ΠΑΝΕΠΙΣΤΗΜΙΟ ΜΑΚΕΔΟΝΙΑΣ
ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΤΜΗΜΑΤΟΣ ΕΦΑΡΜΟΣΜΕΝΗΣ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΔΙΑΓΛΩΣΣΙΚΗ ΑΝΑΛΥΣΗ ΣΥΝΑΙΣΘΗΜΑΤΟΣ ΣΕ ΕΛΛΗΝΙΚΑ TWEETS

Διπλωματική Εργασία

της

Ευαγγελίας Τσουκανάρα

Θεσσαλονίκη, Ιούνιος 2019

ΔΙΑΓΛΩΣΣΙΚΗ ΑΝΑΛΥΣΗ ΣΥΝΑΙΣΘΗΜΑΤΟΣ ΣΕ ΕΛΛΗΝΙΚΑ TWEETS

Ευαγγελία Τσουκανάρα

Πτυχίο Εφαρμοσμένης Πληροφορικής, Πανεπιστήμιο Μακεδονίας, 2012

Διπλωματική Εργασία

υποβαλλόμενη για τη μερική εκπλήρωση των απαιτήσεων του

ΜΕΤΑΠΤΥΧΙΑΚΟΥ ΤΙΤΛΟΥ ΣΠΟΥΔΩΝ ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΠΛΗΡΟΦΟΡΙΚΗ

Επιβλέπουσα Καθηγήτρια
Γεωργία Κολωνιάρη

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την --/06/2019

Γεωργία Κολωνιάρη	Γεώργιος Ευαγγελίδης	Ευκλείδης Κεραμόπουλος
.....	.....	.....

Ευαγγελία Τσουκανάρα

.....

Περίληψη

Η ανάλυση συναισθήματος ή εξόρυξη γνώμης αποτελεί ένα κρίσιμο πεδίο και έχει προσελκύσει το έντονο ενδιαφέρον της επιστημονικής κοινότητας. Η ανάπτυξη του Web 2.0, οδήγησε σε μια νέα παγκόσμια πραγματικότητα με εκρηκτική αύξηση της αλληλεπίδρασης των ανθρώπων είτε με την έκφραση απόψεων μέσω των κοινωνικών δικτύων για οποιοδήποτε ζήτημα ή πρόσωπο απασχολεί την κοινή γνώμη είτε με την διατύπωση κριτικών σε προϊόντα και υπηρεσίες. Επιπρόσθετα, στα πλαίσια της παγκοσμιοποίησης γίνεται επιτακτική η ανάγκη της ανάλυσης του συνόλου των απόψεων, ανεξάρτητα από τη γλώσσα που χρησιμοποιείται. Είναι γεγονός ότι η έρευνα στο πεδίο της ανάλυσης συναισθήματος αναπτύχθηκε κατά βάση στην αγγλική γλώσσα. Η διαγλωσσική ανάλυση συναισθήματος εξετάζει τρόπους αξιοποίησης εργαλείων και μέσων που έχουν αναπτυχθεί σε άλλες γλώσσες. Η παρούσα εργασία εστιάζει στη διαγλωσσική ανάλυση συναισθήματος σε ελληνικά δεδομένα του Twitter, προσεγγίζει το ζήτημα ως πρόβλημα δυαδικής ταξινόμησης και διερευνά το βέλτιστο τρόπο που μπορεί να πραγματοποιηθεί συγκρίνοντας τρεις μεθόδους. Οι δύο πρώτες μέθοδοι χρησιμοποιούν τη μηχανική μετάφραση και στη συνέχεια, την ανάλυση των δεδομένων με εργαλεία για την αγγλική γλώσσα. Η τρίτη μέθοδος αναλύει απευθείας τα ελληνικά δεδομένα, χρησιμοποιώντας εργαλεία που αναπτύχθηκαν για την ελληνική γλώσσα. Η διαδικασία για τις τρεις μεθόδους είναι κοινή και ξεκινά με την προ-επεξεργασία των δεδομένων, σε επόμενο στάδιο γίνεται η εξαγωγή χαρακτηριστικών και τέλος η εκπαίδευση και η αξιολόγηση του μοντέλου. Τα αποτελέσματα των πειραμάτων δείχνουν ότι τα μοντέλα μας μπορούν να επιτύχουν ακρίβεια μέχρι και 84% για το σύνολο των μεθόδων, ενώ παράλληλα η τεχνική του κλαδέματος χαρακτηριστικών λειτουργεί θετικά για τις δύο από τις τρεις μεθόδους.

Λέξεις Κλειδιά: Cross-lingual, sentiment analysis, opinion mining, machine learning, machine translation, Twitter

Abstract

Sentiment analysis or opinion mining is an important subject that has attracted the attention of the research community. The evolution of the Web 2.0 has led to a remarkable increase of people's interaction by expressing their opinions through the social networks or by writing reviews regarding products or services. Today is crucial to analyze these opinions, regardless of the language are written. Unfortunately, sentiment analysis research work has been developed basically in English. Cross-lingual sentiment analysis investigates techniques of leveraging resources from resource-rich languages. The current work focuses on cross-lingual sentiment analysis for Greek tweets. To this purpose, we are employing three methods and we consider the task as a binary classification problem. Two of the methods are using machine translation and then make use of resources developed for English. The last method directly analyzes Greek tweets using tools developed for Greek language. At first, we pre-process the data and then we apply feature extraction techniques. Finally, we train and evaluate our model. The results of the experiments show that our models can achieve up to 84% accuracy for all three methods and that feature pruning works positively for two of the methods.

