ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ΤΜΗΜΑ ΦΙΛΟΛΟΓΙΑΣ
ΤΟΜΕΑΣ ΓΛΩΣΣΟΛΟΓΙΑΣ

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ
ΜΗΧΑΝΙΚΩΝ & ΜΗΧΑΝΙΚΩΝ Η/Υ

ΙΝΣΤΙΤΟΥΤΟ ΕΠΕΞΕΡΓΑΣΙΑΣ ΛΟΓΟΥ

ΔΙΕΠΙΣΤΗΜΟΝΙΚΟ - ΔΙΑΠΑΝΕΠΙΣΤΗΜΙΑΚΟ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ "ΤΕΧΝΟΓΛΩΣΣΙΑ"

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Αυτόματη απόδοση συγγραφικής πατρότητας στα άμεσα μηνύματα του Twitter της Νέας Ελληνικής γλώσσας

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ : ΜΙΚΡΟΣ ΓΕΩΡΓΙΟΣ

ΟΝΟΜΑΤΕΠΩΝΥΜΑ:

ΓΟΥΛΑΣ ΘΕΟΔΩΡΟΣ

ΚΟΥΤΣΙΟΥΚΟΥ ΠΑΝΑΓΙΩΤΑ

ΝΕΟΚΛΕΟΥΣ ΧΡΙΣΤΙΑΝΑ

ΦΕΒΡΟΥΑΡΙΟΣ 2014

Ευχαριστίες

Ευχαριστούμε θερμά τον επιβλέποντα καθηγητή μας κ. Γεώργιο Μικρό, για την υπομονή, το ειλικρινές του ενδιαφέρον και την βοήθεια που προσέφερε κατά την εκπόνηση της διπλωματικής εργασίας μας. Με την καθοδήγηση και τις συμβουλές του καταφέραμε να εμβαθύνουμε στο συγκεκριμένο αντικείμενο και να αναπτύξουμε τους ερευνητικούς μας στόχους.

Γουλάς Θεόδωρος

Κουτσιούκου Παναγιώτα

Νεοκλέους Χριστιάνα

"it would not be altogether wrong to describe stylistics as a 'new rhetoric'

Stephen Ullman

Σύνοψη

Η παρούσα διπλωματική εργασία ασχολείται με τον εντοπισμό της συγγραφικής πατρότητας σε κείμενα του Twitter της Νέας Ελληνικής γλώσσας. Ο αυτόματος εντοπισμός συγγραφέα στηρίζεται στην ορθή επιλογή των κατάλληλων υφολογικών μεταβλητών, δηλαδή αυτών που αρμόζουν στην συγκεκριμένη επικοινωνιακή περίσταση. Στόχος είναι η αναγνώριση της συγγραφικής πατρότητας με υφομετρικό χαρακτηριστικό τα ν- γράμματα χαρακτήρων και λέξεων. Το σώμα κειμένων που συλλέχθηκε αποτελείται από 159 συγγραφείς, εκ των οποίων αναλύθηκε ένα μέρος του συνόλου. Για την ακρίβεια επιλέχθηκαν 32 συγγραφείς με ίσο αριθμό ανδρών και γυναικών ενώ ο συνολικός αριθμός των tweets ανήλθε στα 71000. Η μεθοδολογία που ακολουθήθηκε οδήγησε σε ακριβή αποτελέσματα και έδειξε ότι το υφομετρικό χαρακτηριστικό των ν-γραμμάτων είναι επιτυχές εργαλείο στο γλωσσολογικό γένος που αντιπροσωπεύεται στο Twitter.

Abstract

This thesis deals with the authorship attribution of tweets in the Modern Greek language. Authorship attribution is based on the selection of appropriate features, in order to construct an accurate stylometric profile for each author. The aim is to identify authorship paternity using " N -gram" features. The corpus collected consists of 159 writers of whom were processed 32 with an equal number of men and women. The total number of tweets, for this subset, is around 71000. The methodology followed has led to accurate results and showed that the Author's Multilevel N-gram Profile (AMNP) approach is a successful tool for authorship attribution and gender identification in Modern Greek tweets.

Περιεχόμενα

Εισαγωγή

Οι καινούργιοι δίαυλοι επικοινωνίας που δημιουργήθηκαν με την είσοδο στην ψηφιακή εποχή και το διαδίκτυο έχουν δημιουργήσει την ανάγκη για νέους τρόπους εντοπισμού και απόδοσης της συγγραφικής πατρότητας. Το Web2 και η πληθώρα πληροφοριών που ανταλλάσσονται μεταξύ των χρηστών αποτελεί ένα χώρο που εξελίσσεται και χρήζει μελέτης από πολλούς κλάδους της επιστήμης. Πλέον, ο επιστημονικός χώρος της υπολογιστικής γλωσσολογίας δεν αξιοποιείται μόνο ως κλάδος των φιλολογικών σπουδών και της απόδοσης των έργων στους συγγραφείς αλλά συμβάλλει και στις νέες προοπτικές που δημιουργήθηκαν από την "έκρηξη" του διαδικτύου ως συλλογικής επικοινωνίας και στα νέα προβλήματα που δημιουργήθηκαν όπως είναι η εγκληματικότητα ή η λογοκλοπή.

Ως επιστημονική μέθοδος μπορεί να συμβάλει στην έγκυρη απόδοση της συγγραφικής πατρότητας, επιβεβαιώνοντας ή απορρίπτοντας αν ένα συγκεκριμένο άτομο είναι ή όχι ο συντάκτης του μηνύματος. Στην περίπτωση των κοινωνικών δικτύων και ιδιαίτερα του Twitter, η πρόκληση καθίσταται μεγαλύτερη απ' την στιγμή που τα άμεσα μηνύματα έχουν συγκεκριμένο μήκος, χαρακτηρίζονται από αποσπασματικότητα και συνδέονται με διαφορετικές επικοινωνιακές περιστάσεις.

Η μέχρι τώρα έρευνα αναφέρεται στην απόδοση της συγγραφικής πατρότητας με πληθώρα υφομετρικών χαρακτηριστικών όπως είναι το μήκος της πρότασης ή οι λειτουργικές λέξεις· χαρακτηριστικών, όμως, που προϋποθέτουν όγκους κειμένων ώστε να επιτευχθεί στατιστική σημαντικότητα. Έτσι, η παρούσα ερευνητική προσπάθεια αξιοποιεί τις γνώσεις από τις "παραδοσιακές" μελέτες και εντάσσει τις σύγχρονες μεθόδους και τα εργαλεία ώστε να αποδοθεί με ακρίβεια η πατρότητα των άμεσων κειμένων του Twitter.

Το Twitter αποτελεί ένα κοινωνικό δίκτυο βασισμένο στο διαδίκτυο που επιτρέπει στους ανθρώπους να κάνουν σύντομες δηλώσεις (*posts*). Σε αντίθεση με άλλα κοινωνικά δίκτυα (π.χ. όπως το Facebook) περιορίζει τους χρήστες να εκφράσουν τις σκέψεις τους μέσα σε 140 χαρακτήρες ανά σχόλιο. Το Twitter χρησιμοποιείται κυρίως ως ένα εργαλείο μονόπλευρης (*oneway*) πληροφόρησης (Webster, 2010), γεγονός που οφείλεται στην απλή μορφή του, ευκολύνοντας την μετάδοση της

πληροφορίας παρά την ανταλλαγή και τη δημιουργία συζητήσεων. Συγκριτικά, ο χρήστης στο Twitter, περιορίζεται στην έκθεση των απόψεων του, ενώ ταυτόχρονα μειώνεται η αλληλεπίδραση όσον αφορά τον τρόπο που τις εκθέτει. Αυτό σημαίνει ότι το ύφος του χρήστη παρέχει περισσότερες πληροφορίες για τον ίδιο τον χρήστη και συνεπώς μπορεί να οδηγήσει στην ταυτοποίηση του.

Η παρούσα έρευνα αξιοποιώντας τεχνικές μηχανικής μάθησης, τον τεμαχισμό των tweets σε πακέτα λέξεων (20 έως 200) και την χρήση των ν-γραμμάτων εξάγει ακριβή αποτελέσματα για την συγγραφική πατρότητα. Το σώμα κειμένων που δημιουργήθηκε αποτελεί το μεγαλύτερο - σε όγκο - σώμα για tweets της νέας ελληνικής γλώσσας (ανάλυση για 32 χρήστες).

1. Η υφομετρική ανάλυση και η συμβολή της στον αυτόματο εντοπισμό της συγγραφικής πατρότητας

Εξέλιξη της έννοιας του ύφους ως γλωσσολογικής παραμέτρου

Η μελέτη του ύφους σχετιζόταν με την "αισθητική λειτουργία" του κειμένου και την ανάγκη εξεύρεσης ενός λογοτεχνικού πρότυπου ύφους. Το κείμενο δεν εξετάζεται για την λειτουργικότητά του αλλά για την αξία του, η μελέτη του ύφους δεν στοχεύει στην μορφή αλλά στην ουσία του λόγου. Αναζητούνται, λοιπόν, τα μέσα, οι εκφραστικές δομές που δημιουργούν το αισθητικό αποτέλεσμα. Οι απαρχές της υφολογικής μελέτης προς αυτό τον προσανατολισμό στηρίζονται στην *Περί Ύψους* πραγματεία του Λογγίνου όπου καθορίζονται οι κανόνες, τους οποίους έπρεπε να ακολουθήσει ο κάθε επίδοξος ρήτορας για να επιτύχει το *υψηλό ύφος* (Διονυσίου Λογγίνου, ερμ. έκδ. Μ.Ζ. Κοπιδάκης, 1990)

Με την έλευση του ρομαντισμού επικρατεί η αντίληψη ότι το ύφος είναι το σήμα κατατεθέν του ατόμου, της ομάδας ή του λογοτεχνικού είδους. Αυτές οι παραδοσιακές καταβολές που εντάσσουν την υφολογική μελέτη στην αποκλειστική μελέτη του λογοτεχνικού έργου ανετράπησαν στις αρχές του 20ου αιώνα με την δομική κριτική. Σύμφωνα με τους δομιστές η προσοχή προς τα υλικά οδηγεί σε μετρικο-φωνολογικές αναλύσεις οι οποίες μεταφέρουν νόρμες και στηρίζονται πάνω σε κανονικότητες (Todorov, 1965).

Οι δομιστές μετατοπίζουν το ερευνητικό ενδιαφέρον τους από τις φιλοσοφικές διαστάσεις των έργων στη γλωσσολογική μελέτη. Οι δυνατότητες της Υπολογιστικής Γλωσσολογίας καθώς και η αξιοποίηση των στατιστικών μεθόδων καθιστά δυνατή την απόδειξη της συγγραφικής πατρότητας των κειμένων. Το ζητούμενο στην στατιστική ανάλυση των κειμένων είναι η εφαρμογή μιας αντικειμενικής μεθοδολογίας, η οποία θα αντικαταστήσει τις φιλολογικές τοποθετήσεις που βασίζονταν σε υποκειμενικές θέσεις. Αντιθέτως, οι στατιστικές ποσοτικές μέθοδοι παρέχουν αντικειμενικά στοιχεία (Dabagh, 2007, σελ.149).

Η εκμετάλλευση των ηλεκτρονικών υπολογιστών και η διαθεσιμότητα μεγάλων όγκων κειμένων σε ηλεκτρονική μορφή έχουν επιλύσει πολλά προβλήματα συγγραφικής πατρότητας. Πραγματική άνθηση γνώρισε ο κλάδος στα τέλη του 20ου

αιώνα με τις εργασίες του George Yule και του Gustav Herdan στην Αγγλία, του Wiilhelm Fucks στην Γερμανία, των Federick Mosteller και David L. Wallace στις Η.Π.Α. και του Charles Muller στη Γαλλία.

Η σαφής διάκριση και απεξάρτηση του λόγου από τα κοινωνικά συμφραζόμενα υπήρξε αναγκαία ώστε να αναχθεί η γλωσσολογία σε επιστήμη μελέτης της γλώσσας. Αυτό έγινε εφικτό με την τομή στις αντιλήψεις περί γλώσσας, με τη δυική διάκριση του Saussure (1974). Ωστόσο, η γλώσσα ως σύστημα επικοινωνίας εμβαθύνει στον κοινωνικό ρόλο που επιτελεί με τις έρευνες στους κλάδους της κοινωνιογλωσσολογίας και της πραγματολογίας και έτσι μελετάται και η κοινωνική της μορφή όπως αποτυπώνεται στις δομές της κοινότητας που αντιπροσωπεύει.

Πλέον, η γλωσσολογική επιστήμη χρησιμοποιεί τις έννοιες του ύφους και των υφολογικών μετρήσεων ως μέθοδο για να εξαγάγει επιστημονικά ελέγξιμα συμπεράσματα. Ήδη ο Jakobson αναφέρει ότι η ίδια η γλωσσική ποικιλία μπορεί να ξεπεράσει την ομοιομορφία και την στατικότητα του σωσσυριανού μοντέλου (Jakobson, 1983, σελ.51), αντικαθιστώντας το με τη δυναμική άποψη ενός πολύπλευρου κώδικα που να καθρεφτίζει τις ανθρώπινες λειτουργίες της γλώσσας καθώς και τους χρονικούς και χωρικούς συντελεστές της. Το επικοινωνιακό μοντέλο του Jakobson περιλαμβάνει έξι παραμέτρους:

1. Τον αποστολέα ή πομπό,
2. Τον παραλήπτη ή δέκτη,
3. Το γλωσσικό μήνυμα,
4. Τον επικοινωνιακό κώδικα,
5. Το πλαίσιο αναφοράς,
6. Το κανάλι.

Καθένας από τους παράγοντες αυτούς ορίζει μια επικοινωνιακή λειτουργία. Ειδικότερα, κάθε παράγοντας της γλωσσικής επικοινωνίας ορίζει μία λειτουργία της γλώσσας, η οποία είναι παρούσα σε κάθε γλωσσικό γεγονός αντιστοιχεί σε μια υφομετρική επιλογή, σε ένα υφολογικό πεδίο.

Έτσι η *αναφορική* εστιάζει στο σημασιολογικό πεδίο, η *συγκινησιακή* στο ψυχολογικό, η *βουλητική* στο ρητορικό, η *φατική* στο κοινωνιολογικό, η *μεταγλωσσική* στο γλωσσολογικό και η *ποιητική* στο λογοτεχνικό. Αυτά είναι τα έξι βασικά πεδία στη μελέτη του ύφους, που ο De Vito ορίζει ως την επιλογή και την τοποθέτηση εκείνων των γλωσσολογικών χαρακτηριστικών, τα οποία είναι ανοικτά προς επιλογή (*are open to choice*) και επεξεργασία (De Vito, 1967, σελ.3). Είναι εμφανές ότι δίνεται προτεραιότητα στην μεθοδολογία και στους τρόπους εξεύρεσης των "ανοικτών" και ποσοτικά μετρήσιμων χαρακτηριστικών.

Το ύφος λοιπόν, αποτυπώνει, την γλωσσική επιλογή (*selection*) από τον παραδειγματικό άξονα, την διάταξη του λόγου (*arrangement*) και τα γλωσσολογικά χαρακτηριστικά (*linguistic features*) (De Vito, 1967, σελ.9). Η απόδοση της συγγραφικής πατρότητας με βάση την στατιστική ανάλυση των ιδιαίτερων χαρακτηριστικών (*features*) που χρησιμοποιεί ο συγγραφέας αποτελεί τον κλάδο της υφομετρίας που έχει εξ' αρχής δεδομένη την ιδιαίτερη προσωπική γραφή. Έτσι, προϋποθέτει ότι ο τρόπος γραφής του συγγραφέα είναι μοναδικός, ποσοτικά μετρήσιμος, σταθερός και επαναλαμβανόμενος, και μπορεί να ανιχνευτεί σε όλα τα κείμενα του (Holmes, 1985, σελ.328-341).

