσης διαφόρων βαθμίδων γλωσσικής τεχνολογίας. Έτσι, πραγματοποιήθηκε χωρισμός σε λέξεις (tokenization), λημματοποίηση (lemmatization), μορφολογική επισημείωση (part-of-speech tagging). Με το RapidMiner σχεδιάσαμε και εκτελέσαμε πολλές δεκάδες διεργασιών (processes), για να εξάγουμε υφομετρικά χαρακτηριστικά από τα δύο σώματα κειμένων που δημιουργήσαμε τόσο για την πρόβλεψη των Τύπων προσωπικότητας του Jung όσο και των διαστάσεων προσωπικότητας του μοντέλου των Πέντε Παραγόντων.

4.4. Εξαγωγή υφομετρικών χαρακτηριστικών

Μέσω των υφομετρικών χαρακτηριστικών ποσοτικοποιήθηκε η γλώσσα των εκθέσεων, δηλαδή μετατράπηκαν τα κειμενικά δεδομένα σε αριθμούς προκειμένου να εξετασθεί η αρχική μας υπόθεση ότι η προσωπικότητα των συγγραφέων μπορεί αυτόματα να αναγνωριστεί από τα κείμενά τους. Συγκεκριμένα, τα υφομετρικά χαρακτηριστικά που εξήχθησαν από το ηλεκτρονικό σώμα κειμένων κατηγοριοποιήθηκαν σε τρεις ομάδες, για καθεμία από τις οποίες χρησιμοποιήθηκε διαφορετικά επεξεργασμένο σώμα κειμένων, όπως φαίνεται στον Πίνακα 2 που ακολουθεί.

Πίνακας 2. Υφομετρικά χαρακτηριστικά που χρησιμοποιήθηκαν ανά είδος σώματος κειμένων

Απλό κείμενο	Μορφολογικά επισημειωμένο σώμα κειμένων	Λημματοποιημένο σώμα κειμένων
Μέσο μήκος λέξης σε χαρακτήρες όλων των λέξεων	Σχετική συχνότητα εμφάνισης ρηματικών τύπων	Σχετική συχνότητα εμφάνισης των λειτουργικών λέξεων
Μέσο μήκος πρότασης σε λέξεις όλων των προτάσεων	Σχετική συχνότητα εμφάνισης των ρημάτων ενεργητικής φωνής	Σχετική συχνότητα εμφάνισης των πιο συχνών λέξεων μέσα
Σχετική συχνότητα εμφάνισης των πιο συχνών δίλεκτων	Σχετική συχνότητα εμφάνισης των ρημάτων παθητικής φωνής	Σχετική συχνότητα εμφάνισης των πιο συχνών μη λειτουργικών λέξεων
Σχετική συχνότητα εμφάνισης των πιο συχνών τρίλεκτων	Σχετική συχνότητα εμφάνισης των ουσιαστικών	Σχετική συχνότητα εμφάνισης των πιο σπάνιων λέξεων
Σχετική συχνότητα εμφάνισης των πιο συχνών διγραμμάτων χαρακτήρων	Σχετική συχνότητα εμφάνισης των επιθέτων	Σχετική συχνότητα εμφάνισης των πιο σπάνιων μη λειτουργικών λέξεων
Σχετική συχνότητα εμφάνισης των πιο συχνών τριγραμμάτων χαρακτήρων	Σχετική συχνότητα εμφάνισης των άρθρων	Σχετική συχνότητα μη λειτουργικών λέξεων
Σχετική συχνότητα εμφάνισης των 100 πιο συχνών λέξεων	Σχετική συχνότητα εμφάνισης των αντωνυμιών	Σχετική συχνότητα των άπαξ εμφανιζομένων λέξεων
Σχετική συχνότητα εμφάνισης των 100 πιο συχνών δίλεκτων	Σχετική συχνότητα εμφάνισης των προσωπικών αντωνυμιών	Σχετική συχνότητα των δις εμφανιζομένων λέξεων
Σχετική συχνότητα εμφάνισης των 100 πιο συχνών διγραμμάτων χαρακτήρων	Σχετική συχνότητα εμφάνισης των προσωπικών και κτητικών αντωνυμιών	Λόγος των δις προς άπαξ λεγόμενα.
Σχετική συχνότητα εμφάνισης των 100 πιο συχνών τριγραμμάτων χαρακτήρων	Σχετική συχνότητα εμφάνισης των επιρρημάτων	Σχετική συχνότητα εμφάνισης των μη λειτουργικών λέξεων μέσα
	Σχετική συχνότητα εμφάνισης των προθέσεων	Λόγος των λέξεων περιεχομένου προς τις λειτουργικές
	Σχετική συχνότητα εμφάνισης των συνδέσμων	
	Σχετική συχνότητα εμφάνισης των παρατακτικών συνδέσμων	
	Σχετική συχνότητα εμφάνισης των υποτακτικών συνδέσμων	

