

ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ
ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ ΚΑΙ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ



-Πτυχιακή Εργασία-

**Ανάπτυξη εργαλείου αναγνώρισης μερών του λόγου για την
ελληνική γλώσσα του κοινωνικού ιστού**

Μαρία Νεφέλη Νικηφόρου (Π2015054)

Επιβλέπουσα: Κάτια Λήδα Κερμανίδου

Σεπτέμβριος 2019

Επιβλέπουσα

Κάτια Αήδα Κερμανίδου, Μόνιμη Επίκουρη Καθηγήτρια

Ιόνιο Πανεπιστήμιο, Τμήμα Πληροφορικής

Τριμελής Επιτροπή

Κάτια Αήδα Κερμανίδου, Μόνιμη Επίκουρη Καθηγήτρια

Ιόνιο Πανεπιστήμιο, Τμήμα Πληροφορικής

Φοίβος Μυλωνάς, Αναπληρωτής Καθηγητής

Ιόνιο Πανεπιστήμιο, Τμήμα Πληροφορικής

Θεόδωρος Ανδρόνικος, Αναπληρωτής Καθηγητής

Ιόνιο Πανεπιστήμιο, Τμήμα Πληροφορικής

Περίληψη

Ο μεγάλος όγκος επικοινωνίας μέσω μηνυμάτων τύπου microblogging στα κοινωνικά δίκτυα έχει δημιουργήσει την ανάγκη για αποδοτικά εργαλεία Natural Language Processing (NLP), ειδικά στην επεξεργασία αδόμητου κειμένου. Η εξαγωγή πληροφοριών από social text αποτελεί μία από τις πιο απαιτητικές εργασίες NLP. Στην παρούσα έρευνα αναπτύχθηκε το πρώτο μέχρι στιγμής εργαλείο αναγνώρισης μερών του λόγου για την ελληνική γλώσσα του κοινωνικού ιστού. Η καινοτομία της έρευνας είναι εμφανής σε κάθε στάδιό της, από τη συλλογή και την επισημείωση ως προς τα μέρη του λόγου των δεδομένων (2405 tweets), έως τα αποτελέσματα των δοκιμών με τους αλγορίθμους επιβλεπόμενης μηχανικής μάθησης. Το εργαλείο αυτό, λόγω της υψηλής απόδοσής του στις δοκιμές με τους αλγορίθμους Naive Bayes (accuracy 99.87%) και ID3 (accuracy 99.44%), έχει προοπτική να χρησιμοποιηθεί ως βάση για μελλοντική έρευνα στο πεδίο της αναγνώρισης μερών του λόγου για δεδομένα τύπου microblogging στα ελληνικά από το κοινωνικό δίκτυο Twitter, αλλά και από άλλα κοινωνικά δίκτυα.

Λέξεις-κλειδιά: Αναγνώριση μερών του λόγου, Επεξεργασία φυσικής γλώσσας, Γλώσσα κοινωνικού ιστού, Μηχανική μάθηση, Επιβλεπόμενη μάθηση, Κοινωνικά δίκτυα

Ευχαριστίες

Η παρούσα πτυχιακή εργασία εκπονήθηκε στο Τμήμα Πληροφορικής του Ιονίου Πανεπιστημίου, Σχολή Επιστημών της Πληροφορίας και Πληροφορικής στην Κέρκυρα.

Θα ήθελα να ευχαριστήσω ιδιαίτερα την επιβλέπουσα της πτυχιακής εργασίας, τη Μόνιμη Επίκουρη Καθηγήτρια Κάτια Λήδα Κερμανίδου, για τις πολύτιμες συμβουλές και τον χρόνο που αφιέρωσε στην υποστήριξή μου κατά την επιλογή του θέματος και την συγγραφή της εργασίας.

Θα ήθελα επίσης να ευχαριστήσω τα υπόλοιπα μέλη της επιτροπής αξιολόγησης, τον Αναπληρωτή Καθηγητή Φοίβο Μυλωνά και τον Αναπληρωτή Καθηγητή Θεόδωρο Ανδρόνικο.

Ακόμη, θα ήθελα να εκφράσω ένα μεγάλο ευχαριστώ στην ομάδα των εθελοντών επισημειωτών, των οποίων η συνεισφορά υπήρξε καθοριστικής σημασίας για τη διεξαγωγή αυτής της πτυχιακής εργασίας.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένειά μου, και ιδιαίτερα τον πατέρα μου Στέφανο και την μητέρα μου Ευαγγελία, για την αμέριστη συμπαράσταση, την εποικοδομητική κριτική και τις συμβουλές τους.

Κέρκυρα, Σεπτέμβριος 2019

Μαρία Νεφέλη Νικηφόρου

Περιεχόμενα

Περίληψη.....	4
Ευχαριστίες	5
Κατάλογος πινάκων	8
Κατάλογος εικόνων.....	9
Κατάλογος διαγραμμάτων	10
Κεφάλαιο 1 Εισαγωγή.....	12
1.1 Μέρη του λόγου στην ελληνική γλώσσα	13
1.2 Γλώσσα κοινωνικών δικτύων - Twitter	14
1.3 Ερευνητικό ερώτημα - Σκοπός και στόχοι της έρευνας.....	15
Κεφάλαιο 2 Βιβλιογραφική επισκόπηση	17
2.1 Εργαλεία αναγνώρισης μερών του λόγου για την αγγλική γλώσσα του κοινωνικού ιστού.....	17
2.2 Εργαλεία αναγνώρισης μερών του λόγου για άλλες γλώσσες του κοινωνικού ιστού	19
Κεφάλαιο 3 Μηχανική μάθηση.....	22
3.1 Μάθηση με επίβλεψη - Μάθηση κατά Bayes.....	23
3.2 Μάθηση με επίβλεψη - Δέντρα απόφασης.....	24
Κεφάλαιο 4 Μεθοδολογία έρευνας	26
4.1 Συλλογή δεδομένων.....	26
4.2 Δημιουργία dataset	27
4.2.1 Κανονικοποίηση δεδομένων	27
4.2.2 Δημιουργία tag set	33
4.3 Επισημείωση.....	36
4.3.1 Διαδικασία	37
4.3.2 Παρατηρήσεις	39
Κεφάλαιο 5 Σύγκριση μοντέλων ταξινόμησης και αποτελέσματα	41

5.1 Αλγόριθμος Naive Bayes	43
5.1.1 Μεθοδολογία.....	43
5.1.2 Αποτελέσματα.....	45
5.2 Αλγόριθμος ID3.....	48
5.2.1 Μεθοδολογία.....	49
5.2.2 Αποτελέσματα.....	50
Κεφάλαιο 6 Συμπεράσματα	55
6.1 Συνεισφορά πτυχιακής εργασίας	55
6.2 Προτάσεις για μελλοντική έρευνα.....	62
Πηγές και βιβλιογραφία	64
Δικτυογραφία.....	66
Γραμματικές και λεξικά.....	67
Παράρτημα.....	68
Συντμήσεις	70
Γλωσσάρι ξενικών όρων	71

Κατάλογος πινάκων

Πίνακας 4.2.1.1 Συνολικός αριθμός εμφάνισης κάθε σημείου στίξης και αριθμός των tweets που περιείχαν ένα ή περισσότερα σημεία στίξης	29
Πίνακας 4.2.1.2 Συνολικός αριθμός εμφάνισης κάθε μαθηματικού ή άλλου συμβόλου και αριθμός των tweets που περιείχαν ένα ή περισσότερα σύμβολα	30
Πίνακας 4.2.1.3 Συνολικός αριθμός εμφάνισης κάθε χρονικής, ποσοτικής και αριθμητικής έκφρασης και αριθμός των tweets που περιείχαν μία ή περισσότερες εκφράσεις.....	32
Πίνακας 4.2.2.4 Tag set.....	34
Πίνακας 5.1.2.1 Τιμές των δεικτών precision, recall και F_1 score για κάθε label (Naive Bayes).....	47
Πίνακας 5.2.2.1 Τιμές των δεικτών precision, recall και F_1 score για κάθε label (ID3).....	52

Κατάλογος εικόνων

Εικόνα 4.2.1 Δείγμα τελικού dataset αποτελούμενο από tweets.....	27
Εικόνα 4.3.1 Δείγμα dataset διαμερισμένου σε tokens	36
Εικόνα 4.3.1.1 Δείγμα dataset μετά από τα στάδια της επισημείωσης και του ελέγχου. Στη στήλη A φαίνεται το token και στη στήλη B το tag που του αντιστοιχεί	37
Εικόνα 5.1 Focus word, η θέση του στο tweet και τα συμφραζόμενά του (γειτονικές λέξεις).....	42
Εικόνα 5.2 Δείγμα dataset με attributes και labels. Από αυτό προκύπτουν τα training sets και test sets για τη μηχανική μάθηση	42
Εικόνα 5.1.1.1 Δείγμα dataset μετά τη φόρτωσή του στο RapidMiner Studio Educational	43
Εικόνα 5.1.1.2 Δημιουργία και αξιολόγηση μοντέλου ταξινόμησης με βάση τον αλγόριθμο μηχανικής μάθησης "Naive Bayes" (επιβλεπόμενη μάθηση).....	45
Εικόνα 5.1.2.1 Confusion matrix αποτελεσμάτων για τον αλγόριθμο Naive Bayes	45
Εικόνα 5.2.1.1 Δημιουργία και αξιολόγηση μοντέλου ταξινόμησης με βάση τον αλγόριθμο μηχανικής μάθησης "ID3" (επιβλεπόμενη μάθηση)	50
Εικόνα 5.2.2.1 Confusion matrix αποτελεσμάτων για τον αλγόριθμο ID3.....	51
Εικόνα 6.1.1 Εντοπισμός των λάθος ταξινομημένων παραδειγμάτων (Naive Bayes)	59
Εικόνα 6.1.2a Δείγμα από τον πίνακα που παράγεται από τον operator "Explain Predictions"	59
Εικόνα 6.1.2b Δείγμα από τον πίνακα που παράγεται από τον operator "Explain Predictions"	60
Εικόνα 6.1.3a Δείγμα από τα παραδείγματα που επιστρέφει ο operator "Filter Examples"	60
Εικόνα 6.1.3b Δείγμα από τα παραδείγματα που επιστρέφει ο operator "Filter Examples"	61

Κατάλογος διαγραμμάτων

Διάγραμμα 4.2.1.1 Σημεία στίξης (token). Ποσοστά ανά σημείο στίξης πριν την αντικατάστασή τους από τη λέξη-κωδικό "ΣΗΣΤ"	29
Διάγραμμα 4.2.1.2 Σημεία στίξης (tweet). Ποσοστά tweets ανά σημείο στίξης πριν την αντικατάστασή τους από τη λέξη-κωδικό "ΣΗΣΤ"	30
Διάγραμμα 4.2.1.3 Μαθηματικά και άλλα σύμβολα (token). Ποσοστά ανά σύμβολο πριν την αντικατάστασή τους από τη λέξη-κωδικό "SYMB"	31
Διάγραμμα 4.2.1.4 Μαθηματικά και άλλα σύμβολα (tweet). Ποσοστά tweets ανά σύμβολο πριν την αντικατάστασή τους από τη λέξη-κωδικό "SYMB"	31
Διάγραμμα 4.2.1.5 Χρονικές, ποσοτικές και αριθμητικές εκφράσεις (token). Ποσοστά ανά έκφραση πριν την αντικατάστασή τους από τη λέξη-κωδικό "NUMB"	32
Διάγραμμα 4.2.1.6 Χρονικές, ποσοτικές και αριθμητικές εκφράσεις (tweet). Ποσοστά tweets ανά έκφραση πριν την αντικατάστασή τους από τη λέξη-κωδικό "NUMB"	33
Διάγραμμα 4.3.2.1 Συχνότητα εμφάνισης κάθε tag	40
Διάγραμμα 4.3.2.2 Ποσοστό εμφάνισης κάθε tag	40
Διάγραμμα 5.1.2.1 Γραφική απεικόνιση των τιμών του δείκτη precision για κάθε label (Naive Bayes).....	47
Διάγραμμα 5.1.2.2 Γραφική απεικόνιση των τιμών του δείκτη recall για κάθε label (Naive Bayes).....	48
Διάγραμμα 5.1.2.3 Γραφική απεικόνιση των τιμών του δείκτη F_1 score για κάθε label (Naive Bayes).....	48
Διάγραμμα 5.2.2.1 Γραφική απεικόνιση των τιμών του δείκτη precision για κάθε label (ID3).....	53
Διάγραμμα 5.2.2.2 Γραφική απεικόνιση των τιμών του δείκτη recall για κάθε label (ID3).....	53
Διάγραμμα 5.2.2.3 Γραφική απεικόνιση των τιμών του δείκτη F_1 score για κάθε label (ID3).....	54

Κεφάλαιο 1 Εισαγωγή

Ο μεγάλος όγκος επικοινωνίας μέσω μηνυμάτων τύπου microblogging στα κοινωνικά δίκτυα (όπως το Facebook, το Twitter, τα Blogs και το YouTube), έχει δημιουργήσει την ανάγκη για αποδοτικά εργαλεία Natural Language Processing (NLP), ειδικά στην επεξεργασία αδόμητου κειμένου. Έχει παρατηρηθεί ότι η απόδοση των εργαλείων NLP που έχουν αναπτυχθεί για επεξεργασία δομημένων κείμενων μειώνεται σημαντικά όταν χρησιμοποιούνται σε αδόμητα κείμενα τύπου microblogging (Nand et al., 2014).

Με την αύξηση της επικοινωνίας μέσω των κοινωνικών δικτύων, υπάρχει μια αυξανόμενη ζήτηση για εξόρυξη πληροφοριών από αδόμητο κείμενο για επιχειρηματική ευφυΐα, ασφάλεια, προγραμματισμό και άλλους σκοπούς. Ωστόσο, η εξαγωγή πληροφοριών σε αυτό το πλαίσιο αποτελεί μία από τις πιο απαιτητικές εργασίες NLP, εξαιτίας της ασυνήθιστης δομής του κειμένου (Nand et al., 2014). Επομένως, χρειάζεται είτε να προσαρμοστούν οι υπάρχουσες μέθοδοι για εφαρμογή στο συγκεκριμένο πλαίσιο, είτε να αναπτυχθούν νέες μέθοδοι που να έχουν καλή απόδοση σε αδόμητα κείμενα τύπου microblogging.

Η αναγνώριση μερών του λόγου, γνωστή και ως part-of-speech tagging (POS tagging), αποτελεί θεμελιώδες τμήμα της NLP (Gimpel et al., 2010). Ένα ισχυρό εργαλείο POS tagging διαδραματίζει σημαντικό ρόλο στα περισσότερα προβλήματα και εφαρμογές NLP, όπως η συντακτική και η σημασιολογική ανάλυση, η μηχανική μετάφραση και η ανάλυση αισθήματος (sentiment analysis), (Bach et al., 2018). Οι κύριες προκλήσεις στο POS tagging είναι οι εξής: η δημιουργία του tag set, με βάση το οποίο θα γίνεται η απόδοση του μέρους του λόγου σε κάθε λέξη, το οποίο χρειάζεται να καλύπτει και τις ιδιομορφίες του κοινωνικού δικτύου (βλ. Κεφάλαιο 1, Ενότητα 1.2), η απόδοση του ορθού tag σε αμφιλεγόμενου τύπου λέξεις, και η δημιουργία ενός επισημειωμένου dataset επαρκούς μεγέθους για τα καλύτερα δυνατά αποτελέσματα στη μηχανική μάθηση.

Στόχος της εργασίας αυτής είναι να ανατεθεί σε κάθε λέξη ενός αδόμητου ή ημιδομημένου κειμένου ένα POS tag ("ουσιαστικό", "ρήμα", "επίρρημα" κ.λπ.), ανάλογα με τα συμφραζόμενά της, δημιουργώντας έτσι ένα dataset που θα αποτελέσει τη βάση για την ανάπτυξη ενός αποδοτικού εργαλείου αναγνώρισης μερών του λόγου για την ελληνική γλώσσα του κοινωνικού ιστού.

Στην Ενότητα 1.1 παρουσιάζεται συνοπτικά ο ρόλος των μερών του λόγου στην ελληνική γλώσσα. Στην Ενότητα 1.2 περιγράφονται η μορφή και η δομή της γλώσσας του κοινωνικού ιστού, με έμφαση στο κοινωνικό δίκτυο Twitter. Στην Ενότητα 1.3 τίθεται το ερευνητικό ερώτημα της παρούσας έρευνας, ενώ καθορίζονται και ο σκόπος και οι στόχοι της.

1.1 Μέρη του λόγου στην ελληνική γλώσσα

Η ελληνική γλώσσα αποτελείται από διάφορα μέρη του λόγου, τα οποία χωρίζονται σε κλιτά και άκλιτα (Τριανταφυλλίδης, 1990), (Τζεβελέκου, Κάντζου, & Σταμούλη, 2007). Στα κλιτά μέρη του λόγου ανήκουν τα ουσιαστικά, τα ρήματα, οι αντωνυμίες, τα επίθετα, οι μετοχές και τα άρθρα. Κάθε κλιτό μέρος του λόγου αποτελείται από το θέμα και την κατάληξη, όπου το θέμα παραμένει σταθερά το ίδιο, ενώ η κατάληξη αλλάζει. Για τα ουσιαστικά, τα επίθετα, τα άρθρα και τις αντωνυμίες, οι καταλήξεις αλλάζουν ανάλογα με το γένος, τον αριθμό και την πτώση. Για τα ρήματα, οι καταλήξεις αλλάζουν ανάλογα με το πρόσωπο, το χρόνο, την έγκλιση, τη φωνή και τον αριθμό. Στα άκλιτα μέρη του λόγου ανήκουν οι σύνδεσμοι, τα επιρρήματα, οι προθέσεις, τα μόρια και τα επιφωνήματα. Τα άκλιτα μέρη του λόγου δεν αλλάζουν μορφή.

Τα μέρη του λόγου αποτελούν βασικά δομικά στοιχεία της ελληνικής γλώσσας και έχουν σημαντικό ρόλο στη σύνταξη προτάσεων και κειμένου, γενικότερα. Έχουν αναπτυχθεί ορισμένα εργαλεία για αναγνώριση μερών του λόγου για τη δόκιμη και επίσημη ελληνική γλώσσα. Ένα από αυτά είναι ο "Ελληνικός επισημειωτής μερών του λόγου" (<http://nlp.cs.aueb.gr/software.html>), ο οποίος έχει αναπτυχθεί από την Ομάδα Επεξεργασίας Φυσικής Γλώσσας του Τμήματος Πληροφορικής του Οικονομικού Πανεπιστημίου Αθηνών. Ο επισημειωτής αυτός επιχειρεί να προσδιορίσει αυτόματα το μέρος του λόγου (π.χ. ουσιαστικό, επίθετο, ρήμα κ.λπ.) κάθε λέξης σε ελληνικά κείμενα. Επισημειώνονται, επιπλέον, πρόσθετες πληροφορίες, όπως το γένος, ο αριθμός και η πτώση κάθε ουσιαστικού, η φωνή, ο χρόνος και ο αριθμός κάθε ρήματος κ.λπ. Ένα άλλο εργαλείο είναι ο "Greek part of speech tagger" (<http://php-nlp-tools.com/posts/greek-pos-tag.html>) που βασίζεται στην έρευνα της Ευαγγελίας Κολέλη. Το tag set του αποτελείται από τα εξής tags: verb, noun, adjective, adverb, article, pronoun, numeral, preposition, particle, conjunction, punctuation, other.

Οι Sgarbas et al. (2001) χρησιμοποιούν POS tagging ως κομμάτι της μορφολογικής ανάλυσης για τις δοκιμές τους με δεδομένα στα ελληνικά. Ωστόσο, εκείνοι χρησιμοποιούν πιο λεπτομερή tags, όπου έχει σημασία όχι μόνο το μέρος του λόγου του token, αλλά και η υποκατηγορία-τύπος στην οποία ανήκει, π.χ. η επισημείωση ως άρθρο δεν αρκεί και χρειάζεται να οριστεί και αν είναι οριστικό, αόριστο ή εμπρόθετο. Οι Petasis et al. (2001) σε έρευνά τους αναπτύσσουν έναν μορφολογικό αναλυτή με βάση τα λήμματα των λέξεων, για να χρησιμοποιηθεί ως βάση στην ανάπτυξη ενός εργαλείου ελέγχου ορθογραφίας για τα ελληνικά. Το tag set τους είναι το ίδιο λεπτομερές με των Sgarbas et al. (2001). Πέρα από τα μορφολογικά χαρακτηριστικά κάθε token (λήμμα και κατάληξη), λαμβάνει υπόψιν του και τα συμφραζόμενα. Το accuracy του POS tagger τους είναι 95%. Ακόμη, οι Papageorgiou et al. (2000) ανέπτυξαν έναν POS tagger για δόκιμη και επίσημη ελληνική γλώσσα. Οι δοκιμές έγιναν με 447000 tokens από οικονομικές αναφορές, ειδησεογραφικά πρακτορεία και τύπο, και εγχειρίδια οδηγιών και tag set μεγέθους 584 tags. Ως αποτέλεσμα, έχει accuracy 96.28%

Τέλος, σε μια πιο πρόσφατη έρευνα, ο Keersmaekers (2019) κάνει μια πρώτη απόπειρα αυτόματης επισημείωσης ολόκληρου του σώματος κειμένων που αποτελείται από παπύρους στα Ελληνικά. Αντιμετώπισε ιδιαιτερότητες όπως κείμενα ιστορικής φύσης και ιδιαίτερα κλιτή γλώσσα, ενώ επισημείωσε το σώμα κειμένων με βάση το μέρος του λόγου τους και τη μορφολογία τους, και έκανε δοκιμές συντακτικής ανάλυσης.

Ωστόσο, δεν έχει αναπτυχθεί μέχρι στιγμής κάποιο αντίστοιχο εργαλείο για την ελληνική γλώσσα του κοινωνικού ιστού. Από τη βιβλιογραφία για άλλες γλώσσες πλην της ελληνικής (Gimpel et al., 2010), (Foster et al., 2011), (Liu et al., 2012), (Tommasel & Godoy, 2018), (Zalmout et al., 2018), γίνεται εμφανές ότι τα εργαλεία που είναι φτιαγμένα για δόκιμη και επίσημη γλώσσα, δεν είναι αποδοτικά όταν καλούνται να κάνουν αναγνώριση μερών του λόγου σε ανεπίσημη, αδόμητη ή ημιδομημένη γλώσσα και σε κείμενα κοινωνικών δικτύων (βλ. Κεφάλαιο 1, Ενότητα 1.2 και Ενότητα 1.3).

1.2 Γλώσσα κοινωνικών δικτύων - Twitter

Η ολοένα αυξανόμενη δημοτικότητα των κοινωνικών δικτύων οδηγεί στην ραγδαία ανάπτυξη διαδικτυακού περιεχομένου που δημιουργείται από χρήστες (user-

created web content), έχοντας ως αποτέλεσμα την παραγωγή τεράστιας ποσότητας κειμένου σε ηλεκτρονική μορφή (Gimpel et al., 2010).

To Twitter (<https://twitter.com/>) αποτελεί ένα από τα πιο δημοφιλή κοινωνικά δίκτυα (500 εκατομμύρια tweets ανά ημέρα), συνδυάζοντας το microblogging και την κοινωνική δικτύωση (Foster et al., 2011). Μέσω αυτού, οι χρήστες μπορούν να αναρτούν σύντομα μηνύματα (tweets) - μήκους έως 140 χαρακτήρες - στους ακολούθους τους (followers), οι οποίοι είναι άλλοι χρήστες που έχουν συνδεθεί προηγουμένως (κοινωνική δικτύωση) με τους ακολουθούμενους (followees). Τα tweets εμφανίζονται δημόσια στη σελίδα-προφίλ των followees (microblogging).

Ορισμένες ιδιαιτερότητες αυτού του κοινωνικού δικτύου είναι οι ακόλουθες (Foster et al., 2011): τα tweets προωθούνται εύκολα σε όλους τους followers με ένα retweet. Το σύμβολο "@" πριν από ένα όνομα χρήστη, δημιουργεί έναν σύνδεσμο στη σελίδα προφίλ του χρήστη. Ένα tweet που περιέχει έναν τέτοιο σύνδεσμο ονομάζεται αναφορά (mention). Οι αναφορές χρησιμοποιούνται επίσης για στοχευμένες απαντήσεις σε χρήστη οι οποίες, κατά σύμβαση, αρχίζουν με "@user". Ένα άλλο ειδικό σύμβολο στο Twitter είναι το hashtag (#), το οποίο σηματοδοτεί λέξεις-κλειδιά που κατηγοριοποιούν το tweet σε κάποιο θέμα (topic).