Κάθε δημιουργία κειμένου αποκαλύπτει και ενσωματώνει στοιχεία της προσωπικής και κοινωνικοπολιτισμικής μας ταυτότητας (Γεωργακοπούλου & Γούτσος, 2008[7], σελ.191). Ο λόγος που αναπτύσσει το άτομο αντανακλά τις προσωπικές του πεποιθήσεις και αποτελεί στοιχείο της ηλικίας, του φύλου και της κοινωνικής του θέσης. Ιδιαίτερα στο δεύτερο μισό του 20^{ου} αιώνα οι μελετητές εισήγαγαν στις υφολογικές μελέτες την επιστήμη της στατιστικής, ώστε να καταστεί δυνατή η ποσοτική ανάλυση των γλωσσικών δομών και ακολούθως η ερμηνευτική τους επεξεργασία.

Ο Diller διαφοροποιεί το κειμενικό (*textual*) από το γλωσσολογικό (*linguistic*) ύφος Αναφερόμενος στο *κειμενικό* εννοεί το σύνολο των γλωσσολογικών χαρακτηριστικών που διέπουν ένα κείμενο ή ένα σώμα κειμένων, ενώ θεωρεί οτι το *γλωσσολογικό* είναι αυτό που κατά την επικοινωνιακή χρήση διαφέρει από άνθρωπο σε άνθρωπο. Ο Diller ονομάζει το δεύτερο *γλωσσολογικό ορισμό* γιατί

τονίζει την αξία της γνώσης (*intuitive knowledge*) της υφολογικής ποικιλίας (*stylistic variation*)που διαθέτει μια γλώσσα (Diller, 1998, σελ.155-156).

Η μελέτη του ύφους θεωρεί σαφείς δύο βασικές συνιστώσες, την ιδιαίτερη προσωπική γραφή και ως εκ τούτου την διαφοροποίηση των εκφραστικών μέσων ανάμεσα και σε ομιλούντες της ίδιας γλωσσικής ποικιλίας. Το στίγμα του συγγραφέα είναι διακριτό από τις επιλογές που κάνει. Το ύφος νοείται ως η διαφορετική κωδικοποίηση του ίδιου γλωσσικού περιεχομένου εντός του ίδιου γλωσσικού συστήματος.

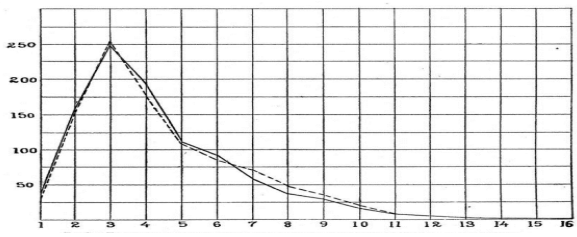
Οι πρώτες υφολογικές μελέτες με στατιστικές παραμέτρους

Η ιστορία της σύγχρονης στατιστικής υφολογίας ανάγεται πίσω στο 1851 όταν ο Augustus de Morgan, Βρετανός μαθηματικός, πρότεινε ότι το μέσο μήκος των λέξεων ενός συγγραφέα μπορεί να αποδειχθεί χαρακτηριστικό γνώρισμα του ύφους του συγγραφέα, το οποίο μάλιστα αποτελεί και διαφοροποιητικό χαρακτηριστικό από άλλον συγγραφέα. Στις επιστολές του, που διασώθηκαν χάριν στην γυναίκα του αναφέρει χαρακτηριστικά:

"Θα περίμενα να βρω ότι τα κείμενα του ίδιου συγγραφέα σε δύο διαφορετικά θέματα συμφωνούν πολύ περισσότερο παρά τα κείμενα δύο διαφορετικών συγγραφέων πάνω στο ίδιο θέμα"

[Πηγή : de Morgan(1851/1882, σελ.216)]

Την υπόθεσή αυτή, ανέλαβε να διερευνήσει ο Thomas Mendenhall (1841-1924), ο οποίος μελέτησε το μήκος των λέξεων ως υφομετρικό χαρακτηριστικό του συγγραφέα. Στο άρθρο του *The Characteristic Curve of Composition* (Mendenhall, 1887), εξέτασε τρεις διαφορετικές υποθέσεις. Πρώτα, συνέκρινε την κατανομή του μήκους των λέξεων στις πρώτες 10.000 λέξεις από τον *Oliver Twist* του Dickens και του *Vanity Fair* του Thackeray και παρατήρησε ότι τα φάσματά τους παρουσίαζαν ομοιότητες.



Εικόνα 1: Δύο ομάδες από 10.000 λέξεις η κάθε μια από τον Oliver Twist, "___" και από το Vanity Fair, "---" [Πηγή: Mendenhall(1887, p. 243)]

Ακολούθως, συνέκρινε δύο εργασίες διαφορετικής χρονικής περιόδου, από τον ίδιο συγγραφέα (John Stuart Mill) με ζητούμενο να εξετάσει τη συνέχεια και την κανονικότητα της κατανομής. Τα δύο φάσματα δεν παρουσίασαν την ομοιότητα που αναμενόταν στη συνολική τους εικόνα, αν και είχαν επιμέρους συγκλίσεις.

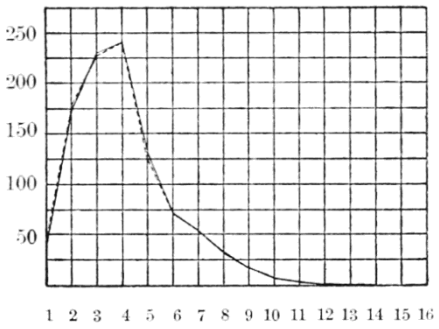
Το τρίτο πείραμα εξέταζε την επίδραση του διαφορετικού ακροατηρίου στον τρόπο με τον οποίο δομήθηκαν οι ομιλίες του Atkinson σε εργατικές ενώσεις και σε απόφοιτους θεολογικής σχολής. Στην ομιλία του στις εργατικές ενώσεις ο Atkinson έκανε ιδιαίτερη προσπάθεια να χρησιμοποιεί απλές, σύντομες λέξεις και φράσεις με απλούστερη και πιο σαφή διατύπωση (Mendenhall, 1901, σελ.99). Συμπερασματικά ο Mendenhall παρατήρησε ότι το λεξικό φάσμα του Atkinson διαφέρει από τους υπολοίπους που εξέτασε, γεγονός που οφείλεται στην αξιοσημείωτη χρήση μικρότερων σε μήκος λέξεων.

Atkinson	4.298
Dickens	4.342
Thackeray	4.481
Mill	4.775

Πίνακας 1: : Το μέσο μήκος λέξεων των 4 συγγραφέων σε μέτρηση 10.000 λέξεων, [Πηγή: Mendenhall (1887, σελ.243)]

Στην δεύτερη δημοσίευση του (Mendehall,1901) συνέκρινε το μέσο μήκος των λέξεων του Shakespeare και του Bacon και ανέλυσε τα φάσματά τους καταλήγοντας στο συμπέρασμα ότι το μέσο μήκος λέξεων του Shakespeare ήταν 4 λέξεις ενώ του

Bacon 3 λέξεις. Στο ίδιο πείραμα συμπεριέλαβε και κείμενα του Ben Jonson, του Addison, του Milton, του Beaumont and του Fletcher, του Goldsmith του Lord Lytton και του Christopher Marlowe. Μάλιστα, ανακαλύφθηκε ότι η καμπύλη από τα έργα του Marlowe συμφωνεί με τον Shakespeare περίπου, όσο και ο ίδιος ο Shakespeare συμφωνεί με τον εαυτό του (Mendehall, 1901, σελ.105).



Εικόνα 2: : Η χαρακτηριστική καμπύλη του Shakespeare, "_____", και του Marlowe, "....."
[Πηγή: Mendenhall (1901, p. 105)]

Η προσέγγιση του Mendenhall, παρά την απλουστευμένη λογική της, εδραίωσε την ανάγκη για μετρήσιμες τεχνικές. Βοήθησε ώστε τα υφομετρικά χαρακτηριστικά του συγγραφέα να λαμβάνονται ως μετρήσιμα, ποσοτικά χαρακτηριστικά με τα οποία μπορεί κανείς να επιλύσει ζητήματα συγγραφικής πατρότητας, να τα αξιοποιήσει στην συγκριτική μελέτη της γλώσσας καθώς και στην κατανόηση της ανάπτυξης του ανθρώπινου λεξιλογίου. Όπως αναφέρει χαρακτηριστικά: "Οι προσωπικές ιδιαιτερότητες στην κατασκευή των προτάσεων, κατά τη χρήση μεγάλων ή μικρών λέξεων μπορούν μακροπρόθεσμα να εκδηλώνονται με τέτοια κανονικότητα ώστε η γραφική παράσταση τους να γίνει ένα μέσο αναγνώρισης" (Mendenhall,1901,σελ.98).

Ο Williams (1975) ασκώντας κριτική στη μεθοδολογία του Mendenhall, τονίζει την ανάγκη για σύγκριση κειμένων που ανήκουν στο ίδιο κειμενικό γένος. Η σύγκριση μεταξύ πεζογραφίας του Bacon και του στίχου του Shakespeare σαφώς και θα παρουσίαζαν διαφορετικές κατανομές. Οι διαφορές που βρήκε ο Mendenhall μπορούν κάλλιστα να εξηγηθούν από την διαφορά της "λογοτεχνικής παρουσίας" (Williams, 1975, σελ. 207). Ήταν πρόδηλο γι' αυτόν ότι στον ποιητικό λόγο θα ήταν πιο δύσκολο να χρησιμοποιηθούν μεγάλες λέξεις εξαιτίας του ρυθμού, ο οποίος

απουσιάζει από το πεζό κείμενο. Κατέδειξε, λοιπόν, συγκρίνοντας τα έργα του Philip Sidney -πεζά και ποιητικά- ότι λιγότερο μοιάζουν μεταξύ τους παρά με έργα σύγχρονών του, του ίδιου λογοτεχνικού γένους.

	Letters per word											
1	2	3	4	5	6	7	8	9	10	11	12	13+
					Bacon: prose							
2·6%	20·0	22·7	17·5	10·3	7·6	7·2	4·8	3·3	2·4	1·3	0·2	0·1
					Shakespeare: verse							
4·8%	17·6	22·5	23·8	12·4	7·1	5·3	3·2	1·8	0·9	0·3	0·2	0·14
					Sidney: prose							
2·4%	17·8	21·0	21·1	12·5	7·9	6·9	4·2	2·4	1·9	1·2	0·45	0·32
					Sidney: verse							
3·4%	15·5	19·3	25·0	16·7	7·8	5·5	4·4	1·03	1·03	0·19	0·19	0·13

Πίνακας 2: Σύγκριση του μήκους των λέξεων στους Bacon (prose), Shakespeare (verse) και Sidney (prose & verse), [Πηγή: Williams (1975, p. 211)]

Η προσφορά του Williams έγκειται στην αναγνώριση και στην εξέταση παραγόντων και κριτηρίων, τα οποία πρέπει να διέπουν μια υφομετρική ανάλυση, ώστε να οδηγούν σε αξιόπιστα αποτελέσματα. Σε κάθε περίπτωση η χρήση του μήκους των λέξεων (*word-length distribution*) αποτελεί ένα παραδοσιακό υφολογικό χαρακτηριστικό, το οποίο σε συνδυασμό με νέα χαρακτηριστικά συνεισφέρει στην εξέταση της πατρότητας των κειμένων.

Πέραν από το μήκος της λέξης, εξετάζεται και η καταμέτρηση και άλλων υφολογικών χαρακτηριστικών από το συντακτικό επίπεδο. Η χρήση του μήκους της πρόταση ως υφολογικού χαρακτηριστικού του συγγραφέα (*author/compositor*) εξετάστηκε από τον Udny Yule (1938/1939). Μπροστά στο ερευνητικό αυτό ερώτημα, ο Yule προέβη σε δύο βασικές διαπιστώσεις:

-ο ορισμός και ο καθορισμός της πρότασης και της λέξης καθίσταται δύσκολος απ' την στιγμή που χρησιμοποιείται για στατιστικούς και μετρήσιμους σκοπούς. Ερωτήματα όπως: "πώς μπορεί να μετρήσει κανείς τις προτάσεις;" "από τελεία σε τελεία;" και "πώς μετράμε λέξεις όπως china-manufactured, χρονολογίες ή αποσπάσματα;", απασχολούν τον Yule. Για το συγκεκριμένο πείραμα θεωρεί ότι ο αριθμός των λέξεων ανάμεσα σε διαδοχικές τελείες είναι πρόταση και οποιαδήποτε

ακολουθία γραμμάτων από το Α ως το Ζ, που χρησιμοποιείται με ακριβές νόημα είναι λέξη (Yule, 1938/1939, σελ.363).

-οι μεγάλες και συντακτικά περίπλοκες προτάσεις φέρουν σε μεγάλο βαθμό το στοιχείο της προσωπικής κρίσης (Yule,1938/1939, σελ.364). Η θεματική του συγγραφέα καθώς και η προσωπικότητά του επιδρούν τόσο στο υλικό που έχει να διαχειριστεί όσο και στο τρόπο που θα το διαχειριστεί. Τα αποσπάσματα όπου ο συγγραφέας επιχειρηματολογεί τείνουν να είναι μακρύτερα παρά σ' αυτά που είναι καθαρά περιγραφικά (Yule, 1938/1939, σελ.367).

Επέλεξε τα έργα *Essays* του Bacon , *Biographia Literaria* του Coleidge, *Elia* και *Last Essays of Elia* του Lamb και *Essays* του Macaulay, και αφού υπολόγισε το μήκος των προτάσεων κατέληξε στο συμπέρασμα ότι το προτασιακό μέγεθος είναι χαρακτηριστικό του συγγραφέα (Yule, 1938/1939, σελ.370). Επιπρόσθετα, στην έρευνα του ασχολήθηκε και με την πατρότητα του *De Imitatione Christi*, έργο του 15^{ου} αιώνα. Ανάμεσα σε δύο υποψήφιους συγγραφείς, του Thomas à Kempis και του Jean Charlier de Gerson κατέληξε στον συμπέρασμα ότι τα γνωστά κείμενα του à Kempis παρουσίαζαν ομοιότητες με το *De Imitatione Christi* και διέφεραν σε σύγκριση με γνωστά κείμενα του de Gerson.