5. Αποτελέσματα

Στην παρούσα ενότητα εξετάζεται η πειραματική διαδικασία που ακολουθήθηκε με σκοπό την αυτόματη κατηγοριοποίηση των εκθέσεων των μαθητών/μαθητριών βάσει της προσωπικότητάς τους, έτσι όπως αποτυπώθηκε στα ερωτηματολόγια προσωπικότητας που συμπλήρωσαν. Ειδικότερα, αξιολογούμε τους αλγορίθμους μηχανικής μάθησης ως προς την προβλεπτική τους ικανότητα χρησιμοποιώντας ως δεδομένα εκπαίδευσης και εν μέρει ελέγχου τις εκθέσεις των μαθητών/μαθητριών, των οποίων η προσωπικότητα έχει ήδη αξιολογηθεί μέσω των ερωτηματολογίων. Παρουσιάζονται τα αποτελέσματα των πειραμάτων ταξινόμησης (Gagiatsou et al. 2021b) με εφαρμογή των εννέα αλγορίθμων μηχανικής μάθησης, οι οποίοι είναι οι: Naive Bayes, Generalized Linear Model, Logistic Regression, Fast Large Margin, Deep Learning, Decision Tree, Random Forest, Gradient Boosted Trees, Support Vector Machine.

5.1. Αποτελέσματα αυτόματης ταξινόμησης μαθητικών εκθέσεων: Ερωτηματολόγιο τύπων προσωπικότητας

Ελέγχοντας όλους τους τύπους του ερωτηματολογίου ο Naive Bayes ήταν ο αλγόριθμος που είχε συνολικά τα καλύτερα αποτελέσματα για την πρόβλεψή τους. 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 των εξωστρεφών, 80% για την πρόβλεψη των διαισθητικών, 69% των συναισθηματικών και 76% των οργανωτικών μαθητών/μαθητριών.