Επιπλέον, η γλώσσα που χρησιμοποιείται στο Twitter παρουσιάζει τα ακόλουθα ιδιόμορφα χαρακτηριστικά (Gimpel et al., 2010), (Liu et al., 2012), (Nand et al., 2014), (Bach et al., 2018), (Tommasel & Godoy, 2018): μη συμβατική ορθογραφία, συντομευμένο κείμενο και συντομογραφίες (λόγω του ορίου των 140 χαρακτήρων), ασυνέπεια στη χρήση κεφαλαίων (χρήση για έμφαση, και όχι απαραίτητα για κύρια ονόματα), παράλειψη τόνων, emojis και emoticons, ξένες λέξεις (συνήθως αγγλικές), greeklish, αργκό, ελλειπτικές προτάσεις.

1.3 Ερευνητικό ερώτημα - Σκοπός και στόχοι της έρευνας

Ως social text (κείμενο κοινωνικού ιστού), θεωρούνται τα tweets στο Twitter, τα μηνύματα SMS, οι δημοσιεύσεις στο Facebook, τα αρχεία καταγραφής συνομιλιών, μηνύματα ηλεκτρονικού ταχυδρομείου κ.λπ. Το social text διαφέρει σημαντικά από το δόκιμο κείμενο, λόγω του ανεπίσημου τρόπου γραφής και τη χρήση καθομιλουμένης γλώσσας. Το social text αποτελεί σημαντική πηγή πληροφοριών για πολλές εφαρμογές NLP, όπως η εξαγωγή, ανάκτηση και σύνοψη πληροφορίας, αλλά και η sentiment analysis (Liu et al., 2012).

Ωστόσο, τα υπάρχοντα συστήματα συχνά δεν έχουν καλή απόδοση στον τομέα αυτό, λόγω της εκτεταμένης χρήσης ανεπίσημης γλώσσας, ελλιπών και ορθογραφικά και συντακτικά μη ορθών προτάσεων, της πολυμορφίας και της συνδεδεμένης φύσης των κοινωνικών δεδομένων (Gimpel et al., 2010), (Liu et al., 2012), (Tommasel & Godoy, 2018), (Zalmout et al., 2018). Ειδικότερα, η ελληνική γλώσσα του κοινωνικού ιστού χαρακτηρίζεται από τη χρήση αγγλικών και άλλων ξένων λέξεων, greeklish, αλλά και την παράλειψη των τόνων.

Το πρόβλημα της προσαρμογής των εργαλείων NLP σε ειδικότερους τομείς είναι ευρέως διαδεδομένο στη βιβλιογραφία της μηχανικής μάθησης (Foster et al., 2011). Μπορεί να συγκεντρωθεί και να επισημειωθεί επαρκής ποσότητα δεδομένων στον ειδικό αυτό τομέα; Είναι δυνατόν να αναπτυχθεί ένα εργαλείο, το οποίο να εκπαιδευτεί με αυτό το επισημειωμένο dataset, και να έχει υψηλή απόδοση;

Επομένως, προκύπτει η ανάγκη για την ανάπτυξη ενός εργαλείου αναγνώρισης μερών του λόγου για την ελληνική γλώσσα του κοινωνικού ιστού, το οποίο να είναι αποδοτικό παρά τις ιδιαιτερότητες της γλώσσας αυτής και να είναι σε θέση να αντιμετωπίζει τις κύριες προκλήσεις στο POS tagging σε αυτό το πλαίσιο. Οι κύριες προκλήσεις στο POS tagging είναι οι εξής: η δημιουργία του tag set, με βάση το οποίο θα γίνεται η απόδοση του μέρους του λόγου σε κάθε λέξη, το οποίο χρειάζεται να καλύπτει και τις ιδιομορφίες του κοινωνικού δικτύου (βλ. Κεφάλαιο 1, Ενότητα 1.2), η απόδοση του ορθού tag σε αμφιλεγόμενου τύπου λέξεις, και η δημιουργία ενός επισημειωμένου dataset επαρκούς μεγέθους για τα καλύτερα δυνατά αποτελέσματα στη μηχανική μάθηση.

Σκοπός της εργασίας είναι η ανάπτυξη ενός εργαλείου αναγνώρισης μερών του λόγου για την ελληνική γλώσσα του κοινωνικού ιστού, όπως περιγράφεται στην προηγούμενη παράγραφο. Στόχος είναι η συλλογή σημαντικής ποσότητας δεδομένων κειμένου του κοινωνικού ιστού στα ελληνικά, και η ανάθεση ενός POS tag ("ουσιαστικό", "ρήμα", "επίρρημα" κ.λπ.) σε κάθε λέξη ανάλογα με τα συμφραζόμενά της, δημιουργώντας έτσι ένα dataset που θα αποτελέσει τη βάση για την ανάπτυξη ενός αποδοτικού εργαλείου αναγνώρισης μερών του λόγου για την ελληνική γλώσσα του κοινωνικού ιστού.

Κεφάλαιο 2 Βιβλιογραφική επισκόπηση

Όπως αναφέρεται και στο προηγούμενο κεφάλαιο, δεν έχει αναπτυχθεί μέχρι στιγμής κάποιο εργαλείο αναγνώρισης μερών του λόγου για την ελληνική γλώσσα του κοινωνικού ιστού. Ωστόσο, έχουν αναπτυχθεί τέτοια εργαλεία για τα αγγλικά, αλλά και για άλλες γλώσσες του κοινωνικού ιστού, όπως τα γερμανικά, τα γαλλικά, τα αραβικά και τα βιετναμέζικα.

Στην Ενότητα 2.1 αναφέρονται συνοπτικά τα κυριότερα εργαλεία αναγνώρισης μερών του λόγου, τα οποία έχουν αναπτυχθεί για την αγγλική γλώσσα του κοινωνικού ιστού. Στην Ενότητα 2.2 αναφέρονται συνοπτικά κάποια από τα εργαλεία αναγνώρισης μερών του λόγου, τα οποία έχουν αναπτυχθεί για άλλες γλώσσες του κοινωνικού ιστού.

2.1 Εργαλεία αναγνώρισης μερών του λόγου για την αγγλική γλώσσα του κοινωνικού ιστού

Ένα από τα πρώτα εργαλεία αναγνώρισης μερών του λόγου για την αγγλική γλώσσα του κοινωνικού ιστού είναι αυτό των Gimpel et al. (2011). Επιχειρούν να αντιμετωπίσουν το πρόβλημα του POS tagging για δεδομένα στα αγγλικά από το Twitter. Δημιούργησαν ένα επισημειωμένο dataset, το οποίο αποτελείται από 1827 tweets ή 26435 tokens (λέξεις) και ένα tag set, το οποίο περιέχει 25 διαφορετικά tags, συμπεριλαμβανομένων των ειδικών tags για τις ιδιαιτερότητες της γλώσσας του Twitter: όνομα χρήστη, υπερσύνδεσμος, διεύθυνση ηλεκτρονικού ταχυδρομείου, hashtag και emoji/ emoticon. Πριν τις δοκιμές τους έκαναν ορισμένες κανονικοποιήσεις για την ελάττωση της ανορθογραφίας και της ασυνεπούς χρήσης των κεφαλαίων, με αποτέλεσμα το εργαλείο τους να αποδίδει καλύτερα σε σύγκριση με τον Stanford tagger για τα συγκεκριμένα δεδομένα (90% accuracy).

Οι Huang et al. (2016) ανέπτυξαν ένα εργαλείο εντοπισμού των γλωσσολογικών παραλλαγών ανά περιοχή των Η.Π.Α., χρησιμοποιώντας μεθοδολογία παρόμοια με αυτή που ακολουθείται στην έρευνα για την αναγνώριση μερών του λόγου με δεδομένα από το Twitter. Το dataset τους αποτελείται από 924 εκατ. tweets ή 7,8 δισ. λέξεις, δημοσιευμένα στο διάστημα ενός έτους. Στόχος τους ήταν η εξαγωγή λεξικολογικών χαρακτηριστικών για τους χρήστες του Twitter, μέσω της συσσώρευσης και εξομάλυνσης κάθε λεξικολογικού χαρακτηριστικού, για την

άντληση γλωσσολογικών μεταβλητών ανά περιοχή. Στο άρθρο τους τονίζεται ως σημαντικό ζήτημα η αντιμετώπιση των συντομογραφιών και των ορθογραφικών λαθών σε τέτοιου είδους περιστασιακά και σύντομα μηνύματα όπως τα tweets.

Σε ένα άλλο άρθρο από τους Liu et al. (2012) προτείνεται ένα γνωστικό σύστημα κανονικοποίησης των μη τυποποιημένων social text tokens πριν την εφαρμογή τεχνικών NLP, το οποίο μετατρέπει αυτόματα τα tokens σε δόκιμες αγγλικές λέξεις. Το σύστημα αυτό αξιολογείται τόσο σε επίπεδο λέξεων, όσο και σε επίπεδο ολόκληρων μηνυμάτων, χρησιμοποιώντας 4 datasets με δεδομένα από SMS και Twitter. Η πρωτοτυπία της προσέγγισής τους έγκειται στην προσπάθεια υλοποίησης μοτίβων μετασχηματισμού γραμμάτων που χρησιμοποιούν οι άνθρωποι για να αποκρυπτογραφήσουν τα tokens, με αποτέλεσμα την κατά περίπου 90% επιτυχή μετατροπή των λέξεων σε όλα τα datasets.

Οι Tommasel & Godoy (2018) προτείνουν μια τεχνική για online feature selection, ειδική για πολυδιάστατα δεδομένα, για την αντιμετώπιση του προβλήματος της μεγάλης κλίμακας κατηγοριοποίησης κειμένου σε κοινωνικά δίκτυα. Τα ζητήματα που επιχειρούν να αντιμετωπίσουν οφείλονται στα σύντομα κείμενα, στο μεγάλο εύρος χαρακτηριστικών και στη συνδεδεμένη φύση των κοινωνικών δεδομένων. Ορίσαν ένα σύνολο από textual features για κάθε ομάδα tweets (500000 συνολικά), το οποίο χρησιμοποιείται για την εκπαίδευση διαφορετικών μοντέλων μάθησης για την ταξινόμηση νέων tweets. Ως αποτέλεσμα, παρατηρείται βελτιωμένη ταξινόμηση σε σύγκριση με τις παραδοσιακές τεχνικές feature selection, με accuracy και F₁ score κοντά στο 90%.

Οι Popescu & Pennacchiotti (2010) αξιοποίησαν εργαλεία αναγνώρισης μερών του λόγου για την αγγλική γλώσσα του κοινωνικού δικτύου Twitter και δημιούργησαν τρία μοντέλα για εντοπισμό αμφιλεγόμενων γεγονότων. Το dataset τους αποτελείται από 73368 στιγμιότυπα (tweets), αναρτημένα από 104713 μοναδικούς λογαριασμούς επωνύμων σε ορισμένη χρονική περίοδο. Τα μοντέλα είναι αποτέλεσμα αλγορίθμων επιβλεπόμενης μηχανικής μάθησης, ενώ για τη δημιουργία τους αξιοποιήθηκαν λεξικά αμφιλεγόμενων λέξεων, αργκό, και ένα λεξικό με 100000, επισημειωμένες ως προς τα μέρη του λόγου, αγγλικές λέξεις. Σε επόμενη έρευνα (Popescu et al., 2011), αξιοποιούν τα μοντέλα αυτά και το εργαλείο αναγνώρισης μερών του λόγου EventAboutness, για αυτόματο εντοπισμό γεγονότων και τις σχέσεις τους με

συγκεκριμένους χρήστες (επώνυμοι) στο Twitter, καθώς και την αντίδρασης του κοινού σε αυτά.

Οι Foster et al. (2011) αξιολόγησαν και βελτίωσαν τον statistical dependency parser, Malt. Χρησιμοποίησαν ένα dataset που αποτελείται από 519 επισημειωμένες προτάσεις από tweets, ωστόσο αντικατέστησαν τα ιδιαίτερα χαρακτηριστικά της γλώσσας του κοινωνικού δικτύου Twitter (ονόματα χρηστών, hashtags, συνδέσμους κ.λπ.) με λέξεις-κωδικούς για την εξομάλυνση των δεδομένων. Στις αρχικές δοκιμές με τον Malt, σημειώθηκε πτώση στην απόδοση (part-of-speech tagging errors) κατά 20% στο accuracy. Έπειτα από την ενίσχυση του Malt με phrase structure parser, ο οποίος έχει αυτο-εκπαιδευτεί σε υλικό του Twitter, το accuracy βελτιώθηκε κατά 4%.

Σημαντική είναι ακόμη και η συνεισφορά των Nand et al. (2014), οι οποίοι ανέπτυξαν ένα εργαλείο αναγνώρισης μερών του λόγου για κείμενα τύπου microblogging, αξιοποιώντας Hidden Markov Models. Για την αξιολόγηση του με cross validation, χρησιμοποίησαν τρία dataset που αποτελούνται από tweets, ενώ ως σημείο αναφοράς για την αξιολόγηση της απόδοσης θέτουν τον Stanford tagger. Υποστηρίζουν πως το εργαλείο αυτό είναι κατάλληλο για εφαρμογές πραγματικού χρόνου, παρά την σημαντικά μειωμένη απόδοση (κάτω από 80% accuracy) για αρκετά tags.

2.2 Εργαλεία αναγνώρισης μερών του λόγου για άλλες γλώσσες του κοινωνικού ιστού

Ένα από τα εργαλεία αναγνώρισης μερών του λόγου για τη γερμανική γλώσσα του κοινωνικού ιστού είναι αυτό της Rehbein (2013), το οποίο χρησιμοποιεί Conditional Random Fields και συσταδοποίηση λέξεων (μη επιβλεπόμενη μάθηση) ως μέθοδο μάθησης. Το επισημειωμένο dataset που δημιουργήθηκε, αποτελείται από 1416 tweets ή 20877 tokens. Επεκτείνοντας το συμβατικό tag set για τα γερμανικά, προσθέτουν ειδικά tags για τις ιδιαιτερότητες της γλώσσας του Twitter (όνομα χρήστη, υπερσύνδεσμος, διεύθυνση ηλεκτρονικού ταχυδρομείου, hashtag και emoji/emoticon), φτάνοντας στα 65 tags. Το εργαλείο αυτό έχει accuracy περίπου 89%.

Ακόμη μία έρευνα για την αναγνώριση μερών του λόγου για τη γερμανική γλώσσα του κοινωνικού ιστού είναι αυτή των Neunerdt et al. (2013). Δημιουργούν ένα νέο σώμα κειμένων (36000 tokens) και συγκρίνουν και αξιολογούν τέσσερις σύγχρονους

POS taggers πάνω σε αυτό, με κριτήριο την απόδοσή τους σε τέσσερα διαφορετικά είδη social text από αυτά που έχουν εκπαιδευτεί (σχόλια ιστολογίου, μηνύματα συνομιλιών, σχόλια στο YouTube και σχόλια σε ιστοσελίδες). Χρησιμοποιούν το συμβατικό tag set για τα γερμανικά (54 tags), και κατηγοριοποιούν τα ιδιαίτερα χαρακτηριστικά της γλώσσας των κοινωνικών δικτύων καθορίζοντας ορισμένους κανόνες κατά την επισημείωση. Οι τιμές accuracy των POS taggers κυμαίνονται από 84% έως 93%.

Οι Nooralahzadeh et al. (2014) αναπτύσσουν ένα μοντέλο για POS tagging για τη γαλλική γλώσσα του κοινωνικού ιστού, το οποίο χρησιμοποιεί Conditional Random Fields και συσταδοποίηση λέξεων (μη επιβλεπόμενη μάθηση) ως μέθοδο μάθησης. Χρησιμοποιούν ένα υπάρχον dataset με 1700 προτάσεις ή 38000 λέξεις social text στα γαλλικά (Facebook, Twitter, ιατρικά forums), ενώ έχουν ως σημείο αναφοράς για τον θόρυβο των δεδομένων το συμβατικό σώμα κειμένων για τα γαλλικά. Ως tag set χρησιμοποιείται το συμβατικό, με την προσθήκη ορισμένων tags για τις ιδιαιτερότητες του social text. Το accuracy για τις λέξεις κυμαίνεται μεταξύ 88% και 92%, ενώ για τις προτάσεις μεταξύ 45% και 52%.

Οι Zalmout et al. (2018) αναπτύσσουν ένα neural morphological tagging and disambiguation model για την αιγυπτιακή αραβική γλώσσα, χρησιμοποιώντας προηγούμενα μοντέλα για τη Σύγχρονη Πρότυπη Αραβική Γλώσσα ως βάση. Χρησιμοποιούν ένα αιγυπτιακό μονογλωσσικό σώμα κειμένων, μεγέθους 410 εκατ. λέξεων, που προέρχονται είτε από σχόλια σε ιστολόγια, είτε από κοινωνικά δίκτυα, καθώς και ένα χειρωνακτικά επισημειωμένο αιγυπτιακό αραβικό σώμα κειμένων, μεγέθους 160000 λέξεων. Η επισημείωση των λέξεων γίνεται με βάση τα μορφολογικά τους χαρακτηριστικά (λεξικοποιημένα, μη λεξικοποιημένα (π.χ. μέρη του λόγου, αδύνατοι τύποι αντωνυμιών) και τα συμφραζόμενά τους, ενώ εφαρμόζουν διάφορες τεχνικές κανονικοποίησης για τη μείωση του θορύβου. Υποστηρίζουν ότι τα κύρια ζητήματα που αυξάνουν την σποραδικότητα στα μοντέλα NLP και μειώνουν το accuracy είναι οι πολλές λεξικολογικές και ορθογραφικές ασυνέπειες, αλλά και οι μορφολογικές, συντακτικές και φωνητικές παραλλαγές κάθε διαλέκτου ή γεωγραφικής περιοχής. Ως αποτέλεσμα έχουν 22% relative error reduction (στην αναγνώριση μερών του λόγου), και ταιριάζουν με την απόδοση μιας εκδοχής του συστήματος που έχει εκπαιδευτεί και δοκιμαστεί σε ένα χειροκίνητα κανονικοποιημένο (ορθογραφικά) αντίγραφο των δεδομένων.

Η έρευνα των Bach et al. (2018) αφορά το POS tagging για κείμενο στα βιετναμέζικα που προέρχεται από κοινωνικά δίκτυα. Δημιούργησαν ένα επισημειωμένο σώμα κειμένων για POS tagging, το οποίο αποτελείται από 4150 προτάσεις από το Facebook ή 38498 tokens. Χρησιμοποιούν Conditional Random Fields ως μέθοδο μάθησης και συγκρίνουν την απόδοση με συμβατικούς taggers για βιετναμέζικα. Επεκτείνοντας το συμβατικό tag set, προσθέτουν πέντε emoticon tags, 1 tag για όλα τα σημεία στίξης, 1 tag για όλες τις ξένες λέξεις και 1 tag για όλες τις άγνωστες λέξεις. Οι δοκιμές τους γίνονται για μοντέλα μονογράμμου, διγράμμου και τριγράμμου και 10-fold cross validation. Τα αποτελέσματα είναι 88.26% tagging accuracy σε σενάρια επιβλεπόμενης μάθησης, 88.92% tagging accuracy σε σενάρια ημι-επιβλεπόμενης μάθησης, ενώ σημειώνεται βελτίωση κατά 12% συγκριτικά με τον vnTagger (ο πιο σύγχρονος και ευρέως χρησιμοποιούμενος part-of-speech tagger για συμβατικό κείμενο στα βιετναμέζικα).

Κεφάλαιο 3 Μηχανική μάθηση

Το ερευνητικό πεδίο της μηχανικής μάθησης περιλαμβάνει τη μελέτη αλγορίθμων και στατιστικών μοντέλων, τα οποία χρησιμοποιούνται από υπολογιστικά συστήματα για την εκτέλεση εργασιών, χωρίς να είναι αυστηρά προγραμματισμένα, αλλά δημιουργώντας ένα μαθηματικό μοντέλο βάση ενός training set. Διακρίνεται σε επιβλεπόμενη (supervised learning) και μη επιβλεπόμενη (unsupervised learning) μάθηση.

Η μάθηση με επίβλεψη ή επιβλεπόμενη μάθηση εφαρμόζεται για επίλυση προβλημάτων ή λήψη αποφάσεων, χρησιμοποιώντας ένα σύνολο επισημειωμένων παραδειγμάτων εκπαίδευσης (training set) ως είσοδο. Κάθε παράδειγμα του training set αποτελείται από ένα διάνυσμα χαρακτηριστικών (attributes) και μια επιθυμητή τιμή εξόδου (label). Στην παρούσα έρευνα, η έξοδος λαμβάνει διακριτές τιμές (βλ. tag set, Πίνακας 4.2.2.4), οπότε το πρόβλημα στο οποίο εφαρμόζονται οι αλγόριθμοι επιβλεπόμενης μηχανικής μάθησης ονομάζεται ταξινόμηση (classification).

Οι αλγόριθμοι επιβλεπόμενης μηχανικής μάθησης αναλύουν το training set και παράγουν ένα μοντέλο, το οποίο μπορεί να εφαρμοστεί για την ταξινόμηση νέων, μη επισημειωμένων, δεδομένων (test set). Στόχος αυτής της διαδικασίας είναι να μπορεί ο αλγόριθμος να αποδίδει το σωστό label στα δεδομένα του test set. Οι αλγόριθμοι επιβλεπόμενης μηχανικής μάθησης που εφαρμόστηκαν σε αυτή την εργασία είναι ο Naive Bayes (Μάθηση κατά Bayes) και ο ID3 (Δέντρα Απόφασης).

Η απόδοση των αλγορίθμων επιβλεπόμενης μηχανικής μάθησης αξιολογείται με βάση ορισμένους δείκτες αξιολόγησης (Stehman, 1997), (Juba & Le, 2019). Οι δείκτες αξιολόγησης που υπολογίζονται για τις δοκιμές της παρούσας έρευνας (βλ. Κεφάλαιο 5) είναι οι ακόλουθοι:

- *Accuracy*. Αναφέρεται στην εγγύτητα των ταξινομήσεων σε μια συγκεκριμένη τιμή. Το accuracy είναι υψηλό όταν ο μέσος όρος των τιμών είναι κοντά στην πραγματική τιμή της προς ταξινόμηση ποσότητας.
- *Kappa statistics*. Κανονικοποιημένη τιμή του accuracy.

- *Precision*. Ο αριθμός των σωστά ταξινομημένων τιμών ως προς ένα label, προς τον αριθμό των τιμών που ταξινομήθηκαν είτε σωστά είτε λάθος σε αυτό το label. Αναφέρεται στην εγγύτητα των ταξινομήσεων μεταξύ τους.
- *Recall* ή *Sensitivity*. Ο αριθμός των σωστά ταξινομημένων τιμών ως προς ένα label, προς τον αριθμό των τιμών που θα έπρεπε να έχουν ταξινομηθεί σε αυτό το label.
- *F₁ score*. Προκύπτει από τον υπολογισμό του αρμονικού μέσου των precision και recall.

Στην Ενότητα 3.1 περιγράφεται συνοπτικά ο αλγόριθμος επιβλεπόμενης μηχανικής μάθησης Naive Bayes (Μάθηση κατά Bayes). Στην Ενότητα 3.2 περιγράφεται συνοπτικά ο αλγόριθμος επιβλεπόμενης μηχανικής μάθησης ID3 (Δέντρα απόφασης).

3.1 Μάθηση με επίβλεψη - Μάθηση κατά Bayes

Η συμπερασματολογία του Bayes είναι μια στατιστική μέθοδος στην οποία χρησιμοποιείται το θεώρημα του Bayes για τον επαναπροσδιορισμό της πιθανότητας μιας υπόθεσης, καθώς περισσότερες πληροφορίες ή δεδομένα γίνονται διαθέσιμα.

Ο αλγόριθμος επιβλεπόμενης μηχανικής μάθησης Naive Bayes είναι ένας ταξινομητής ισχυρής προκατάληψης (high-bias) και χαμηλής διακύμανσης (low-variance) που μπορεί να δημιουργήσει ένα καλό μοντέλο ταξινόμησης, ακόμα και με ένα μικρό σύνολο δεδομένων (https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/bayesian/naive_bayes.html). Χρησιμοποιείται κυρίως για την κατηγοριοποίηση κειμένων, την ανίχνευση ανεπιθύμητων μηνυμάτων (spam), το sentiment analysis και τα recommender systems.

Η θεμελιώδης παραδοχή του Naive Bayes είναι ότι, δεδομένης της τιμής του label, η τιμή οποιουδήποτε attribute είναι ανεξάρτητη από την τιμή οποιουδήποτε άλλου attribute. Η παροδοχή αυτή είναι "αφελής", αλλά ο ταξινομητής Naive Bayes έχει συχνά καλή απόδοση. Η υπόθεση της ανεξαρτησίας απλοποιεί κατά πολύ τους υπολογισμούς που απαιτούνται για τη δημιουργία του πιθανοτικού μοντέλου Naive Bayes.