Ο Yule αναγνωρίζει την δυσκολία στον ορισμό της πρότασης αλλά θεωρεί ότι αξίζει να αξιοποιηθεί το προτασιακό μέγεθος ως υφολογικό χαρακτηριστικό:

> "Δεδομένου ότι το μήκος μιας πρότασης μπορεί να μετρηθεί εύκολα, για πρακτικούς καθαρά σκοπούς, από τον αριθμό των λέξεων, μου φαίνεται ότι θα ήταν ενδιαφέρον να υποβληθεί αυτή η υπόθεση στην στατιστική έρευνα"
>
> *[Πηγή:Yule G.U (1938/1939, σελ. 363)]*

Στις Η.Π.Α. η έρευνα στρέφει το ενδιαφέρον της στα Ομοσπονδιακά Κείμενα τα οποία εκδόθηκαν ανωνύμως (με ψευδώνυμο Publius) σε εφημερίδες με σκοπό να ωθήσουν τους Αμερικανούς πολίτες στην επικύρωση του Αμερικανικού Συντάγματος (1787-1788). Τα κείμενα γράφτηκαν από τους Alexander Hamilton, John Jay και James Madison. Η ανάγκη για την πατρότητα των κειμένων υπήρξε έντονη όταν τα τρία αυτά πρόσωπα αναδείχθηκαν σε ηγετικές πολιτικές μορφές και

η πολιτική διαδρομή τους δεν συμβάδιζε πλήρως με τις προκηρύξεις των *Federalist Papers*.

Το ερευνητικό ερώτημα των Mosteller και Wallace (1963) για την συγγραφική πατρότητα στηρίζεται πάνω στις συχνότητες των λέξεων και το προσωπικό συγγραφικό ύφος ανάμεσα στους τρεις πιθανούς συγγραφείς. Όσον αφορά στα 12 αμφισβητούμενα κείμενα (*disputed papers*), οι Mosteller & Wallace βασίστηκαν στις λειτουργικές λέξεις (*functional words*).

1	*a*	15	*do*	29	*is*	43	*or*	57	*this*
2	*all*	16	*down*	30	*it*	44	*our*	58	*to*
3	*also*	17	*even*	31	*its*	45	*shall*	59	*up*
4	*an*	18	*every*	32	*may*	46	*should*	60	*upon*
5	*and*	19	*for*	33	*more*	47	*so*	61	*was*
6	*any*	20	*from*	34	*must*	48	*some*	62	*were*
7	*are*	21	*had*	35	*my*	49	*such*	63	*what*
8	*as*	22	*has*	36	*no*	50	*than*	64	*when*
9	*at*	23	*have*	37	*not*	51	*that*	65	*which*
10	*be*	24	*her*	38	*now*	52	*the*	66	*who*
11	*been*	25	*his*	39	*of*	53	*their*	67	*will*
12	*but*	26	*if*	40	*on*	54	*then*	68	*with*
13	*by*	27	*in*	41	*one*	55	*there*	69	*would*
14	*can*	28	*into*	42	*only*	56	*things*	70	*your*

Table I. Function Words and Their Code Numbers

Πίνακας 3: Οι λειτουργικές λέξεις και κωδικός αριθμός τους για την έρευνα των Federalist Papers, [Πηγή: Mosteller & Wallace (1963, p.280)]

Οι λειτουργικές λέξεις της γλώσσας όπως το *as*, *an*, *of*, *upon* και γενικότερα, τα άρθρα, οι προθέσεις και οι σύνδεσμοι παρουσιάζουν αρκετά σταθερές τιμές, ενώ οι λέξεις με νόημα όπως war, executive και legislature δεν παρουσιάζουν την ίδια σταθερότητα στη συχνότητα εμφάνισης (Mosteller & Wallace, 1963, σελ.275). Έγινε κατανοητή η ανάγκη για υφομετρικές λεξιλογικές μεταβλητές πέραν των παραδοσιακών υφομετρικών μεταβλητών, όπως το μέσο μήκος πρότασης.

Στις παρατηρήσεις τους (Mosteller & Wallace, 1963, σελ.306) τονίζουν ότι οι λειτουργικές λέξεις παρουσιάζονται ως διαφοροποιόν στοιχείο για τη διάκριση ανάμεσα στους συγγραφείς, ενώ βασικό πρόβλημα αποτελεί η συμφραστικότητα κυρίως στα βοηθητικά ρήματα (π.χ.*do*) και στις αντωνυμίες (π.χ. *him*) που εξαρτώνται από την συντακτική δομή της πρότασης. Οι συγκεκριμένες λέξεις έχουν περιορισμένη λεξική σημασία και χρησιμοποιούνται για να εκφράσουν γραμματικές

σχέσεις με άλλες λέξεις μέσα σε μια πρόταση (Μικρός, 2012, σελ. 81). Πράγματι, οι λειτουργικές λέξεις αποτελούν προσιτά στοιχεία για μια αντικειμενική ανάλυση αφού επιλέγονται μη συνειδητά από τον άνθρωπο.

Στηριζόμενοι ακολούθως στο θεώρημα του Bayes, κατάφεραν αφού προσδιόρισαν από τη μια τις συχνότητες των λειτουργικών λέξεων των γνωστών κειμένων και από την άλλη την αναμενόμενη θεωρητική συχνότητα στα κείμενα αγνώστου πατρότητας απέδωσαν τα 12 κείμενα στον Madison με επιφυλάξεις για το κείμενο με αύξοντα αριθμό 55.

Η προσφορά των εργασιών του Mendenhall, του Yule και των Mosteller και Wallace υπήρξε σημαντική και θεμελίωσε τις βάσεις για την υφομετρική ανάλυση: το μήκος της λέξης εντάσσεται στο γραφηματικό επίπεδο, το μήκος της πρότασης στο συντακτικό επίπεδο και οι λειτουργικές λέξεις στο λεξιλογικό επίπεδο. Πλέον αυτά τα υφομετρικά χαρακτηριστικά ανήκουν στα βασικά "παραδοσιακά" χαρακτηριστικά, τα οποία σε συνδυασμό με άλλα χαρακτηριστικά προσδιορίζουν το συγγραφέα ενός κειμένου.

Ο Bailey (1979) υποδηλώνει ότι τα χαρακτηριστικά θα πρέπει να είναι εξέχοντα, με διαρθρωτική δομή, να εμφανίζονται με συχνότητα, ευκολοδιάκριτα, ποσοτικά μετρήσιμα και σχετικά ανεπηρέαστα από συνειδητό έλεγχο (Bailey, 1979, σελ.10). Η αποτελεσματικότητα των χαρακτηριστικών εξαρτάται από την ακριβή καταμέτρησή τους. Έτσι μπορούν να παράγουν τα ουσιώδη χαρακτηριστικά, στα οποία θα στηρίζεται η διάκριση των συγγραφέων και θα επικυρώνουν την ταυτότητα του δημιουργού τους.

2. Η εγκληματικότητα όπως αυτή ασκείται μέσα από το Twitter

Τον 21^{ο} αιώνα, τον οποίο διανύουμε, τον αιώνα της πληροφορίας και της επικοινωνίας, μια άλλη μορφή βίας έχει κάνει την αποφασιστική εμφάνισή της , η οποία αυξάνεται και εξαπλώνεται με ραγδαίους ρυθμούς. Η εγκληματικότητα μέσω των μέσων κοινωνικής επικοινωνίας και δικτύωσης είναι πλέον πραγματικότητα. Πρόκειται για τη βία που ασκείται μέσα από τα μέσα κοινωνικής δικτύωσης όπως το Facebook, το Twitter, το IRC chat και άλλα room chats, τα e-mails, διάφορα φόρουμ και άλλες ιστοσελίδες.

Η αύξηση της διαδικτυακής εγκληματικότητας (*cybercrime*), στηρίζεται στην ανωνυμία πίσω από την οποία καλύπτεται ο κάθε χρήστης. Η ανωνυμία αυτή του εξασφαλίζει την ατιμωρησία καθώς είναι δύσκολο όχι μόνο να ανακαλυφθεί ο χρήστης ο οποίος στέλνει τα μηνύματα αλλά κυρίως να ταυτοποιηθεί ότι είναι αυτός ο ίδιος και όχι κάποιος άλλος που έχει υποκλέψει τα στοιχεία του, και γενικότερα την ηλεκτρονική του ταυτότητα, και προσπαθεί να τον ενοχοποιήσει.

Ωστόσο, η ταυτοποίηση του χρήστη με το αποσταλθέν μήνυμα έχει σημειώσει πολύ ενθαρρυντικά βήματα χάρη στις μελέτες που έχουν γίνει τα τελευταία χρόνια στο πεδίο της αναγνώρισης της συγγραφικής πατρότητας. Πλέον, δε μελετούνται μόνο οι λέξεις και οι φράσεις αλλά και κάποια χαρακτηριστικά γνωρίσματα όπως ο τονισμός, τα σημεία στίξης, οι συντομογραφίες, τα εικονίδια που εκφράζουν συναισθήματα, τα γνωστά ως emoticons, και κάποια άλλα χαρακτηριστικά σημάδια γραφής τα οποία συμβάλλουν στη καλύτερη και ακριβέστερη περιγραφή του συγγραφικού στυλ του κάθε συγγραφέα. Επιπλέον, ένα πρόσθετο τροχοπέδη, το οποίο συμβάλλει στην χωρίς όρια διάδοση της διαδικτυακής εγκληματικότητας είναι ο μικρός αριθμός των χαρακτήρων από τους οποίους αποτελείται ένα μήνυμα.

Συγκεκριμένα, για παράδειγμα, στο Twitter, ο μέγιστος αριθμός χαρακτήρων περιορίζεται στους 140. Ωστόσο, τα περισσότερα μηνύματα είναι δυνατό να περιορίζονται στους 10 με 15 χαρακτήρες ή ακόμα και λιγότερους. Σαν αποτέλεσμα, οι συνήθεις μέθοδοι αναγνώρισης της συγγραφικής πατρότητας όπως η επανάληψη λέξεων ή οι στερεότυπες και επαναλαμβανόμενες συντακτικές δομές, οι οποίες

εφαρμόζονται σε κείμενα με μεγαλύτερο γλωσσικό εύρος και με περισσότερες λέξεις όπως 250-500 λέξεις πέφτουν στο κενό και θεωρούνται αναξιόπιστες σε κείμενα με μικρότερο αριθμό λέξεων και χαρακτήρων.

Χαρακτηριστική περίπτωση μη ταυτοποίησης του μηνύματος με το συγγραφέα λόγω της ανωνυμίας ή της ψεύτικης δήλωσης στοιχείων είναι η περίπτωση μιας δολοφονίας η οποία δημοσιεύτηκε τον Ιανουάριο του 2010 στη New York Daily News (Silva et al., 2011, σελ.161). Έτσι, είναι προφανές ότι το πρόβλημα που ανακύπτει είναι η ταυτοποίηση του αληθινού συγγραφέα ενός μηνύματος με το εκάστοτε μήνυμα ανάμεσα σε ένα πολύ ευρύ φάσμα υποψηφίων συγγραφέων.

Σύμφωνα με στοιχεία των αστυνομικών αρχών, τα κρούσματα της διαδικτυακής εγκληματικότητας που σχετίζονται με τα μέσα κοινωνικής δικτύωσης, και κυρίως το Facebook και το Twitter φαίνεται ότι έχουν μέσα σε τέσσερα χρόνια οκταπλασιαστεί. Το αισιόδοξο μήνυμα εκ μέρους της αστυνομίας είναι ότι στις περισσότερες περιπτώσεις οι διαδικτυακές κοινότητες μπορούν να προστατεύουν τα μέλη τους χάρη στην αυτάρκεια που τις χαρακτηρίζει, στον περιορισμένο αριθμό μελών τους καθώς και στην εξακρίβωση των στοιχείων των μελών κατά την εγγραφή τους στην κοινότητα.

Τα πεδία στα οποία ειδικεύεται η διαδικτυακή εγκληματικότητα είναι τα απειλητικά μηνύματα που μπορούν να καταλήξουν σε πραγματικά βίαιες επιθέσεις, οι σεξουαλικές παρενοχλήσεις και παρακολουθήσεις, η διακίνηση πορνογραφικού υλικού, οι ρατσιστικές επιθέσεις όπως και οι απάτες κυρίως για υπεξαίρεση χρηματικών ποσών (Layton, Watters & Dazeley, 2010, σελ.1).Το διαδίκτυο προσφέρει πρόδηλη ανωνυμία καθώς και την ευκολία με την οποία μπορεί κανείς να δημιουργήσει πολλαπλές ταυτότητες ψεύτικες ή μη.

Συμπερασματικά, είναι κατανοητό ότι η απόδοση συγγραφικής πατρότητας μέσω του Twitter, το οποίο αποτελεί και το αντικείμενο της έρευνάς μας, είναι μια αναδυόμενη ερευνητική προσπάθεια για την οποία γίνονται τα τελευταία χρόνια αρκετές έρευνες καταλήγοντας σε πολύ αισιόδοξα αποτελέσματα. Ωστόσο, παρά τις επίμονες και επίπονες προσπάθειες των ερευνητών, η ακριβής απόδοση της συγγραφικής πατρότητας απέχει πολύ ακόμα από το να θεωρείται δεδομένο εργαλείο στα χέρια των διωκτικών αρχών.

3. Σύγχρονες έρευνες υφομετρικής ανάλυσης και συγγραφικής πατρότητας

Εφαρμογή του Source Code Author Profiles (SCAP)

Όπως προαναφέρθηκε σε προηγούμενο κεφάλαιο η εγκληματικότητα στο διαδίκτυο με οποιαδήποτε μορφή, στηρίζεται στην ανωνυμία ή τη δήλωση ψευδών στοιχείων στο Twitter. Το έργο των διωκτικών αρχών με σκοπό να εντοπίσουν τη συγγραφική πατρότητα (*Authorship Attribution*) των σχολίων φάνταζε πολύ δύσκολο. Ωστόσο, τα τελευταία χρόνια έχουν γίνει πολλές έρευνες για να ταυτοποιηθεί η συγγραφική πατρότητα στο Twitter.

Η σύγχρονη υπολογιστική έρευνα και η επεξεργασία φυσικής γλώσσας οδήγησε στην δημιουργία κώδικα με στόχο τον εντοπισμό της συγγραφικής πατρότητας και την ακριβή και μετρήσιμη παρουσίαση αποτελεσμάτων. Η προσέγγιση SCAP (Source Code Author Profiles) έχει εφαρμοσθεί επιτυχώς (Peng et al, 2004). Ακολούθως, η έρευνα των Frantzeskou, Stamatatou, Gritzali, Chaski και Howald (2007) διαφοροποιήθηκε χρησιμοποιώντας τη γλώσσα προγραμματισμού C++ και Java και βρίσκοντας τις συχνότητες των ν-γραμμάτων σε επίπεδο χαρακτήρων (περιλαμβανομένων των κενών, της αλλαγής γραμμής) και σχηματίζοντας το προφίλ 6 συγγραφέων. Το προφίλ έχει διαφορετικά μήκη ανάλογα με το μήκος των δεδομένων προγραμματισμού και το μήκος των ν-γραμμάτων (Frantzeskou et al., 2007, σελ. 6). Το συμπέρασμα από την έρευνα έδειξε ότι η εξόρυξη των ν-γραμμάτων και η κατάταξη είναι πλήρως αυτοματοποιημένη διαδικασία, αλλά η επιλογή του μεγέθους των ν-γραμμάτων και το μήκος του προφίλ, δεν είναι και, ως εκ τούτου επιδέχεται παρεμβάσεις (Frantzeskou et al., 2007).

	bigram	trigram	4-gram	5-gram	6-gram	7-gram
author 1	250	400	650	890	1000	1300
author 2	475	680	980	1200	1700	1982
test document	100	223	447	589	793	874

Πίνακας 4: Παράδειγμα διαφορετικού μήκους ν-γραμμάτων ανά συγγραφέα [Πηγή: Frantzeskou et al. (2007, p. 6)]

Συγγραφική πατρότητα στα άμεσα κείμενα του Twitter

Το 2010 οι Layton, Watters και Dazeley (Layton et al., 2010) χρησιμοποιώντας το SCAP στόχευσαν την έρευνα τους στην εξακρίβωση της συγγραφικής πατρότητας σε γνήσια κείμενα του Twitter χρησιμοποιώντας ως υλικό τα 200 πιο πρόσφατα tweets (χωρίς τα retweets) από 14.000 χρήστες του Twitter ενός συγκεκριμένου μήνα, του Φεβρουαρίου.