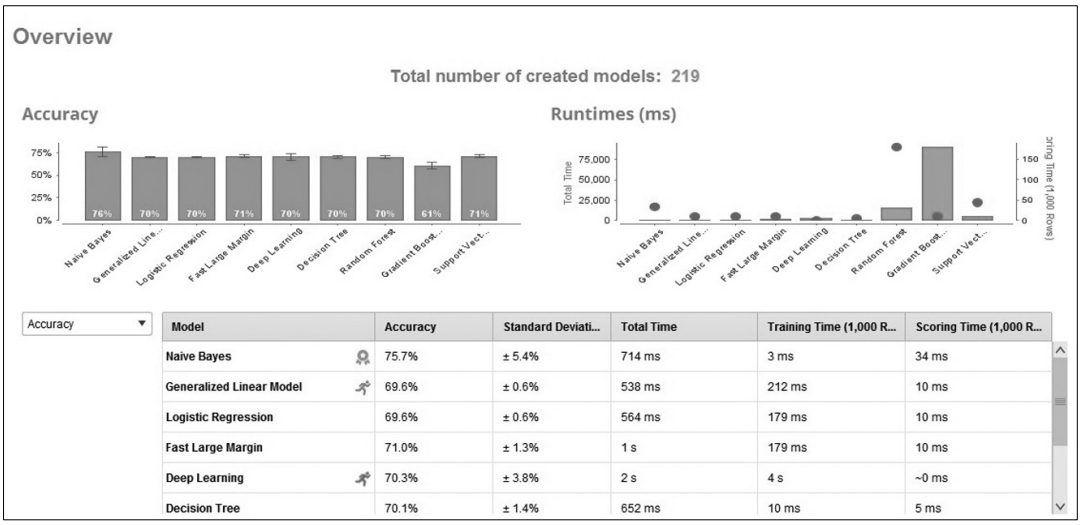
Ο Πίνακας 3 παρουσιάζει την απόδοση όσον αφορά στις μετρήσεις ακρίβειας, ορθότητας και ανάκλησης αυτού του ταξινομητή για κάθε τύπο προσωπικότητας.

Πίνακας 3: Η απόδοση του αλγορίθμου Naive Bayes

Τύπος προσωπικότητας	*Ακρίβεια*	*Ορθότητα*	*Ανάκληση*
Εξωστρεφής	80,67%	80,54%	100%
Διαισθητικός	79,85%	81,31%	92,55%
Συναισθηματικός	68,75%	67,69%	96,70%
Οργανωτικός	75,68%	76,15%	95,19%

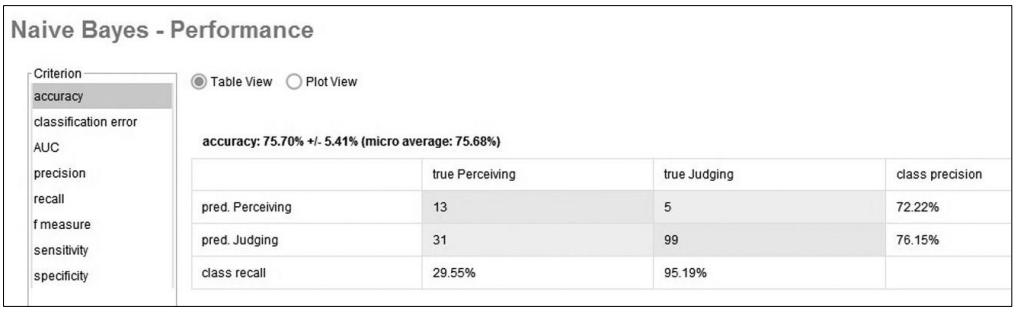
Στην Εικόνα 1 παραθέτουμε ενδεικτικά τα αποτελέσματα της πρόβλεψης σε εκατοστιαίο ποσοστό των 219 μοντέλων του RapidMiner για την οργανωτικότητα των μαθητών. Παρατηρούμε ότι ο Naive Bayes πέτυχε το καλύτερο αποτέλεσμα, 76%. Οι άλλοι οκτώ αλγόριθμοι πέτυχαν ποσοστό ακρίβειας που κυμαίνεται από 61% και δεν ξεπερνά το 71%.

Εικόνα 1. Η απόδοση όλων των αλγορίθμων του RapidMiner για τους οργανωτικούς



Τα μέτρα αξιολόγησης του αποτελεσματικότερου αλγόριθμου φαίνονται στην Εικόνα 2. Με ορθότητα 76,15% και ανάκληση 95,19% η ακρίβεια του αλγόριθμου ήταν 75,70%.