Μια αδυναμία αυτού του αλγορίθμου είναι η εξής: εάν μέσα στα δεδομένα εκπαίδευσης (training set) μια δεδομένη τιμή κάποιου attribute δεν εμφανίζεται ποτέ στο πλαίσιο μιας δεδομένης κλάσης, τότε η δεσμευμένη πιθανότητα (conditional probability) τίθεται στο μηδέν. Όταν αυτή η μηδενική τιμή πολλαπλασιάζεται μαζί με άλλες πιθανότητες, αυτές οι τιμές είναι επίσης μηδενικές και τα αποτελέσματα θα είναι παραπλανητικά. Η αδυναμία αυτή αντιμετωπίζεται με το τέχνασμα Laplace correction, όπου προστίθεται ένα (+1) σε κάθε μέτρηση για να αποφευχθεί η εμφάνιση μηδενικών τιμών. Για τα περισσότερα training sets, η πρόσθεση ενός σε κάθε μέτρηση έχει αμελητέα επίδραση στις εκτιμώμενες πιθανότητες.

3.2 Μάθηση με επίβλεψη - Δέντρα απόφασης

Η μάθηση με δέντρα απόφασης χρησιμοποιεί ένα δέντρο απόφασης ως μοντέλο πρόβλεψης, για να μεταβεί από παρατηρήσεις για ένα attribute - το οποίο αναπαρίσταται από τα κλαδιά - σε συμπεράσματα σχετικά με την τιμή-στόχο (label) του attribute - το οποίο αναπαρίσταται από τα φύλλα. Τα μοντέλα δέντρων όπου η μεταβλητή-στόχος λαμβάνει διακριτές τιμές ονομάζονται δέντρα ταξινόμησης. Τα φύλλα αναπαριστούν τα labels και τα κλαδιά αναπαριστούν τις συζεύξεις των attributes που οδηγούν στα φύλλα.

Ο αλγόριθμος επιβλεπόμενης μηχανικής μάθησης ID3 (Iterative Dichotomiser 3) εφευρέθηκε από τον Ross Quinlan, είναι προκάτοχος του αλγορίθμου C4.5, και είναι ένας ταξινομητής που μπορεί να δημιουργήσει ένα δέντρο απόφασης ως μοντέλο ταξινόμησης από τα παραδείγματα ενός προκαθορισμένου training set (<https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/trees/id3.html>). Τα παραδείγματα αυτά έχουν διάφορα attributes και κάθε παράδειγμα έχει επισημειωθεί με ένα label. Οι κόμβοι των φύλλων του δέντρου αποφάσεων περιέχουν το όνομα του label, ενώ οι υπόλοιποι κόμβοι ονομάζονται κόμβοι απόφασης. Από τον κόμβο απόφασης ξεκινούν τόσα κλαδιά, όσες οι πιθανές τιμές (labels) που μπορεί να πάρει το attribute. Ο ID3 χρησιμοποιεί μια ευριστική παράμετρο για την επιλογή χαρακτηριστικών (feature selection heuristic), ώστε να αποφασίσει ποιο attribute θα εξεταστεί από έναν κόμβο απόφασης.

Ο αλγόριθμος ID3 μπορεί να συνοψιστεί ως εξής:

- Λήψη όλων των μη χρησιμοποιημένων attributes
- Υπολογισμός του κριτηρίου επιλογής τους (π.χ. information gain)

- Επιλογή του attribute για το οποίο το κριτήριο επιλογής έχει την καλύτερη τιμή (π.χ. ελάχιστη εντροπία ή μέγιστο information gain)

Ο ID3 αναζητά στα attributes των παραδειγμάτων εκπαίδευσης και εξάγει εκείνο το attribute που διαχωρίζει καλύτερα τα παραδείγματα. Εάν το attribute ταξινομεί τέλεια τα training sets, τότε ο αλγόριθμος τερματίζει. Αλλιώς, λειτουργεί αναδρομικά στα n (όπου n ο αριθμός των πιθανών τιμών ενός attribute) υποτιμήματα του training set, για να βρει το καλύτερο attribute τους. Ο αλγόριθμος χρησιμοποιεί μια άπληστη αναζήτηση, που σημαίνει ότι επιλέγει το καλύτερο attribute και ποτέ δεν επανεξετάζει τις προηγούμενες επιλογές του.

Μερικά σημαντικά πλεονεκτήματα του ID3 είναι τα εξής: δημιουργεί κατανοητούς κανόνες πρόβλεψης από τα δεδομένα εκπαίδευσης, μπορεί να δημιουργήσει ένα μικρό δέντρο σε σχετικά μικρό χρονικό διάστημα και η αναζήτηση τελειώνει μόλις βρει αρκετά attributes ώστε να ταξινομηθούν όλα τα δεδομένα. Ο ID3 μπορεί να έχει κάποια μειονεκτήματα σε ορισμένες περιπτώσεις, π.χ. τα δεδομένα μπορεί να είναι over-fitted ή over-classified, εάν το test set είναι μικρό, ενώ σημαντικό μειονέκτημα αποτελεί το γεγονός ότι κάθε φορά ελέγχεται μόνο ένα attribute για τη λήψη απόφασης.

Κεφάλαιο 4 Μεθοδολογία έρευνας

Στην Ενότητα 4.1 παρουσιάζεται αναλυτικά η διαδικασία της συλλογής των δεδομένων και η δημιουργία του αρχικού dataset. Στην Ενότητα 4.2 αναλύεται η διαδικασία της δημιουργίας του τελικού dataset, και, πιο συγκεκριμένα, η κανονικοποίηση των δεδομένων και η δημιουργία του tag set. Στην Ενότητα 4.3 περιγράφεται εκτενώς η διαδικασία της επισημείωσης του dataset, καθώς και ορισμένες παρατηρήσεις που έγιναν κατά τη διαδικασία αυτή.

4.1 Συλλογή δεδομένων

Τα δεδομένα που επιλέχθηκαν για τους σκοπούς της παρούσας έρευνας και την εξασφάλιση της ομαλής διεξαγωγής της είναι ένα σύνολο από tweets γραμμένα στην ελληνική γλώσσα και δημοσιευμένα στο Twitter τον Απρίλιο του 2019. Για τη συλλογή των δεδομένων δημιουργήθηκε μία εφαρμογή (app) στο Standard search API (<https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets.html>) του Twitter. Κατά τη δημιουργία της εφαρμογής αποδόθηκαν μοναδικά API keys και access tokens, τα οποία είναι απαραίτητα για την αναζήτηση, τη συλλογή και την αποθήκευση των tweets.

Προκειμένου να εξασφαλιστεί ότι τα δεδομένα που συλλέγονται είναι κατάλληλα για την εκπλήρωση των σκοπών της έρευνας, ορίστηκαν οι ακόλουθες παραμέτροι και φίλτρα κατά την αναζήτηση των tweets μέσω του app:

- Φίλτρο για αγνόηση των retweets, ώστε να μην υπάρχουν διπλότυπα δεδομένα στο τελικό dataset
- Παράμετρος για καθορισμό της γλώσσας στην οποία είναι γραμμένα τα tweets, ώστε το τελικό dataset να αποτελείται αποκλειστικά από δεδομένα γραμμένα στα ελληνικά
- Παράμετρος για καθορισμό του χρονικού διαστήματος δημοσίευσης των tweets (1^η - 30^η Απριλίου 2019), ώστε το τελικό dataset να έχει μέγεθος ορισμένης κλίμακας

Τα tweets αναζητήθηκαν στο Standard search API του Twitter, συλλέχθηκαν και αποθηκεύτηκαν σε αρχείο CSV, τρέχοντας ειδικό python script (βλ. Παράτημα). Έτσι, δημιουργήθηκε το αρχικό dataset, το οποίο αποτελούνταν από 2578 tweets.

4.2 Δημιουργία dataset

Πριν γίνει κάποια ενέργεια επεξεργασία ή καθαρισμού, το αρχικό μέγεθος του dataset ανερχόταν στα 2578 tweets. Το τελικό μέγεθος του dataset ανέρχεται συνολικά στα 2405 tweets, ή στα 31697 tokens (μεμονωμένες λέξεις). Προέκυψε έπειτα από:

- καθαρισμό των διπλότυπων. Αφαιρέθηκαν συνολικά 88 tweets (3.41% επί του συνολικού αρχικού dataset)
- καθαρισμό των κατεστραμμένων (corrupted) δεδομένων. Αφαιρέθηκαν συνολικά 85 tweets (3.29% επί του συνολικού αρχικού dataset)

Επιπλέον, προκειμένου να αναγνωρίζονται σωστά τα συμφραζόμενα της κάθε λέξης (token) κατά την επισημείωση, χρειάζεται να είναι εμφανές πότε τελειώνει το κάθε tweet. Επομένως, τοποθετήθηκε η συμβολοσειρά "&&" στο τέλος κάθε tweet, ώστε να γίνει εμφανές. Οπότε, το token που βρίσκεται αμέσως μετά από αυτή τη συμβολοσειρά είναι η πρώτη λέξη του επόμενου tweet κ.ο.κ. (βλ. Εικόνα 4.2.1).

	A	B
1	@ersikal Σε ευχαριστώ ΣΗΖΤ &&	
2	Συνέχεια βλέπουμε κινητά με πολύ δυνατές μπαταρίες ΣΗΖΤ αφού υπάρχουν και αυτοί οι χρήστες που θέλουν να αισθάνονται HTTP &&	
3	Περισσότεροι από 140 επιχειρηματίες θα συνοδεύσουν τον Πρωθυπουργό Αλέξη Τσίπρα στα Σκόπια HTTP μέσω @enikos_gr &&	
4	Ulefone Power 5s ΣΗΖΤ Ένα κινητό με μπαταρία 13000mAh και δυνατά χαρακτηριστικά ΣΗΖΤ τώρα δικό σας με μόλις 173 SYMB SYMB κάρτα 6 HTTP &&	
5	@proneusgr @alexandris2019 Μια και είσαι χρυσαιγίτης θα έπρεπε να το περιμένω ΣΗΖΤ αλλά παρ ΣΗΖΤ όλα αυτά δεν μπορώ παρά να HTTP &&	
6	@theFAQgr Με την κατηγορία περι εσχάτης προδοσίας πρέπει να δικάσουν ΣΗΖΤ Αυτοί είναι ικανοί να πουλήσουν στρατιωτικά HTTP &&	
7	Ερώτημα ΣΗΖΤ HTTP &&	
8	@alex_rise_kid πρότειναν σ όλους ΣΗΖΤ απλά δεν δέχθηκαν όλοι &&	
9	@prototHEMA Αυτό γίνεται καιρό τώρα κ με θύματα φίλους κ γνωστούς μου ΣΗΖΤ Η αρχή είναι ενήμερες γίνονται συνεχώς κατα HTTP &&	
10	@bemyspecial2 @JamesBluz Κι μένα ανοίγει αν και γράφει ότι είναι περιορισμένο ΣΗΖΤ Τι του κάνανε του φίλου μου ΣΗΖΤ EMOT EMOT &&	
11	Συνάντηση Κατήχολαση με διοίκηση ΕΥΑΘ για τα θέματα του δήμου Θεμαϊκού HTTP &&	
12	@KoutsopoulosLeo Ασε τον Μέλι και δώσε ενα σουτ δούτ στην Σπυριδούλα ΣΗΖΤ EMOT &&	
13	Πανικός επικράτησε σήμερα στο συννινor αφού παίκτης λιποθύμησε και οδηγήθηκε επευνωμένα στο νοσοκομείο ΣΗΖΤ Πανικός HTTP &&	
14	@Sia_MM2 Χαχαχαχαχαχα &&	
15	Brexit ΣΗΖΤ Τέσσερις Νέες Ψηφοφορίες στο Βρετανικό Κοινοβούλιο HTTP &&	
16	Δυναμική αρχή για τη ΣΗΖΤ Λαϊκή Συμπέριωση ΣΗΖΤ Δήμου Ξηρομέρου ΣΗΖΤ με υπ ΣΗΖΤ Δήμαρχο τον Δημήτρη Κομπίληση στον ΑΣΤΑΚΟ ΣΗΖΤ ΦΩΤΟ HTTP &&	
17	@aistinelki Θα σου έφερα εγώ αλλά είμαι εκτός Αθήνας &&	
18	@Alissel_Bristow Εγώ πιστεύω πως θα έχει εξαφανιστεί ο δύσμοιρος ΣΗΖΤ αποκλείεται να γίνει γάμος ΣΗΖΤ &&	
19	@Paskalitsa14 @Nefeligeron Δε γίνεται χωρίς λόγο ΣΗΖΤ Η Κατερίνα εκπέμπει έναν εγωισμό που εκνευρίζει πολλούς ΣΗΖΤ Σχεδόν HTTP &&	
20	Θα διαλέξουν και οι παίκτες 1 ποδια ΣΗΖΤ #MasterChefGR &&	
21	Το σχολίο της ημέρας ΣΗΖΤ Πρωταθλητικά ΣΗΖΤ HTTP &&	
22	Αγρίνιο ΣΗΖΤ Παρουσίασε το ψηφοδέλτιο Αιτωλοακαρνανίας η Αντίσταση Πολιτών Δυτικής Ελλάδας ΣΗΖΤ Φωτο ΣΗΖΤ HTTP &&	
23	Ημερίδες της ΠΕΔ για τις επικείμενες εκλογικές διαδικασίες HTTP &&	
24	Brexit ΣΗΖΤ Στίβεν Μπάρκλεϊ ΣΗΖΤ Μπορούμε να αποχωρήσουμε από την ΕΕ σε 11 ημέρες HTTP &&	
25	αν είναι ο ιδανικός τελειο HTTP &&	
26	@alex_rise_kid όσοι θέλουν πανε ΣΗΖΤ &&	
27	Ήταν 9 ΣΗΖΤ γίναν 12 ΣΗΖΤ θα γίνουν 13 ΣΗΖΤ παίζει να ξεκινήσουν πάλι τα bootcamps με όσους δεν μπόηκαν να φτάσουν καμία 30ρια ΣΗΖΤ HTTP &&	
28	#MasterChefGR ως δεν πήρε ο Τάσος τον Σταμάτη ΣΗΖΤ ε ρε τι έχει να γίνει αύριο &&	
29	@Ninina_Nicole ΜΠΑΟΚ έχεις αργήσει &&	

Εικόνα 4.2.1 Δείγμα τελικού dataset αποτελούμενο από tweets

Στην υποενότητα 4.2.1 αναλύεται η διαδικασία κανονικοποίησης των δεδομένων. Στην υποενότητα 4.2.2 περιγράφονται αναλυτικά η δημιουργία του tag set και οι παραδοχές που έγιναν.

4.2.1 Κανονικοποίηση δεδομένων

Έχοντας ως στόχο την καλή απόδοση των αλγορίθμων μηχανικής μάθησης σε ένα dataset με μεγάλο βαθμό πολυμορφίας (Papageorgiou et al., 2000), (Gimpel et al.,

2010), (Liu et al., 2012), ορισμένες κατηγορίες tokens αντικαταστάθηκαν με λέξεις-κωδικούς πριν την έναρξη της επισημείωσης. Σύμφωνα με τους Zalmout et al. (2018), τα κύρια ζητήματα που αυξάνουν την σποραδικότητα στα μοντέλα NLP και μειώνουν το accuracy είναι οι πολλές λεξικολογικές και ορθογραφικές ασυνέπειες, αλλά και οι μορφολογικές, συντακτικές και φωνητικές παραλλαγές κάθε διαλέκτου ή γεωγραφικής περιοχής. Η διαδικασία της κανονικοποίησης των δεδομένων περιγράφεται αναλυτικά παρακάτω.

Οι υπερσύνδεσμοι (links και hyperlinks) και οι διαδικτυακές διευθύνσεις (URLs) αντικαταστάθηκαν από τη λέξη-κωδικό "HTTP". Υπήρχαν συνολικά 1356 tokens τα οποία αντικαταστάθηκαν από τη συγκεκριμένη λέξη-κωδικό. 1197 tweets περιείχαν έναν ή περισσότερους υπερσυνδέσμους.

Τα emojis και τα emoticons αντικαταστάθηκαν από τη λέξη-κωδικό "EMOT". Όταν εμφανίζεται το ίδιο emoji ή emoticon πάνω από μία φορά στη σειρά (π.χ. :/ :/ :/), χρησιμοποιείται μόνο μία λέξη-κωδικός για να αντικαταστήσει την ακολουθία των ίδιων tokens. 342 tokens αντικαταστάθηκαν από τη συγκεκριμένη λέξη-κωδικό. 287 tweets που περιείχαν ένα ή περισσότερα emoticons.

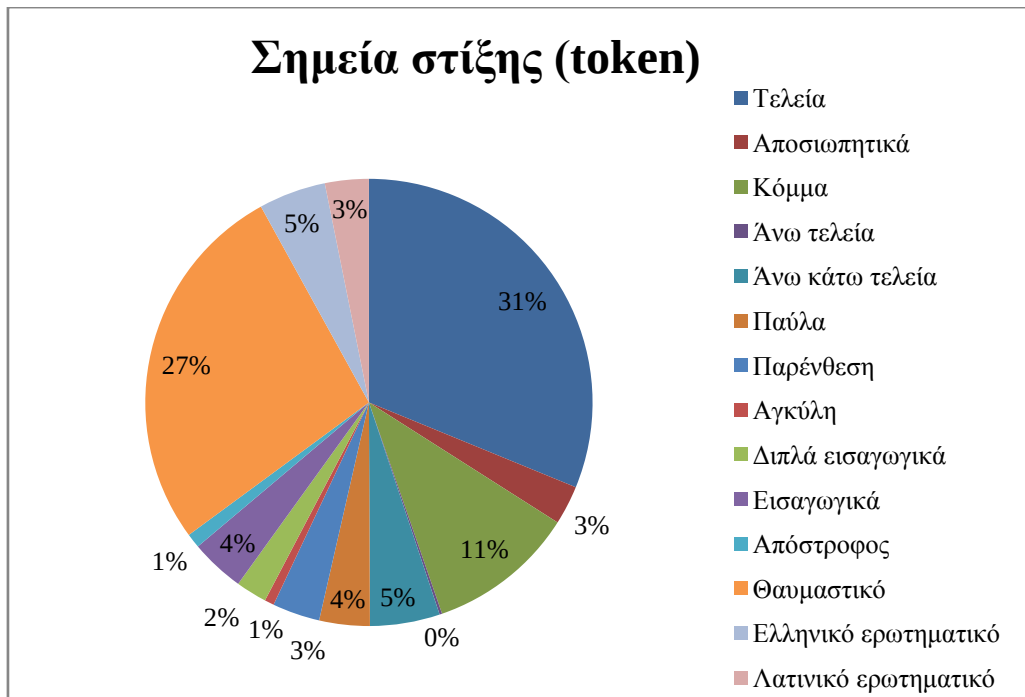
Τα σημεία στίξης αντικαταστάθηκαν από τη λέξη-κωδικό "ΣΗΣΤ" (βλ. και Διάγραμμα 4.2.1.1, Διάγραμμα 4.2.1.2). Όταν εμφανίζεται το ίδιο σημείο στίξης πάνω από μία φορά στη σειρά (π.χ. ??, !!!), χρησιμοποιείται μόνο μία λέξη-κωδικός για την ακολουθία των ίδιων tokens. Ο συνολικός αριθμός εμφάνισης κάθε σημείου στίξης και ο αριθμός των tweets που περιείχαν ένα ή περισσότερα σημεία στίξης, παρουσιάζονται αναλυτικά στον Πίνακα 4.2.1.1.

Σημείο στίξης	Συνολικός αριθμός	Αριθμός tweets
Τελεία (.)	1735	864
Αποσιωπητικά (...)	157	155
Κόμμα (,)	594	453
Άνω τελεία (·)	10	5
Άνω κάτω τελεία (:)	281	263
Παύλα (-)	203*	171
Παρένθεση (())	191	90
Αγκύλη ([])	38	22
Διπλά εισαγωγικά ("")	125	64
Εισαγωγικά (« »)	218	102
Απόστροφος (')	58	35
Θαυμαστικό (!)	1506	382
Ελληνικό ερωτηματικό	270	195

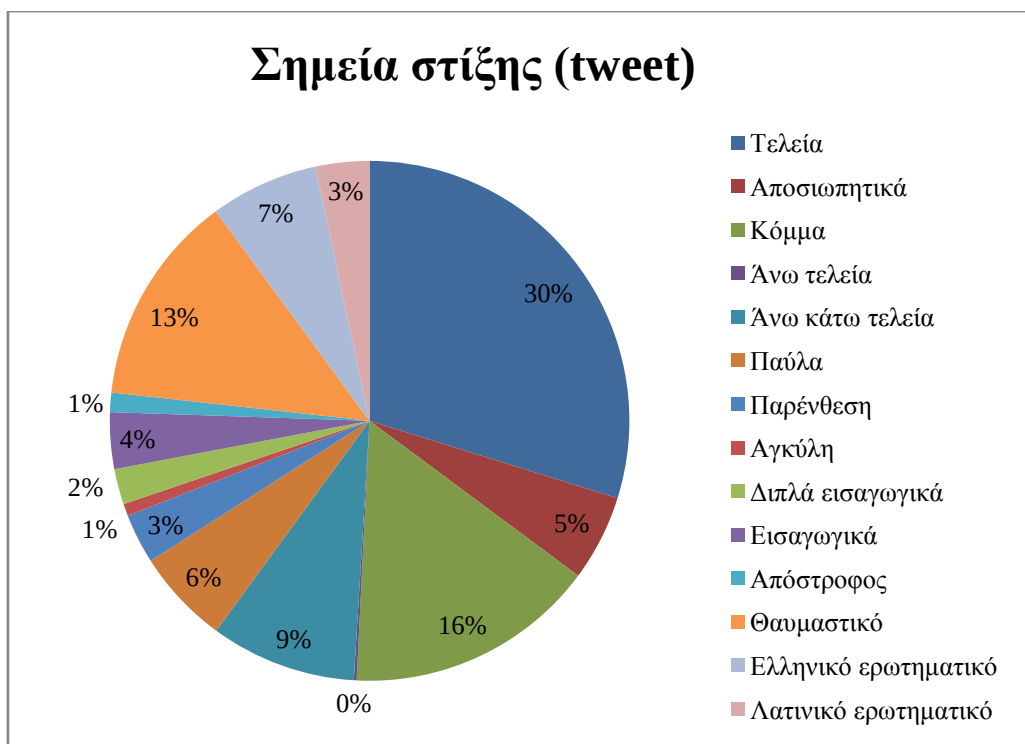
Λατινικό ερωτηματικό	176	97
----------------------	-----	----

Πίνακας 4.2.1.1 Συνολικός αριθμός εμφάνισης κάθε σημείου στίξης και αριθμός των tweets που περιείχαν ένα ή περισσότερα σημεία στίξης

*27 παύλες αφαιρέθηκαν, καθώς λειτουργούσαν ως ενωτικά μεταξύ δύο λέξεων (π.χ. παιδί-θαύμα, ωμέγα-3).



Διάγραμμα 4.2.1.1 Σημεία στίξης (token). Ποσοστά ανά σημείο στίξης πριν την αντικατάστασή τους από τη λέξη-κωδικό "ΣΗΣΤ"



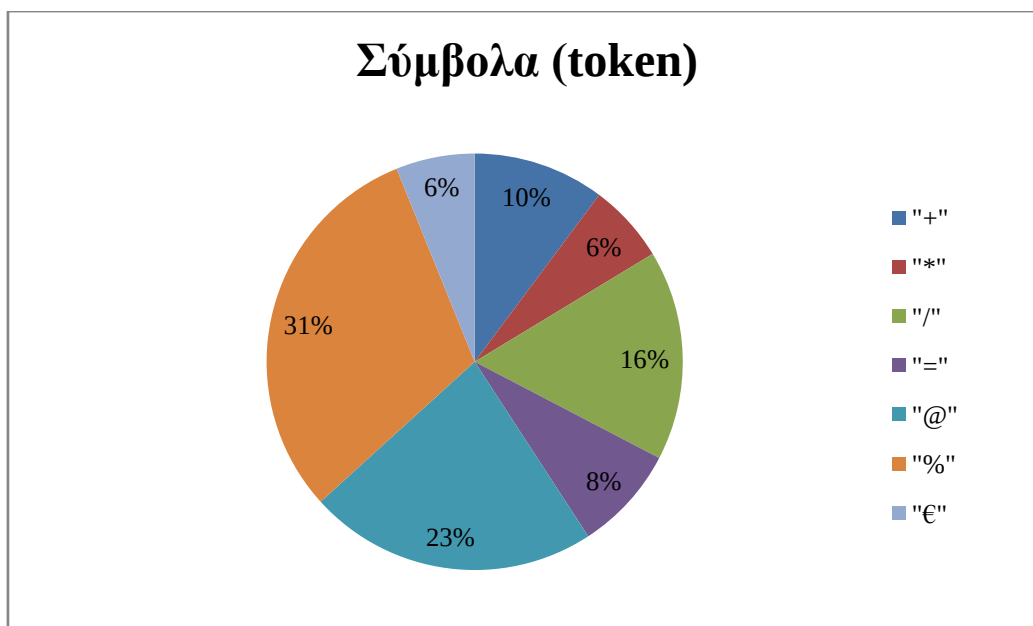
Διάγραμμα 4.2.1.2 Σημεία στίξης (tweet). Ποσοστά tweets ανά σημείο στίξης πριν την αντικατάστασή τους από τη λέξη-κωδικό "ΣΗΣΤ"

Μαθηματικά και άλλα σύμβολα αντικαταστάθηκαν από τη λέξη-κωδικό "SYMB" (βλ. και Διάγραμμα 4.2.1.3, Διάγραμμα 4.2.1.4). Ο συνολικός αριθμός εμφάνισης κάθε συμβόλου και ο αριθμός των tweets που περιείχαν ένα ή περισσότερα σύμβολα, παρουσιάζονται αναλυτικά στον Πίνακα 4.2.1.2.