Η εφαρμογή της μεθόδου SCAP στοχεύει στο να εξακριβωθεί κατά πόσο είναι αποτελεσματική αυτή η μέθοδος απόδοσης συγγραφικής πατρότητας αποδίδοντας tweets σε ένα συγκεκριμένο συγγραφέα.

Η μέθοδος SCAP (Layton,Watters, Dazeley, 2010, σελ.2) χρησιμοποίησε το εργαλείο SPI (Simplified Profile Intersection), το οποίο αξιοποιεί τη σχετική απόσταση των ν-γραμμάτων χαρακτήρων των γνωστών tweets που αποδίδονται σε κάποιο συντάκτη σε σύγκριση με το άγνωστο tweet καταδεικνύοντας αν είναι πράγματι ο συγγραφέας και του άγνωστου tweet.

Σύμφωνα με το εργαλείο αυτό, τεμαχίζονται τα προς εξέταση δεδομένα του κάθε συγγραφέα σε πιο απλές μορφές, υπολογίζεται στη συνέχεια ο αριθμός των πιο συχνά εμφανιζόμενων λέξεων-κλειδιών για κάθε δεδομένο και αυτή η λίστα που συγκεντρώνεται αποτελεί το SPI για κάθε συγγραφέα. Πιο συγκεκριμένα, για κάθε συγγραφέα συλλέγονται όλα τα δεδομένα του και στη συνέχεια εξάγεται ο αριθμός των πιο συχνά εμφανιζόμενων λέξεων-κλειδιών.

Οι χρήστες επιλέχθησαν τυχαία ανάμεσα σε 56.000 πλήθος χρηστών με βάση μια συγκεκριμένη λέξη-κλειδί, με αποτέλεσμα η μηχανή αναζήτησης του Twitter να επιστρέψει τα usernames των χρηστών που είχαν κάνει τις πιο πρόσφατες δημοσιεύσεις που εμπεριείχαν τη συγκεκριμένη λέξη-κλειδί. Συνελέγησαν μόνο τα δημοσιευμένα tweets και κάθε δεδομένο συνοδευόταν από το username του χρήστη, την ημερομηνία δημοσίευσης του tweet καθώς και το περιεχόμενο του tweet. Πάραυτα, η ημερομηνία δε συμπεριελήφθη στα πειράματα.

Σημαντικό εύρημα της έρευνας (Layton et al., 2010, σελ.7) είναι το γεγονός ότι αν και το ανώτατο όριο των συλεχθέντων tweets ήταν 200 tweets ανά συγγραφέα

ωστόσο φάνηκε ότι και από τα 120 tweets, κατά προσέγγιση, ανά συγγραφέα η έρευνα είχε φτάσει σε ικανοποιητικό αποτέλεσμα.

Το ποσοστό της ακρίβειας απόδοσης συγγραφικής πατρότητας μειώθηκε κατά 27% όταν αφαιρέθηκαν από τα προφίλ των συντακτών πληροφορίες όπως τα @replies και τα #hashtags τα οποία εμπεριέχουν πολύ σημαντικές πληροφορίες για τον εκάστοτε συντάκτη. Τα @replies είναι η ένδειξη της απευθείας κατεύθυνσης ενός μηνύματος σε ένα χρήστη με ένα συγκεκριμένο όνομα χρήστη και συμπεριλαμβάνεται σε ένα tweet και τα #hashtags χρησιμοποιούνται για να αποσταλούν ομάδες παρόμοιων μηνυμάτων σε διαφορετικούς ανθρώπους και συμπεριλαμβάνονται και αυτά στα tweets. (Layton et al.,2010, σελ.4).

Η έρευνα των Layton, Watters και Dazeley κατέληξε σε τέσσερα σημαντικά συμπεράσματα (Layton, Watters, Dazeley, 2010, σελ.6-7). Αρχικά, η μέθοδος SCAP είναι πολύ ακριβής στο να αποφασιστεί ο συγγραφέας ενός συγκεκριμένου tweet. Η ακρίβεια που αποδόθηκε είναι πάνω από 70% σε ν-γράμματα με 3, 4, 5 και 6 χαρακτήρες.

Επιπροσθέτως, η δημιουργία υπό-προφίλ από το αρχικό προφίλ ενός συντάκτη φάνηκε ότι δίνει μια μικρή αλλά όχι σημαντική αύξηση στην ακρίβεια της απόδοσης συγγραφικής πατρότητας ενώ τα 120 tweets ανά συγγραφέα αποτελούν ένα πολύ σημαντικό και αντιπροσωπευτικό δείγμα όσον αφορά την ακρίβεια. Η αύξηση κατά 20 tweets ανά χρήστη αυξάνει την ακρίβεια αλλά όχι σε σημαντικό βαθμό.

Η έρευνα αυτή κατέδειξε ότι το να συμπεριλάβεις τα @replies τα οποία συμπεριλαμβάνουν ολόκληρο το username συμβάλλει στην ακριβέστερη απόδοση συγγραφικής πατρότητας. Είναι κατανοητό ότι το δίκτυο επικοινωνίας ενός συγκεκριμένου συντάκτη είναι αποφασιστικής σημασίας για την ακρίβεια στην απόδοση συγγραφικής πατρότητας. Υποδηλώνεται, λοιπόν ότι η «γλώσσα» που χρησιμοποιεί ο κάθε χρήστης στη διαδικτυακή του επικοινωνία είναι αντιπροσωπευτική της ταυτότητάς του και χαρακτηριστικό γνώρισμα όσον αφορά τον εντοπισμό του. Η πληροφορία αυτή αρχικά αφαιρέθηκε επειδή σε ένα δίκτυο εγκληματικότητας μέσω του διαδικτύου (*cybercrime*) είναι φυσικό ότι κάποιο πρόσωπο στην προσπάθεια του να κρατήσει την ανωνυμία του δε θα έχει επαφές με το κύκλο των φίλων του για να μην ανακαλυφθεί.

Σε μια μελλοντική έρευνα, η μελέτη δε θα αφορά μόνο τους συντάκτες των tweets αλλά και τη μελέτη και του κύκλου των επαφών τους αν και κάτι τέτοιο μπορεί να δημιουργήσει προβλήματα κυρίως λόγω της χρήσης κοινών ν-γραμμάτων. Επίσης, ένα άλλο στοιχείο που θα χρησιμοποιηθεί σε μια μελλοντική έρευνα είναι η εφαρμογή αυτών των μεθόδων και σε άλλα δίκτυα κοινωνικής δικτύωσης όπως το facebook και το IRC chat για τη μελέτη του διαδικτυακού εγκλήματος (cybercrime). Τέλος, σε μια μελλοντική έρευνα, η βελτίωση της μεθόδου των υπό-προφίλ θα συμβάλλει καθοριστικά στην ακριβέστερη απόδοση συγγραφικής πατρότητας.

Η έρευνα τους αποτελεί μια διερευνητική εξέταση στην απόδοση της συγγραφικής πατρότητας των tweets, η οποία στοχεύει στην ανακάλυψη της βιωσιμότητας και της ακρίβειας κυρίως της συγγραφικής πατρότητας και σε μικρότερα σε μήκος μηνύματα.

Το 2011 η Silva (Silva et al. 2011, σελ.164) δημιούργησε μια βάση δεδομένων από συγγραφείς που έγραφαν tweets στα πορτογαλικά ενώ το κύριο σώμα των κειμένων προερχόταν από 200.000 χρήστες και πάνω από 4.000.000 tweets, τα οποία συνελέγησαν την περίοδο 12 Ιανουαρίου με 1 Οκτωβρίου του 2010. Από αυτό το σώμα κειμένων συνελέγησαν οι 120 πιο ενεργοί χρήστες-συγγραφείς tweets, οι οποίοι είχαν τουλάχιστον 2.000 διακριτά και αυθεντικά μηνύματα αποκλείοντας τα retweets (*RT*).

Το πρωτότυπο στοιχείο της έρευνας είναι το γεγονός ότι επικεντρώθηκε σε μηνύματα ανάμεσα σε τρείς συγγραφείς και χώρισε τους 120 χρήστες σε 40 ομάδες των 3 συγγραφέων, οι οποίοι επιλέχθησαν τυχαία και όχι με κάποια μέθοδο, και διατήρησε αυτές τις ομάδες κατά τη διάρκεια των πειραμάτων. Αυτές οι ομάδες των τριών συγγραφέων αποτελούσαν το προς εξέταση δείγμα της έρευνας.

Ο αλγόριθμος (Silva et al., 2010, σελ. 164) που χρησιμοποιήθηκε ονομάζεται Support Vector Machine (SVM) και επιλέχθηκε ως αλγόριθμος κατηγοριοποίησης χάρη στην αποτελεσματικότητα του στην κατηγοριοποίηση κειμένων και στην αξιοπιστία του όσον αφορά τη διαχείριση μεγάλου αριθμού χαρακτηριστικών (*features*). Ο SVM κρίθηκε κατάλληλος στην αναγνώριση του υφομετρικού μοντέλου κάθε συγγραφέα διακρίνοντας ποια μηνύματα ανήκουν σε κάθε συγγραφέα. Με τη χρήση του SVM (όπως συμβαίνει και με όλους τους αλγόριθμους ταξινόμησης-

classification algorithms) δίνοντας ένα «ύποπτο» μήνυμα για κάθε συγγραφέα είναι δυνατό να προβλέπεται ο βαθμός της πιθανότητας ο εκάστοτε συγγραφέας να είναι ο πραγματικός συγγραφέας

Η έρευνα διεξήχθη σε δυο επίπεδα πειραμάτων (Silva et al., 2010.σελ. 165). Αρχικά, στο πρώτο επίπεδο, στη διαδικασία κατηγοριοποίησης χρησιμοποιήθηκαν όλες οι πιθανές ομάδες χαρακτηριστικών που περιγράφουν τα μηνύματα. Έτσι, δημιουργήθηκαν ομάδες δεδομένων (*data sets*) μεγέθους 75, 250, 1.250 και 2.000 μηνυμάτων ανά συγγραφέα με σκοπό να εξεταστεί η επίδραση του μεγέθους των δεδομένων (*dataset size*) ή του αριθμού των κειμένων στην αποτελεσματικότητα κατά τη διαδικασία απόδοσης συγγραφικής πατρότητας. Στο δεύτερο επίπεδο, κατά τη διαδικασία κατηγοριοποίησης χρησιμοποιήθηκε μια ομάδα χαρακτηριστικών κάθε φορά. Σε αυτή την περίπτωση χρησιμοποιήθηκε η μέγιστη ομάδα δεδομένων (2.000 μηνύματα ανά συγγραφέα) του προηγούμενου πειράματος.

Είναι αξιοσημείωτο ότι στην έρευνα της Silva χρησιμοποιήθηκε και ένας μεγάλος αριθμός μη λεξικών χαρακτηριστικών (*non-lexical features*) (Silva 2010, σελ. 163-164). Τέτοια είναι τα quantitative markers όπως το μέγεθος μιας λέξης, τα διπλά σύμφωνα, οι ημερομηνίες, λέξεις που δεν υπάρχουν στο λεξικό αλλά χρησιμοποιούνται σκόπιμα και κάποια σύμβολά όπως τα #,@. Επίσης, σύμβολα με εικονίδια (*emoticons*) τα οποία χωρίζονται σε τρείς κατηγορίες: smileys, LOLs και τεμάχια λέξεων αποτελούμενα από δυο διαφορετικά γράμματα όπως το αγγλικό γέλιο ‘hahaha’, το τυπικό βραζιλιάνικο γέλιο ‘rsrsrs’και το ισπανικό γέλιο ‘jejeje’ (*interjections*). Τέλος, χρησιμοποιήθηκαν συντομογραφίες (*abbreviations*) και τα σημεία στίξης όπως οι τρείς τελείες (…), τα διπλά θαυμαστικά (!!) κλπ.

Η έρευνα, όσον αφορά τα δύο επίπεδα στα οποία διεξήχθη, κατέδειξε ότι η παρουσίαση όλων των ομάδων των χαρακτηριστικών ταυτόχρονα θεωρείται καλύτερη από την παρουσίαση της κάθε ομάδας χαρακτηριστικών μεμονωμένα. Ωστόσο, τα σύμβολα των εικονιδίων παρουσιάστηκαν σε μεγαλύτερο βαθμό από τις άλλες κατηγορίες των quantitative markers. Ενώ, η χρήση των μη λεξικών χαρακτηριστικών όπως τα quantitative markers, τα emoticons, τα abbreviations και τα punctuations, τα οποία χρησιμοποιήθηκαν για πρώτη φορά, έδωσαν πολύ καλά αποτελέσματα και θεωρήθηκαν αποτελεσματικά στην ταυτοποίηση του συγγραφέα

ενός tweet, παρά τους περιορισμούς ως προς το μήκος των μηνυμάτων, «κουβαλώντας» και πληροφορίες σχετικές με τις ιδιολέκτους που αντιπροσωπεύουν. Τέλος, (Silva et al. 2011, σελ.167) η έρευνα κατέδειξε ότι η αξιόπιστη απόδοση συγγραφικής πατρότητας μπορεί να γίνει με τουλάχιστον 100 tweets ανά συγγραφέα.

Η Boutwell (2011) ανέπτυξε έναν πολυτροπικό ταξινομητή-κατηγοριοποιητή για tweets με σκοπό να συνδυάσει ένα κινητό τηλέφωνο με το χρήστη του. Το αρχικό δείγμα του σώματος των tweets συμπεριλάμβανε 4.045 tweets από 53 ενεργούς χρήστες και, με μια ακόλουθη επαύξηση που έγινε, ο αριθμός των tweets αυξήθηκε δραματικά στα 114.000 tweets. (Boutwell, S.R. 2011, σελ.89-93). Η προεπεξεργασία περιελάμβανε tweets μικρότερα από 3 λέξεις και τη μετακίνηση των quantitative markers όπως τα @ και # από το κείμενο. Στους χαρακτήρες ν-γραμμάτων υπολογίζονταν αυτοί με ν από 2 έως 6. Επίσης, έγινε μια απόπειρα να μοντελοποιηθεί το πρότυπο ζωής του κάθε χρήστη χρησιμοποιώντας σαν ξεχωριστά γνωρίσματα τις χρονικές ενδείξεις (*timestamps*) των tweets και το quantitative marker @.

Στην έρευνα της Boutwell χρησιμοποιήθηκαν Separate Naïve Bayes (NB) ταξινομητές οι οποίοι δημιουργήθηκαν για συντάκτες και τηλέφωνα. Σε πρώτη φάση, τα tweets αποδόθηκαν σε συγκεκριμένους χρήστες με τη χρήση του NB ταξινομητή και το χαρακτήρα ν-γραμμάτων και τα τηλέφωνα συνδέθηκαν με τους χρήστες τους χρησιμοποιώντας μια σειρά χαρακτηριστικών. Στο δεύτερο και τελευταίο στάδιο, οι δυο ταξινομητές συνδέθηκαν και τα αποτελέσματα κατέδειξαν ότι ο συνδυασμός ανάμεσα στη φυσική γλώσσα και στα χαρακτηριστικά από τους ταξινομητές ταυτοποιεί το χρήστη με το τηλέφωνό του καλύτερα από όταν οι δυο ταξινομητές χρησιμοποιούνταν ανεξάρτητα ο ένας από τον άλλο.