Εικόνα 2: Η απόδοση του Naive Bayes για τους Οργανωτικούς



Ακολουθούν τα υφομετρικά χαρακτηριστικά που επέδρασαν σημαντικά στην πρόβλεψη του αλγορίθμου για όλους τους τύπους προσωπικότητας. Για την εξωστρέφεια επέδρασε σημαντικά η χρήση ρηματικών τύπων ενεργητικής φωνής. Ακολουθούν το μέσο μήκος πρότασης σε λέξεις όλων των προτάσεων, οι λέξεις που απαντούν δύο μόνο φορές σε ένα κείμενο, οι πιο συχνές λέξεις περιεχομένου και οι προσωπικές αντωνυμίες.

Τη σημαντικότερη επίδραση στην πρόβλεψη των διαισθητικών μαθητών είχε το μέσο μήκος λέξης σε χαρακτήρες. Έπονται τα συχνότερα τριγράμματα χαρακτήρων, οι άπαξ εμφανιζόμενες λέξεις, οι προσωπικές αντωνυμίες, οι λέξεις περιεχομένου, τα συχνότερα δίλεκτα, οι πιο σπάνιες λέξεις, τα συχνότερα τρίλεκτα και όλες οι λέξεις περιεχομένου.

Τα υφομετρικά χαρακτηριστικά που επηρέασαν το αποτέλεσμα της ταξινόμησης των εκθέσεων ως προς το συναίσθημα είναι τα ρήματα, τα επίθετα, οι συχνότερες μη λειτουργικές λέξεις, οι προσωπικές και κτητικές αντωνυμίες, τα ουσιαστικά και τα επιρρήματα.

Τέλος, τα οκτώ υφομετρικά χαρακτηριστικά που συνέβαλαν στην πρόβλεψη των οργανωτικών μαθητών ήταν κατά σειρά φθίνουσα: τα συχνότερα τρίλεκτα, τα συχνότερα δίλεκτα, το μέσο μήκος πρότασης σε λέξεις, τα συχνότερα διγράμματα χαρακτήρων και τα συχνότερα τριγράμματα χαρακτήρων, οι προσωπικές και κτητικές αντωνυμίες, τα άρθρα και το μέσο μήκος λέξης σε χαρακτήρες.

5.2. Αποτελέσματα αυτόματης ταξινόμησης μαθητικών εκθέσεων: Ερωτηματολόγιο προσωπικότητας των Πέντε Παραγόντων

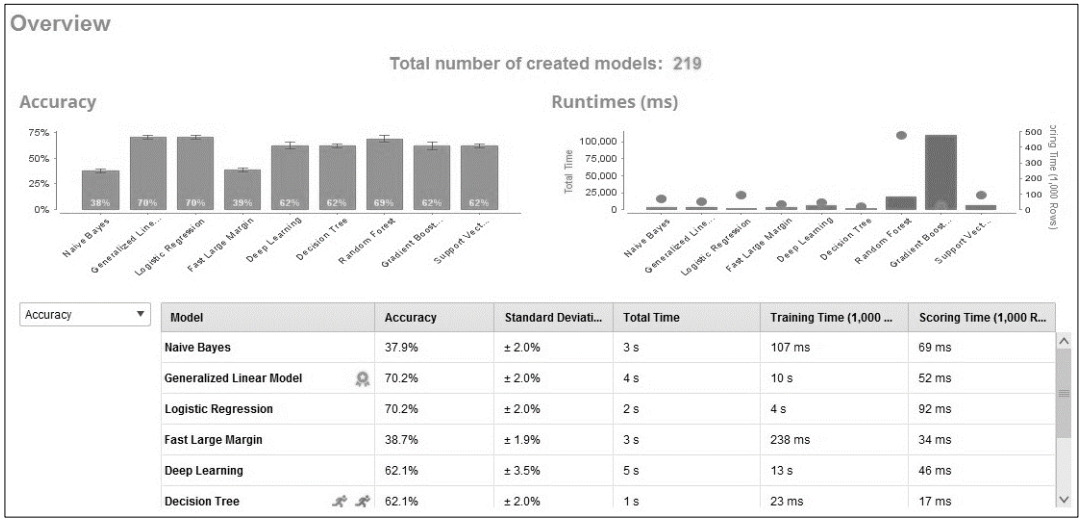
Για την πρόβλεψη όλων των διαστάσεων του ερωτηματολογίου των Πέντε Παραγόντων ο αλγόριθμος που είχε τα καλύτερα αποτελέσματα ήταν ο Generalized Linear Model. Τα ποσοστά ακρίβειας της πρόβλεψης του Generalized Linear Model κυμαίνονται από 66% έως 86%, με μέσο όρο 72,2%. Συγκεκριμένα, πέτυχε ακρίβεια 86% για την πρόβλεψη της μη δεκτικότητας των μαθητών/μαθητριών σε νέες εμπειρίες, 71% για την πρόβλεψη της έλλειψης οργάνωσης, 68% της εσωστρέφειας, 70% της προσήνειας και 66% της συναισθηματικής σταθερότητας. Ο Πίνακας 4 παρουσιάζει την απόδοση όσον αφορά στις μετρήσεις ακρίβειας, ορθότητας και ανάκλησης αυτού του ταξινομητή για όλες τις διαστάσεις της προσωπικότητας.