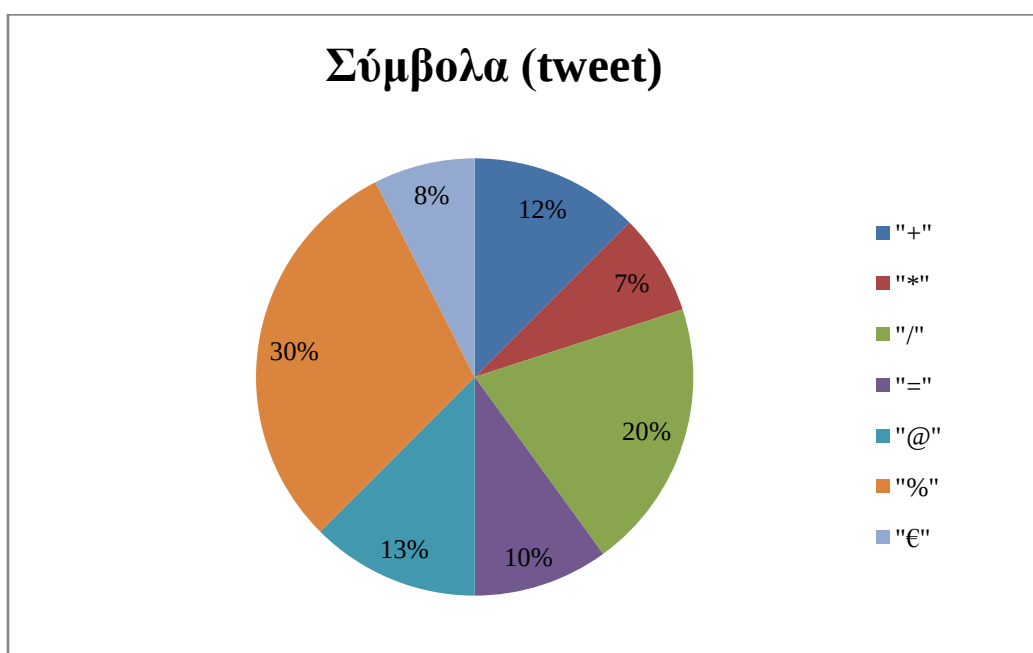
Σύμβολο	Συνολικός αριθμός	Αριθμός tweets
Πρόσθεσης (+)	5	5
Πολλαπλασιασμού (*)	3	3
Διαίρεσης (/)	8	8
Ισότητας (=)	4	4
@	11*	5
%	15	12
€	3	3

Πίνακας 42.1.2 Συνολικός αριθμός εμφάνισης κάθε μαθηματικού ή άλλου συμβόλου και αριθμός των tweets που περιείχαν ένα ή περισσότερα σύμβολα

*Αντικαταστάθηκαν όλα, εκτός από αυτά που αποτελούσαν κομμάτι λέξης, π.χ. Κ@@@Α.



Διάγραμμα 4.2.1.3 Μαθηματικά και άλλα σύμβολα (token). Ποσοστά ανά σύμβολο πριν την αντικατάστασή τους από τη λέξη-κωδικό "SYMB"



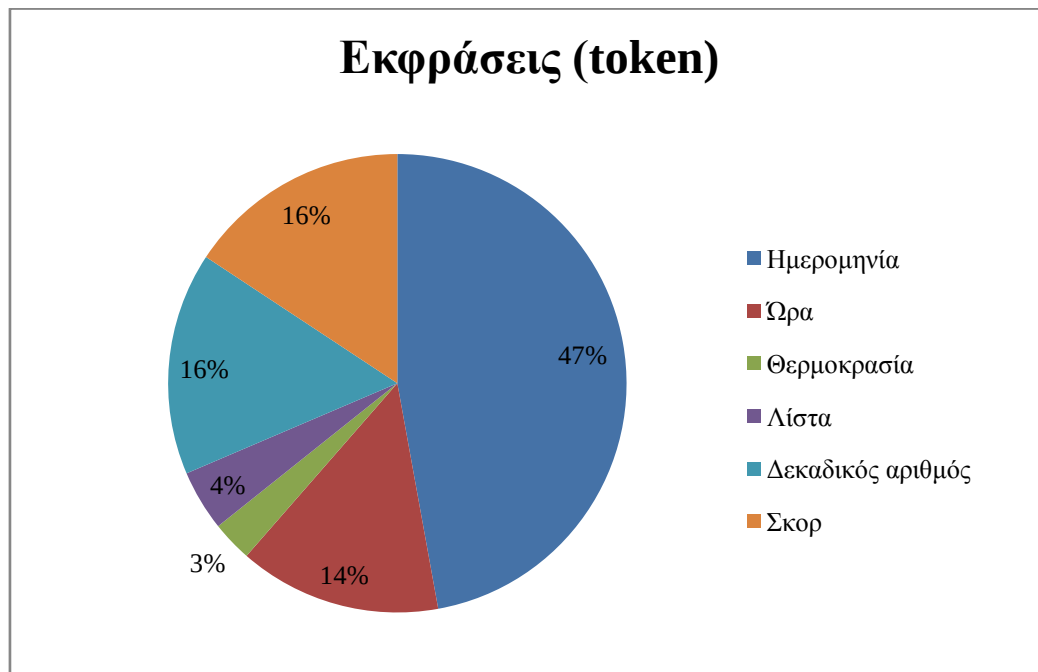
Διάγραμμα 4.2.1.4 Μαθηματικά και άλλα σύμβολα (tweet). Ποσοστά tweets ανά σύμβολο πριν την αντικατάστασή τους από τη λέξη-κωδικό "SYMB"

Οι χρονικές, ποσοτικές και αριθμητικές εκφράσεις αντικαταστάθηκαν από τη λέξη-κωδικό "NUMB" (βλ. και Διάγραμμα 4.2.1.5, Διάγραμμα 4.2.1.6). Ως τέτοιες εκφράσεις ορίζονται οι ημερομηνίες, η ώρα, οι τιμές, οι θερμοκρασίες, οι αποστάσεις και τα σκορ κάθε μορφής (π.χ. 1., 4/10, 2018-2019, 23:30, 2,5, 14.7°C, 68-66, 0.0), (Αναστασιάδη-Συμεωνίδη & Κυριακοπούλου, 2015). Ο συνολικός αριθμός

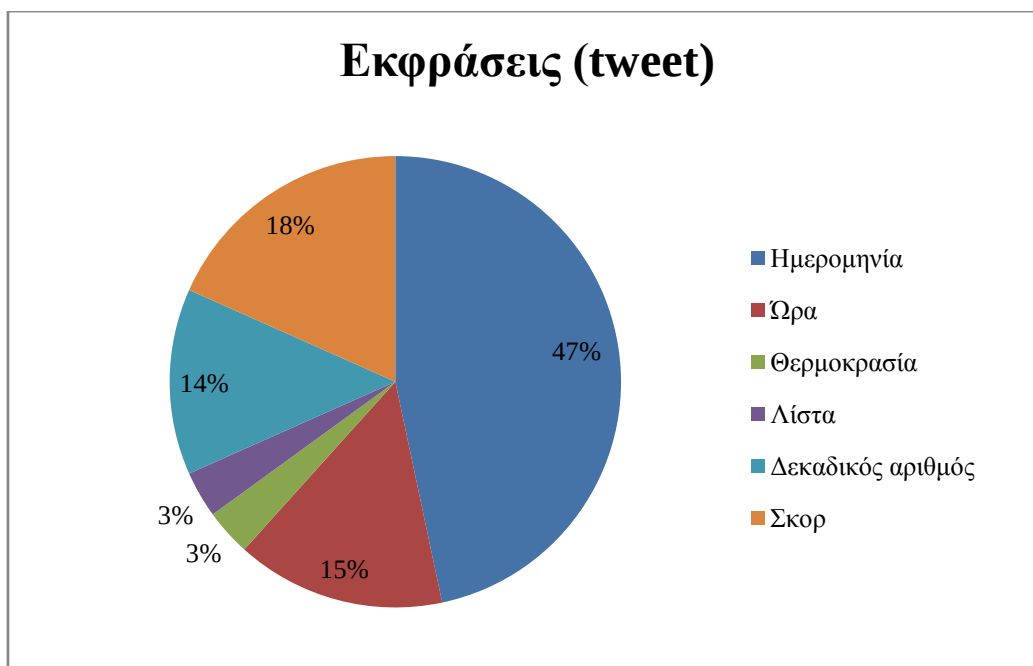
εμφάνισης κάθε έκφρασης και ο αριθμός των tweets που περιείχαν μία ή περισσότερες εκφράσεις, παρουσιάζονται αναλυτικά στον Πίνακα 4.2.1.3.

Έκφραση	Συνολικός αριθμός	Αριθμός tweets
Ημ/νία ή χρονική περίοδος	33	28
Ωρα	10	9
Θερμοκρασία	2	2
Λίστα (π.χ. 1. 2.)	3	2
Δεκαδικός αριθμός (π.χ. 0.0, 2,5)	11	8
Σκορ (π.χ. 68-66)	11	11

Πίνακας 4.2.1.3 Συνολικός αριθμός εμφάνισης κάθε χρονικής, ποσοτικής και αριθμητικής έκφρασης και αριθμός των tweets που περιείχαν μία ή περισσότερες εκφράσεις



Διάγραμμα 4.2.1.5 Χρονικές, ποσοτικές και αριθμητικές εκφράσεις (token). Ποσοστά ανά έκφραση πριν την αντικατάστασή τους από τη λέξη-κωδικό "NUMB"



Διάγραμμα 4.2.1.6 Χρονικές, ποσοτικές και αριθμητικές εκφράσεις (tweet). Ποσοστά tweets ανά έκφραση πριν την αντικατάστασή τους από τη λέξη-κωδικό "NUMB"

4.2.2 Δημιουργία tag set

Μετά την ολοκλήρωση της κανονικοποίησης των δεδομένων, δημιουργήθηκε το tag set, το οποίο θα χρησιμοποιηθεί ως βάση για τη διαδικασία της επισημείωσης (βλ. Κεφάλαιο 4, Ενότητα 4.3). Το tag set αυτό αποτελείται από συνολικά 22 tags, και παρουσιάζεται πλήρως παρακάτω, στον Πίνακα 4.2.2.1:

Tag	Περιγραφή	Παραδείγματα
KNO	Κοινό ουσιαστικό	γραμματέας τιμες αποχωρηση νερό
KPO	Κύριο ουσιαστικό	Θήβα Πρωταπριλιά Μπαμπινιώτη Τόνια
P	Ρήμα ή ενεργητική μετοχή (σε οποιοδήποτε χρόνο, φωνή, έγκλιση, πρόσωπο, αριθμό κ.τ.λ.)	ειναι κρύβεις λες παμε μιλώντας
EΘ	Επίθετο ή παθητική μετοχή	ωραίος μοναχικό απλη περιορισμένος
EP	Επίρρημα (τοπικό, χρονικό, τροπικό, ποσοτικό, ή βεβαιωτικό/ διστακτικό/ αρνητικό)	κάτω αύριο πολυ καλά μαζί πιο ναι οπωσδήποτε
AP	Άρθρο (οριστικό, αόριστο, ή εμπρόθετο)	ο η τους την μια ένα ένας στα
AN	Αντωνυμία (προσωπική, κτητική,	μου εσυ της αυτο

	αυτοπαθής, αόριστη, οριστική, δεικτική, αναφορική, ή ερωτηματική)	εκείνος τι τετοια τόσος όποιος που κάποια κάτι κάθε κανένα αλλος
ΣΥΝ	Σύνδεσμος (συμπλεκτικός, αντιθετικός, ειδικός, χρονικός, συμπερασματικός, επεξηγηματικός, τελικός, διαχωριστικός, αιτιολογικός, διστακτικός, υποθετικός, αποτελεσματικός, ή συγκριτικός)	κ όταν ή οτι να και αν γιατί άρα
ΠΡ	Πρόθεση	από έως επι με αντί για χωρίς υπό
Μ	Μόριο	σαν ως ας θα να μα για πάρα δεν μην
ΕΦ	Επιφώνημα	ε α ω αμέ μακάρι μπα χμ μπράβο χαχα άντε
ΣΣ	Σημεία στίξης (. , · : - () [] " « » ' ! ; ... ?)	ΣΗΣΤ
ΣΥΜ	Μαθηματικό ή άλλο σύμβολο (+ * / = @ € %)	SYMB
ΑΡΘΜ	Αριθμητικό ουσιαστικό, αριθμητικό επίθετο, ή αριθμός	τέσσερα 2 τριπλή μία όγδοος δεκάδα δεκαριά
ΕΚ	Χρονική, ποσοτική ή αριθμητική έκφραση (ημερομηνία, ώρα, τιμή, θερμοκρασία, απόσταση, σκορ)	NUMB
Ξ	Ξένη λέξη	reunion ακαου οκ
ΣΥΝΤΜ	Συντομογραφία	ΕΕ πχ ΗΥ κλπ
Η (λατινικό)	Hashtag (υποδηλώνει θέμα/ κατηγορία για το tweet)	#πρωταπριλια #Greece
U	@at-mention (υποδηλώνει έναν άλλον χρήστη ως παραλήπτη του tweet)	@PrimeministerGR
Ε (λατινικό)	Emoji ή Emoticon	EMOT (λατινικά)
ΥΣ	Υπερσύνδεσμος (link και hyperlink) ή διαδικτυακή διεύθυνση (URL)	HTTP (λατινικά)
Δ	Διάφορα, οτιδήποτε δεν ανήκει σε καμία από τις παραπάνω κατηγορίες	σαγαπω σκκσκκς άστα H31007244

Πίνακας 4.2.2.4 Tag set

Το tag set προέκυψε έπειτα από μελέτη γραμματικών της ελληνικής γλώσσας (Τριανταφυλλίδης, 1990), (Τζεβελέκου, Κάντζου, & Σταμούλη, 2007), και λεξικών (Βικιλεξικό, <https://bit.ly/2ylRINE>). Τα tags τα οποία αφορούν ειδικές κατηγορίες tokens που συναντώνται αποκλειστικά στο πλαίσιο του κοινωνικού δικτύου Twitter ("ΥΣ", "Η" (λατινικό), "Ε" (λατινικό), "U") δημιουργήθηκαν με βάση το tag set των Gimpel et al. (2010) για την αγγλική γλώσσα του κοινωνικού ιστού.

Για την αποφυγή φαινομένων μειωμένης απόδοσης κατά τις διαδικασίες της επισημείωσης και της μηχανικής μάθησης, τα οποία θα μπορούσαν να προκύψουν εξαιτίας ενός υπερβολικά εκτεταμένου tag set, έγιναν ορισμένες παραδοχές για να περιοριστεί το πλήθος των tags (Gimpel et al., 2010). Οι παραδοχές αυτές περιγράφονται εκτενώς στις επόμενες παραγράφους.

Καθώς το πλήθος των μετοχών είναι μικρό κατ' αναλογία με τον όγκο του dataset, αποφασίστηκε να μην οριστεί ξεχωριστό tag για τις μετοχές. Επομένως, οι παθητικές μετοχές κατηγοριοποιήθηκαν ως επίθετα και οι ενεργητικές ως ρήματα. Κατά συνέπεια, στις παθητικές μετοχές αποδίδεται το tag "ΕΘ" και στις ενεργητικές το tag "Ρ".

Εφόσον στα πλαίσια της συγκεκριμένης έρευνας δεν εξετάζεται η σημασιολογία των δεδομένων, έγιναν οι ακόλουθες παραδοχές:

- Όλα τα σημεία στίξης είναι ισοδύναμα και, επομένως, κατηγοριοποιούνται σε κοινό tag "ΣΣ"
- Όλα τα άρθρα - οριστικά, αόριστα και εμπρόθετα - είναι ισοδύναμα και, επομένως, κατηγοριοποιούνται σε κοινό tag "ΑΡ"
- Όλες οι αντωνυμίες - προσωπικές, κτητικές, αυτοπαθείς, αόριστες, οριστικές, δεικτικές, αναφορικές ή ερωτηματικές - είναι ισοδύναμες και, επομένως, κατηγοριοποιούνται σε κοινό tag "ΑΝ"
- Όλα τα επιρρήματα - τοπικά, χρονικά, τροπικά, ποσοτικά, ή βεβαιωτικά/διστακτικά/ αρνητικά - είναι ισοδύναμα και, επομένως, κατηγοριοποιούνται σε κοινό tag "ΕΡ"

Επίσης, έγινε η παραδοχή ότι τα επιρρήματα "σαν" και "ως", κατηγοριοποιούνται σε κοινό tag "Μ", μαζί με τα μόρια. Τέλος, ως εξαίρεση σε σχέση με τη γραμματική του Τριανταφυλλίδη (1990), τα "δε(ν)" και "μη(ν)" αντιμετωπίζονται ως μόρια και όχι ως επιρρήματα, καθώς η πλειοψηφία των πιο σύγχρονων γραμματικών (Τζεβελέκου, Κάντζου, & Σταμούλη, 2007) για τα νέα ελληνικά τα κατατάσσουν στα μόρια.

4.3 Επισημείωση

Μετά από την κανονικοποίηση των δεδομένων και τη δημιουργία του tag set, ξεκίνησε η διαδικασία της επισημείωσης. Το dataset διαμερίστηκε κατάλληλα ώστε να μην αποτελείται από ολόκληρα tweets, αλλά από τα 31697 tokens (μεμονωμένες λέξεις) που τα συνθέτουν (βλ. Εικόνα 4.3.1). Από το στάδιο αυτό και έπειτα, το dataset στο οποίο γίνεται αναφορά είναι αυτό.

	A	B	C	D	E	F	G	H	I
1	@ersikal								
2	Σε								
3	ευχαριστώ								
4	ΣΗΣΤ								
5									
6									
7									
8	Συνέχεια								
9	βλέπουμε								
10	κινητά								
11	με								
12	πολύ								
13	δυνατές								
14	μπαταρίες								
15	ΣΗΣΤ								
16	αφού								
17	υπάρχουν								
18	και								
19	αυτοί								
20	οι								
21	χρήστες								
22	που								
23	θέλουν								
24	να								
25	αισθάνονται								
26	HTTP								
27									
28									
29									
30	Περισσότεροι								

Εικόνα 4.3.1 Δείγμα dataset διαμερισμένου σε tokens

Τα συμφοραζόμενα κάθε token είναι εμφανή κατά την επισημείωση, καθώς έχει τοποθετηθεί ειδική συμβολοσειρά στο τέλος κάθε tweet, ώστε να είναι σαφές ποιες λέξεις (tokens) το συνθέτουν (βλ. Κεφάλαιο 4, Ενότητα 4.2). Αυτή η ενέργεια είναι απαραίτητη, καθώς παρατηρήθηκε ότι τα συμφοραζόμενα λειτουργούν βοηθητικά για τον επισημειωτή στην αναγνώριση των μερών του λόγου για μεμονωμένες λέξεις, με αποτέλεσμα να αποδίδουν το σωστό tag πιο εύκολα και με λιγότερη αμφιβολία, και, συνεπώς, με λιγότερα λάθη προς διόρθωση στο στάδιο του ελέγχου. Επίσης, τα

συμφραζόμενα χρησιμοποιούνται ως attributes στα training set και test set για τις δοκιμές με αλγορίθμους μηχανικής μάθησης (βλ. Κεφάλαιο 5).

Στην υποενότητα 4.3.1 αναλύεται η διαδικασία επισημείωσης των δεδομένων και οι παραδοχές που έγιναν. Στην υποενότητα 4.3.2 αναφέρονται ορισμένες παρατηρήσεις που προέκυψαν κατά την επισημείωση.

4.3.1 Διαδικασία

Αρχικά, συγκροτήθηκε μια ομάδα από εθελοντές - ειδικούς στον τομέα, οι οποίοι είχαν τον ρόλο του επισημειωτή στα πλαίσια της παρούσας έρευνας. Ο καθέ επισημειωτής ανέλαβε να αποδώσει tags στα tokens που του είχαν ανατεθεί από την ερευνήτρια με βάση το μέρος του λόγου στο οποίο ανήκουν, έχοντας ως υποστηρικτικό υλικό τις γραμματικές: Τζεβελέκου, Κάντζου, & Σταμούλη (2007), Τριανταφυλλίδης (1990) και το διαδικτυακό λεξικό: Βικιλεξικό (<https://bit.ly/2ylRINe>).

	A	B	C	D	E	F	G	H	I
1	@ersikal	U							
2	Σε	AN							
3	ευχαριστώ	P							
4	ΣΗΣΤ	ΣΣ							
5									
6									
7									
8	Συνέχεια	EP							
9	βλέπουμε	P							
10	κινητά	KNO							
11	με	ΠP							
12	πολύ	EP							
13	δυνατές	ΕΘ							
14	μπαταρίες	KNO							
15	ΣΗΣΤ	ΣΣ							
16	αφού	ΣΥΝ							
17	υπάρχουν	P							
18	και	ΣΥΝ							
19	αυτοί	AN							
20	οι	AP							
21	χρήστες	KNO							
22	που	AN							
23	θέλουν	P							
24	να	ΣΥΝ							
25	αισθάνονται	P							
26	HTTP	ΥΣ							
27									
28									
29									
30	Περισσότεροι	ΕΘ							

Εικόνα 4.3.1.1 Δείγμα dataset μετά από τα στάδια της επισημείωσης και του ελέγχου. Στη στήλη A φαίνεται το token και στη στήλη B το tag που του αντιστοιχεί

Για την αποφυγή συχνών σφαλμάτων που παρατηρήθηκαν κατά την επισημείωση, έγιναν ορισμένες παραδοχές, ώστε να υπάρχουν κοινές οδηγίες ως σημείο αναφοράς για τους επισημειωτές σε αμφιλεγόμενες και ειδικές περιπτώσεις. Οι παραδοχές και οι οδηγίες αυτές περιγράφονται εκτενώς στις επόμενες παραγράφους.

Τα ρήματα σε χρόνο που σχηματίζεται με βοηθητικό ρήμα, όπως ο παρακείμενος, ο υπερσυντέλικος κ.λπ., είναι της μορφής "έχω κάνει". Επομένως, δόθηκε η οδηγία ότι οι δύο αυτές λέξεις (βοηθητικό και κύριο ρήμα) αντιμετωπίζονται και επισημειώνονται ως δύο ξεχωριστά ρήματα ("P").

Ο τελικός σύνδεσμος "για να" αποτελείται από δύο tokens (λέξεις). Οπότε, δόθηκε η οδηγία να αντιμετωπίζονται ως ανεξάρτητα και να επισημειώνονται ως πρόθεση ("PP") και σύνδεσμος ("SYN"), αντίστοιχα.

Ως ξένες λέξεις ("Ξ") επισημειώνονται όσες λέξεις είναι γραμμένες σε άλλη γλώσσα πέρα από τα ελληνικά, π.χ. "reunion", "Trump", αλλά και όσες λέξεις είναι γραμμένες στα ελληνικά, αλλά έχουν ξενική προέλευση, π.χ. ακαου, οκ, μπαρ. Εξαίρεση αποτελούν τα κύρια ουσιαστικά, π.χ. Ντράγκι, Λίβερπουλ, Τομ, Μπρέντι κ.λπ.

Τα ονοματεπώνυμα επισημειώνονται ως κύρια ουσιαστικά ("KPO") με δύο tags, ένα για το όνομα και ένα για το επώνυμο. Εξαίρεση αποτελούν τα ονοματεπώνυμα της μορφής "Α. Σισιλιάνος", όπου το όνομα επισημειώνεται ως συντομογραφία ("SYNTM") και το επώνυμο ως κύριο ουσιαστικό ("KPO").