Τα αποτελέσματα κατέδειξαν ότι το γνώρισμα ν-γραμμάτων χρησιμοποιώντας ξεχωριστά tweets σα βασική μονάδα κειμένου εμφάνισε κακή απόδοση (40,3% σε δείγμα 50 συντακτών). Αντίθετα, αν κάθε κειμενική ενότητα περιελάμβανε πολλαπλά tweets η ακρίβεια στην απόδοση συγγραφικής πατρότητας ανερχόταν στο 99,6% σε μονάδες κειμένου με 23 tweets. Μεγάλα ποσοστά ακρίβειας στην απόδοση συγγραφικής πατρότητας (94%) σημειώθηκαν σε περιπτώσεις που

χρησιμοποιήθηκαν χαρακτηριστικά γνωρίσματα (quantitative markers) όπως το @ ενώ οι χρονικές ενδείξεις εμφάνισαν χαμηλά ποσοστά στην ακρίβεια απόδοσης (35%).

Τα μεγάλα ποσοστά ακρίβειας στην απόδοση συγγραφικής πατρότητας που σημειώθηκαν στην περίπτωση χρήσης του quantitative marker @ επιβεβαίωσαν και προηγούμενα ευρήματα όπως στην έρευνα Layton et al., 2010.

Οι έρευνες των Layton et al., (2010) και Boutwell (2011) έχουν καταδείξει ότι το quantitative marker @ αποτελεί μια παντοδύναμη ένδειξη συγγραφικής πατρότητας και δυναμικό γνώρισμα κάθε συντάκτη αφού οι περισσότεροι συντάκτες μοιράζονται σκέψεις και πληροφορίες με συγκεκριμένους χρήστες, οι οποίοι αποτελούν έναν αυστηρό κύκλο κοινωνικού δικτύου. Φυσικά αυτό δεν είναι ένα σταθερό φαινόμενο αφού βέβαια υπάρχουν οι χρήστες που έχουν ένα σταθερό κύκλο επαφών αλλά υπάρχουν και άλλοι χρήστες που χρησιμοποιούν το twitter σαν ένα κανάλι μονόδρομης επικοινωνίας, γεγονός που δυσχεραίνει την κατάσταση. Σε αυτές τις περιπτώσεις, για την απόδοση συγγραφικής πατρότητας χρησιμοποιούνται μέθοδοι απόδοσης συγγραφικής πατρότητας, οι οποίοι μπορούν να ταυτοποιήσουν τους συντάκτες χρησιμοποιώντας πληροφορίες που αφορούν μόνο το γλωσσολογικό κομμάτι του κάθε tweet (Mikros & Perifanos 2013, σελ.3).

Οι Koppel et al. (Koppel et al. 2011, σελ. 83-84) παρατήρησαν ότι όλες οι προηγούμενες έρευνες στο πεδίο της απόδοσης συγγραφικής πατρότητας είχαν επικεντρωθεί στην απόδοση ενός συγκεκριμένου και ανώνυμου δεδομένου σε μια μικρή και κλειστή ομάδα από υποψήφιους συντάκτες. Κάτι τέτοιο στον πραγματικό κόσμο και όχι στο ερευνητικό πεδίο είναι αρκετά δύσκολο καθώς ανακύπτει το πρόβλημα ότι ο αριθμός των υποψηφίων συντακτών είναι πολύ μεγάλος (πολλές φορές χιλιάδες), επίσης μπορεί να μη συμπεριλαμβάνεται ο πραγματικός συντάκτης και τέλος πιθανότατα είτε τα γνωστά κείμενα είτε τα ανώνυμα είτε και τα δυο μπορεί να είναι περιορισμένα. Έτσι, οι Koppel et al. (Koppel et al. 2011, σελ .83-84) προσπάθησαν, χρησιμοποιώντας μια δική τους τεχνική στη ανεύρεση της συγγραφικής πατρότητας, να δείξουν, ότι και σε αυτές τις δύσκολες περιπτώσεις και παρά τους περιορισμούς, μπορούν να χρησιμοποιηθούν μέθοδοι οι οποίοι μπορούν να επιτύχουν μεγάλη ακρίβεια ακόμα και όταν ο αριθμός των υποψηφίων

συντακτών είναι χιλιάδες. Επίσης, προσπάθησαν να αποδείξουν τη σχέση που υπάρχει ανάμεσα στην ακρίβεια απόδοσης συγγραφικής πατρότητας και στο μέγεθος της ομάδας των υποψηφίων συντακτών, την ποσότητα των γνωστών κειμένων που αποδίδονται σε υποψήφιους συντάκτες και στο μήκος των ανώνυμων δεδομένων.0

Οι Koppel et al. (Koppel et al. 2011, σελ .83-84) χρησιμοποιώντας ένα δείγμα 10.000 υποψήφιων bloggers συντακτών στόχευσαν στο να αποδώσουν σε κάποιο συντάκτη ένα απόσπασμα *(snippet*) 500 λέξεων. Αυτή η προσέγγιση στοχεύει να αποδοθεί σε ένα συγκεκριμένο συντάκτη ένα συγκεκριμένο απόσπασμα το οποίο περιλαμβάνει γλωσσολογικά γνωρίσματα. Αυτή η μέθοδος απέδειξε ότι μπορεί να επιτευχθεί αξιόπιστη απόδοση συγγραφικής πατρότητας αποσπασμάτων σε περιπτώσεις όπου υπάρχουν χιλιάδες υποψήφιοι συντάκτες.

Σύμφωνα με τον Koppel et al. (Koppel et al. 2011, σελ. 84) οι υπάρχουσες μέθοδοι απόδοσης συγγραφικής πατρότητας ταξινομούνται σε δυο παραδείγματα: το παράδειγμα που βασίζεται στην ομοιότητα (*the similarity based paradigm*) και το παράδειγμα της μηχανικής μάθησης (*the machine learning paradigm*). Στην περίπτωση του παραδείγματος που βασίζεται στην ομοιότητα η απόσταση ανάμεσα σε δυο δεδομένα και σε ένα ανώνυμο δεδομένο υπολογίζεται και η απόδοση συγγραφικής πατρότητας αποδίδεται στο συντάκτη του οποίου η γραφή έχει πιο πολλά κοινά με το κείμενο που αναζητείται. Στην περίπτωση του παραδείγματος της μηχανικής μάθησης η γραφή ενός υποψήφιου συντάκτη χρησιμοποιείται για να κατασκευαστεί ένας ταξινομητής, ο οποίος θα χρησιμοποιηθεί για να κατηγοριοποιηθούν ανώνυμα δεδομένα. Οι Koppel et al. (Koppel et al. 2011, σελ. 84) πιστεύουν ότι οι μέθοδοι που βασίζονται στο παράδειγμα που βασίζεται στην ομοιότητα είναι πιο κατάλληλοι σε περιπτώσεις μεγάλου αριθμού υποψηφίων συντακτών. Η έρευνα στο παράδειγμα που βασίζεται στην ομοιότητα έχει επικεντρωθεί στην επιλογή των χαρακτηριστικών για την εκπροσώπηση ενός εγγράφου, σε μεθόδους μείωσης του κενού ενός γνωρίσματος-χαρακτηριστικού και στην επιλογή μιας μετρικής απόστασης. Η έρευνα στο παράδειγμα της μηχανικής μάθησης (Koppel et al. 2011, σελ. 84) έχει επικεντρωθεί στην επιλογή των

χαρακτηριστικών για την εκπροσώπηση ενός εγγράφου και στην επιλογή μαθησιακών αλγορίθμων.

Οι Koppel et al. χρησιμοποίησαν σα βάση της ανάλυσής τους τα τετραγράμματα (4-grams) δηλαδή ακολουθίες χαρακτήρων με μήκος τεσσάρων χαρακτήρων χωρίς κενά ή ακολουθίες τεσσάρων ή λιγότερων χαρακτήρων μαζί με κενά. Η χρήση των ν-γραμμάτων αποδείχτηκε αποτελεσματική στην απόδοση συγγραφικής πατρότητας

Οι Koppel et al. επισήμαναν ότι ένα σημαντικό πλεονέκτημα τους είναι το γεγονός της προσαρμοστικότητας της μεθόδου τους σε οποιαδήποτε γλώσσα ανεξάρτητα από τις προηγούμενες γλωσσικές γνώσεις. Η μέθοδος που χρησιμοποιήθηκε από τους Koppel et al. σημείωσε επιτυχία στο 46% των υποθέσεων η οποία ανήλθε στο 93,2 % σε ακρίβεια μετά την εισαγωγή της επιλογής «Δε Γνωρίζω» (Don't Know) Η μέθοδος, που χρησιμοποίησαν, αποδεικνύεται αποτελεσματική σε περιπτώσεις διαχείρισης ενός μεγάλου αριθμού υποψηφίων συντακτών, όπου οι παραδοσιακές μέθοδοι δεν ήταν ιδιαίτερα αποτελεσματικές, ενώ σε περιπτώσεις μικρού αριθμού υποψήφιων συντακτών και περιορισμένου αριθμού ανώνυμων κειμένων φαίνεται ότι η μέθοδος αυτή δε δίνει ικανοποιητική λύση.

Οι Mikros και Perifanos (2011) στην προσπάθεια για συγγραφική ταυτοποίηση των ηλεκτρονικών μηνυμάτων, επικεντρώθηκαν σε μια ομάδα χαρακτηριστικών (*features*), τα οποία καλύπτουν ένα ευρύ φασμά γλωσσολογικών επιπέδων και την ίδια ώρα είναι εύκολο να εφαρμοστούν ενώ παράλληλα είναι ανεξάρτητα από οποιαδήποτε γλώσσα. Σε πρώτο στάδιο χρησιμοποίησαν πέντε ομάδες διαφορετικών χαρακτηριστικών και σε δεύτερο στάδιο συνδύασαν αυτές τις ομάδες δημιουργώντας μια ομάδα χαρακτηριστικών με το όνομα "All", μέθοδος που απέδωσε πολύ καλά αποτελέσματα ενώ παράλληλα είναι γενικά αποδεκτή ως η καλύτερη στρατηγική. Σε όλες τις ομάδες χαρακτηριστικών κανονικοποίησαν τις συχνότητες σε σύγκριση με το μήκος του κειμένου. Με σκοπό να δημιουργηθεί η ομάδα χαρακτηριστικών "All" χρησιμοποιήθηκαν τα 1000 πιο συχνά χαρακτηριστικά από κάθε ομάδα του σώματος κειμένων με αποτέλεσμα να συγκεντρωθούν 5000 χαρακτηριστικά. Οι πέντε ομάδες χαρακτηριστικών που χρησιμοποιήθηκαν ήταν διγράμματα, τριγράμματα και μονογράμματα χαρακτήρων και διγράμματα και τριγράμματα λέξεων.

Τα αποτελέσματα από την έρευνα απόδοσης συγγραφικής πατρότητας ήταν ενθαρρυντικά και επιβεβαίωσαν την άποψή τους οτι η αναγνώριση συγγραφικής πατρότητας στηρίζεται σε κειμενικά χαρακτηριστικά τα οποία είναι διάσπαρτα σε ένα ευρύ φάσμα γλωσσολογικών επιπέδων.

Σε μια μελλοντική έρευνα , οι προσπάθειες απόδοσης συγγραφικής πατρότητας θα κατευθυνθούν και στην ανίχνευση χαρακτηριστικών από άλλα γλωσσολογικά πεδία στα οποία θα λαμβάνεται υπόψη όχι μόνο η συχνότητα αλλά και η δύναμη διάκρισης σε μικρότερες κατηγορίες με σκοπό να αυξηθεί η απόδοση σε μακροεπίπεδο. Επιπλέον, θα χρησιμοποιηθεί διαφορετικός αλγόριθμος και περισσότερες ομάδες χαρακτηριστικών.

Συγγραφική πατρότητα σε ελληνικό σώμα κειμένων

Οι Mikros και Perifanos (2013) ανέπτυξαν το πρώτο σώμα κειμένων από ελληνικά tweets. Υπήρξε η πρώτη προσπάθεια να συγκεντρωθεί όγκος κειμένων με σκοπό την ανίχνευση της πατρότητας στα κοινωνικά δίκτυα. Επέλεξαν χρήστες που χαρακτηρίζονταν από έντονη παρουσία στα δρώμενα (ηθοποιοί, δημοσιογράφοι, πολιτικοί)· γι' αυτό και αφαιρέθηκαν τα @replies, τα #hashtags και τα retweets ως εξωγλωσσικά χαρακτηριστικά που οδηγούν απευθείας στον συγγραφέα. Ακολούθως χρησιμοποίησαν ως χαρακτηριστικό τις ακολουθίες ν-γραμμάτων λαμβάνοντας υπόψη τα σημεία στίξης και χαρακτηριστικά που εμφανίζονται με συχνότητα στο Twitter (όπως τα emoticons). Επιπρόσθετα, ο τεμαχισμός του σώματος σε πακέτα λέξεων (από 20- 100) έδειξε ότι το μέγεθος του κειμένου επηρεάζει την ακρίβεια των αποτελεσμάτων με τα πακέτα των 75 και 100 λέξεων να εμφανίζουν πιο ακριβή αποτελέσματα.

4. Μεθοδολογία

Επιλογή των χαρακτηριστικών (Features Selection)

Ο περιορισμός που θέτει το Twitter ως μέσο κοινωνικής δικτύωσης δημιουργεί την ανάγκη στους χρήστες (*users*) να χρησιμοποιούν διαφοροποιημένο τρόπο επικοινωνίας από τον συμβατικό λόγο της καθημερινότητας. Η παρατήρηση δείχνει ότι οι χρήστες "εφευρίσκουν" διάφορους τρόπους ώστε να μεταφέρονται οι σκέψεις και τα συναισθήματά τους μέσω του διαύλου των άμεσων μηνυμάτων του Twitter στα πλαίσια των 140 χαρακτήρων.

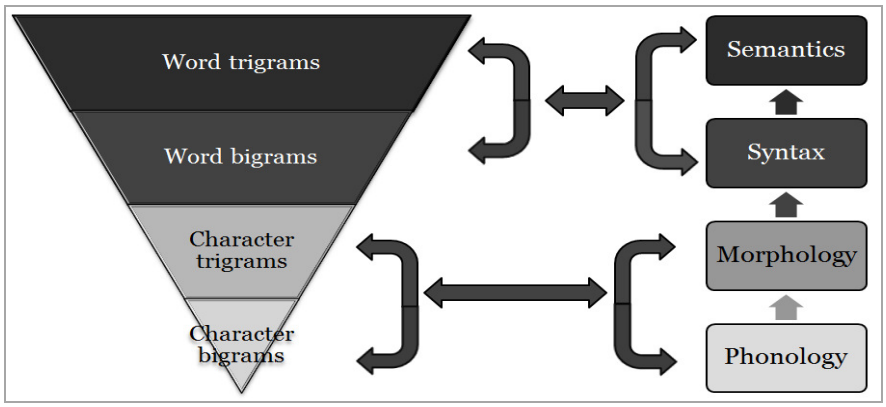
Λίστα ορισμένων Χαρακτηριστικών (Features List)		
Χαρακτηριστικό	Περιγραφή	Παράδειγμα
Emoticons	σειρές χαρακτήρων που σχηματίζουν πρόσωπα και εκφράζουν συναισθήματα	:-) ;)
Prosodic Emphasisers	αποτύπωση της προφορικής προφοράς μέσω της ορθογραφίας	Άντεεεεεε αγάπηηηηηηη
Punctuation Marks	σημεία στίξης που χρησιμοποιούνται επαυξημένα	!!!!!!!!!!!!!!!!!!!!!!!!! .....