Keywords: Cross-lingual, sentiment analysis, opinion mining, machine learning, machine translation, Twitter

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά την επιβλέπουσα καθηγήτριά μου για τη βοήθεια και την υποστήριξή της καθ' όλη τη διάρκεια της εκπόνησης της διπλωματικής μου εργασίας. Επίσης, θα ήθελα να ευχαριστήσω την οικογένειά μου και τον Γιώργο για την αμέριστη συμπαράστασή τους.

Περιεχόμενα

Κατάλογος Εικόνων

Κατάλογος Πινάκων

1 Εισαγωγή

1.1 Πρόβλημα – Σημαντικότητα του θέματος

Καθώς το διαδίκτυο έχει φτάσει σχεδόν σε κάθε γωνιά του πλανήτη, όλο και μεγαλύτερος αριθμός ανθρώπων χρησιμοποιούν τις υπηρεσίες του. Συγκεκριμένα, ο αριθμός των χρηστών του διαδικτύου παγκοσμίως ανέρχεται για το 2018 σχεδόν στα 4 δις (Statista). Η ανάπτυξη του Web 2.0, που βασίζεται στην αλληλεπίδραση και το διαμοιρασμό των πληροφοριών μεταξύ των χρηστών, οδήγησε στην αλματώδη αύξηση του αριθμού των ανθρώπων που επικοινωνούν, εκφράζονται και γράφουν τη γνώμη τους μέσω διαδικτύου, γεγονός που οφείλεται μεταξύ άλλων στην αμεσότητα, αλλά και στην ευκολία στη χρήση του. Το παραπάνω εκφράζεται ως επί το πλείστον με τη χρήση των κοινωνικών δικτύων και το σχολιασμό σε ποικίλα θέματα, όπως κοινωνικά, πολιτικά ή με τη δημιουργία κριτικών και την αξιολόγηση προϊόντων και υπηρεσιών (βιβλία, ταινίες, τουριστικές μονάδες κ.ά.). Αυτός ο τεράστιος όγκος πληροφοριών, αποτελεί ουσιαστικά αδόμητη γνώση. Το πρόβλημα που προκύπτει σε πρώτη φάση λοιπόν, αφορά στην εύκολη αξιολόγηση αυτού του τεράστιου όγκου δεδομένων τόσο από την πλευρά του χρήστη, όσο και από την πλευρά μιας εταιρίας ή και του κράτους, προκειμένου να αξιοποιηθεί και να αναλυθεί συνολικά και ολοκληρωμένα η γνώμη του κοινού για κάποιο δημόσιο πρόσωπο ή για κάποιο προϊόν ή υπηρεσία που διατίθεται στην αγορά. Στα προηγούμενα έρχεται να προστεθεί το ετερόκλητο χαρακτηριστικό των δεδομένων, καθώς τα πεδία για τα οποία υπάρχουν σχόλια και συζητήσεις ποικίλουν, όπως επίσης και το γλωσσικό εμπόδιο, λόγω της πληθώρας των γλωσσών και τις έλλειψης εργαλείων για κάθε γλώσσα. Όλα τα παραπάνω συνδυαστικά αποτελούν μια πραγματική πρόκληση, την οποία καλείται να φέρει εις πέρας η εξόρυξη γνώμης ή η ανάλυση συναισθήματος. Σε ό,τι αφορά την ελληνική γλώσσα, όπως και πολλές άλλες γλώσσες πέραν των αγγλικών, οι πόροι είναι περιορισμένοι και τίθεται το ερώτημα αν είναι απαραίτητη η χρήση των διαθέσιμων πόρων για ανάλυση συναισθήματος ελληνικών δεδομένων ή αν η χρήση τεχνικών μηχανικής μετάφρασης και η εκμετάλλευση εργαλείων για τα αγγλικά είναι επαρκής.

1.2 Σκοπός – Στόχοι

Η εκπόνηση αυτής της διπλωματικής εργασίας έχει ως στόχο να ερευνήσει σε βάθος το πεδίο της διαγλωσσικής εξόρυξη γνώμης και να συγκρίνει τρεις μεθόδους επιβλεπόμενης μάθησης (supervised learning) για τη δυαδική ανάλυση συναισθήματος σε ελληνικά δεδομένα τα οποία προέρχονται από το Twitter. Οι δύο από τις προσεγγίσεις που εξετάζονται, έχουν να κάνουν με μηχανική μετάφραση από την ελληνική στην αγγλική γλώσσα και στη συνέχεια, ανάλυση συναισθήματος με εργαλεία για τα αγγλικά και η τρίτη ερευνά την απευθείας ανάλυση ελληνικών δεδομένων με εργαλεία που έχουν αναπτυχθεί αποκλειστικά για την ελληνική γλώσσα. Σε πρώτο στάδιο, γίνεται η συγκέντρωση των δεδομένων στην ελληνική γλώσσα, τα οποία έχουν εκ των προτέρων την κατάλληλη ετικέτα (label) και έπειτα για τις δύο πρώτες μεθόδους, γίνεται η μηχανική μετάφρασή τους στην αγγλική γλώσσα. Στη συνέχεια και για το σύνολο των μεθόδων, ακολουθεί η επεξεργασία των δεδομένων και η εξαγωγή σημαντικών χαρακτηριστικών των tweets, τα οποία συμβάλλουν στην αυτόματη κατηγοριοποίησή τους σε θετικού ή αρνητικού περιεχομένου. Έπειτα, τα δεδομένα διαχωρίζονται σε πέντε ίσα μέρη (5 folds) προκειμένου να εφαρμοστεί cross-validation και ακολουθεί η εκπαίδευση (training) του μοντέλου με τους αλγόριθμους Naive Bayes και Random Forest και τέλος, η δοκιμή (testing) και η αξιολόγηση των αποτελεσμάτων.