Στα αριθμητικά συμπεριλαμβάνονται επιπρόσθετα τα επίθετα και τα ουσιαστικά τα οποία σχετίζονται με την ηλικία, τον χρόνο ή την ποσότητα, π.χ. "16χρονος", "εξηντάχρονος", "δεκάλεπτο", "20λεπτο", "10άδα", "ντουζίνα". Επομένως, τους αποδίδεται το tag "APΘM".

Τέλος, οι λέξεις "όλος, -η, -ο" κατηγοριοποιούνται κατά κανόνα ως επίθετα ("EΘ"), καθώς όταν το "όλος, -η, -ο" προσδιορίζει ένα ουσιαστικό, υπάρχει πάντα οριστικό άρθρο πριν από το ουσιαστικό, π.χ. "όλος ο κόσμος" (Τζεβελέκου, Κάντζου, & Σταμούλη, 2007). Σε κάθε διαφορετική περίπτωση κατηγοριοποιούνται ως αντωνυμίες ("AN").

Μόλις ολοκληρώθηκε η επισημείωση των επί μέρους τμημάτων του dataset από κάθε επισημειωτή, αυτά συγκεντρώθηκαν για τον τελικό έλεγχο από την ερευνήτρια, κατά τον οποίο διορθώθηκαν ορισμένα λάθη. Έτσι, προκύπτει το τελικό, πλήρως και

ορθώς επισημειωμένο, dataset, το οποίο αποτελείται από 31697 tokens συνολικά (βλ. Εικόνα 4.3.1.1).

4.3.2 Παρατηρήσεις

Κατά τη διαδικασία της επισημείωσης έγιναν ορισμένες παρατηρήσεις για το dataset, σχετικά με τη φύση της γλώσσας του κοινωνικού ιστού και τις ιδιαιτερότητές της. Οι παρατηρήσεις αυτές αναφέρονται στις παρακάτω παραγράφους.

Παρατηρήθηκε ότι τα tweets περιέχουν συχνά δείγματα προφορικού λόγου, π.χ. "ναιuuu", "Γουστάρωωωω", "Χααααχαχα", ιδιαίτερα όταν προέρχονται από χρήστες οι οποίοι δεν είναι δημόσια πρόσωπα (ειδησεογραφικά πρακτορεία, κυβερνητικά στελέχη, διάσημοι κ.λπ.). Ακόμη, είναι σύνηθης η συγκόλληση δύο ή περισσότερων λέξεων, π.χ. "Γιαυτό", "αστο", με αποτέλεσμα ένα token να αποτελείται από δύο ή περισσότερες λέξεις αντί για μία (αποδίδεται tag "Δ"). Αντίθετα, τα δημόσια πρόσωπα τείνουν να χρησιμοποιούν επίσημη και δόκιμη γλώσσα.

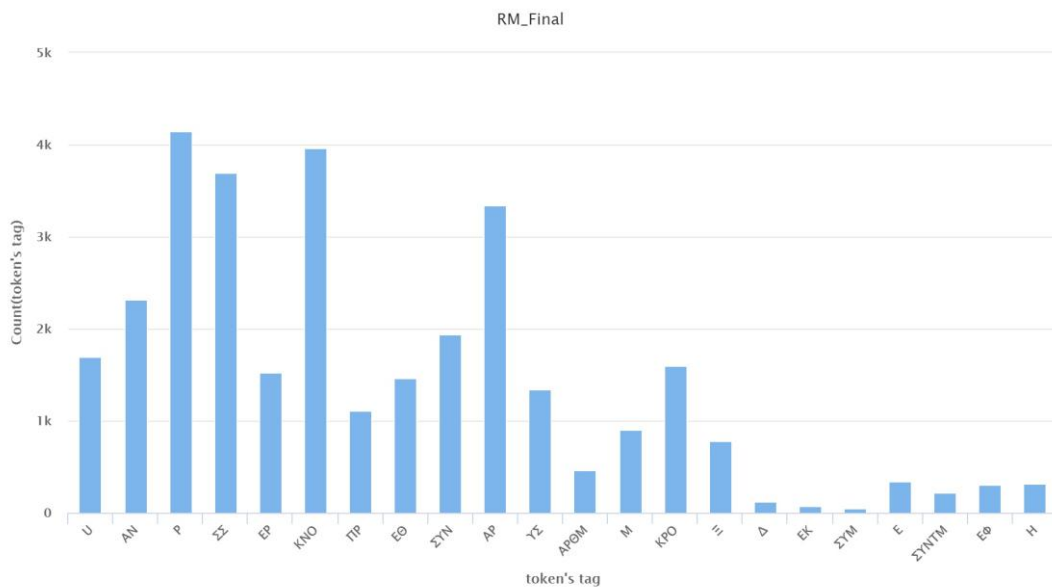
Συχνή είναι, επίσης, και η παρουσία ιδιωμάτων και διαλέκτων, όπως η κρητική, η κυπριακή και η ποντιακή. Τα tokens που είτε ανήκουν σε αυτές τις διαλέκτους ή είναι ιδιώματα αντιμετωπίζονται ως ξένες λέξεις ("Ξ"), για να μετριάσει ο βαθμός πολυμορφίας του dataset (Papageorgiou et al., 2000), (Gimpel et al., 2010), (Liu et al., 2012), (Zalmout et al., 2018), ώστε να επιτευχθεί η καλή απόδοση των αλγορίθμων μηχανικής μάθησης.

Αναφορικά με τα μέρη του λόγου στα οποία κατηγοριοποιήθηκαν τα tokens παρατηρήθηκαν τα εξής:

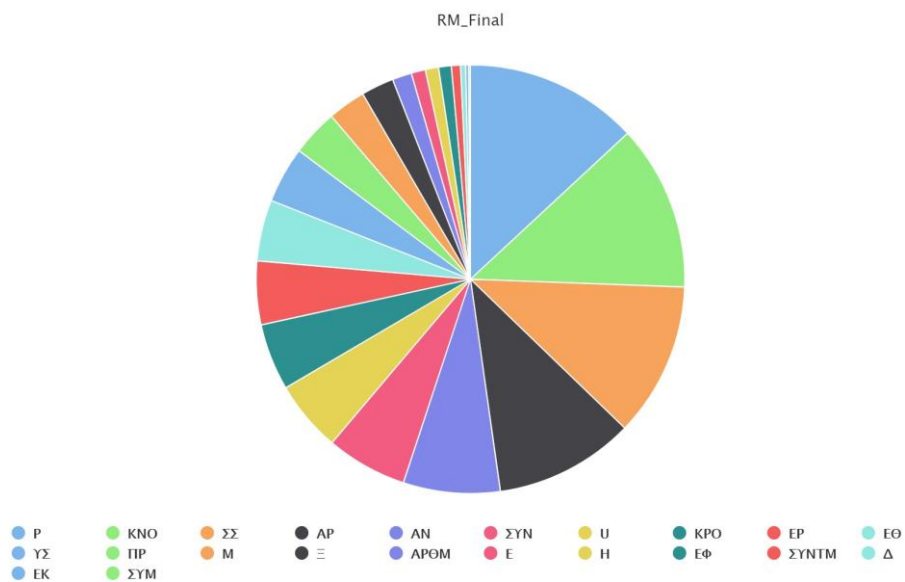
- Λέξεις όπως οι "καληνύχτα", "καλημέρα" κ.λπ. χρησιμοποιούνται κατά κανόνα ως ουσιαστικά ("ΚΝΟ") αντί ως επιφωνήματα ("ΕΦ")
- Το πιο συχνό tag είναι το "P" (ρήμα), το οποίο αποδόθηκε συνολικά σε 4149 tokens
- Το λιγότερο συχνό tag είναι το "ΣΥΜ" (μαθηματικό ή άλλο σύμβολο), το οποίο αποδόθηκε συνολικά σε 43 tokens

- Το πιο συχνό token, με 3692 εμφανίσεις, είναι η λέξη-κωδικός "ΣΗΣΤ" (σημείο στίξης)

Η συχνότητα εμφάνισης κάθε tag φαίνεται αναλυτικά στα παρακάτω διαγράμματα (Διάγραμμα 4.3.2.1, Διάγραμμα 4.3.2.2).



Διάγραμμα 4.3.2.1 Συχνότητα εμφάνισης κάθε tag



Διάγραμμα 4.3.2.2 Ποσοστό εμφάνισης κάθε tag

Κεφάλαιο 5 Σύγκριση μοντέλων ταξινόμησης και αποτελέσματα

Με την ολοκλήρωση της διαδικασίας της επισημείωσης, το dataset αποτελείται από 31697 tokens και τα tags τους (βλ. Εικόνα 4.3.1.1). Προκειμένου να γίνουν οι δοκιμές με τους αλγορίθμους μηχανικής μάθησης, χρειάζεται να δημιουργηθούν τα training set και test set, με βάση το τρέχον dataset. Το training set/ test set αποτελείται από 31697 παραδείγματα (ένα παράδειγμα για κάθε token). Για το εκάστοτε focus word (token) ορίζονται τα εξής attributes, με βάση τα συμφραζόμενά του (βλ. Εικόνα 5.1):

- *-3 word*. Η τρίτη προηγούμενη γειτονική λέξη του focus word
- *-3 tag*. Το tag της τρίτης προηγούμενης γειτονικής λέξης του focus word

- *-2 word*. Η δεύτερη προηγούμενη γειτονική λέξη του focus word
- *-2 tag*. Το tag της δεύτερης προηγούμενης γειτονικής λέξης του focus word

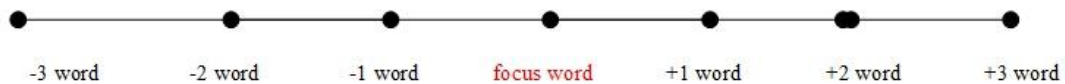
- *-1 word*. Η αμέσως προηγούμενη γειτονική λέξη του focus word
- *-1 tag*. Το tag της αμέσως προηγούμενης γειτονικής λέξης του focus word

- *+1 word*. Η αμέσως επόμενη γειτονική λέξη του focus word
- *+1 tag*. Το tag της αμέσως επόμενης γειτονικής λέξης του focus word

- *+2 word*. Η δεύτερη επόμενη γειτονική λέξη του focus word
- *+2 tag*. Το tag της δεύτερης επόμενης γειτονικής λέξης του focus word

- *+3 word*. Η τρίτη επόμενη γειτονική λέξη του focus word
- *+3 tag*. Το tag της τρίτης επόμενης γειτονικής λέξης του focus word

- *3-char suffix*. Η κατάληξη του focus word (οι 3 τελευταίοι χαρακτήρες)



Εικόνα 5.1 Focus word, η θέση του στο tweet και τα συμφραζόμενά του (γειτονικές λέξεις)

Για όλα τα παραπάνω attributes, εκτός του "3-char suffix", ισχύει ότι μπορούν να παίρνουν την τιμή "NULL" (tags) ή το κενό (words) εάν το focus word δεν έχει συμφραζόμενα σε εκείνες τις θέσεις. Αυτό συμβαίνει συνήθως για tokens (λέξεις) που βρίσκονται κοντά στην αρχή ή στο τέλος του tweet, τα οποία δεν έχουν αρκετές γειτονικές λέξεις για να αποδοθεί τιμή στα attributes που συνδέονται με αυτές.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	token	-3 word	-3 tag	-2 word	-2 tag	-1 word	-1 tag	+1 word	+1 tag	+2 word	+2 tag	+3 word	+3 tag	3-char suffix	token's tag
2	@ersikal		NULL		NULL		NULL	Σε	AN	ευχαριστώ	P	ΣΗΤ	ΣΣ	kal	U
3	Σε		NULL		NULL	@ersikal	U	ευχαριστώ	P	ΣΗΤ	ΣΣ		NULL	Σε	AN
4	ευχαριστώ		NULL	@ersikal	U	Σε	AN	ΣΗΤ	ΣΣ		NULL		NULL	στώ	P
5	ΣΗΤ	@ersikal	U	Σε	AN	ευχαριστώ	P		NULL		NULL		NULL	ΗΤ	ΣΣ
6	Συνέχεια		NULL		NULL		NULL	βλέπουμε	P	κινητά	KNO	με	PR	εια	EP
7	βλέπουμε		NULL		NULL	Συνέχεια	EP	κινητά	KNO	με	PR	πολύ	EP	υμε	P
8	κινητά		NULL	Συνέχεια	EP	βλέπουμε	P	με	PR	πολύ	EP	δυνατές	EO	ητά	KNO
9	με	Συνέχεια	EP	βλέπουμε	P	κινητά	KNO	πολύ	EP	δυνατές	EO	μπαταρίες	KNO	με	PR
10	πολύ	βλέπουμε	P	κινητά	KNO	με	PR	δυνατές	EO	μπαταρίες	KNO	ΣΗΤ	ΣΣ	ολύ	EP
11	δυνατές	κινητά	KNO	με	PR	πολύ	EP	μπαταρίες	KNO	ΣΗΤ	ΣΣ	αφού	ΣYN	τές	EO
12	μπαταρίες	με	PR	πολύ	EP	δυνατές	EO	ΣΗΤ	ΣΣ	αφού	ΣYN	υπάρχουν	P	ίες	KNO
13	ΣΗΤ	πολύ	EP	δυνατές	EO	μπαταρίες	KNO	αφού	ΣYN	υπάρχουν	P	και	ΣYN	ΗΤ	ΣΣ
14	αφού	δυνατές	EO	μπαταρίες	KNO	ΣΗΤ	ΣΣ	υπάρχουν	P	και	ΣYN	αυτοί	AN	φού	ΣYN
15	υπάρχουν	μπαταρίες	KNO	ΣΗΤ	ΣΣ	αφού	ΣYN	και	ΣYN	αυτοί	AN	οι	AP	ουν	P
16	και	ΣΗΤ	ΣΣ	αφού	ΣYN	υπάρχουν	P	αυτοί	AN	οι	AP	χρήστες	KNO	και	ΣYN
17	αυτοί	αφού	ΣYN	υπάρχουν	P	και	ΣYN	οι	AP	χρήστες	KNO	που	AN	τοί	AN
18	οι	υπάρχουν	P	και	ΣYN	αυτοί	AN	χρήστες	KNO	που	AN	θέλουν	P	οι	AP
19	χρήστες	και	ΣYN	αυτοί	AN	οι	AP	που	AN	θέλουν	P	να	ΣYN	τες	KNO
20	που	αυτοί	AN	οι	AP	χρήστες	KNO	θέλουν	P	να	ΣYN	αισθάνονται	P	που	AN
21	θέλουν	οι	AP	χρήστες	KNO	που	AN	να	ΣYN	αισθάνονται	P	HTTP	ΥΣ	ουν	P
22	να	χρήστες	KNO	που	AN	θέλουν	P	αισθάνονται	P	HTTP	ΥΣ		NULL	να	ΣYN
23	αισθάνονται	που	AN	θέλουν	P	να	ΣYN	HTTP	ΥΣ		NULL		NULL	ται	P
24	HTTP	θέλουν	P	να	ΣYN	αισθάνονται	P		NULL		NULL		NULL	TTP	ΥΣ
25	Περισσότεροι		NULL		NULL		NULL	από	PR	140	APΘM	επιχειρηματίες	KNO	ροι	EO
26	από		NULL		NULL	Περισσότεροι	EO	140	APΘM	επιχειρηματίες	KNO	θα	M	από	PR
27	140		NULL	Περισσότεροι	EO	από	PR	επιχειρηματίες	KNO	θα	M	συνοδεύσουν	P	140	APΘM
28	επιχειρηματίες	Περισσότεροι	EO	από	PR	140	APΘM	θα	M	συνοδεύσουν	P	τον	AP	ίες	KNO
29	θα	από	PR	140	APΘM	επιχειρηματίες	KNO	συνοδεύσουν	P	τον	AP	Πρωθυπουργό	KPO	θα	M
30	συνοδεύσουν	140	APΘM	επιχειρηματίες	KNO	θα	M	τον	AP	Πρωθυπουργό	KPO	Δλέφν	KPO	ουν	P

Εικόνα 5.2 Δείγμα dataset με attributes και labels. Από αυτό προκύπτουν τα training sets και test sets για τη μηχανική μάθηση

Η στήλη "token's tag" περιέχει το tag που έχει αποδοθεί στο κάθε token στο στάδιο της επισημείωσης (βλ. Κεφάλαιο 4, Ενότητα 4.3). Αυτή αποτελεί και την κλάση (label ή predicted class) του κάθε παραδείγματος για τους αλγορίθμους μηχανικής μάθησης. Από το στάδιο αυτό και έπειτα, το dataset στο οποίο γίνεται αναφορά είναι αυτό περιγράφεται παραπάνω (βλ. Εικόνα 5.2), και από αυτό προκύπτουν σε κάθε δοκιμή τα training sets και test sets.

Το εργαλείο που χρησιμοποιήθηκε για την διεξαγωγή των δοκιμών με αλγορίθμους μηχανικής μάθησης, τη συλλογή των αποτελεσμάτων, καθώς και τη σύγκριση των μοντέλων είναι το RapidMiner Studio Educational (<https://rapidminer.com/products/studio/>).

Στην Ενότητα 5.1 αναλύεται η μεθοδολογία που ακολουθήθηκε κατά τις δοκιμές με τον αλγόριθμο μηχανικής μάθησης Naive Bayes, καθώς και τα αποτελέσματα από τις δοκιμές αυτές. Στην Ενότητα 5.2 περιγράφεται αναλυτικά η μεθοδολογία που ακολουθήθηκε κατά τις δοκιμές με τον αλγόριθμο μηχανικής μάθησης ID3, ενώ παρουσιάζονται και τα αποτελέσματα από τις δοκιμές αυτές.

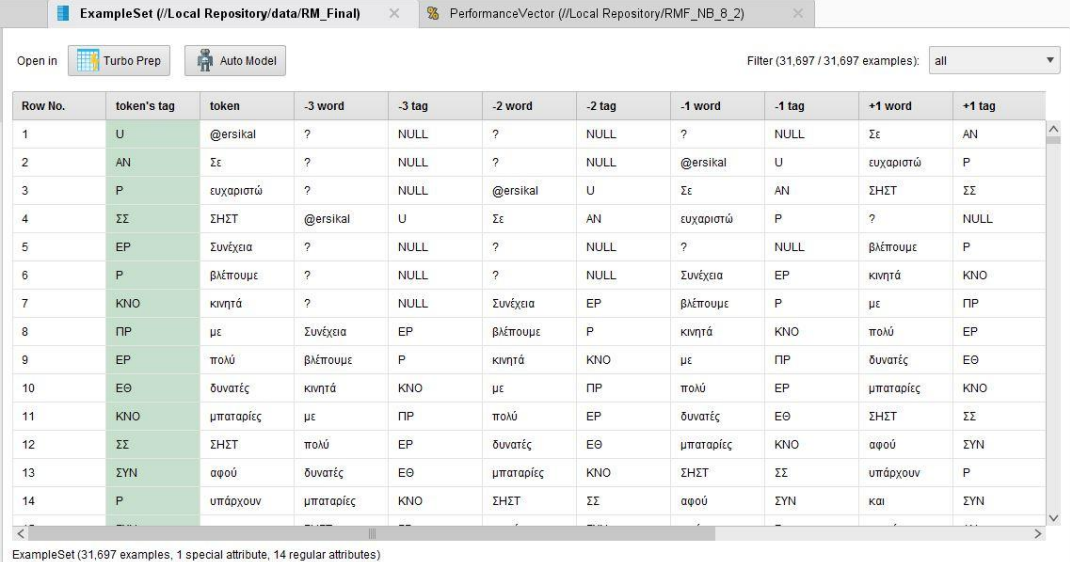
5.1 Αλγόριθμος Naive Bayes

Οι πρώτες δοκιμές έγιναν με τον αλγόριθμο μηχανικής μάθησης Naive Bayes (supervised learning). Ο αλγόριθμος αυτός περιγράφεται αναλυτικά στο Κεφάλαιο 3, Ενότητα 3.1.

Στην υποενότητα 5.1.1 αναλύεται η μεθοδολογία που ακολουθήθηκε κατά τις δοκιμές με τον συγκεκριμένο αλγόριθμο. Στην υποενότητα 5.1.2 παρουσιάζονται εκτενώς τα αποτελέσματα από τις δοκιμές αυτές.

5.1.1 Μεθοδολογία

Αρχικά, το dataset φορτώθηκε στο εργαλείο RapidMiner Studio Educational, όπου ορίστηκε το attribute "token's tag" ως label για καθένα από τα παραδείγματα μάθησης (βλ. Εικόνα 5.1.1.1).



Row No.	token's tag	token	-3 word	-3 tag	-2 word	-2 tag	-1 word	-1 tag	+1 word	+1 tag
1	U	@ersikal	?	NULL	?	NULL	?	NULL	Σε	AN
2	AN	Σε	?	NULL	?	NULL	@ersikal	U	ευχαριστώ	P
3	P	ευχαριστώ	?	NULL	@ersikal	U	Σε	AN	ΣΗΣΤ	ΣΣ
4	ΣΣ	ΣΗΣΤ	@ersikal	U	Σε	AN	ευχαριστώ	P	?	NULL
5	EP	Συνέχεια	?	NULL	?	NULL	?	NULL	βλέπουμε	P
6	P	βλέπουμε	?	NULL	?	NULL	Συνέχεια	EP	κινητά	KNO
7	KNO	κινητά	?	NULL	Συνέχεια	EP	βλέπουμε	P	με	ΠΡ
8	ΠΡ	με	Συνέχεια	EP	βλέπουμε	P	κινητά	KNO	πολύ	EP
9	EP	πολύ	βλέπουμε	P	κινητά	KNO	με	ΠΡ	δυνατές	EΘ
10	EΘ	δυνατές	κινητά	KNO	με	ΠΡ	πολύ	EP	μπαταρίες	KNO
11	KNO	μπαταρίες	με	ΠΡ	πολύ	EP	δυνατές	EΘ	ΣΗΣΤ	ΣΣ
12	ΣΣ	ΣΗΣΤ	πολύ	EP	δυνατές	EΘ	μπαταρίες	KNO	αφού	ΣΥΝ
13	ΣΥΝ	αφού	δυνατές	EΘ	μπαταρίες	KNO	ΣΗΣΤ	ΣΣ	υπάρχουν	P
14	P	υπάρχουν	μπαταρίες	KNO	ΣΗΣΤ	ΣΣ	αφού	ΣΥΝ	και	ΣΥΝ

Εικόνα 5.1.1.1 Δείγμα dataset μετά τη φόρτωσή του στο RapidMiner Studio Educational

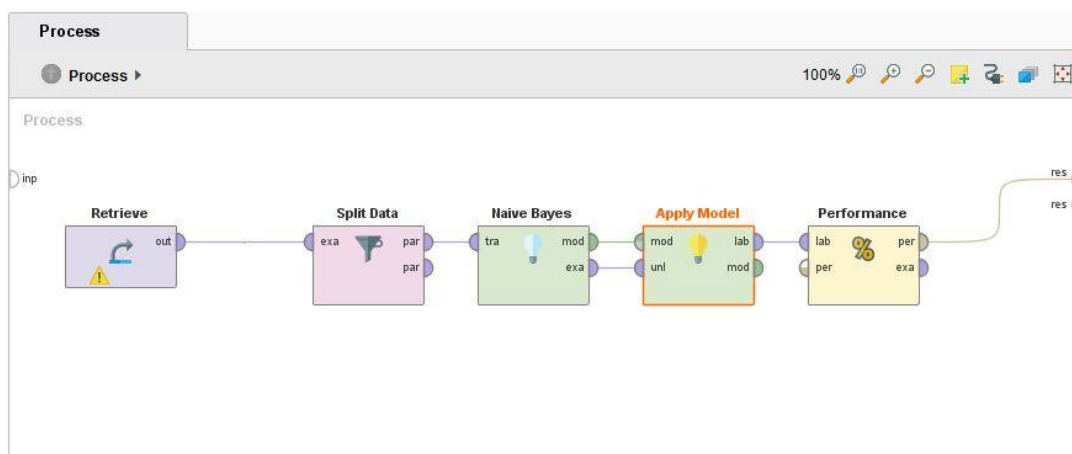
Στη συνέχεια, το dataset χωρίζεται από τον operator "Split Data" (<https://docs.rapidminer.com/latest/studio/operators/blending/examples/sampling/split>

[_data.html](#)), ώστε το 80% να αποτελεί το training set και το 20% το test set. Το training set χρησιμοποιείται για την εκπαίδευση του αλγορίθμου, ενώ το test set για την αξιολόγηση του μοντέλου ταξινόμησης.