Πίνακας 4: Η απόδοση του Generalized Linear Model

Διάσταση προσωπικότητας	*Ακρίβεια*	*Ορθότητα*	*Ανάκληση*
Μη Δεκτικότητα σε νέες εμπειρίες	85,94%	85,37%	100%
Έλλειψη οργάνωσης	71,19%	68,57%	80%
Εσωστρέφεια	67,62%	66,67%	86,67%
Προσήνεια	70,16%	67,86%	98,70%
Συναισθηματική σταθερότητα	65,60%	64,79%	71,88%

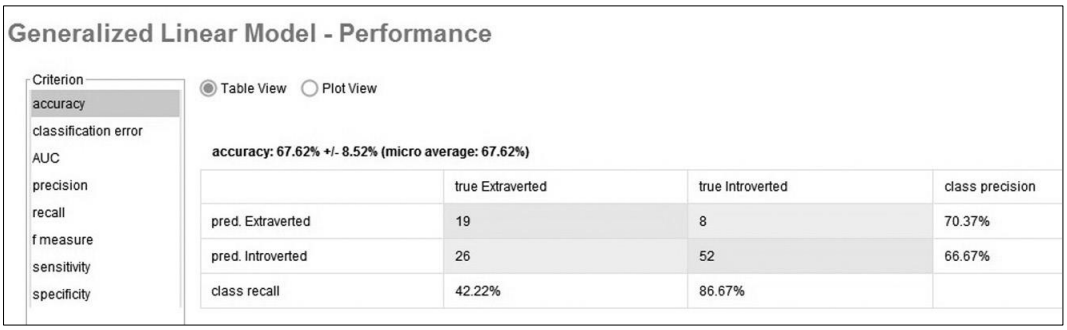
Ενδεικτικά, στην Εικόνα 3 αποτυπώνεται η γενική απόδοση των εννέα αλγορίθμων για την πρόβλεψη της διάστασης της εσωστρέφειας. Η εσωστρέφεια, που επικράτησε έναντι της εξωστρέφειας στο συγκεκριμένο δείγμα, προβλέφθηκε από τον αλγόριθμο με ακρίβεια 68%. Το αμέσως χαμηλότερο ποσοστό ήταν 65%, ενώ ακολουθούν τα υπόλοιπα μοντέλα με ακρίβεια από 60% έως 53% το ελάχιστο.