1.3 Διάρθρωση της μελέτης

Η εργασία στα επόμενα κεφάλαια δομείται ως εξής: στο κεφάλαιο 2 γίνεται βιβλιογραφική επισκόπηση, όπου περιγράφονται έννοιες και στοιχεία σχετικά με το αντικείμενο της εργασίας. Συγκεκριμένα, στην Ενότητα 2.1 αναλύεται η ανάλυση συναισθήματος και δίνονται παραδείγματα που εφαρμόζεται, ενώ αναφέρονται και οι δύο βασικές προσεγγίσεις. Στη συνέχεια, στην Ενότητα 2.2 περιγράφεται η διαδικασία της μηχανικής μάθησης, παρουσιάζονται οι κατηγορίες αλγορίθμων μηχανικής μάθησης, καθώς και οι εφαρμογές της. Στην Ενότητα 2.3 γίνεται εκτενής αναφορά στην διαγλωσσική ανάλυση συναισθήματος και τέλος, στην 2.4 παρουσιάζονται τα κοινωνικά δίκτυα και γίνεται ειδική αναφορά στο Twitter. Στο Κεφάλαιο 3 γίνεται θεωρητική περιγραφή των μεθόδων και των σταδίων της μεθοδολογίας. Στην Ενότητα 3.1 αναλύεται η μέθοδος της μηχανικής μετάφρασης, στην Ενότητα 3.2 παρουσιάζονται τα βήματα και οι τεχνικές για την προ-επεξεργασία των δεδομένων, στην 3.3. αναλύονται οι

τρόποι εξαγωγής χαρακτηριστικών, στην Ενότητα 3.4 περιγράφονται οι αλγόριθμοι ταξινόμησης που χρησιμοποιούνται στη μεθοδολογία και στην 3.5 παρουσιάζονται οι τρόποι αξιολόγησης του μοντέλου. Στο κεφάλαιο 4 παρουσιάζεται η διαδικασία και τα βήματα που ακολουθούνται στη μεθοδολογία. Συγκεκριμένα, η Ενότητα 4.1 επικεντρώνεται στην περιγραφή των δεδομένων και τον τρόπο συλλογής τους και οι Ενότητες 4.2, 4.3 και 4.4 παρουσιάζουν τη διαδικασία και τις μεθόδους που υιοθετούνται στα στάδια του pre-processing, της εξαγωγής χαρακτηριστικών και των αλγόριθμων ταξινόμησης. Στο Κεφάλαιο 5, παρουσιάζονται τα αποτελέσματα των μετρήσεων για τα διάφορα πειράματα που εκτελέστηκαν. Τέλος, το Κεφάλαιο 6 περιλαμβάνει τα συμπεράσματα για τη συγκεκριμένη εργασία καθώς και κάποιες πιθανές επεκτάσεις της.

2 Βιβλιογραφική Επισκόπηση

Στο παρόν κεφάλαιο γίνεται ανάλυση των εννοιών, πληροφοριών και βιβλιογραφικών στοιχείων που θα συμβάλουν στην καλύτερη κατανόηση της εργασίας. Αρχικά, γίνεται μια περιγραφή της ανάλυσης συναισθήματος, των εφαρμογών που έχει σήμερα, καθώς και των βασικών προσεγγίσεων από πλευράς έρευνας. Στη συνέχεια, δίνεται ο ορισμός της μηχανικής μάθησης, γίνεται αναφορά των κατηγοριών των αλγορίθμων και περιγράφονται πρακτικές εφαρμογές των αλγορίθμων αυτών. Επίσης, επεξηγείται η διαγλωσσική ανάλυση συναισθήματος και η σχετική με το πεδίο έρευνα και τέλος, δίνεται ο ορισμός των μέσων κοινωνικής δικτύωσης και γίνεται ξεχωριστή μνεία στο Twitter.

2.1 Ανάλυση Συναισθήματος

Η ανάλυση συναισθημάτων ή διαφορετικά εξόρυξη γνώμης, ασχολείται με την ανάλυση της γνώμης ή του συναισθήματος σχετικά με κάποιο προϊόν, υπηρεσία, κάποιο δημόσιο πρόσωπο ή ενδεχομένως, κάποιο γεγονός. Πρόκειται για μια υποκατηγορία της ανάλυσης κειμένου και είναι η διαδικασία κατά την οποία διερευνάται η ύπαρξη συναισθήματος σε κάποιο σύνολο δεδομένων κειμένου και προκειμένου αυτό να επιτευχθεί, χρησιμοποιούνται τεχνικές επεξεργασίας της φυσικής γλώσσας και μέθοδοι μηχανικής μάθησης για την ταξινόμηση του κειμένου σε κλάσεις συναισθήματος, συνήθως σε θετική και αρνητική κλάση. Αν και η επεξεργασία φυσικής γλώσσας έχει εμφανιστεί ως όρος και ερευνάται εδώ και κάποιες δεκαετίες, περίπου από τα τέλη της δεκαετίας του 1940, το πεδίο της ανάλυσης συναισθήματος απασχολεί την έρευνα την τελευταία εικοσαετία και παρουσιάζει εξαιρετικό ενδιαφέρον λόγω της πληθώρας των εφαρμογών του. Σημαντικό ρόλο στην περαιτέρω έρευνα του πεδίου έχει διαδραματίσει και ο τεράστιος πλέον όγκος δεδομένων, λόγω της ύπαρξης των κοινωνικών δικτύων. Κατά συνέπεια, η έρευνα στο πεδίο της ανάλυσης συναισθήματος επηρεάζει με τη σειρά της σε μεγάλο βαθμό τις επιστήμες που συνδέονται με την άποψη του κόσμου, δηλαδή τις κοινωνικές επιστήμες, την πολιτική και οικονομική επιστήμη (Liu, 2012).