Έπειτα, εφαρμόζεται ο operator "Naive Bayes" (https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/bayesian/naive_bayes.html) για τη δημιουργία του μοντέλου ταξινόμησης. Με την ενεργοποίηση της παραμέτρου laplace_correction, αντιμετωπίζεται η εξής αδυναμία του αλγορίθμου: εάν μέσα στα δεδομένα εκπαίδευσης (training set) μια δεδομένη τιμή κάποιου attribute δεν εμφανίζεται ποτέ στο πλαίσιο μιας δεδομένης κλάσης, τότε η δεσμευμένη πιθανότητα (conditional probability) τίθεται στο μηδέν. Όταν αυτή η μηδενική τιμή πολλαπλασιάζεται μαζί με άλλες πιθανότητες, αυτές οι τιμές είναι επίσης μηδενικές και τα αποτελέσματα θα είναι παραπλανητικά. Το Laplace correction είναι ένα απλό τέχνασμα για την αποφυγή αυτού του προβλήματος, προσθέτοντας ένα (+1) σε κάθε μέτρηση για να αποφευχθεί η εμφάνιση μηδενικών τιμών. Για τα περισσότερα training sets, η πρόσθεση ενός σε κάθε μέτρηση έχει αμελητέα επίδραση στις εκτιμώμενες πιθανότητες. Έτσι, δημιουργείται το μοντέλο ταξινόμησης.

Μετά, χρησιμοποιείται ο operator "Apply Model" (https://docs.rapidminer.com/latest/studio/operators/scoring/apply_model.html) για την εφαρμογή του μοντέλου ταξινόμησης στο test set και, τέλος, ο operator "Performance" (<https://docs.rapidminer.com/latest/studio/operators/validation/performance/performance.html>) εφαρμόζεται για την αξιολόγηση της απόδοσης του μοντέλου ταξινόμησης, υπολογίζοντας τις τιμές διάφορων κριτηρίων απόδοσης (βλ. Πίνακας 5.1.2.1).

Η υλοποίηση της μεθοδολογίας, χρησιμοποιώντας το εργαλείο RapidMiner Studio Educational, όπως περιγράφεται παραπάνω, φαίνεται συνοπτικά στην Εικόνα 5.1.1.2.



Εικόνα 5.1.1.2 Δημιουργία και αξιολόγηση μοντέλου ταξινόμησης με βάση τον αλγόριθμο μηχανικής μάθησης "Naive Bayes" (επιβλεπόμενη μάθηση)

5.1.2 Αποτελέσματα

Τα αποτελέσματα των δοκιμών, που περιγράφονται στην προηγούμενη υποενότητα, φαίνονται πλήρως στην Εικόνα 5.1.2.1, όπου απεικονίζονται το confusion matrix, καθώς και οι τιμές των precision και recall για κάθε label. Το μοντέλο έχει **accuracy 99.87%** και **kappa statistics 0.999**, τα οποία είναι πολύ υψηλά σε σύγκριση με τη βιβλιογραφία (βλ. Κεφάλαιο 6, Ενότητα 6.1).

	U	AN	P	Σ	EP	KNO	ΠP	EΘ	ΣYN	AP	YΣ	APOM	M	KPO	Ξ	Δ	EK	SYM	E	SYNTM	EΦ	H	class
	1349	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	precision
pred. U	0	1850	0	0	1	0	0	0	0	2	1	0	0	0	1	0	0	0	0	0	0	0	099.73%
pred. AN	0	0	3318	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	099.94%
pred. P	0	0	0	2954	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0100.00%
pred. Σ	0	0	0	0	1213	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	099.84%
pred. EP	0	1	0	0	0	3166	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	099.91%
pred. KNO	0	0	0	0	1	0	891	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	099.89%
pred. ΠP	0	0	0	0	0	1	0	1168	0	0	0	1	0	0	0	0	0	0	0	0	0	0	099.83%
pred. EΘ	0	0	0	0	0	0	0	0	1547	0	0	0	0	0	0	0	0	0	0	0	0	0	0100.00%
pred. ΣYN	0	1	0	0	0	0	0	0	0	2671	0	0	0	0	1	0	0	0	0	0	0	0	099.93%
pred. AP	0	1	0	0	0	0	0	0	0	0	1075	0	0	0	0	0	0	0	0	0	0	0	099.91%
pred. YΣ	0	0	0	0	0	0	0	0	0	0	0	372	0	0	0	0	0	0	0	0	0	0	0100.00%
pred. APOM	0	1	0	0	1	0	1	0	1	0	0	0	726	0	0	0	0	0	0	0	0	0	099.45%
pred. M	0	0	0	0	0	0	0	0	0	1	0	0	0	1274	0	0	0	0	0	0	0	0	099.92%
pred. KPO	0	0	1	0	0	0	0	0	0	0	0	0	0	0	625	0	0	0	0	1	0	0	099.68%
pred. Ξ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	98	0	0	0	0	0	0	0100.00%
pred. Δ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	54	0	0	0	0	0	0100.00%
pred. EK	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	34	0	0	0	0	0100.00%
pred. SYM	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	270	0	0	0	0100.00%
pred. E	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	173	0	0	0100.00%
pred. SYNTM	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	241	0	0100.00%
pred. EΦ	8	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	25696.60%	
pred. H																							
class																							
recall	99.41%	99.78%	99.97%	100.00%	99.67%	99.94%	99.89%	99.91%	99.81%	99.89%	99.91%	99.73%	100.00%	99.77%	99.84%	100.00%	100.00%	100.00%	100.00%	99.43%	100.00%	100.00%	

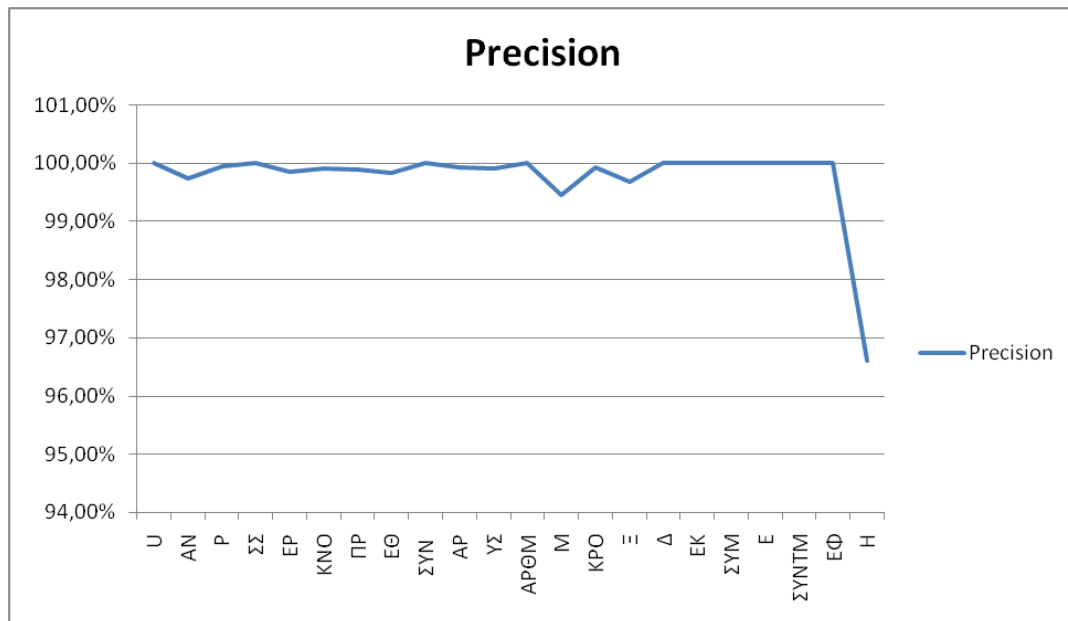
Εικόνα 5.1.2.1 Confusion matrix αποτελεσμάτων για τον αλγόριθμο Naive Bayes

Οι δείκτες precision, recall και F_1 score έχουν επίσης υψηλές τιμές και παρουσιάζονται αναλυτικά για κάθε label στον Πίνακα 5.1.2.1.

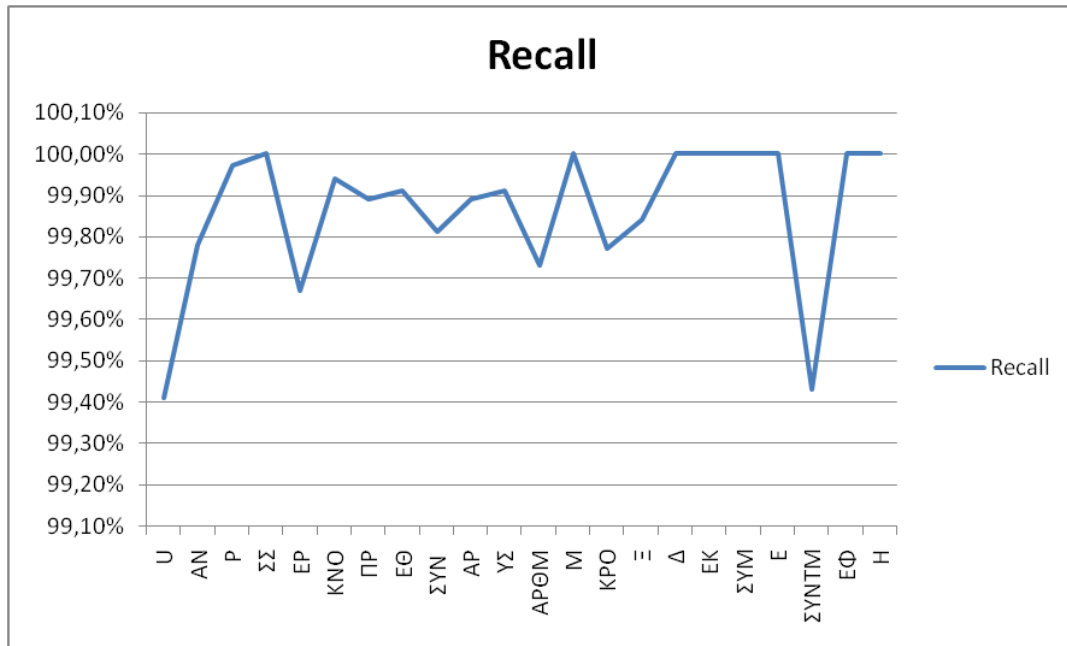
Label	Precision	Recall	F_1 score
U	100.00%	99.41%	99.70%
AN	99.73%	99.78%	99.75%
P	99.94%	99.97%	99.95%
ΣΣ	100.00%	100.00%	100.00%
EP	99.84%	99.67%	99.75%
KNO	99.91%	99.94%	99.92%
ΠΠ	99.89%	99.89%	99.89%
ΕΘ	99.83%	99.91%	99.86%
ΣΥΝ	100.00%	99.81%	99.90%
ΑΡ	99.93%	99.89%	99.90%
ΥΣ	99.91%	99.91%	99.91%
ΑΡΘΜ	100.00%	99.73%	99.86%
Μ	99.45%	100.00%	99.72%
ΚΡΟ	99.92%	99.77%	99.84%
Ξ	99.68%	99.84%	99.75%
Δ	100.00%	100.00%	100%
ΕΚ	100.00%	100.00%	100%
ΣΥΜ	100.00%	100.00%	100%
Ε (λατινικό)	100.00%	100.00%	100%
ΣΥΝΤΜ	100.00%	99.43%	99.71%
ΕΦ	100.00%	100.00%	100%
Η (λατινικό)	96.60%	100.00%	98.27%

Πίνακας 5.1.2.1 Τιμές των δεικτών *precision*, *recall* και F_1 score για κάθε label (Naive Bayes)

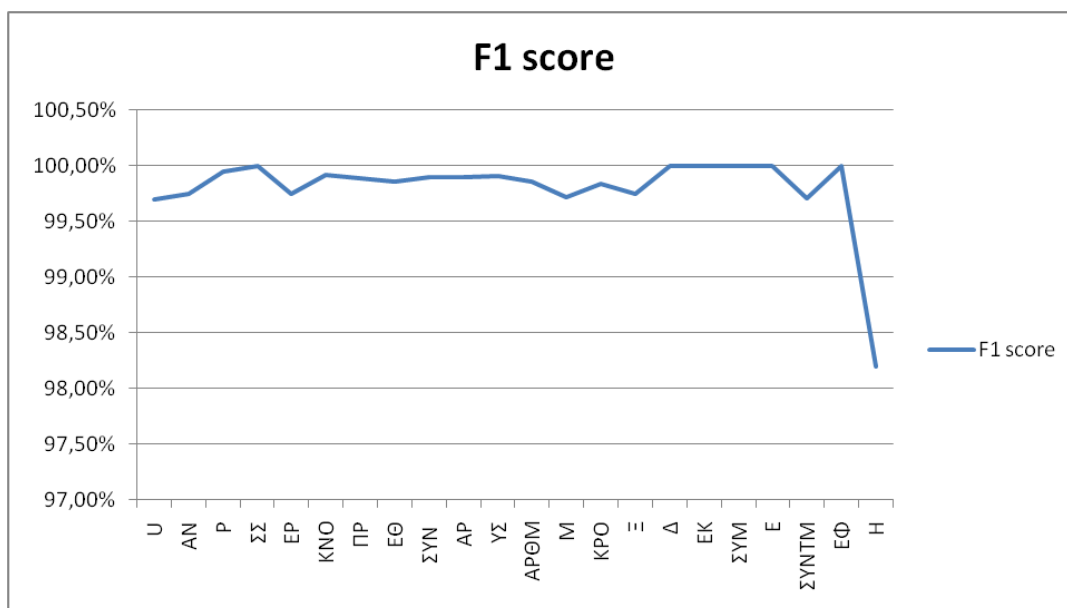
Στα Διαγράμματα 5.1.2.1, 5.1.2.2 και 5.1.2.3 απεικονίζονται γραφικά οι τιμές των δεικτών *precision*, *recall* και F_1 score για κάθε label. Οι δείκτες και τα κριτήρια απόδοσης που παρουσιάζονται σε αυτό το κεφάλαιο, περιγράφονται αναλυτικά στο Κεφάλαιο 3.



Διάγραμμα 5.1.2.1 Γραφική απεικόνιση των τιμών του δείκτη *precision* για κάθε label (Naive Bayes)



Διάγραμμα 5.1.2.2 Γραφική απεικόνιση των τιμών του δείκτη recall για κάθε label (Naive Bayes)



Διάγραμμα 5.1.2.3 Γραφική απεικόνιση των τιμών του δείκτη F_1 score για κάθε label (Naive Bayes)

5.2 Αλγόριθμος ID3

Οι επόμενες δοκιμές έγιναν με τον αλγόριθμο μηχανικής μάθησης ID3 (supervised learning). Ο αλγόριθμος αυτός περιγράφεται αναλυτικά στο Κεφάλαιο 3, Ενότητα 3.2.

Στην υποενότητα 5.2.1 αναλύεται η μεθοδολογία που ακολουθήθηκε κατά τις δοκιμές με τον συγκεκριμένο αλγόριθμο. Στην υποενότητα 5.2.2 παρουσιάζονται εκτενώς τα αποτελέσματα από τις δοκιμές αυτές.

5.2.1 Μεθοδολογία

Αρχικά, το dataset φορτώθηκε στο εργαλείο RapidMiner Studio Educational, όπου ορίστηκε το attribute "token's tag" ως label για καθένα από τα παραδείγματα μάθησης (βλ. Εικόνα 5.1.1.1), όπως ακριβώς έγινε και για τις δοκιμές με τον αλγόριθμο μηχανικής μάθησης Naive Bayes (βλ. Κεφάλαιο 5, Ενότητα 5.1).

Στη συνέχεια, τα attributes του dataset μετατρέπονται από τύπου text σε nominal από τον operator "Text to Nominal" (https://docs.rapidminer.com/latest/studio/operators/blending/attributes/types/text_to_nominal.html). Κάθε τιμή τύπου text χρησιμοποιείται απλά ως τιμή τύπου nominal του νέου attribute. Αυτό γίνεται διότι ο αλγόριθμος μηχανικής μάθησης ID3 απαιτεί attributes τύπου nominal για τη δημιουργία του μοντέλου ταξινόμησης.

Έπειτα, το dataset χωρίζεται από τον operator "Split Data" (https://docs.rapidminer.com/latest/studio/operators/blending/examples/sampling/split_data.html), ώστε το 80% να αποτελεί το training set και το 20% το test set. Το training set χρησιμοποιείται για την εκπαίδευση του αλγορίθμου, ενώ το test set για την αξιολόγηση του μοντέλου ταξινόμησης.

Μετά, εφαρμόζεται ο operator "ID3" (<https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/trees/id3.html>) για τη δημιουργία του μοντέλου ταξινόμησης. Ορίζοντας την τιμή της παραμέτρου criterion σε gain_ratio, ρυθμίζεται το information gain για κάθε attribute, ώστε να καθορίζει το εύρος και την ομοιομορφία των τιμών των attributes.

Η τιμή της παραμέτρου minimal_size_for_split ορίζεται σε 4. Το μέγεθος ενός κόμβου είναι ο αριθμός των παραδειγμάτων στο υποσύνολο του. Το μέγεθος του κόμβου ρίζας είναι ίσο με τον συνολικό αριθμό παραδειγμάτων (examples) του dataset. Με την παράμετρο αυτή καθορίζεται ότι θα χωρίζονται μόνο οι κόμβοι των οποίων το μέγεθος είναι μεγαλύτερο από ή ίσο με 4.

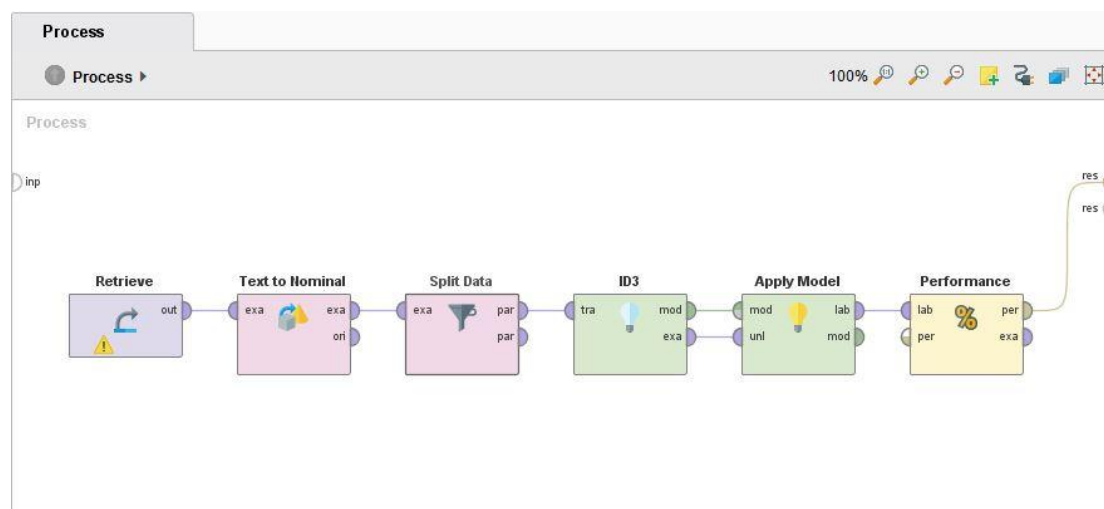
Η τιμή της παραμέτρου minimal_leaf_size ορίζεται σε 2. Το μέγεθος ενός κόμβου-φύλλο είναι ο αριθμός των παραδειγμάτων στο υποσύνολο του. Το δέντρο

δημιουργείται με τέτοιον τρόπο, ώστε το κάθε υποσύνολο κόμβου-φύλλου να έχει τουλάχιστον 2 παραδείγματα.

Η τιμή της παραμέτρου `minimal_gain` ορίζεται σε 0.01. Το information gain ενός κόμβου υπολογίζεται και, εάν είναι μεγαλύτερο από 0.01, ο κόμβος χωρίζεται. Όσο πιο υψηλή είναι η τιμή της συγκεκριμένης παραμέτρου, τόσο λιγότερες διαιρέσεις απαιτούνται και, κατά συνέπεια, δημιουργείται μικρότερο δέντρο (με λιγότερους κόμβους). Μια υπερβολικά υψηλή τιμή αποτρέπει εντελώς τη διαίρεση και έχει ως αποτέλεσμα τη δημιουργία ενός δέντρου με έναν μόνο κόμβο.

Έπειτα, χρησιμοποιείται ο operator "Apply Model" (https://docs.rapidminer.com/latest/studio/operators/scoring/apply_model.html) για την εφαρμογή του μοντέλου ταξινόμησης στο test set και, τέλος, ο operator "Performance" (<https://docs.rapidminer.com/latest/studio/operators/validation/performance/performance.html>) εφαρμόζεται για την αξιολόγηση της απόδοσης του μοντέλου ταξινόμησης, υπολογίζοντας τις τιμές διάφορων κριτηρίων απόδοσης (βλ. Πίνακας 5.2.2.1).

Η υλοποίηση της μεθοδολογίας, χρησιμοποιώντας το εργαλείο RapidMiner Studio Educational, όπως περιγράφεται παραπάνω, φαίνεται συνοπτικά στην Εικόνα 5.2.1.1



Εικόνα 5.2.1.1 Δημιουργία και αξιολόγηση μοντέλου ταξινόμησης με βάση τον αλγόριθμο μηχανικής μάθησης "ID3" (επιβλεπόμενη μάθηση)

5.2.2 Αποτελέσματα

Τα αποτελέσματα των δοκιμών, που περιγράφονται στην προηγούμενη υποενότητα, φαίνονται πλήρως στην Εικόνα 5.2.2.1, όπου απεικονίζονται το

confusion matrix, καθώς και οι τιμές των precision και recall για κάθε label. Το μοντέλο έχει **accuracy 99.44%** και **kappa statistics 0.994**, τα οποία είναι πολύ υψηλά σε σύγκριση με τη βιβλιογραφία (βλ. Κεφάλαιο 6, Ενότητα 6.1).

[illegible]

Εικόνα 5.2.2.1 Confusion matrix αποτελεσμάτων για τον αλγόριθμο ID3

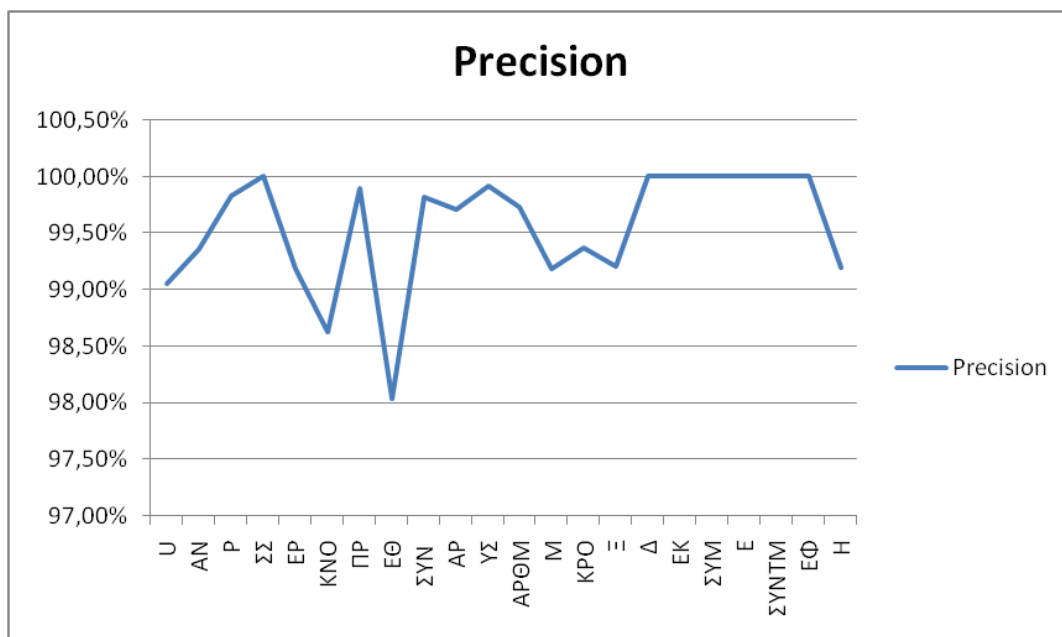
Οι δείκτες precision, recall και F₁ score έχουν επίσης υψηλές τιμές και παρουσιάζονται αναλυτικά για κάθε label στον Πίνακα 5.2.2.1.