Πίνακας 5: Λίστα επιλεγμένων χαρακτηριστικών

Η έρευνα της Silva (2011) έδειξε ότι είναι σημαντική η συνεισφορά της χρήσης των emoticons για την εξαγωγή της συγγραφικής πατρότητας. Τα smileys αναπαριστούν ανθρώπινα συναισθήματα με τους διαφορετικούς συνδυασμούς ανάμεσα στα σύμβολα (όπως "(", ")", "-", ";"). Επιπρόσθετα, έχει συμπεριληφθεί η χρήση των τραβηγμένων φωνηέντων, τα οποία χρησιμοποιεί ένας χρήστης στο τέλος της λέξης ώστε να εκφράσει συναισθήματα όπως ο θαυμασμός ή η έκπληξη. Τέλος, λήφθηκε υπόψη και η χρήση των σημείων στίξης που δεν συμβαδίζουν με τον γραπτό λόγο και αποτελούν χαρακτηριστικό όχι μόνο του συγγραφέα αλλά γενικότερα της "γλώσσας" που χρησιμοποιείται στο Twitter. Στην παρούσα έρευνα έγινε προσπάθεια να ληφθούν υπόψη αυτά τα υφολογικά χαρακτηριστικά. Ωστόσο,

αρκετά από αυτά δεν εμφανίζονταν με συχνότητα και δεν καταμετρήθηκαν αφού δεν μπήκαν στη λίστα με τα 1000 πιο συχνά εμφανιζόμενα ν-γράμματα (σε χαρακτήρες και λέξεις). Όσα από αυτά εμφανίζονται μέσα στη λίστα των χιλίων συχνότερων συχνοτήτων λήφθηκαν υπόψη για να κατασκευαστούν οι τελικοί συγκεντρωτικοί πίνακες με τις εμφανίσεις τους.

Ως γλωσσολογικό χαρακτηριστικό λαμβάνονται στην υφομετρική ανάλυση τα ν-γράμματα (*n-grams*). Τα ν-γράμματα είναι ακολουθίες χαρακτήρων ή λέξεων που ανταποκρίνονται στα αντίστοιχα γλωσσολογικά επίπεδα. Π.χ. για ν=1, έχουμε word-unigrams και characters-unigram, για ν=2, έχουμε διγράμματα λέξεων(*word-bigrams*) και διγράμματα χαρακτήρων (*character-bigrams*) Οι λέξεις αναφέρονται στις σχέσεις σε σημασιολογικό και συντακτικό επίπεδο, ενώ οι χαρακτήρες, τα φωνήματα δηλαδή, ανταποκρίνονται στο μορφολογικό και φωνολογικό επίπεδο. Αυτός είναι και ο λόγος που αποτελούν ένα ισχυρό υφολογικό χαρακτηριστικό για τον κάθε συγγραφέα.



Εικόνα 3: AMNP: ιεραρχική των ν-γραμμάτων σε αντιστοιχία με τα γλωσσολογικά επίπεδα [Πηγή: Mikros & Perifanos (2013, σελ.3)]

Συλλογή και επεξεργασία δεδομένων

Η συλλογή των δεδομένων έγινε χρησιμοποιώντας το tweeter API για τη γλώσσα προγραμματισμού Python. Το πακέτο αυτό είναι ελεύθερο προς χρήση και διαθέσιμο online.

Στην προσπάθεια για συλλογή δεδομένων με απώτερο σκοπό τη δημιουργία σώματος κειμένου και επεξεργασία αυτού, δεδομένου ότι δεν υπάρχει διαθέσιμη συλλογή από tweets στην νέα ελληνική έπρεπε να δημιουργηθεί κάτι απο το μηδέν.

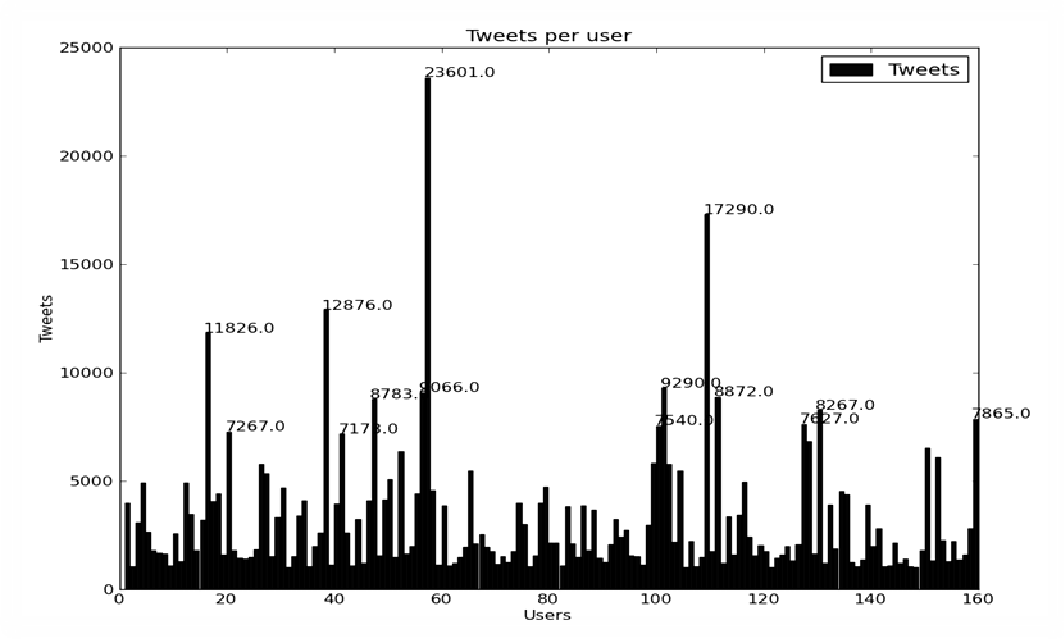
Ως προς την επιλογή των χρηστών στην αρχή λάβαμε υπ όψη το ελληνικό site treding.gr. Είναι ένα site το οποίο παρέχει πληροφορίες και στατιστικά ως προς τους followers, trends top users, ποιός είναι δηλαδή ο χρήστης με τα περισότερα tweets κ.α.

Χρησιμοποιώντας το API του tweeter και καθώς κάναμε συλλογή δεδομένων αποδείχθηκε ότι το σύνολο των tweets που εμφάνιζαν οι χρήστες στους λογαριασμούς τους δεν ήταν δικά τους μόνο αλλά το σύνολο όλων των χρηστών οι οποίοι αποστέλουν μηνύματα στο λογαριασμό του κάθε χρήστη. Μετά απο αυτό η επιλογή των χρηστών έγινε με κριτήρια κυρίως τη δημοσιότητας τους. Ψάχναμε δηλαδή για χρήστες δημοσιογράφους, πολιτικούς, καλλιτέχνες καθώς και τους followers αυτών. Δεδομένου ότι υπάρχει και ο περιορισμός απο το tweeter για συγκεκριμένο αριθμό tweets που μπορεί να κατεβάσει ο καθένας, χρειάσθηκε αρκετός καιρός έτσι ώστε να συλλεχθεί ένας ικανοποιητικός αριθμός tweets.

Τα συνολικά tweets τα οποία συλλέχθησαν για τη δημιουργία του σώματος κειμένων προήλθαν από 235 χρήστες και είναι στον αριθμό περί τις εννιακόσιες χιλιάδες (927798). Το σώμα κειμένων το οποίο θέλαμε να δημιουργήσουμε έπρεπε να πληροί τα παρακάτω κριτήρια: δεδομένα στην ελληνική γλώσσα, με αριθμό λέξεων μεγαλύτερο των 4 λέξεων ανά tweet και το κάθε tweet, το οποίο συλλέγαμε δεν θέλαμε να περιέχει retweet (*RT*). Επιπλέον ήταν επιθυμητό ο κάθε χρήστης να έχει συνολικά tweets άνω των χιλίων (1000). Κατόπιν τούτου μετά την εφαρμογή των κριτηρίων καταλήξαμε στο τωρινό σώμα κειμένων.

Συγκεκριμένα στο σύνολο όλων των χρηστών έχουμε συγκεντρώσει 552.040 tweets και για χρήστες με tweets άνω των 1000 έχουμε συνολικά 159 χρήστες με 512376

tweets. Οι χρήστες στο σύνολο τους είναι δημοσιογράφοι, πολιτικοί, παρουσιαστές, εκδότες περιοδικών, μηχανικοί, καλλιτέχνες και άλλοι.



Εικόνα 4: Οι χρήστες συνολικά με αριθμό tweets ανά χρήστη

Στην πραγματικότητα ο μέσος όρος των λέξεων ανά tweet είναι ελαφρώς μεγαλύτερος διότι το σύνολο των λέξεων για κάθε χρήστη υπολογίστηκε αφού κάναμε tokenization και υπολογίζοντας ως λέξεις ακόμα και τα σημεία στίξης όπως «:-))», «..», «!!...» διότι είναι τυπικά για το επικοινωνιακό περιβάλλον του Twitter και αποτελούν υφολογικό χαρακτηριστικό του κάθε συγγραφέα. Με τον αναλυτή των λεξικών μονάδων (*tokenizer*) καταφέραμε να αναλύσουμε τα κείμενα του σώματος που δημιουργήσαμε ενώ ταυτόχρονα αναγνωρίσαμε σ' αυτά τις βασικές λεξικές μονάδες, τις οποίες εξαγάγαμε με την χρήση τυποποιημένων εκφράσεων (*regular expressions*) ώστε να προσθέσουμε επιπλέον κειμενικά στοιχεία, τα οποία κατά την διαδικασία της επεξεργασίας θεωρήσαμε για σκοπούς μέτρησης ως λέξεις.

Ο μέσος όρος των λέξεων ανά tweet καθώς και το μέγεθος των μέγιστων αλλά και ελάχιστων λέξεων ανά tweet και κατ΄ επέκταση η τυπική απόκλιση υπολογίστηκε άμεσα από τη βάση με μοναδικό κριτήριο το κενό μεταξύ των λέξεων. Ακόμα υπολογίστηκαν ως λέξεις mentions (@userA) καθώς και τα hashtags (Π.χ. #agapwTaZwa).

Οι στατιστικές πληροφορίες σχετικά με τους χρήστες και τα tweets τους παρατίθενται στον παρακάτω πίνακα

Authors	Tweets	Words	average word	max words	min words	Standard deviation
1	3995	46337	104.365	28	5	49.352
2	1048	13408	114.294	27	5	51.577
3	3070	46121	144.033	28	5	55.361
4	4920	64077	120.130	41	5	56.139
5	2638	45967	148.760	29	5	55.255
…						
158	2815	40141	130.160	29	5	55.431
159	7865	100034	120.134	30	5	50.321
Total	512376	7335560				

Πίνακας 6: Στατιστικά στοιχεία του σώματος κειμένων

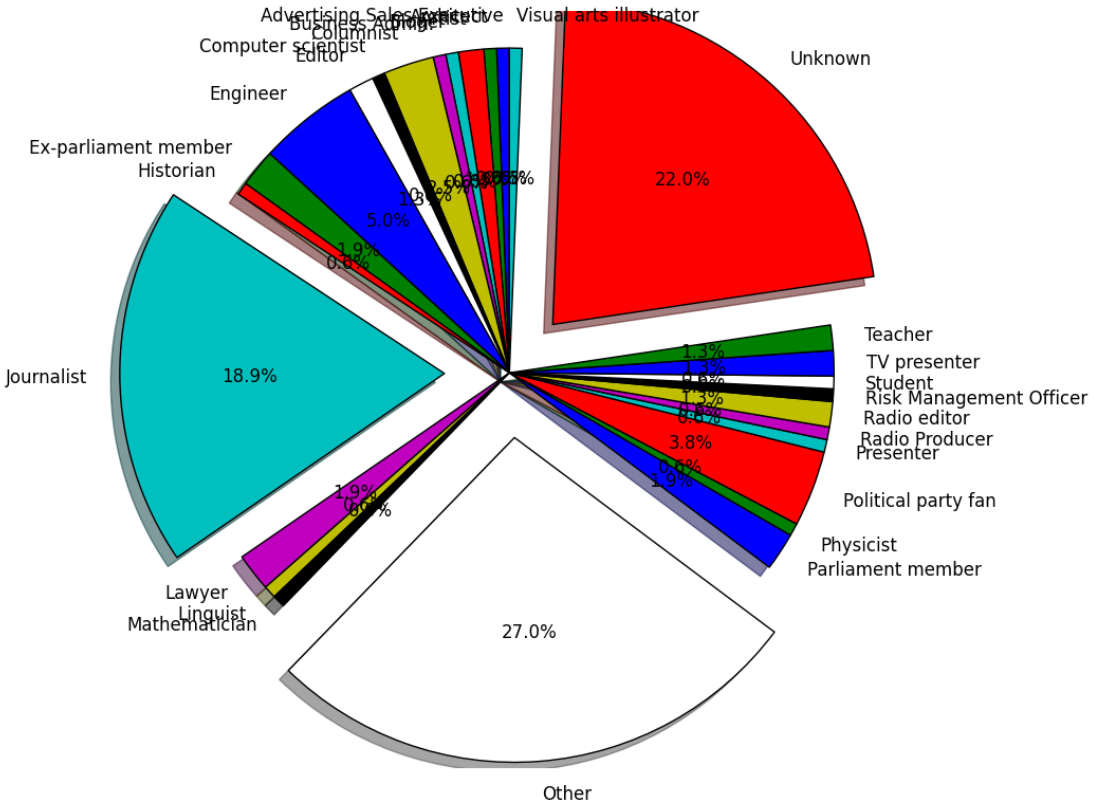
Επιλογή αλγορίθμου

Ο αλγόριθμος μηχανικής μάθησης ο οποίος επιλέχθηκε για την κατηγοριοποίηση ήταν ο αλγόριθμος SVM (Support Vector Machine). Επιπλέον για να εξαγάγουμε τα αποτελέσματα (scores) της ακρίβεια της κατηγοριοποίησης με αυτόν τον αλγόριθμο και να κάνουμε μια αποτίμηση του μοντέλου της κατηγοριοποίησης που χρησιμοποιήσαμε ακολουθήσαμε τεχνικές cross validation με αριθμό τεμαχισμού σε πακέτα των 10 (10-Fold cross validation, cv=10) για καθένα πακέτο δεδομένων μας (200, 180, … ,20) και για κάθε τύπο δεδομένων που συλλέξαμε (διγράμματα χαρακτήρων και λέξεων καθώς και τριγράμματα χαρακτήρων και λέξεων). Δεδομένου ότι τα features του κάθε πίνακα μας ήταν 1000 (οι χίλιες πιο συχνές εμφανίσεις στο συνολικό σώμα κειμένων μας) και δεδομένου ότι τα instances του κάθε πίνακα ήταν μεταξύ 5000 - 51.000, με τον ταυτόχρονο τεμαχισμό των πακέτων σε 10 ακολουθώντας την τεχνική cross validation, η όλη διαδικασία ήταν υπολογιστικά και χρονικά επίπονη.