2.1.1 Εφαρμογές

Κατά κανόνα, οι απόψεις τρίτων επηρεάζουν τη συμπεριφορά των ανθρώπων σε μεγάλο βαθμό. Σε ό,τι αφορά το κομμάτι των επιχειρήσεων, είναι ζωτικής σημασίας για

μια εταιρία ή έναν οργανισμό να γνωρίζει τη γνώμη του κόσμου για το προϊόν ή την υπηρεσία που προσφέρει. Όπως επίσης, εξίσου σημαντικό είναι για τον ίδιο τον καταναλωτή / υποψήφιο αγοραστή να γνωρίζει την άποψη εκείνων που χρησιμοποιούν το προϊόν που τους ενδιαφέρει να αποκτήσουν. Δεν αφορά όμως αποκλειστικά τις επιχειρήσεις, καθώς ενδιαφέρον για την άποψη του κόσμου, έχουν για παράδειγμα και τα πολιτικά πρόσωπα. Παλαιότερα, η απόκτηση αυτής της πολύτιμης γνώσης γινότανε με τη βοήθεια γκάλοπ, συμπλήρωσης ερωτηματολογίων και παρόμοιων μεθόδων. Σήμερα, με την εκτεταμένη χρήση του διαδικτύου και των μέσων κοινωνικής δικτύωσης και τον τεράστιο όγκο δεδομένων (σχόλια, κριτικές κλπ.), η πληροφορία αυτή είναι εύκολα προσβάσιμη στα ενδιαφερόμενα μέρη. Ωστόσο, προκειμένου όλη αυτή η πληροφορία να μετατραπεί σε γνώση και να μπορέσει να αξιοποιηθεί και να λάβει μέρος στη λήψη αποφάσεων, είναι απαραίτητο να γίνει με αυτοματοποιημένο τρόπο. Για όλους τους παραπάνω λόγους, έχουν αναπτυχθεί διάφορες εφαρμογές σε οποιοδήποτε κάθε πεδίο (domain) για το οποίο υπάρχει πληροφορία στο διαδίκτυο, είτε πρόκειται για feedback για προϊόντα και υπηρεσίες είτε για την άποψη που εκφράζεται για κοινωνικές εκδηλώσεις και τη στάση της κοινής γνώμης στις εκλογές (Liu, 2012). Παρακάτω γίνεται περιγραφή των δύο κύριων προσεγγίσεων στο πεδίο της ανάλυσης συναισθήματος, την lexicon-based και την ανάλυση συναισθήματος με τεχνικές μηχανικής μάθησης.

2.1.2 Lexicon-based προσέγγιση

Η lexicon-based προσέγγιση υπολογίζει την πολικότητα ενός κειμένου βασιζόμενη στη χρήση λεξικών και το σημασιολογικό προσανατολισμό των λέξεων (Maite Taboada, 2011). Τα λεξικά αποδίδουν το βαθμό συναισθήματος ανά λέξη ή φράση και κατ' επέκταση καθορίζουν και την κατάταξη του κειμένου στην αρνητική ή τη θετική κλάση. Μία από τις πρώτες έρευνες στο αντικείμενο προτείνει μία μέθοδο για την αυτόματη ταξινόμηση των επιθέτων όταν σε αυτά παρεμβάλλονται σύνδεσμοι, όπως «και», «ενώ», «αλλά» κλπ. (Vasileios Hatzivassiloglou, 1997). Μία ακόμη έρευνα έχει να κάνει με τη γλωσσολογική προσέγγιση σε κριτικές διαφόρων domains και ερευνά κατά πόσο φράσεις που περιέχουν επίθετα ή επιρρήματα σχετίζονται συχνότερα με τις λέξεις «excellent» ή «poor», το οποίο τελικά θα δώσει στην εκάστοτε φράση θετική ή αρνητική χροιά (Turney, 2002). Μία πρόσφατη έρευνα σχετική με τα παραπάνω είναι η ανάλυση συναισθήματος με γλωσσολογική προσέγγιση σε κριτικές ταινιών, κατά την οποία δοκιμάζονται επτά διαφορετικοί συνδυασμοί στα μέρη του λόγου (POS) που

λαμβάνουν μέρος στην εξαγωγή χαρακτηριστικών (feature extraction) για τέσσερις διαφορετικούς αλγόριθμους, από όπου προκύπτει ότι ο συνδυασμός επίθετο, επίρρημα και ρήμα αποτελεί τον καλύτερο δυνατό (Najma Sultana, 2019).

Η παραπάνω προσέγγιση έχει ωστόσο κάποια μειονεκτήματα, τα οποία γίνονται εμφανή σε περιπτώσεις όπου το συναίσθημα υπονοείται ή εκφράζεται σιωπηρά, σε κείμενο που υπάρχει σαρκασμός ή ειρωνεία και σε κείμενο που περιέχει λέξεις με ερμηνευτική ασάφεια. Μία ακόμα περίπτωση είναι οι λέξεις που εξαρτώνται από το πεδίο στο οποίο χρησιμοποιούνται (domain dependency), δηλαδή λέξεις που έχουν διαφορετικό συναίσθημα και συνεπώς, διαφορετική πολικότητα ανά domain. Για παράδειγμα, η λέξη *απρόβλεπτος* μπορεί να έχει θετικό πρόσημο για μια ταινία, αλλά αρνητικό για τη απόκριση του τιμονιού ενός αυτοκινήτου (Abhijit Mishra, 2016).