Label	Precision	Recall	F ₁ score
U	99.05%	99.78%	99.41%
AN	99.35%	99.30%	99.32%
P	99.82%	99.85%	99.83%
$\Sigma\Sigma$	100.00%	100.00%	100.00%
EP	99.18%	99.59%	99.38%
KNO	98.62%	99.46%	99.03%
PP	99.89%	99.78%	99.83%

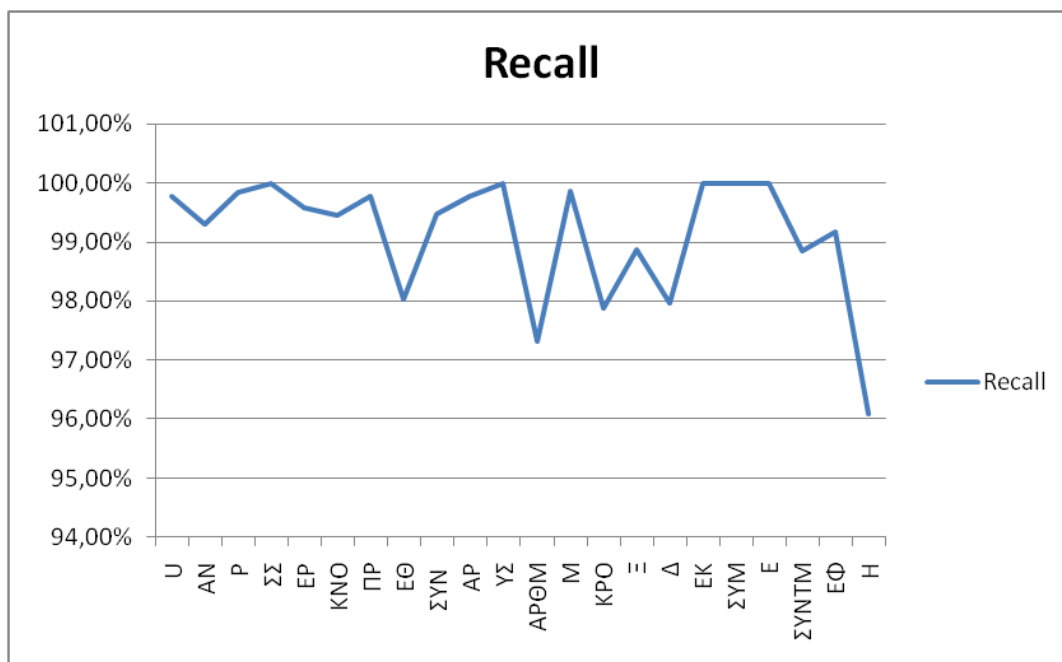
ΕΘ	98.03%	98.03%	98.03%
ΣΥΝ	99.81%	99.48%	99.64%
ΑΡ	99.70%	99.78%	99.73%
ΥΣ	99.91%	100.00%	99.95%
ΑΡΘΜ	99.73%	97.32%	98.51%
Μ	99.18%	99.86%	99.51%
ΚΡΟ	99.36%	97.89%	98.61%
Ξ	99.20%	98.88%	99.03%
Δ	100.00%	97.96%	98.96%
ΕΚ	100.00%	100.00%	100%
ΣΥΜ	100.00%	100.00%	100%
Ε (λατινικό)	100.00%	100.00%	100%
ΣΥΝΤΜ	100.00%	98.85%	99.42%
ΕΦ	100.00%	99.17%	99.58%
Η (λατινικό)	99.19%	96.09%	97.61%

***Πίνακας 5.2.2.1** Τιμές των δεικτών precision, recall και F_1 score για κάθε label (ID3)*

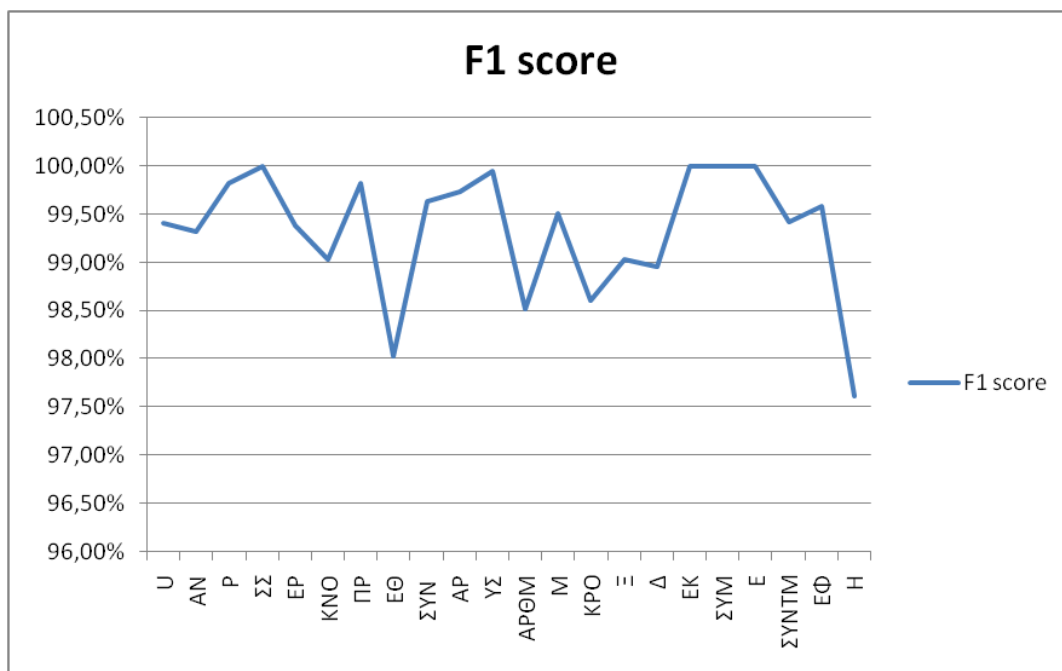
Στα Διαγράμματα 5.2.2.1, 5.2.2.2 και 5.2.2.3 απεικονίζονται γραφικά οι τιμές των δεικτών precision, recall και F_1 score για κάθε label. Οι δείκτες και τα κριτήρια απόδοσης που παρουσιάζονται σε αυτό το κεφάλαιο, περιγράφονται αναλυτικά στο Κεφάλαιο 3.



Διάγραμμα 5.2.2.1 Γραφική απεικόνιση των τιμών του δείκτη precision για κάθε label (ID3)



Διάγραμμα 5.2.2.2 Γραφική απεικόνιση των τιμών του δείκτη recall για κάθε label (ID3)



Διάγραμμα 5.2.2.3 Γραφική απεικόνιση των τιμών του δείκτη F_1 score για κάθε label (ID3)

Κεφάλαιο 6 Συμπεράσματα

Στην παρούσα έρευνα αναπτύχθηκε το πρώτο μέχρι στιγμής εργαλείο αναγνώρισης μερών του λόγου για την ελληνική γλώσσα του κοινωνικού ιστού. Η καινοτομία της έρευνας είναι εμφανής σε κάθε στάδιό της, από τη συλλογή των δεδομένων έως τα αποτελέσματα των δοκιμών με τους αλγορίθμους μηχανικής μάθησης. Το εργαλείο αυτό, λόγω της υψηλής απόδοσής του στις δοκιμές μηχανικής μάθησης, έχει προοπτική να χρησιμοποιηθεί ως βάση για μελλοντική έρευνα στο πεδίο του POS tagging για δεδομένα τύπου microblogging στα ελληνικά από το κοινωνικό δίκτυο Twitter, αλλά και από άλλα κοινωνικά δίκτυα γενικότερα.

Στην Ενότητα 6.1 συνοψίζεται η συνεισφορά της πτυχιακής εργασίας. Στην Ενότητα 6.2 γίνονται ορισμένες προτάσεις για μελλοντική έρευνα που προκύπτουν από την παρούσα εργασία.

6.1 Συνεισφορά πτυχιακής εργασίας

Αρχικά, σημαντική συνεισφορά αποτελεί η συλλογή, η κανονικοποίηση και η δημιουργία του πρώτου μέχρι στιγμής επισημειωμένου (POS tagging) dataset για την ελληνική γλώσσα του κοινωνικού ιστού, μεγέθους 2405 tweets ή 31697 tokens (βλ. Κεφάλαιο 4). Το νέο tag set (22 tags), που χρησιμοποιήθηκε για την επισημείωση των δεδομένων, αναπτύχθηκε με βάση τα μέρη του λόγου της ελληνικής γλώσσας (Τζεβελέκου, Κάντζου, & Σταμούλη, 2007), (Τριανταφυλλίδης, 1990) και τα ιδιαίτερα γλωσσολογικά χαρακτηριστικά της γλώσσας του κοινωνικού δικτύου Twitter (όνομα χρήστη, υπερσύνδεσμος, διεύθυνση ηλεκτρονικού ταχυδρομείου, hashtag και emoji/ emoticon).

Χρησιμοποιώντας το επισημειωμένο dataset που αναφέρεται στην προηγούμενη παράγραφο, αναπτύχθηκε το πρώτο μέχρι στιγμής εργαλείο αναγνώρισης μερών του λόγου για την ελληνική γλώσσα του κοινωνικού ιστού (βλ. Κεφάλαιο 5). Το dataset χωρίστηκε σε training set (80%) και test set (20%) για τις δοκιμές στο εργαλείο RapidMiner Studio Educational με τους αλγορίθμους μηχανικής μάθησης Naive Bayes και ID3 (επιβλεπόμενη μάθηση). Παρατηρείται υψηλή απόδοση στις δοκιμές και με τους δύο αλγορίθμους σε σύγκριση με τη βιβλιογραφία (περίπου 90% accuracy στην πλειονότητα των εργαλείων για άλλες γλώσσες), με accuracy 99.87% και kappa

statistics 0.999 για τον αλγόριθμο Naive Bayes, και accuracy 99.44% και kappa statistics 0.994 για τον αλγόριθμο ID3 (βλ. Εικόνα 5.1.2.1, Εικόνα 5.2.2.1).

Η απόδοση του εργαλείου αναγνώρισης μερών του λόγου για την ελληνική γλώσσα του κοινωνικού ιστού είναι υψηλή ακόμα και συγκριτικά με αντίστοιχα εργαλεία για δόκιμη και επίσημη ελληνική γλώσσα. Συγκεκριμένα, οι Petasis et al. (2001) σε έρευνά τους αναπτύσσουν έναν μορφολογικό αναλυτή με βάση τα λήμματα των λέξεων, για να χρησιμοποιηθεί ως βάση στην ανάπτυξη ενός εργαλείου ελέγχου ορθογραφίας για τα ελληνικά. Το tag set τους είναι το ίδιο λεπτομερές με των Sgarbas et al. (2001). Πέρα από τα μορφολογικά χαρακτηριστικά κάθε token (λήμμα και κατάληξη), λαμβάνει υπόψιν του και τα συμφραζόμενα. Το accuracy του POS tagger τους είναι 95%, αρκετά μικρότερο από αυτό της παρούσας έρευνας (99.87% και 99.44%). Ακόμη, ο POS tagger των Papageorgiou et al. (2000), ο οποίος δοκιμάστηκε με 447000 tokens από οικονομικές αναφορές, ειδησεογραφικά πρακτορεία και τύπο, και εγχειρίδια οδηγιών και tag set μεγέθους 584 tags, δεν πετυχαίνει υψηλότερο accuracy από 96.28%

Για τον αλγόριθμο Naive Bayes, οι δείκτες αξιολόγησης precision, recall και F_1 score για κάθε label είναι υψηλοί (βλ. Πίνακας 5.1.2.1). Πιο συγκεκριμένα, παρατηρούνται τα εξής: το precision κυμαίνεται από 99% έως 100% για όλα τα labels, εκτός από αυτό του "H" που έχει την χαμηλότερη τιμή 96.60% (βλ. Διαγράμμα 5.1.2.1), το recall κυμαίνεται από 99% έως 100% για όλα τα labels, με αυτά των "U" και "ΣΥΝΤΜ" να έχουν τις χαμηλότερες τιμές 99.41% και 99.43%, αντίστοιχα (βλ. Διαγράμμα 5.1.2.2), και το F_1 score κυμαίνεται από 99% έως 100% για όλα τα labels, εκτός από αυτό του "H" που έχει την χαμηλότερη τιμή 98.27% (βλ. Διαγράμμα 5.1.2.3).

Από τις τιμές των δεικτών αξιολόγησης που περιγράφονται παραπάνω για τον Naive Bayes, φαίνεται πως τα μέρη του λόγου που είναι πιο εύκολο να αναγνωριστούν είναι τα σημεία στίξης (label "ΣΣ"), τα παραδείγματα που δεν ανήκουν σε κανένα μέρος του λόγου (label "Δ"), οι εκφράσεις (label "ΕΚ"), τα σύμβολα (label "ΣΥΜ"), τα emoji/ emoticons (label "Ε") και τα επιφωνήματα (label "ΕΦ"), καθώς οι τιμές όλων των δεικτών αξιολόγησης για αυτά είναι οι μέγιστες (precision, recall και F_1 score 100%). Αυτό πιθανώς οφείλεται στο γεγονός ότι τα labels "ΣΣ", "ΕΚ", "ΣΥΜ" και "Ε" αποδίδονται συνεχώς και αποκλειστικά στα ίδια tokens ("ΣΗΣΤ", "NUMB", "SYMB", "EMOT"), λόγω της κανονικοποίησης των

δεδομένων (βλ. Κεφάλαιο 4, Ενότητα 4.2, Υποενότητα 4.2.1). Για τα labels "Δ" και "ΕΦ", πιθανώς να οφείλεται στο γεγονός ότι αποδίδονται συχνά στα ίδια ή σε μορφολογικά παρόμοια tokens ή/ και έχουν παρόμοια μέρη του λόγου ως συμφραζόμενα.

Το μέρος του λόγου hashtag (label "H") είναι το πιο δύσκολο να αναγνωριστεί από τον συγκεκριμένο αλγόριθμο, καθώς οι τιμές των δεικτών αξιολόγησης για αυτό είναι λίγο χαμηλότερες (precision 96.60%, F_1 score 98.27%) από αυτές των άλλων (precision και F_1 score μεταξύ 99% και 100%). Τα hashtag του Twitter είναι της μορφής "#text", όπου το text μπορεί να είναι token που να ανήκει σε οποιοδήποτε μέρος του λόγου, ενώ είναι συχνή και η τοποθέτηση των hashtags στο εσωτερικό των tweets, ώστε να αντικαθιστούν τα αντίστοιχα tokens χωρίς το "#", π.χ. "μεγάλη #απογοήτευση", αντί για "μεγάλη απογοήτευση". Αυτό έχει παρατηρηθεί και από άλλους ερευνητές με εργαλεία για άλλες γλώσσες του κοινωνικού ιστού (Gimpel et al., 2010), (Rehbein, 2013), (Nand et al., 2014), (Nooralahzadeh, 2014), (Bach et al., 2018). Ορισμένοι μάλιστα επιλέγουν να αποδίδουν στο token το label που θα είχε κανονικά, αγνοώντας την παρουσία του "#". Οι υπόλοιποι αποδίδουν το label "H" σε όλα τα tokens που ξεκινούν με "#", όπως γίνεται και στην παρούσα έρευνα.

Για τον αλγόριθμο ID3, οι δείκτες αξιολόγησης precision, recall και F_1 score για κάθε label είναι επίσης υψηλοί (βλ. Πίνακας 5.2.2.1). Πιο συγκεκριμένα, παρατηρούνται τα εξής: το precision κυμαίνεται από 98% έως 100% για όλα τα labels, με αυτά των "ΕΘ" και "ΚΝΟ" να έχουν τις χαμηλότερες τιμές 98.03% και 98.62%, αντίστοιχα (βλ. Διαγράμμα 5.2.2.1), το recall κυμαίνεται από 96% έως 100% για όλα τα labels, με αυτά των "H" και "ΑΡΘΜ" να έχουν τις χαμηλότερες τιμές 96.09% και 97.32%, αντίστοιχα (βλ. Διαγράμμα 5.2.2.2), και το F_1 score κυμαίνεται από 97.50% έως 100% για όλα τα labels, με αυτά των "H" και "ΕΘ" να έχουν τις χαμηλότερες τιμές 97.61% και 98.03%, αντίστοιχα (βλ. Διαγράμμα 5.2.2.3).

Από τις τιμές των δεικτών αξιολόγησης που περιγράφονται παραπάνω για τον ID3, φαίνεται πως τα μέρη του λόγου που είναι πιο εύκολο να αναγνωριστούν είναι τα σημεία στίξης (label "ΣΣ"), οι εκφράσεις (label "ΕΚ"), τα σύμβολα (label "ΣΥΜ") και τα emoji/ emoticons (label "Ε"), καθώς οι τιμές όλων των δεικτών αξιολόγησης για αυτά είναι οι μέγιστες (precision, recall και F_1 score 100%). Αυτό πιθανώς οφείλεται στο γεγονός ότι τα labels "ΣΣ", "ΕΚ", "ΣΥΜ" και "Ε" αποδίδονται συνεχώς και

αποκλειστικά στα ίδια tokens ("ΣΗΣΤ", "NUMB", "SYMB", "EMOT"), λόγω της κανονικοποίησης των δεδομένων (βλ. Κεφάλαιο 4, Ενότητα 4.2, Υποενότητα 4.2.1).

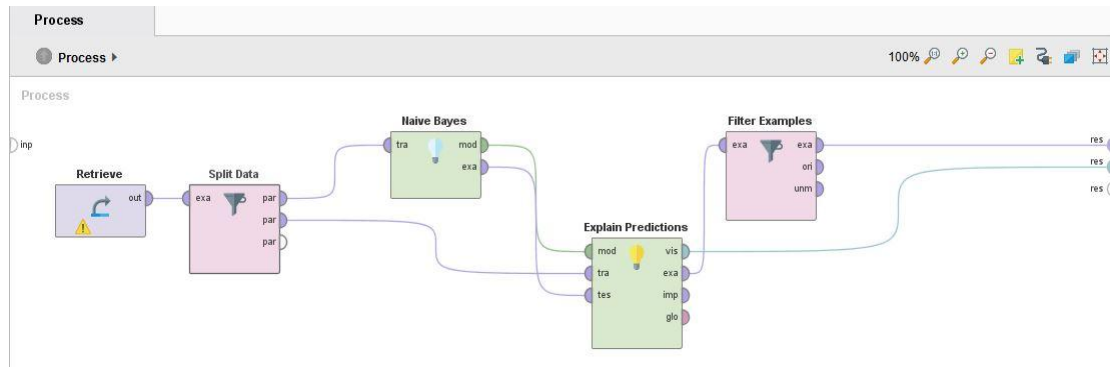
Το μέρος του λόγου hashtag (label "H") είναι το πιο δύσκολο προς αναγνώριση από τον συγκεκριμένο αλγόριθμο, καθώς οι τιμές των δεικτών αξιολόγησης για αυτό είναι λίγο χαμηλότερες (recall 96.09% και F_1 score 97.61%) από αυτές των άλλων (precision, recall και F_1 score μεταξύ 97% και 100%). Οι τιμές των δεικτών αξιολόγησης για τα επίθετα (label "ΕΘ") και τα αριθμητικά ουσιαστικά και επίθετα (label "ΑΡΘΜ") παρουσιάζουν ελάχιστες αποκλίσεις από αυτές των υπόλοιπων μερών του λόγου. Για την απόδοση στα hashtags ισχύουν οι ίδιοι λόγοι που περιγράφονται στις προηγούμενες παραγράφους για τον Naive Bayes.

Αξίζει να σημειωθεί ότι δεν παρατηρείται μεγάλη πτώση στις τιμές των δεικτών αξιολόγησης για τα περισσότερα labels που δεν είναι ιδιαιτερότητες του Twitter. Για παράδειγμα, δεν φαίνεται να υπάρχουν συχνά σφάλματα στην ταξινόμηση όταν πρόκειται για ουσιαστικά (labels "ΚΝΟ" και "ΚΡΟ") και επίθετα (label "ΕΘ"), ενώ ο POS tagger των Petasis et al. (2001) αποδίδει συχνά tag επιθέτου σε ουσιαστικό, και το αντίστροφο.

Ο εντοπισμός των λάθος ταξινομημένων παραδειγμάτων έγινε με το εργαλείο RapidMiner Studio Educational. Ακολουθώντας την ίδια μεθοδολογία με τις δοκιμές μηχανικής μάθησης με τον αλγόριθμο Naive Bayes (βλ. Κεφάλαιο 5, Ενότητα 5.1), χρησιμοποιήθηκαν οι operators "Explain Predictions" και "Filter Examples", στη θέση των operators "Apply Model" και "Performance" (βλ. Εικόνα 6.1.1). Αυτή η διαδικασία δεν ήταν δυνατόν να γίνει και για τον αλγόριθμο ID3, καθώς ο όγκος των δεδομένων και το μεγάλο πλήθος των labels δημιουργούν εξαιρετικά μεγάλες απαιτήσεις σε πόρους και υπολογιστική ισχύ, καθιστώντας πρακτικά αδύνατη την εφαρμογή των εν λόγω τελεστών στις δοκιμές αυτές.

Ο operator "Explain Predictions" (https://docs.rapidminer.com/latest/studio/operators/scoring/explain_predictions.html) λαμβάνει ως είσοδο το μοντέλο ταξινόμησης του Naive Bayes, το training set και το test set, και παράγει έναν πίνακα που επισημαίνει τα attributes που υποστηρίζουν (πράσινο χρώμα) ή διαψεύδουν (κόκκινο χρώμα) κάθε πρόβλεψη (βλ. Εικόνα 6.1.2a, Εικόνα 6.1.2b). Ουσιαστικά, βοηθάει στον εντοπισμό των attributes που έχουν σημαντικότερο ρόλο στη διαμόρφωση της πρόβλεψης. Ο operator "Filter Examples"

(https://docs.rapidminer.com/latest/studio/operators/blending/examples/filter/filter_examples.html) επιστρέφει εκείνα τα παραδείγματα που ικανοποιούν μια καθορισμένη συνθήκη. Προκειμένου να επιστρέφονται μόνο τα λάθος ταξινομημένα παραδείγματα, καθορίστηκε η τιμή της παραμέτρου condition class σε wrong_predictions. Το αποτέλεσμα φαίνεται στις Εικόνες 6.1.3a και 6.1.3b.



Εικόνα 6.1.11 Εντοπισμός των λάθος ταξινομημένων παραδειγμάτων (Naive Bayes)

ExplainPredictionsIOObject (/Local Repository/NB_RES 2)											
Row No. ↑	token's tag	prediction(t...	confidence(U	confidence(...	confidence(P	confidence(...	confidence(...	confidence(...	confidence(...	confidence(...	confidence...
1	U	U	1	0	0	0	0	0	0	0	0
2	AN	AN	0.000	1.000	0	0	0.000	0	0.000	0	0
3	P	P	0	0.000	1.000	0.000	0.000	0.000	0	0.000	0
4	ΣΣ	ΣΣ	0	0	0	1	0	0	0	0	0
5	EP	EP	0.000	0.000	0	0	1.000	0.000	0	0.000	0
6	P	P	0	0	1	0	0	0	0	0	0
7	KNO	KNO	0	0	0	0	0.000	1.000	0	0	0
8	ΠP	ΠP	0	0	0	0	0	0	1	0	0
9	EP	EP	0	0	0	0	1	0	0	0	0
10	EΘ	EΘ	0	0	0	0	0	0	0	1	0
11	KNO	KNO	0	0	0	0	0	1	0	0	0
12	ΣΣ	ΣΣ	0	0.000	0	1.000	0	0	0	0	0
13	AN	AN	0	1.000	0.000	0	0	0	0	0.000	0
14	AP	AP	0	0	0	0	0	0	0	0	0
15	KNO	KNO	0	0	0	0	0	1.000	0	0.000	0
16	P	P	0	0	1	0	0	0	0	0	0
17	P	P	0	0.000	1.000	0.000	0	0	0	0	0

Εικόνα 6.1.12a Δείγμα από τον πίνακα που παράγεται από τον operator "Explain Predictions"

ExplainPredictionsIOObject (/Local Repository/NB_RES 2)											
token	-3 word	-3 tag	-2 word	-2 tag	-1 word	-1 tag	+1 word	+1 tag	+2 word	+2 tag	+3 word
@ersikal	?	NULL	?	NULL	?	NULL	Σε	AN	ευχαριστώ	P	ΣΗΤ
Σε	?	NULL	?	NULL	@ersikal	U	ευχαριστώ	P	ΣΗΤ	ΣΣ	?
ευχαριστώ	?	NULL	@ersikal	U	Σε	AN	ΣΗΤ	ΣΣ	?	NULL	?
ΣΗΤ	@ersikal	U	Σε	AN	ευχαριστώ	P	?	NULL	?	NULL	?
Συνέχεια	?	NULL	?	NULL	?	NULL	βλέπουμε	P	κινητά	KNO	με
βλέπουμε	?	NULL	?	NULL	Συνέχεια	EP	κινητά	KNO	με	ΠΡ	πολύ
κινητά	?	NULL	Συνέχεια	EP	βλέπουμε	P	με	ΠΡ	πολύ	EP	δυνατές
με	Συνέχεια	EP	βλέπουμε	P	κινητά	KNO	πολύ	EP	δυνατές	ΕΘ	μπαταρίες
πολύ	βλέπουμε	P	κινητά	KNO	με	ΠΡ	δυνατές	ΕΘ	μπαταρίες	KNO	ΣΗΤ
δυνατές	κινητά	KNO	με	ΠΡ	πολύ	EP	μπαταρίες	KNO	ΣΗΤ	ΣΣ	αφού
μπαταρίες	με	ΠΡ	πολύ	EP	δυνατές	ΕΘ	ΣΗΤ	ΣΣ	αφού	ΣΥΝ	υπάρχουν
ΣΗΤ	πολύ	EP	δυνατές	ΕΘ	μπαταρίες	KNO	αφού	ΣΥΝ	υπάρχουν	P	και
αυτοί	αφού	ΣΥΝ	υπάρχουν	P	και	ΣΥΝ	οι	AP	χρήστες	KNO	που
οι	υπάρχουν	P	και	ΣΥΝ	αυτοί	AN	χρήστες	KNO	που	AN	θέλουν
χρήστες	και	ΣΥΝ	αυτοί	AN	οι	AP	που	AN	θέλουν	P	να
θέλουν	οι	AP	χρήστες	KNO	που	AN	να	ΣΥΝ	αισθάνονται	P	HTTP
αισθάνονται	που	AN	θέλουν	P	να	ΣΥΝ	HTTP	ΥΣ	?	NULL	?