5. Αποτελέσματα

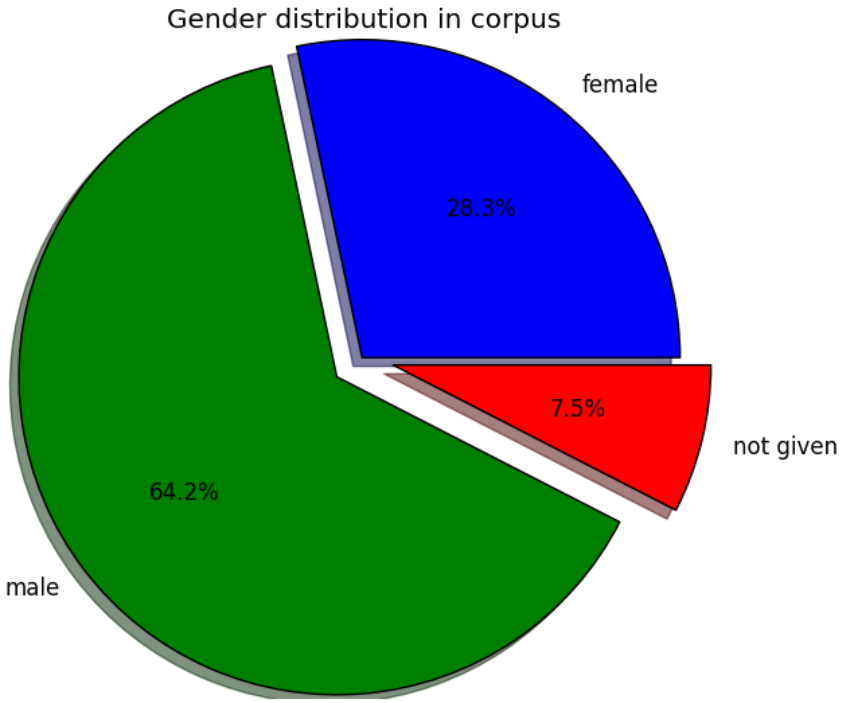
Τα αποτελέσματα έδειξαν ότι οι χρήστες που χρησιμοποιούν το Twitter ως το βασικό εργαλείο επικοινωνίας τους στοχεύουν στην επαγγελματική τους προώθηση. Ήδη από τα αποτελέσματα των 3-γράμματων σε επίπεδο λέξεων φαίνεται ότι οι χρήστες προωθούν τις απόψεις τους αλλά αποσκοπούν και στην αξιοποίηση του Twitter ως μέσο προώθησης της εργασίας τους. Η ανάλυση των 32 χρηστών που προέρχονται κυρίως από τον χώρο της πολιτικής, της δημοσιογραφίας και της τηλεόρασης δείχνει ότι η παραπομπή σε ιστοσελίδα και η μετάδοση πληροφορίας αποτελεί σημαντική παράμετρο για την χρήση του Twitter.

Στην παρακάτω εικόνα βλέπουμε στο σύνολο του σώματος μας την κατανομή ανά ιδιότητα των χρηστών, πληροφορία η οποία εμφανίζεται ανάλογα με το τι έχει δηλώσει ο χρήστης, όπως π.χ. δημοσιογράφος.



Εικόνα 5: Ιδιότητα των χρηστών στο συνολικό σώμα

Τα κείμενα στο συνολικό σώμα κειμένων μας προέρχονται από 45 γυναίκες, 102 άνδρες και από 12 χρήστες οι οποίοι δε δίνουν πληροφορίες για το γένος τους. Στην παρακάτω εικόνα βλέπουμε την κατανομή του γένους στο σύνολο του σώματος κειμένων που συγκεντρώσαμε.



Εικόνα 6: Ποσοστό ανδρών και γυναικών στο συνολικό σώμα κειμένων

Τα tweets στο σύνολο τους συγκεντρώθηκαν σε κειμενικά αρχεία ανά συγγραφέα. Π.χ. για τον συγγραφέα με το όνομα «a_morellas» δημιουργήθηκε το αρχείο «a_morellas.txt» και εκεί συγκεντρώθηκαν όλα τα tweets στο σύνολο τους από τον ίδιο συγγραφέα. Στην συνέχεια τα κείμενα αυτά καθαρίστηκαν από mentions, hashtags συνδέσμους (URLs) καθώς και "html-tags" τα οποία υπήρχαν σε διάφορα tweets. Οι σύνδεσμοι αντικαταστήθηκαν με τη λέξη «httpaddr» ενώ όλα τα υπόλοιπα διαγράφηκαν από τα κείμενα.

Στην συνέχεια αφού καθαρίστηκαν τα κείμενα τα χωρίσαμε σε word unigrams (λέξεις) υπολογίσαμε τις συχνότητες εμφάνισης τους ανά χρήστη και επιπλέον

υπολογίσαμε και εξαγάγαμε τα στατιστικά στοιχεία τα οποία εμφανίζονται στον παραπάνω πίνακα Τα κείμενα του κάθε συγγραφέα χωρίστηκαν σε πακέτα (batches) των 20 – 200 λέξεων με βήματα των 20 λέξεων. Κατά αυτόν τον τρόπο καταλήξαμε να έχουμε κείμενα διαφορετικού μεγέθους λέξεων. Από όλα τα tweets συνολικά κατασκευάσαμε στην συνέχεια ένα πολύ μεγάλο κειμενικό αρχείο, το σώμα tweet. Από αυτό υπολογίσαμε τις 1000 πιο συχνές εμφανίσεις των unigrams, bigrams και three-grams (λέξεις, διγράμματα και τριγράμματα) και κατασκευάσαμε πίνακες που περιέχουν τις συχνότητες εμφάνισης ανά συγγραφέα στο κάθε πακέτο-αρχείο. Για παράδειγμα το αρχείο «a_morellas_batch_wpt_20_1.txt» περιέχει 4 φορές τη «λέξη» «..» και 1 φορά τη λέξη «να», οι οποίες βρίσκονται στις πρώτες 1000 πιο συχνές λέξεις στο συνολικό σώμα κειμένων (*corpus*).

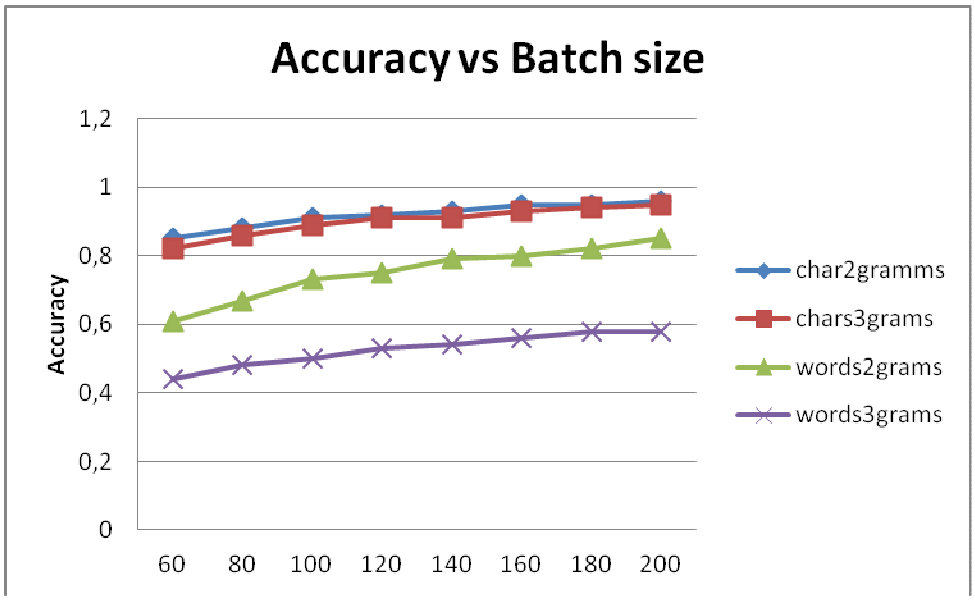
Κατά αυτόν τον τρόπο κατασκευάσαμε ένα μεγάλο συγκεντρωτικό πίνακα με όλα αυτά τα στοιχεία που θα χρησιμοποιήσουμε στην συνέχεια για την κατηγοριοποίηση τους χρησιμοποιώντας τεχνικές μηχανικής μάθησης.

Τα αποτελέσματα αυτής της αποτίμησης που ακολουθήσαμε παρουσιάζονται στον παρακάτω πίνακα καθώς και τα παρακάτω γραφήματα.

Batch size	char2gramms	chars3grams	words2grams	words3grams
60	0,85(+/-0.02)	0,82(+/-0.02)	0,61(+/-0.02)	0,44(+/-0.01)
80	0,88(+/-0.01)	0,86(+/-0.02)	0,67(+/-0.03)	0,48(+/-0.03)
100	0,91(+/-0.02)	0,89(+/-0.02)	0,73(+/-0.02)	0,5(+/-0.02)
120	0,92(+/-0.01)	0,91(+/-0.01)	0,75(+/-0.03)	0,53(+/-0.02)
140	0,93(+/-0.01)	0,91(+/-0.02)	0,79(+/-0.02)	0,54(+/-0.03)
160	0,95(+/-0.02)	0,93(+/-0.02)	0,8(+/-0.03)	0,56(+/-0.02)
180	0,95(+/-0.02)	0,94(+/-0.02)	0,82(+/- 0.03)	0,58(+/-0.03)
200	0,96 (+/- 0.02)	0,95 (+/- 0.02)	0,85(+/- 0.03)	0,58(+/-0.05)

Πίνακας 7: Η ακρίβεια (accuracy) σε σχέση με το μέγεθος των κειμένων ανά τύπο n-grams

Στο παρακάτω γράφημα φαίνεται ότι καθώς αυξάνει το μέγεθος των κειμένων αυξάνει και η ακρίβεια του μοντέλου της κατηγοριοποίησης.



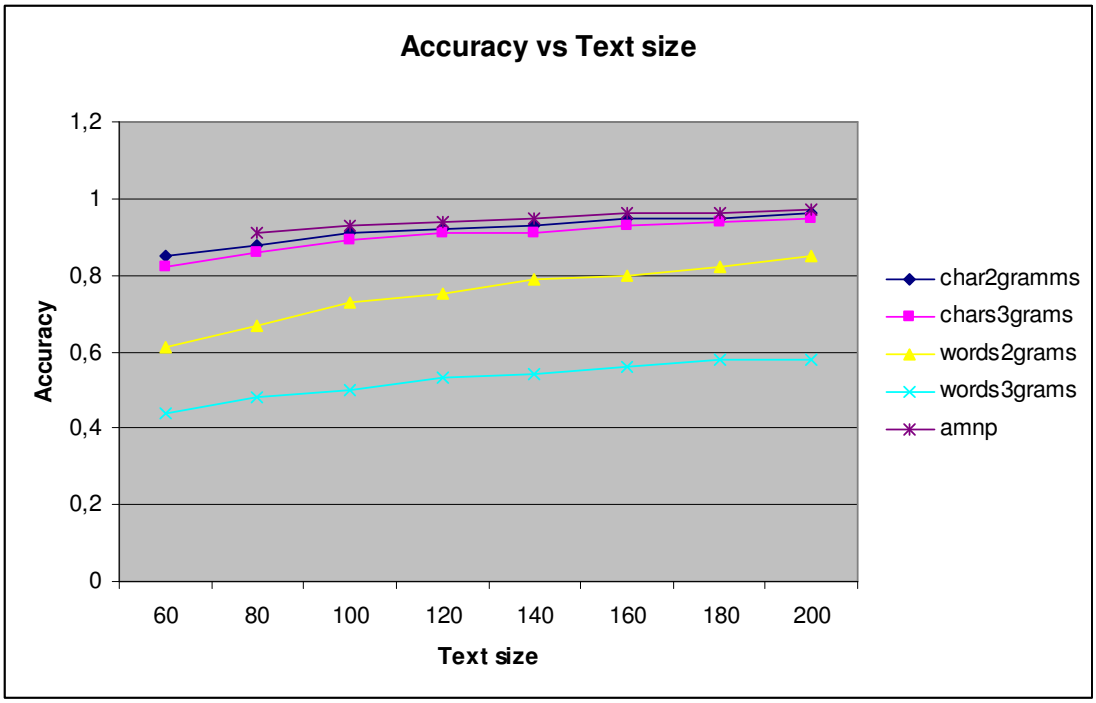
Εικόνα 7: Η ακρίβεια (accuracy) σε σχέση με το μέγεθος των κειμένων και τον τύπο του n-gram

Επιπλέον, στα πειράματα μας δημιουργήσαμε το AMNP (*Author Multilevel N-gramm Profile*) δηλαδή χαρακτηριστικά (*features*) ν-γραμμάτων με αυξανόμενο κειμενικό μέγεθος και γλωσσολογική μονάδα. Συγκεκριμένα εξαγάγαμε τα 1000 πιο συχνά διγράμματα και τριγράμματα χαρακτήρων και λέξεων και δημιουργήσαμε ένα διάνυσμα με 4000 χαρακτηριστικά. Αυτό το διάνυσμα στην συνέχεια το τροφοδοτήσαμε στο SVM- αλγόριθμο κατηγοριοποίησης και για να αποτιμήσουμε την απόδοση του χρησιμοποιήσαμε 10 fold cross validation.

Αξίζει να σημειωθεί ότι για κειμενικό μέγεθος των 200 λέξεων με τη μέγιστη τιμή ακρίβειας 0,97 για 10-fold cross validation είναι μια πολύ καλή ένδειξη ότι η γλωσσολογική δομή των tweets φέρει πληροφορία για την πατρότητα τους. Ακόμα για κειμενικό μέγεθος των 100 λέξεων και πάνω μπορούμε να πούμε ότι έχουμε μια ικανοποιητική ακρίβεια δεδομένου ότι χρησιμοποιούμε μόνο ν-γράμματα και δεν κάνουμε χρήση άλλης πληροφορίας όπως θέματος (*topic*), χρόνου αποστολής (*time*) - γενικότερα timestamp - και κοινωνικού δικτύου (*social network*) του κάθε χρήστη. Τα αποτελέσματα αυτά παρατίθενται παρακάτω:

Batch size	AMNP
80	0,91(+/-0.01)
100	0,93(+/-0.02)
120	0,94(+/-0.02)
140	0,95(+/-0.01)
160	0,96(+/-0.01)
180	0,96(+/-0.02)
200	0,97(+/- 0.02)

Πίνακας 8: Η Ακρίβεια (*Accuracy*) για το AMNP



Εικόνα 8: Επίδραση των γλωσσικών χαρακτηριστικών και κειμενικού μεγέθους στην ακρίβεια της πατρότητας χρησιμοποιώντας 10-fold cross validation

6. Μεθοδολογικές βελτιώσεις

Αρκετές φορές θεωρούμε κάποια πράγματα δεδομένα και δεν υπολογίζουμε τον ανθρώπινο παράγοντα. Από παρατηρήσεις που έγιναν διαπιστώθηκε ότι υπάρχουν χρήστες, οι οποίοι αρκετές φορές γράφουν κάτι λανθασμένα ή με ελλείψεις. Παρατηρήθηκε ότι ορισμένοι χρήστες στις παραπομπές τους σε URLs είτε δεν παρέθεταν τον σύνδεσμο όπως θα έπρεπε στην πλήρη μορφή, είτε κατά την διαδικασία της επικόλλησης στο μήνυμα του tweet αντέγραφαν το σύνδεσμο με λιγότερα γράμματα.