2.1.3 Machine learning προσέγγιση

Η προσέγγιση αυτή έχει να κάνει με την εξαγωγή χαρακτηριστικών και την ταξινόμηση του κειμένου συνήθως σε δυαδική κλάση, δηλαδή σε θετική ή αρνητική. Συνήθως, αυτό επιτυγχάνεται με τη χρήση τεχνικών επιβλεπόμενης μάθησης και την ανάθεση ετικετών στο σύνολο δεδομένων προς εκπαίδευση (Pollyanna Gonçalves, 2013). Σπανιότερα, χρησιμοποιούνται τεχνικές μη επιβλεπόμενης μάθησης, οι οποίες ανακαλύπτουν τον τρόπο που είναι δομημένα τα δεδομένα και στη συνέχεια, τα ομαδοποιούν σε συστάδες (clusters). Αναλυτικότερη περιγραφή της μηχανικής μάθησης γίνεται στην επόμενη ενότητα.

Στην παρούσα εργασία, χρησιμοποιείται επιβλεπόμενη μάθηση με τη βοήθεια δύο από τους πλέον δημοφιλείς αλγόριθμους, τον Naive Bayes και τον Random Forest, για τους οποίους γίνεται εκτενής αναφορά στο επόμενο κεφάλαιο.

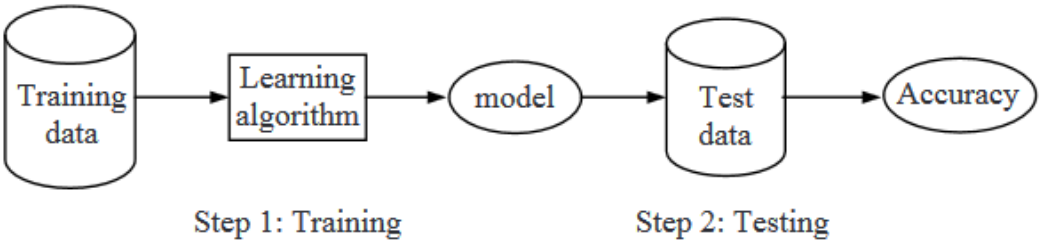
2.2 Μηχανική Μάθηση

2.2.1 Ορισμός

Η μηχανική μάθηση αποτελεί μια υπο-περιοχή του πεδίου της Τεχνητής Νοημοσύνης και με τη χρήση διάφορων αλγόριθμων και στατιστικών μοντέλων, αναζητά πρότυπα (patterns) και τον τρόπο που δομούνται τα δεδομένα. Έτσι καταφέρνει να τα μοντελοποιήσει και να κάνει προβλέψεις με βάση αυτά, χωρίς ταυτόχρονα να

απαιτεί σαφείς οδηγίες και μεγάλο αριθμό γραμμών κώδικα. Η ονομασία της δόθηκε αρκετά χρόνια πριν από τον Arthur Lee Samuel το 1959 (Judith Hurwitz, 2018).

Σε γενικές γραμμές, η διαδικασία της μηχανικής μάθησης περιλαμβάνει τη δημιουργία μαθηματικών μοντέλων πάνω σε ένα σύνολο δεδομένων με χρήση διαφόρων αλγόριθμων. Η διαδικασία αυτή ονομάζεται εκπαίδευση (training) και με αυτό τον τρόπο το σύστημα αποκτά εμπειρία και διαμορφώνει τα μοντέλα δεδομένων. Στη συνέχεια, τροφοδοτείται με δεδομένα δοκιμής (testing data), τα οποία δεν έχουν συμμετάσχει στη διαδικασία εκπαίδευσης και με βάση την επίδοση σε αυτά, αξιολογείται και η ακρίβεια του συστήματος.



Εικόνα 1 Διαδικασία μηχανικής μάθησης

(Liu, Web Data Mining – Exploring Hyperlinks, Contents and Usage Data, 2011)

2.2.2 Κατηγορίες αλγορίθμων

Υπάρχουν τρεις τύποι αλγόριθμων μηχανικής μάθησης: οι αλγόριθμοι επιβλεπόμενης μάθησης (supervised learning), οι αλγόριθμοι μη επιβλεπόμενης μάθησης (unsupervised learning) και εκείνοι της ενισχυτικής μάθησης (reinforcement learning). Στην πρώτη περίπτωση, τα δεδομένα συνοδεύονται από τα επιθυμητά αποτελέσματα (labels). Επομένως, το σύστημα προσπαθεί να βρει το μοτίβο που συνδέει τα δεδομένα εισόδου με τα δεδομένα εξόδου, ώστε να δημιουργήσει το μοντέλο με το οποίο θα γίνονται οι προβλέψεις για τα δεδομένα δοκιμής. Κάποιοι από τους πιο δημοφιλείς αλγόριθμους επιβλεπόμενης μάθησης είναι οι Naive Bayes, Random Forest και Support Vector Machine. Στη δεύτερη περίπτωση, το σύστημα δέχεται τα δεδομένα εισόδου χωρίς να αντιστοιχίζονται με δεδομένα εξόδου και προσπαθεί να διαπιστώσει τον τρόπο που είναι δομημένα, χωρίζοντάς τα σε ομάδες ή συστάδες (clusters). Χαρακτηριστικός αλγόριθμος μη επιβλεπόμενης μάθησης είναι ο k-means (Ozgur, 2004).