Εικόνα 3: Απόδοση όλων των αλγορίθμων του RapidMiner για την εσωστρέφεια



Ο Generalized Linear Model, που είναι ο αποτελεσματικότερος αλγόριθμος, αξιολογήθηκε με ορθότητα 66,67% και ανάκληση 86,67%, και επομένως ακρίβεια 67,62%, όπως παρουσιάζεται στην ακόλουθη εικόνα:

Εικόνα 4: Η Απόδοση του Generalized Linear Model για την εσωστρέφεια



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

Τα υφομετρικά χαρακτηριστικά που επιλέχθηκαν, καθώς διαπιστώθηκε ότι ο συνδυασμός τους είχε το καλύτερο αποτέλεσμα, αξιολογήθηκαν από το RapidMiner ως προς την επίδραση που έχουν στο μοντέλο για την πρόβλεψη της μη δεκτικότητας στην εμπειρία. Έτσι, το πιο σημαντικό κρίθηκε η χρήση προσωπικών αντωνυμιών. Ακολουθούν τα ρήματα και οι λέξεις που εμφανίζονται δύο φορές σε κάθε κείμενο.

Το πιο σημαντικό υφομετρικό χαρακτηριστικό για την πρόβλεψη της έλλειψης οργάνωσης είναι η λειτουργική πυκνότητα, οι μη λειτουργικές λέξεις, οι λειτουργικές λέξεις, οι εμφανίσεις των μη λειτουργικών λέξεων, οι λέξεις που απαντούν δύο φορές, οι πιο συχνές μη λειτουργικές λέξεις, ο λόγος των δις προς άπαξ λεγόμενα.

Για την εσωστρέφεια επέδρασαν σημαντικά κατά φθίνουσα σειρά το μέσο μήκος πρότασης σε λέξεις, ο λόγος των δις προς άπαξ λεγόμενα, οι προσωπικές και κτητικές αντωνυμίες, τα συχνότερα δίλεκτα, τα επιρρήματα.

Τα χαρακτηριστικά που συνέβαλαν στην απόδοση του αλγορίθμου για την πρόβλεψη της προσήνειας, έτσι όπως αποτυπώνονται στο RapidMiner, είναι τα ρήματα, ο λόγος των δις προς άπαξ λεγόμενα, οι λέξεις που απαντούν δύο μόνο φορές, οι ρηματικοί τύποι ενεργητικής φωνής.

Για την πρόβλεψη της συναισθηματικής σταθερότητας 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 διάσταση προσωπικότητας, στα οποία μετρήθηκαν όλα τα υπόλοιπα 31 υφομετρικά χαρακτηριστικά. Έτσι, εκτός από τα πιο συχνά τριγράμματα χαρακτήρων όλου του σώματος κειμένων, επηρέασαν και οι υποτακτικοί σύνδεσμοι, τα επιρρήματα, τα ουσιαστικά και τα πιο συχνά τρίλεκτα.

6. Συμπεράσματα

Συνοψίζοντας, στην παρούσα μελέτη παρουσιάσαμε τα αποτελέσματα της έρευνάς μας στο πεδίο της αυτόματης πρόβλεψης της προσωπικότητας. Εφαρμόσαμε υπολογιστικές τεχνικές ανίχνευσης της προσωπικότητας για πρώτη φορά τόσο στα ελληνικά όσο και σε άλλες γλώσσες σε γραπτά κείμενα μαθητών/μαθητριών Λυκείου. Το νέο ηλεκτρονικό σώμα κειμένων συντέθηκε από τις εκθέσεις που έγραψαν, αφού προηγουμένως το ψυχολογικό προφίλ τους είχε αποτυπωθεί μέσω δύο προτυποποιημένων ερωτηματολογίων προσωπικότητας. Τα χαρακτηριστικά που εξήχθησαν από τις εκθέσεις και συνέβαλαν στην ταξινόμηση ήταν αποκλειστικά υφομετρικά και διαφοροποιούνται ανάλογα με το είδος της επισημείωσης του σώματος κειμένων. Το ζητούμενο ήταν να προσδιοριστεί ο βαθμός συμφωνίας του αλγορίθμου ταξινόμησης των μαθητών/μαθητριών με τα αποτελέσματα των ψυχομετρικών δοκιμών στις ίδιες κατηγορίες που έχουν προκύψει από τις δοκιμές σε ένα σύνολο άγνωστων ως προς τον αλγόριθμο εκθέσεων, ερευνητική υπόθεση που επιβεβαιώθηκε από τα αποτελέσματα.