Υπήρχαν δηλαδή περιπτώσεις όπου ο σύνδεσμος αντί, για παράδειγμα, να είναι “http://news.in.gr/world/article/?aid=1231297117” ήταν της μορφής «ttp://news.in.gr/world/article/?aid=1231297117” · έλειπε δηλαδή το πρώτο γράμμα από το πρωτόκολλο. Άλλες φόρες παρέθεταν συνδέσμους σε μορφή για παράδειγμα:”eirinika.gr” ή “eirinika .gr” αφήνοντας δηλαδή κενό πριν το top-level domain. Ακόμα κάποιοι παρέθεταν τους συνδέσμους αυτούς χωρίς κενά με τις προτάσεις που προηγούνταν. .Σε αυτές τις περιπτώσεις δε λάβαμε υπόψη τέτοιου είδους λάθη με αποτέλεσμα να μην καθαριστεί το κείμενο μας από αυτές τις παραπομπές στον επιθυμητό βαθμό.

Εάν ληφθεί υπόψη ότι στο παρόν σώμα κειμένων που εξετάσαμε των 32 χρηστών τα URLs που αντικαταστήσαμε ήταν κοντά στις είκοσι χιλιάδες με έναν έλεγχο με regular expressions τέτοιου είδους λάθη, τα οποία δεν αφαιρέθηκαν είναι περίπου 700 στον αριθμό. Πιθανό αποτέλεσμα της αφαίρεσης αυτής θα μπορούσε να ήταν η αλλαγή στη συχνότητα των 1000 πρώτων εμφανίσεων των ν-γραμμάτων.

7. Συμπερασματική Επισκόπηση

Η συγγραφική πατρότητα αποτελεί μια νέα επιστημονική μέθοδο που αποσκοπεί στην απόδοση ενός κειμένου στον συγγραφέα του. Ως μέθοδος οφείλει να εξετάζει και να παρέχει επιστημονικά αποτελέσματα και ακρίβεια. Η αξιοποίηση της μεθόδου στα κοινωνικά δίκτυα και γενικότερα στο χώρο του διαδικτύου μπορεί να συμβάλει στην πρόληψη εγκληματικών ενεργειών σε κοινωνικό επίπεδο. Σε επιστημονικό επίπεδο μπορεί να εναρμονίσει τη μεταβολή στον τομέα της επικοινωνίας με την εξέλιξη της υφομετρίας και της χρήσης της στατιστικής. Η βιβλιογραφική ανασκόπηση προσφέρει πληθώρα γνώσεων για τα υφομετρικά χαρακτηριστικά που χρησιμοποιούνται στις έρευνες. Ωστόσο, το ελλιπές θεωρητικό υπόβαθρο δημιουργεί την αδυναμία για γλωσσολογική ανάλυση των δεδομένων και την αναγωγή τους σε ένα γλωσσολογικό μοντέλο που να λειτουργεί με τρόπο ερμηνευτικό για τα στατιστικά αποτελέσματα.

Στη συγκεκριμένη ερευνητική προσπάθεια δόθηκε έμφαση στα νέα υφομετρικά χαρακτηριστικά που έχουν ήδη εξεταστεί στη βιβλιογραφία ενώ προστέθηκαν χαρακτηριστικά τα οποία εξάγονται μέσω της γλωσσολογικής παρατήρησης της ελληνικής γλώσσας. Παράλληλα, ο κατακερματισμός του όγκου των δεδομένων σε μικρότερες ομάδες των 20 έως 200 λέξεων έδειξε ότι η ακρίβεια αυξάνει όταν αυξάνουν τα πακέτα λέξεων ανά συγγραφέα. Είναι σημαντικό, λοιπόν, σε κάθε έρευνα συγγραφικής απόδοσης, να αναγνωρίζουμε τη θετική επίδραση των κειμενικών τεμαχίων. Επιπρόσθετα, ο καθορισμός της ιδιότητας του κάθε χρήστη συμβάλλει στην αξιολόγηση των αποτελεσμάτων των ν-γραμμάτων καθώς είναι προφανές ότι η χρήση του Twitter αποτελεί μέθοδο επικοινωνίας με σκοπό την προώθηση ιδεών, ιστοσελίδων και προσωπικών επιδιώξεων.

Η ακρίβεια των αποτελεσμάτων είναι ενθαρρυντική για μελλοντικά πειράματα, στα οποία θα ληφθούν υπόψη οι διορθώσεις στο ήδη υπάρχον σώμα κειμένων, ενώ θα προστεθούν και νέα υφολογικά χαρακτηριστικά από τον λεξιλογικό και σημασιολογικό τομέα.

8. Μελλοντικές Προεκτάσεις

Τα αποτελέσματα της παρούσας εργασίας είναι ενθαρρυντικά για την απόδοση της συγγραφικής πατρότητας που στηρίζεται μόνο σε γλωσσολογικά χαρακτηριστικά. Η μεθοδολογία που ακολουθήθηκε με τον κατακερματισμό του σώματος κειμένων σε πακέτα λέξεων και με τη χρήση των ν-γραμμάτων έδειξε ότι για τα κλειστού τύπου ερωτήματα, δεν υφίσταται ανάγκη για ανάλυση εξωγλωσσικών παραγόντων. Ωστόσο, για την βέλτιστη απόδοση και σε προβλήματα ανοικτού τύπου (απόδοση συγγραφικής πατρότητας χωρίς να παραπέμπουμε σε συγκεκριμένη ομάδα συγγραφέων) μπορούμε μελλοντικά να εντάξουμε στην μεθοδολογία τις ακόλουθες προτάσεις.

- Η δημιουργία νέων υφομετρικών χαρακτηριστικών που να λαμβάνουν υπόψη τις κοινωνιογλωσσολογικές δομές που αναπτύσσονται στον χώρο του διαδυκτίου μπορούν να συντελέσουν στην απόδοση της συγγραφικής πατρότητας. Τα υφομετρικά χαρακτηριστικά που παρουσιάζει η μέχρι τώρα βιβλιογραφία δεν μπορούν να ανταποκριθούν στις ανάγκες της ανάλυσης του λόγου που αναπτύσσεται στα κοινωνικά δίκτυα. Λαμβάνοντας υπόψη την ανάλυση των περιστάσεων επικοινωνίας του Hymes (1974), οφείλουμε σε κάθε πραγμάτωση του λόγου να εξετάζουμε παράγοντες όπως το φυσικό περιβάλλον, την σκηνή, τους μετόχους, τον σκοπό, το κλειδί, τα κανάλια, τις νόρμες αλληλεπίδρασης και τα είδη του λόγου. Οι νέοι ρόλοι που καλείται να υποστηρίξει ο χρήστης στο περιβάλλον της διεπαφής και η ανάπτυξη ενός μεστού και συνοπτικού λόγου συμβάλλουν στη μετατροπή και στην εξέλιξη των γλωσσικών δομών που χρησιμοποιεί. Οι λεξιλογικές επιλογές που κάνει ο χρήστης εντάσσονται στην επικοινωνιακή περίσταση στην οποία πραγματώνεται ο λόγος. Επομένως, η ανάπτυξη ενός λεξικού με θέματα (*topics*) και ρόλους (*actors*) μπορεί να συντελέσει στην αύξηση της ακρίβειας.
- Αν και η παρούσα εργασία θέλησε να εξετάσει την ακρίβεια γλωσσολογικών και μόνο χαρακτηριστικών, σε οποιαδήποτε άλλη περίπτωση είναι ορθό να λαμβάνονται υπόψη και εξωγλωσσικά στοιχεία, τα οποία σχετίζονται με την ιδιοσυγκρασία του συγγραφέα. Σε οποιοδήποτε επικοινωνιακή περίσταση, η

οποία λαμβάνει χώρα στο χώρο του διαδυκτίου, ο χρήστης μπορεί να καταθέσει το "ίδιον" στίγμα του, είτε με την χρήση μορφοποιητικών ενεργειών όπως η χρήση της γραμματοσειράς ή η στοίχιση του κειμένου, ή το χρώμα που χρησιμοποιεί κ.α. Σε μελλοντικές έρευνες στο κλάδο της υφομετρίας τα συγκεκριμένα χαρακτηριστικά θα μπορέσουν να ενταχθούν στα μορφοποιητικά χαρακτηριστικά ενός κειμένου.

Χαρακτηριστικά	
Κειμενικά	γλωσσολογικά
	μορφοποιητικά
Εξωκειμενικά	στοιχεία για το συγγραφέα(φύλο, ηλικία, κύκλος επαφών, ώρα)

- Η γενίκευση της ερευνητικής διαδικασίας και σε άλλες ιστοσελίδες και σε άλλα μέσα κοινωνικής δικτύωσης όπως το Facebook και IRC chat εκτός από το Twitter θα συνέβαλλε αποφασιστικά στην προσπάθεια που γίνεται εκ μέρους των διωκτικών αρχών για την καταπολέμηση του ηλεκτρονικού εγκλήματος, το οποίο λαμβάνει ανεξέλεγκτες διαστάσεις.
- Απώτερος στόχος όλων των ερευνητικών προσπαθειών θα είναι η δημιουργία ενός καθολικού πίνακα χαρακτηριστικών, που ανάλογα με το κειμενικό γένος στο οποίο εξετάζεται η συγγραφική πατρότητα, να προσδιορίζονται τα χαρακτηριστικά, τα οποία θα συντελέσουν στην απόδοση της συγγραφικής πατρότητας. Ένας καθολικός πίνακας που θα αξιοποιεί και κοινωνιογλωσσολογικά δεδομένα μπορεί να εφαρμοσθεί σε όλες τις γλώσσες και να αξιοποιηθεί και για ιστορικο-μεθοδολογικούς σκοπούς δείχνοντας τη συγγένεια των γλωσσών και την ομοιόμορφη εξέλιξή τους.

Η εξέλιξη της υφομετρίας και η χρήση της στατιστικής μπορούν να αναγάγουν την απόδοση της συγγραφικής πατρότητας σε σημαντικό παράγοντα, όπου η ανωνυμία δεν θα είναι πια δυνατή, ενώ ταυτόχρονα θα μπορεί να αξιοποιηθεί και στη μελέτη ιστορικών έργων στα οποία δεν έχει αποδοθεί ακόμα ο συγγραφέας τους.

Βιβλιογραφία

Bailey, R. W. (1979), *Author attribution in a forensic setting*, ιn D. E. Ager, F. E. Knowles & J.Smith (Eds.), *Advances in Computer-aided Literary and Linguistic Research*, Birmingham: AMLC

Boutwell, S. R. (2011), *Authorship attribution of short messages using multimodal features.* .D. Dissertation, Naval Postgraduate School, Monterey, California.

Dabagh R. M. (2007), *Authorship Attribution and Statistical Text Analysis*, Metodoloski zvezki, 4, 149-163.

de Morgan, A. (1851/1882) Letter to Rev. Heald 18/08/1851. In Elizabeth & de Morgan (Eds.), *Memoirs of Augustus de Morgan by his wife Sophia Elizabeth de Morgan with selections from his Letters*, London: Longman' s Green and Co

DeVito Joseph A. (1967), *Style and stylistics: An attempt at definition*, Quarterly Journal of Speech.

Diller H. (1998), Stylistics: Linguistic and textual, European Journal of English Studies, σελ.155-156.

Frantzeskou, G., E. Stamatatos, S. Gritzalis, C.E. Chaski, and B.S. Howald (2007*), Identifying Authorship by Byte-Level N-Grams: The Source Code Author Profile (SCAP) Method*, Int. Journal of Digital Evidence, 6(1)

Holmes, D. I. (1985), The analysis of literary style: A review, Journal of the Royal Statistical Society, Series A, 148

Koppel, M., Schler, J. & Argamon, S. (2011) *Authorship attribution in the Wild*. Language Resources & Evaluation 45.

Layton, R., Watters, P., Dazeley, R. (2010) *Authorship attribution for twitter in 140characters or less*, In Workshop Cybercrime and Trustworthly Computing Workshop (CTC), Ballarat, Australia, 1-8.

Mendenhall, T.C. (1887), *The characteristic curves of composition*, Science, 9(214), 237-249. Mendenhall, T.C. (1901), *A mechanic solution of a literary problem*, The popular Science Monthly, 9, 97-105.

Mikros, George K. & Perifanos, K. (2013) *Authorship Attribution in Greek Tweets Using Author's Multilevel N-Gram Profiles*, AAAI Spring Symposium: Analyzing Microtext

Mikros, George K., & Perifanos, K. (2011) *Authorship identification in large email collections: Experiments using features that belong to different linguistic levels Proceedings of PAN 2011 Lab*, Uncovering Plagiarism, Authorship, and Social Software Misuse held in conjunction with the CLEF 2011 Conference on Multilingual and Multimodal Information Access Evaluation, 19-22 September 2011, Amsterdam

Mosteller Fre. & Wallace L. Dav. (1963) *Inference in an Authorship Problem*, Journal of the American Statistical Association, Volume 58, Issue 302, 275-309.

Peng, F., D., Shuurmans, and S., Wang. (2004) *Augmenting naive Bayes classifiers with statistical language models,* Information Retrieval Journal, 7(1): 317-345.

Saussure, Ferdinand de ([1915] 1974): *Course in General Linguistics* London: Fontana/Collins

Silva, S., Sarmento, R., Grant, L., Oliveira, T., E.C., Maia, B. (2012) *Comparing sentence-level features for authorship analysis in Portuguese*, in: PROPOR, 51-54.

Todorov, T. (1964), Theorie de la litterature, Textes der formalistes russes, Παρίσι, Seuil.

Yule G.U. (1938/1939) *On sentence length as a statistical characteristic of style in prose : With application to two cases of disputed authorship*, Biometrika, 30, 363-369.

Webster T (2010) *Twitter usage in America*. Edison Research. Available at: http://www.edisonresearch.com/twitter_usage_2010.php (accessed 7 March 2011).

Williams, C.B. (1975) *Mendenhall's studies of word-length distribution in the work of Shakespeare and Bacon*, Biometrika, 62, 207-212.

Διονυσίου Λογγίνου, *Περί Ύψους*: ερμηνευτική έκδοση Μ.Ζ. Κοπιδάκης (1990), Βικελαία Δημοτική Βιβλιοθήκη: Ηράκλειο

Γεωργακοπούλου Αλ. & Γούτσος Διον.(2008[7]), *Κείμενο και Επικοινωνία*, Ελληνικά Γράμματα: Αθήνα

Μικρός Κ. Γ. (2012), *Αυτόματος Εντοπισμός Συγγραφέα: Μέθοδοι και εργαλεία υφομετρικής απόδοσης συγγραφικής πατρότητας*.

Jakobson, R.(1983), *Τα Μεγάλα Ρέματα της Γλωσσολογίας*, (μτφρ) Δ. Σωτηρόπουλος, Εκδόσεις Περιοδικού "Θεσσαλική Εστία": Αθήνα

Natural language processing with Python, *Analyzing Text with the Natural Language Toolkit* , Steven Bird, Ewan Klein, and Edward Loper, O'Reily Media 2009, ISBN: 978-0-596-51649-9, Διαθέσιμο online: http://www.nltk.org/book/

Mining the social web, Matthew A. Russell, O'Reilly Media October 2013