2.2.3 Εφαρμογές

Σήμερα, η μηχανική μάθηση έχει ευρεία εφαρμογή σε διάφορους τομείς, όπως για παράδειγμα το News Feed του Facebook[1], όπου με τη βοήθεια της μηχανικής μάθησης επιτυγχάνεται η εξατομίκευση των ενημερώσεων του εκάστοτε χρήστη, σύμφωνα με τις επιλογές που έχει κάνει στο παρελθόν. Σε ό,τι αφορά στις επιχειρήσεις, ο τομέας του Business Intelligence (BI) αποτελεί αναπόσπαστο εργαλείο, καθώς επεξεργάζεται και αναλύει τα δεδομένα, κάνει προβλέψεις και συμβάλει στη διαδικασία λήψης αποφάσεων. Ένας ακόμα τομέας που μοιάζει να έρχεται από το μέλλον, αλλά δεν είναι και τόσο μακρινός, είναι η αυτόνομη οδήγηση, της οποίας δείγματα βλέπουμε ήδη στα σημερινά οχήματα με τα διάφορα συστήματα υποβοήθησης οδηγού. Ακόμη, η τεχνολογία των εικονικών βοηθών (virtual assistants) χρησιμοποιεί επίσης μηχανική μάθηση, προκειμένου να επεξεργαστεί τη φυσική γλώσσα και να προχωρήσει σε κάποια ενέργεια. Ιδιαίτερο ρόλο έχει διαδραματίσει η μηχανική μάθηση στον τραπεζικό – οικονομικό τομέα υπηρεσιών, κάνοντας έξυπνες προτάσεις, προσαρμοσμένες στη συμπεριφορά των πελατών και βοηθώντας τις τράπεζες να λάβουν τις κατάλληλες αποφάσεις. Τέλος, εντυπωσιακή είναι η συμβολή της μηχανικής μάθησης και στον τομέα της υγείας, όπου αναλύει τα κλινικά δεδομένα και βοηθά στην πρόγνωση, την αποτελεσματικότερη θεραπεία και την παρακολούθηση των ασθενών (Judith Hurwitz, 2018).

2.3 Cross-lingual ανάλυση συναισθήματος

2.3.1 Ορισμός

Η cross-lingual (διαγλωσσική) είναι η ανάλυση συναισθήματος που πραγματοποιείται σε πολλαπλές γλώσσες. Η ανάγκη για cross-lingual ανάλυση συναισθήματος προκύπτει από τον τεράστιο όγκο πληροφορίας που μαζικά παράγεται από τα διάφορα μέσα και το γεγονός ότι αυτή η πληροφορία είναι πολυγλωσσική. Επομένως, δημιουργείται η ανάγκη για την κατασκευή συστημάτων ανάλυσης συναισθήματος για γλώσσες πέρα από την αγγλική. Όπως είναι φυσικό, η έρευνα στο πεδίο της ανάλυσης συναισθήματος έχει γίνει κατά κύριο λόγο στα αγγλικά, χρησιμοποιώντας τεράστιες ποσότητες δεδομένων με τις αντίστοιχες ετικέτες (labels)

[1] https://www.facebook.com/

συναισθήματος, με αποτέλεσμα η πλειονότητα των εργαλείων και των διαθέσιμων πόρων να περιορίζονται στην αγγλική γλώσσα (Mohamed Abdalla, 2017). Εύκολα αντιλαμβάνεται κανείς, ότι το χτίσιμο αντίστοιχων εργαλείων στις υπόλοιπες γλώσσες θα ήταν αρκετά χρονοβόρο. Συνεπώς, η έρευνα στράφηκε κυρίως σε τρόπους που παρέχουν οικονομία χρόνου και χρήματος, όπως είναι η αξιοποίηση της μηχανικής μετάφρασης, των εργαλείων και των πόρων που είναι διαθέσιμα στα αγγλικά, προκειμένου να δημιουργήσει συστήματα ανάλυσης συναισθήματος και σε άλλες γλώσσες. Τα συστήματα αυτά βρίσκουν εφαρμογή και αποτελούν το πλέον χρήσιμο εργαλείο στον εμπορικό τομέα, καλύπτοντας την επιτακτική ανάγκη των εταιριών για αξιολόγηση του συνόλου των κριτικών ενός προϊόντος ή μιας υπηρεσίας, ανεξάρτητα από τη γλώσσα στην οποία είναι διατυπωμένες (Liu, Sentiment Analysis and Opinion Mining, 2012).

2.3.2 Έρευνα

Από πλευράς έρευνας, οι περισσότερες προσεγγίσεις αφορούν στην ανάλυση συναισθήματος ενός σώματος κειμένου και όχι τόσο στην ανάλυση συναισθήματος βασισμένη σε χαρακτηριστικά (Aspect Based Sentiment Analysis). Οι Faqeeh et al. στην έρευνά τους ασχολούνται με τη διαγλωσσική κατηγοριοποίηση σύντομου κειμένου (όπως αυτό των κοινωνικών δικτύων), χρησιμοποιώντας δεδομένα σε αγγλικά και αραβικά, για τα οποία συγκρίνεται η απόδοση τεσσάρων αλγορίθμων με χρήση του εργαλείου WEKA. (Mosab Faqeeh, 2014). Στην έρευνά του ο Xiaojun Wan για να αντιμετωπίσει το πρόβλημα της διαγλωσσικής ανάλυσης συναισθήματος, προτείνει μια μέθοδο συνεκπαίδευσης για κριτικές σε αγγλικά και κινέζικα, όπου για τη δημιουργία των ταξινομητών αξιοποιούνται από κοινού αγγλικές κριτικές που φέρουν ετικέτες και κινέζικες κριτικές που δε διαθέτουν ετικέτες. (Wan, 2009). Οι Honglei Guo et al. προτείνουν μια μέθοδο εξόρυξης γνώμης βασισμένη σε χαρακτηριστικά, για την διερεύνηση των διαφορετικών απόψεων που εκφράζονται σε διαφορετικές γλώσσες. Αρχικά, κατηγοριοποιούνται τα χαρακτηριστικά των προϊόντων σημασιολογικά και στη συνέχεια, συνοψίζονται οι διαφορές ανά χαρακτηριστικό (Honglei Guo, 2010). Οι Balahur και Turchi διερευνούν την απόδοση των συστημάτων μηχανικής μετάφρασης και καταλήγουν στο συμπέρασμα ότι μπορούν να παραγάγουν καλής ποιότητας δεδομένα εκπαίδευσης για άλλες γλώσσες, πέρα από τα αγγλικά (Alexandra Balahur, 2012). Στο ίδιο μήκος κύματος, οι Adel Al-Shabi et al. προτείνουν ένα απλό μοντέλο για διαγλωσσική ανάλυση συναισθήματος με μηχανική μετάφραση από τα αραβικά στα