Ο επικρατέστερος αλγόριθμος για την αναγνώριση των τύπων προσωπικότητας, όπως αποδίδονται από το ερωτηματολόγιο τύπων προσωπικότητας είναι ο Naive Bayes με μέσο όρο ακρίβειας 76,5%, ποσοστό αυξημένο σε σχέση με το 68,62% που καταγράφηκε για την ολλανδική γλώσσα (Luyckx & Daelemans 2008). Για την πρόβλεψη των διαστάσεων της 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 παρούσα έρευνα το ποσοστό είναι 72,2%.

Είναι σημαντικό πως υπάρχει πλέον για την ελληνική γλώσσα ένα σύνολο υφομετρικών δεικτών που έχουν δοκιμαστεί με εννέα αλγόριθμους, από τους οποίους έχουν προκύψει συγκεκριμένα αποτελέσματα. Καθίσταται σαφές από τα αποτελέσματα της έρευνας ότι οι υφομετρικές μεταβλητές μπορούν να χρησιμοποιηθούν ως αξιόπιστοι δείκτες πρόβλεψης του ψυχολογικού προφίλ ενός/μιας συγγραφέα. Η πολυεπίπεδη υφομετρική επεξεργασία των μαθητικών εκθέσεων και η ποσοτικοποίηση της γλώσσας τους επιβεβαίωσαν ότι η διαφορά στο ύφος του γράφοντος ανάλογα με την ιδιοσυγκρασία του ισχύει και για την ελληνική γλώσσα.

Βιβλιογραφία

Argamon, S., Dhawle, S., Koppel, M. & Pennebaker, J. W. 2005. Lexical predictors of personality type. *Proceedings of the 2005 Joint Annual Meeting of the Interface and the Classification Society of North America, (June 2005)*, 1-16.

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.

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

Gagiatsou, S., Markopoulos, G. & Mikros, G. 2021a. Using stylometric features for the prediction of author personality types in Modern Greek essays. *Proceedings of the Fifteenth International Conference on Digital Society, (Nice, 18-22 July 2021*), 34-39.

Gagiatsou, S., Markopoulos, G. & Mikros, G. 2021b. Prediction of Authors' Personality Types and Traits in Modern Greek Essays Using Stylometric Features. *International Journal on Advances in Life Sciences* 13(1&2), 124-133.

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

Juola, P. 2008. Authorship attribution. *Foundations and Trends in Information Retrieval* 1(3), 233-334.

Luyckx, K. & Daelemans, W. 2008. Personae: A corpus for author and personality prediction from text. *Proceedings of the 6th International Language Resources and Evaluation Conference, (Morocco 28-30 May 2008)*, 2981-2987.

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, (Vancouver 26-29 July 2006)*, 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.

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 (Sorrento 17-20 November 2020)*, 1184-1189.

Mierswa, I., & Klinkenberg, R. 2018. RapidMiner Studio (9.1) Data science, machine learning, predictive analytics. Available at: https://rapidminer.com.

Myers-Briggs, I. 1962. *The Myers-Briggs Type Indicator*. Palo Alto, California: Consulting Psychologists Press.

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., Boyd, R. L., Jordan, K. & Blackburn, K. 2015. *The Development and Psychometric Properties of LIWC2015.* Austin, TX: University of Texas at Austin.

Wilson, M. 1988. MRC Psycholinguistic database: Machine usable dictionary, Version 2.00. *Behavioural Research Methods, Instruments and Computers* 20, 6-11.