Εικόνα 6.1.2b Δείγμα από τον πίνακα που παράγεται από τον operator "Explain Predictions"

ExampleSet (/Local Repository/NB_WrongPred)											
Open in		Turbo Prep		Auto Model		Filter (34 / 34 examples): all					
Row No.	token's tag	prediction(L...	confidence(U	confidence(...	confidence(P	confidence(...	confidence(...	confidence(...	confidence(...	confidence(...	confidence(...
1	AP	AN	0	0.982	0	0	0	0	0	0	0
2	AP	AN	0	0.994	0.000	0	0	0	0	0	0
3	KNO	ΕΘ	0	0.000	0	0.000	0.000	0.007	0	0.993	0
4	AN	AP	0.000	0.035	0	0	0.000	0	0	0	0
5	AN	M	0.000	0.000	0.000	0.000	0.000	0	0	0.000	0.028
6	EP	AN	0	0.998	0	0	0.002	0	0	0	0.000
7	U	H	0.000	0	0	0	0	0	0	0	0
8	U	H	0.000	0	0	0	0	0	0	0	0
9	U	H	0.000	0	0	0	0	0	0	0	0
10	U	H	0.000	0	0	0	0	0	0	0	0
11	U	H	0.000	0	0	0	0	0	0	0	0
12	APΘM	ΕΘ	0	0	0	0	0.000	0.000	0	0.595	0
13	U	H	0.000	0	0	0	0	0	0	0	0
14	ΣΥΝ	M	0.000	0.000	0	0	0	0	0	0	0.001

ExampleSet (34 examples, 26 special attributes, 14 regular attributes)

Εικόνα 6.1.3a Δείγμα από τα παραδείγματα που επιστρέφει ο operator "Filter Examples"

token	-3 word	-3 tag	-2 word	-2 tag	-1 word	-1 tag	+1 word	+1 tag	+2 word	+2 tag	+3 word	+3 tag	3-char suffix
Στο	?	NULL	@bemyspeci...	U	@JamesBluz	U	βγάει	P	κι	ΣΥΝ	εσένα	AN	Στο
τα	ΣΗΤ	ΣΣ	Μα	M	γιατί	ΣΥΝ	λένε	P	αυτά	AN	ΣΗΤ	ΣΣ	τα
νέα	ΣΗΤ	ΣΣ	Ειδήσεις	KNO	ΣΗΤ	ΣΣ	HTTP	ΥΣ	?	NULL	?	NULL	νέα
Το	?	NULL	?	NULL	?	NULL	είπαν	P	και	ΣΥΝ	το	AN	Το
Να	?	NULL	?	NULL	?	NULL	γιατί	ΣΥΝ	αυτό	AN	το	AP	Να
που	#οδικομουπ...	H	δε	M	ξέρω	P	είναι	P	και	ΣΥΝ	ούτε	ΣΥΝ	που
@aegaeonews	Κω	KPO	HTTP	ΥΣ	via	Ξ	?	NULL	?	NULL	?	NULL	ews
@aegaeonews	γατονική	EΘ	Αλκαρνασό	KPO	HTTP	ΥΣ	?	NULL	?	NULL	?	NULL	ews
@aegaeonews	Ελάτε	P	συμπέρασμα	KNO	HTTP	ΥΣ	?	NULL	?	NULL	?	NULL	ews
@aegaeonews	ΣΗΤ	ΣΣ	HTTP	ΥΣ	via	Ξ	?	NULL	?	NULL	?	NULL	ews
@aegaeonews	Καταγγένια	KNO	HTTP	ΥΣ	via	Ξ	?	NULL	?	NULL	?	NULL	ews
τελευταία	va	ΣΥΝ	είναι	P	η	AP	μέρα	KNO	σου	AN	στη	AP	αία
@aegaeonews	ΣΗΤ	ΣΣ	HTTP	ΥΣ	via	Ξ	?	NULL	?	NULL	?	NULL	ews
Να	?	NULL	?	NULL	?	NULL	κάνω	P	σέση	KNO	ΣΗΤ	ΣΣ	Να

Εικόνα 6.1.3b Δείγμα από τα παραδείγματα που επιστρέφει ο operator "Filter Examples"

Σχετικά με τα παραδείγματα, αυτά που συνήθως κατηγοριοποιούνται λάθος και συγχέονται εμφανώς με μόνο ένα άλλο συγκεκριμένο μέρος του λόγου παρουσιάζονται στην επόμενη παράγραφο.

Τα άρθρα (label "AP") κατηγοριοποιούνται λανθασμένα ως αντωνυμίες (label "AN"). Αυτό φαίνεται πως συμβαίνει για άρθρα τα οποία μοιάζουν μορφολογικά (ίδιο attribute "3-char suffix") με αδύνατους τύπους αντωνυμιών. Επιπλέον, τα ονόματα χρηστών (label "U") κατηγοριοποιούνται λανθασμένα ως hashtags (label "H"). Αυτό φαίνεται πως συμβαίνει όταν το όνομα χρήστη είναι η τελευταία λέξη του tweet και δεν έχει επόμενες γειτονικές λέξεις, δηλαδή τα attributes "+1 word", "+2 word" και "+3 word" δεν έχουν τιμή, και τα "+1 tag", "+2 tag" και "+3 tag" έχουν τιμή NULL.

Τα παραδείγματα τα οποία είναι πιο δύσκολο να αναγνωριστούν, γιατί κατηγοριοποιούνται λάθος, αλλά συγχέονται με πάνω από ένα άλλα μέρη του λόγου περιγράφονται στις επόμενες παραγράφους.

Οι σύνδεσμοι (label "ΣΥΝ") κατηγοριοποιούνται λανθασμένα είτε ως επιρρήματα (label "EP"), είτε ως μόρια (label "M"). Αυτό φαίνεται πως συμβαίνει όταν ο σύνδεσμος είναι η πρώτη λέξη του tweet και δεν έχει προηγούμενες γειτονικές λέξεις, δηλαδή τα attributes "-1 word", "-2 word" και "-3 word" δεν έχουν τιμή, και τα "-1 tag", "-2 tag" και "-3 tag" έχουν τιμή NULL.

Τα επιρρήματα (label "EP") κατηγοριοποιούνται λανθασμένα είτε ως μόρια (label "M"), ή ως αντωνυμίες (label "AN"), ή ως προθέσεις (label "ΠΡ"), ή ως ρήματα (label

"P"). Αυτό φαίνεται πως συμβαίνει για επιρρήματα τα οποία έχουν την ίδια κατάληξη (ίδιο attribute "3-char suffix") με ορισμένα μόρια, αντωνυμίες, προθέσεις και ρήματα.

Οι αντωνυμίες (label "AN") κατηγοριοποιούνται λανθασμένα είτε ως μόρια (label "M"), ή ως άρθρα (label "AP"), ή ως προθέσεις (label "IP"), ή ως κοινά ουσιαστικά (label "KNO"). Αυτό φαίνεται πως συμβαίνει για αντωνυμίες οι οποίες έχουν την ίδια κατάληξη (ίδιο attribute "3-char suffix") με ορισμένα μόρια και άρθρα, ή η αντωνυμία είναι η πρώτη λέξη του tweet και δεν έχει προηγούμενες γειτονικές λέξεις, δηλαδή τα attributes "-1 word", "-2 word" και "-3 word" δεν έχουν τιμή, και τα "-1 tag", "-2 tag" και "-3 tag" έχουν τιμή NULL, οπότε συγχέεται με υπερσύνδεσμο ή κοινό ουσιαστικό.

6.2 Προτάσεις για μελλοντική έρευνα

Ένα ζήτημα που χρειάζεται αντιμετώπιση από τους μελλοντικούς ερευνητές που θα αξιοποιήσουν το εργαλείο που αναπτύχθηκε σε αυτή την εργασία είναι αυτό της υπερπροσαρμογής (overfitting). Το ζήτημα αυτό τίθεται όταν το αποτέλεσμα μιας ανάλυσης προκύπτει και εξαρτάται σε μεγάλο βαθμό από (υπερπροσαρμόζεται σε) ένα συγκεκριμένο σύνολο δεδομένων, έχοντας ως πιθανή συνέπεια την αδυναμία προσαρμογής σε περισσότερα ή λιγότερο ειδικά δεδομένα. Επομένως, εκκρεμεί η εξέταση των αλγορίθμων μηχανικής μάθησης σε διαφορετικά datasets και η εκπαίδευσή τους με περισσότερα δεδομένα, ώστε να αντιμετωπιστεί το εν λόγω ζήτημα.

Το εργαλείο αναγνώρισης μερών του λόγου για την ελληνική γλώσσα του κοινωνικού ιστού που αναπτύχθηκε στα πλαίσια της παρούσας έρευνας θα είχε ενδιαφέρον να χρησιμοποιηθεί με μικρές τροποποιήσεις για POS tagging για άλλα datasets που να αποτελούνται κείμενο τύπου microblogging στα ελληνικά, και ενδεχομένως να προέρχονται από άλλα κοινωνικά δίκτυα πέρα από το Twitter. Η αξιολόγηση της απόδοσής του σε άλλα datasets και η σύγκρισή της με το παρόν θα μπορούσαν να έχουν μεγάλο ερευνητικό ενδιαφέρον.

Σε μελλοντικές έρευνες, το εργαλείο αυτό θα μπορούσε να αποτελέσει τη βάση για την ανάπτυξη περισσότερων και νέων εργαλείων συντακτικής και σημασιολογικής ανάλυσης για την ελληνική γλώσσα του κοινωνικού ιστού. Για παράδειγμα, ορισμένες αντίστοιχες έρευνες για τα αγγλικά χρησιμοποιούν αντίστοιχα εργαλεία για sentiment analysis για εμπορικές, πολιτικές, κοινωνικές και άλλες εφαρμογές, με τη

μόνη διαφορά ότι διατηρούν ξεχωριστά tags για κάθε διαφορετικό σημείο στίξης και για κάθε emoji/ emoticon.

Τέλος, το εργαλείο που αναπτύχθηκε στα πλαίσια της παρούσας εργασίας είναι δυνατόν να χρησιμοποιηθεί για πιο εξειδικευμένη γλωσσολογική ανάλυση των tweets. Τα αντίστοιχα εργαλεία αναγνώρισης μερών του λόγου για την αγγλική γλώσσα του κοινωνικού ιστού έχουν εφαρμογή σε έρευνες που αναλύουν τα γλωσσολογικά χαρακτηριστικά των κειμένων που δημιουργούνται από τους χρήστες του Twitter, σε συνάρτηση με τη γεωγραφική περιοχή (Huang et al., 2016), ή που συνδέουν τη σύνταξη και τη θέση των μερών του λόγου στα tweets, ώστε να τα κατατάξουν σε συγκεκριμένες κατηγορίες events ή topics (Popescu et al., 2011).

Πηγές και βιβλιογραφία

- Bach. N. X., Linh. N. D., & Phuong. T. M. (2018). An empirical study on POS tagging for Vietnamese social media text. *Computer Speech & Language*. 50. 1-15. <https://doi.org/10.1016/j.csl.2017.12.004>
- Foster. J., Çetinoglu. Ö., Wagner. J., Le Roux. J., Hogan. S., Nivre. J., ... & Van Genabith. J. (2011). # hardtoparse: POS Tagging and Parsing the Twittiverse. In *AAAI 2011 workshop on analyzing microtext* (pp. 20-25). Ανάκτηση από <https://www.aaai.org/ocs/index.php/WS/AAAIW11/paper/viewPaper/3912>
- Gimpel. K., Schneider. N., O'Connor. B., Das. D., Mills. D., Eisenstein. J., ... & Smith. N. A. (2010). *Part-of-speech tagging for twitter: Annotation, features, and experiments*. Carnegie-Mellon Univ Pittsburgh Pa School of Computer Science. Ανάκτηση από <https://apps.dtic.mil/docs/citations/ADA547371>
- Huang. Y., Guo. D., Kasakoff. A., & Grieve. J. (2016). Understanding US regional linguistic variation with Twitter data analysis. *Computers, Environment and Urban Systems*. 59. 244-255. <https://doi.org/10.1016/j.compenvurbsys.2015.12.003>
- Juba, B., & Le, H. S. (2019). Precision-recall versus accuracy and the role of large data sets. *Proc. 33rd AAAI*. Ανάκτηση από https://cpb-us-w2.wpmucdn.com/sites.wustl.edu/dist/c/1120/files/2018/11/AAAI_2019_Imbalance_1_earning_camera_ready-1t6nmry.pdf
- Keersmaekers. A. (2019). Creating a richly annotated corpus of papyrological Greek: The possibilities of natural language processing approaches to a highly inflected historical language. *Digital Scholarship in the Humanities*. 1-16. <https://doi.org/10.1093/llc/fqz004>
- Liu. F., Weng. F., & Jiang. X. (2012. July). A broad-coverage normalization system for social media language. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1* (pp. 1035-1044). Association for Computational Linguistics. Ανάκτηση από <https://dl.acm.org/citation.cfm?id=2390662>
- Nand, P., Perera, R., & Lal, R. (2014, December). A HMM POS tagger for micro-blogging type texts. In *Pacific Rim International Conference on Artificial*

Intelligence (pp. 157-169). Springer, Cham. https://doi.org/10.1007/978-3-319-13560-1_13

Neunerdt, M., Trevisan, B., Reyer, M., & Mathar, R. (2013). Part-of-speech tagging for social media texts. In *Language Processing and Knowledge in the Web* (pp. 139-150). Springer, Berlin, Heidelberg. DOI: 10.1007/978-3-642-40722-2.

Nooralahzadeh, F., Brun, C., & Roux, C. (2014, August). Part of speech tagging for french social media data. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers* (pp. 1764-1772). Ανάκτηση από <https://www.aclweb.org/anthology/C14-1166>

Papageorgiou, H., Prokopidis, P., Giouli, V., & Piperidis, S. (2000, May). A Unified POS Tagging Architecture and its Application to Greek. In *LREC*. Ανάκτηση από <http://www.lrec-conf.org/proceedings/lrec2000/pdf/181.pdf>

Petasis, G., Karkaletsis, V., Farmakiotou, D., Samaritakis, G., Androutsopoulos, I., & Spyropoulos, C. (2001, November). A Greek Morphological Lexicon and its Exploitation by a Greek Controlled Language Checker. In *Proceedings of the 8th Panhellenic Conference on Informatics* (pp. 8-10). Ανάκτηση από http://www2.aueb.gr/users/ion/docs/epy_morph.pdf

Popescu. A. M.. & Pennacchiotti. M. (2010. October). Detecting controversial events from twitter. In *Proceedings of the 19th ACM international conference on Information and knowledge management* (pp. 1873-1876). ACM. Ανάκτηση από http://www.marcopennacchiotti.com/pro/publications/CIKM_2010.pdf

Popescu. A. M.. Pennacchiotti. M.. & Paranjpe. D. (2011. March). Extracting events and event descriptions from twitter. In *Proceedings of the 20th international conference companion on World wide web* (pp. 105-106). ACM. Ανάκτηση από http://www.marcopennacchiotti.com/pro/publications/WWW_2011_3.pdf

Rehbein, I. (2013). Fine-grained pos tagging of german tweets. In *Language Processing and Knowledge in the Web* (pp. 162-175). Springer, Berlin, Heidelberg. DOI: 10.1007/978-3-642-40722-2.

Sgarbas, K. N., Fakotakis, N. D., & Kokkinakis, G. K. (2001). A straightforward approach to morphological analysis and synthesis. *arXiv preprint cs/0112010*. Ανάκτηση από <https://arxiv.org/abs/cs/0112010>

Stehman, S. V. (1997). Selecting and interpreting measures of thematic classification accuracy. *Remote sensing of Environment*, 62(1), 77-89.

[https://doi.org/10.1016/S0034-4257\(97\)00083-7](https://doi.org/10.1016/S0034-4257(97)00083-7)

Tommasel. A.. & Godoy. D. (2018). A Social-aware online short-text feature selection technique for social media. *Information Fusion*. 40. 1-17.

<https://doi.org/10.1016/j.inffus.2017.05.003>

Zalmout. N.. Erdmann. A.. & Habash. N. (2018). Noise-robust morphological disambiguation for dialectal Arabic. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Volume 1 (Long Papers)* (Vol. 1. pp. 953-964). Ανάκτηση από <https://www.aclweb.org/anthology/N18-1087/>

Αναστασιάδη-Συμεωνίδη. Α.. Κυριακοπούλου. Π. (2015). *Αυτόματη επεξεργασία λέξεων*. [Κεφάλαιο Συγγραμματος]. Στο Τάντος, Α.. Μαρκαντωνάτου. Σ.. Αναστασιάδη-Συμεωνίδη. Α.. Κυριακοπούλου. Π. (2015). *Υπολογιστική γλωσσολογία*. [ηλεκτρ. βιβλ.] Αθήνα: Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών. κεφ 2. Διαθέσιμο στο: <http://hdl.handle.net/11419/2207>

Δικτυογραφία

Apply Model operator. Ανάκτηση από: https://docs.rapidminer.com/latest/studio/operators/scoring/apply_model.html.

Τελευταία ανάκτηση: 09/09/2019

"Greek part of speech tagger". Ευαγγελία Κολέλη. Ανάκτηση από <http://php-nlp-tools.com/posts/greek-pos-tag.html>. Τελευταία ανάκτηση: 11/09/2019

ID3 operator. Ανάκτηση από <https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/trees/id3.html>

. Τελευταία ανάκτηση: 10/09/2019

Naive Bayes operator. Ανάκτηση από https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/bayesian/naive_bayes.html. Τελευταία ανάκτηση: 09/09/2019

Performance operator. Ανάκτηση από <https://docs.rapidminer.com/latest/studio/operators/validation/performance/performance.html>. Τελευταία ανάκτηση: 09/09/2019

RapidMiner Studio Educational. Ανάκτηση από <https://rapidminer.com/products/studio/>. Τελευταία ανάκτηση: 05/09/2019

Split Data operator. Ανάκτηση από https://docs.rapidminer.com/latest/studio/operators/blending/examples/sampling/split_data.html. Τελευταία ανάκτηση: 09/09/2019

Standard search API - Twitter Developers. Ανάκτηση από <https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets.html>. Τελευταία ανάκτηση: 20/04/2019.

Twitter. It's what's happening. Ανάκτηση από <https://twitter.com/>. Τελευταία ανάκτηση: 20/04/2019.

Text to Nominal operator. Ανάκτηση από https://docs.rapidminer.com/latest/studio/operators/blending/attributes/types/text_to_nominal.html. Τελευταία ανάκτηση: 10/09/2019

Ελληνικός επισημειωτής μερών του λόγου. Ομάδα Επεξεργασίας Φυσικής Γλώσσας, Τμήμα Πληροφορικής, Οικονομικό Πανεπιστήμιο Αθηνών. Ανάκτηση από <http://nlp.cs.aueb.gr/software.html>. Τελευταία ανάκτηση: 11/09/2019

Γραμματικές και λεξικά

Βικιλεξικό. Ανάκτηση από <https://bit.ly/2ylRINe>

Τζεβελέκου. Μ.. Κάντζου. Β.. Σταμούλη. Σ. (2007). *Βασική γραμματική της ελληνικής*. Αθήνα: Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών. Ανάκτηση από http://www.ilsp.gr/files/Basic_Greek_Grammar.pdf. Τελευταία ανάκτηση: 08/08/2019.

Τριανταφυλλίδης. Μ. (1990). *Νεοελληνική Γραμματική. Αναπροσαρμογή της Μικρής Νεοελληνικής Γραμματικής*. Αθήνα: Οργανισμός Εκδόσεως Διδακτικών βιβλίων.

Παράρτημα

Κώδικας συλλογής δεδομένων από το Standard search API του Twitter και αποθήκευσής τους σε αρχείο CSV

Python script:

```
#!/usr/bin/python

# This Python file uses the following encoding: utf-8

import os, sys

import tweepy

import csv #Import csv

auth = tweepy.auth.OAuthHandler('YOUR_API_KEY'.
'YOUR_API_SECRET_KEY')

auth.set_access_token('YOUR_ACCESS_TOKEN'.
'YOUR_SECRET_ACCESS_TOKEN')

api = tweepy.API(auth)

# Open/create a file to append data to
csvFile = open('result.csv', 'a')

#Use csv writer
csvWriter = csv.writer(csvFile)

for tweet in tweepy.Cursor(api.search.
q = "-filter:retweets".      #don't include retweets (to avoid duplicates)
lang = "el". #tweets written in greek
since = "2019-04-01".      #tweets posted during April. 2019
```

```
until = "2019-04-30").items():  
  
# Write a row to the CSV file. I use encode UTF-8  
csvWriter.writerow([tweet.text.encode('utf-8')])  
  
print tweet.text  
  
csvFile.close()
```

Συντμήσεις

API	Application Programming Interface
app	Application
CSV	Comma-Separated Values
NLP	Natural Language Processing
SMS	Short Message Service
URL	Uniform Resource Locator

Γλωσσάρι ξενικών όρων

Access token	Διακριτικό πρόσβασης
Accuracy	Ακρίβεια
API	Διεπαφή/ Διασύνδεση Προγραμματισμού Εφαρμογών
API key	Κλειδί πρόσβασης σε API
App	Εφαρμογή
Blog	Ιστολόγιο
Classification	Ταξινόμηση
Conditional probability	Δεσμευμένη πιθανότητα
Confusion matrix	Πίνακας σύγχυσης
Corrupted	Κατεστραμμένο (αρχείο ή δεδομένα)
Criterion	Κριτήριο
Cross validation	Διασταυρωτική αξιολόγηση
CSV	Τιμές διαχωρισμένες με κόμμα
Dataset	Σετ δεδομένων
Emoji/ Emoticon	Ιδεόγραμμα/ Φατσούλα
Entity	Οντότητα
Event	Γεγονός
Facebook	Κοινωνικό δίκτυο
Focus word	Λέξη εστίασης
Followee	Ακολουθούμενος
Follower	Ακόλουθος
Hashtag	Δίεση (#)
Hyperlink	Υπερσύνδεσμος
Information gain	Όφελος πληροφορίας
Label	Ετικέτα

Laplace correction	Διόρθωση Laplace
Link	Σύνδεσμος
Machine learning	Μηχανική μάθηση
Mention	Αναφορά
Microblog	Μικροϊστολόγιο
Natural Language Processing	Επεξεργασία φυσικής γλώσσας
Nominal	Ονομαστικός
Operator	Τελεστής
Overfitting	Υπερπροσαρμογή
Precision	Εξειδίκευση
Predicted class	Κλάση προς πρόβλεψη
Python script	"Σενάριο" γραμμένο στη γλώσσα προγραμματισμού Python
Recall	Ευαισθησία/ Ανάκληση
Recommender systems	Συστήματα συστάσεων
Retweet	Προώθηση/ Επαναδημοσίευση ενός tweet
Sentiment analysis	Ανάλυση αισθήματος
SMS	Υπηρεσία Σύντομου Μηνύματος
Social text	Κείμενο κοινωνικού ιστού
Spam	Ανεπιθύμητο μήνυμα
Supervised learning	Επιβλεπόμενη μάθηση
Tag	Ετικέτα
Tag set	Σετ ετικετών
Test set	Σετ ελέγχου
Token	Λεκτική μονάδα
Topic	Θέμα
Training set	Σετ εκπαίδευσης
Tweet	Ανάρτηση/ Σύντομο μήνυμα στο Twitter

Twitter	Κοινωνικό δίκτυο αποστολής και ανάγνωσης σύντομων μηνυμάτων (tweets)
Unsupervised learning	Μη επιβλεπόμενη μάθηση
URL	Ενιαίος Εντοπιστής Πόρων
YouTube	Κοινωνικό δίκτυο