αγγλικά και δοκιμάζοντας διαφορετικές τεχνικές εξαγωγής χαρακτηριστικών σε τέσσερα διαφορετικά domains και ενισχύουν την παραπάνω άποψη ότι τα σύγχρονα συστήματα μηχανικής μετάφρασης είναι αρκετά ώριμα ώστε να παραγάγουν αξιόπιστα δεδομένα εκπαίδευσης για γλώσσες περιορισμένων πόρων (Adel Al-Shabi, 2017). Στις παραπάνω έρευνες βασίζεται και η πρόταση που γίνεται στην παρούσα εργασία για την αξιολόγηση της αποδοτικότητας της μηχανικά μεταφρασμένης ανάλυσης συναισθήματος για δεδομένα στην ελληνική γλώσσα.

2.4 Κοινωνικά Δίκτυα

2.4.1 Ορισμός

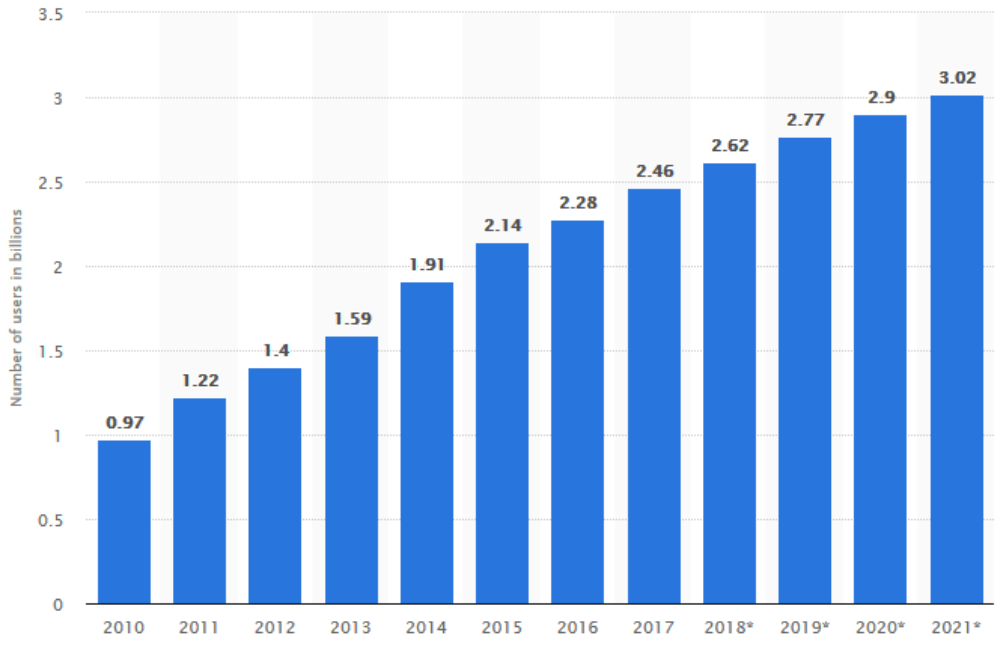
Πρόκειται για μια κοινωνική δομή η οποία περιγράφει τις σχέσεις μεταξύ διαφόρων οντοτήτων (φυσικά πρόσωπα, ομάδες, οργανισμοί). Οι οντότητες απεικονίζονται με κόμβους και οι μεταξύ τους σχέσεις και αλληλεπιδράσεις δηλώνονται με τις συνδέσεις αυτών των κόμβων. Ο όρος κοινωνικά δίκτυα τα τελευταία χρόνια χρησιμοποιείται κυρίως για να περιγράψει τις ιστοσελίδες οι οποίες επιτρέπουν την αλληλεπίδραση και την ανταλλαγή πληροφοριών μεταξύ των χρηστών τους, όπως είναι το Facebook, το YouTube[2], το LinkedIn[3], το Instagram[4] και το Twitter[5] για το οποίο γίνεται περαιτέρω ανάλυση παρακάτω. Όπως απεικονίζεται και στην Εικόνα 2, αριθμός των χρηστών των μέσων κοινωνικής δικτύωσης αυξάνεται χρόνο με το χρόνο και αυτό καταδεικνύει την κυριαρχία και την εδραίωσή τους ως μέσα έκφρασης της κοινής γνώμης για οποιοδήποτε αντικείμενο απασχολεί τους πολίτες, είτε πρόκειται για νέα και πολιτικές εξελίξεις είτε πρόκειται για προϊόντα, υπηρεσίες και δημόσιες υπηρεσίες είτε πρόκειται για δημόσια πρόσωπα και εκδηλώσεις (Manning, 2014).

[2] https://www.youtube.com/

[3] https://www.linkedin.com/

[4] https://www.instagram.com/

[5] https://twitter.com/

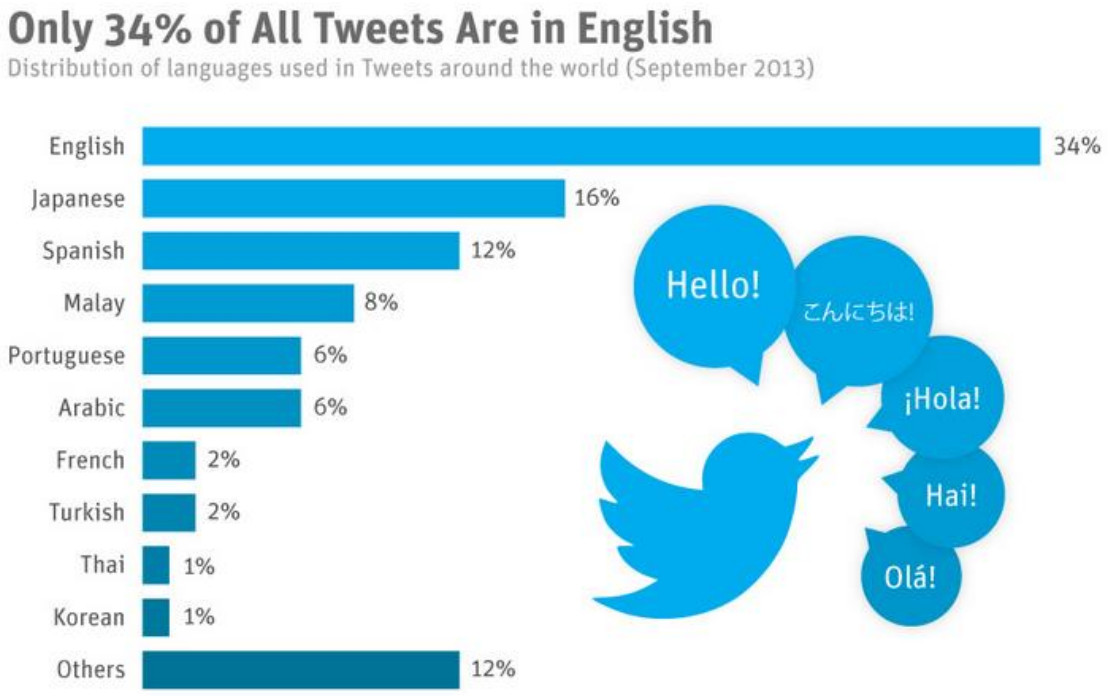


Εικόνα 2 Αριθμός χρηστών κοινωνικών δικτύων παγκοσμίως ανά έτος

(Statista, Number of social media users worldwide from 2010 to 2021 (in billions))

2.4.2 Twitter

Πρόκειται για μια δημοφιλή micro-blogging υπηρεσία, όπου οι χρήστες της επικοινωνούν μέσω σύντομων μηνυμάτων, των tweets και εκφράζουν τις απόψεις τους πάνω σε διάφορα ζητήματα (Alec Go, 2009). Ιδρύθηκε το 2006 και σήμερα αριθμεί περί τα 321 εκατομμύρια ενεργών χρηστών παγκοσμίως. Το γεγονός ότι καθημερινά παράγεται ένας τεράστιος αριθμός tweets σε συνδυασμό με την έκφραση απόψεων για μια ποικίλα θέματα και από ανθρώπους διαφόρων καταβολών και ιδιοτήτων, κάνουν το Twitter μια σημαντική πηγή αναζήτησης της γνώμης των ανθρώπων. (Alexander Pak, 2010) Ένα μεγάλο ποσοστό των tweets διατυπώνεται σε γλώσσες πέραν των αγγλικών (Εικόνα 3), πράγμα που ενισχύει την ανάγκη για έρευνα και στις υπόλοιπες γλώσσες που εκφράζονται τα διάφορα μηνύματα.



Εικόνα 3 Κατανομή των γλωσσών που χρησιμοποιούνται στο Twitter

(Statista, Only 34% of All Tweets Are in English)

Τα δεδομένα του Twitter παρουσιάζουν ιδιαίτερα και μοναδικά χαρακτηριστικά. Καταρχάς, τα tweets διαφοροποιούνται ως προς τις κριτικές, καθώς χαρακτηρίζονται από αυθορμητισμό, εκφράζονται με πιο πρόχειρο λόγο και επίσης, έχουν περιορισμένο μήκος χαρακτήρων (Alec Go, 2009). Παρουσιάζουν όμως διαφορές και με τα δεδομένα των άλλων κοινωνικών δικτύων. Συγκεκριμένα, το μέγεθος των tweets περιορίζονταν στους 140 χαρακτήρες, ενώ από το Νοέμβριο του 2017 το όριο αυτό διπλασιάστηκε. Επίσης, περιέχουν αρκετές συντομεύσεις, λάθη στη σύνταξη λόγω της προχειρότητας στη διατύπωσή τους, όπως επίσης και κάποια tokens, όπως τα hashtags που ξεκινούν με το σύμβολο «#» και δηλώνουν το θέμα του tweet ή τη συναισθηματική του κατάσταση, τα URLs, τα usernames που ξεκινούν με το σύμβολο «@», τα retweets (RT) με οποία κοινοποιείται κάποιο άλλο tweet και τα emoticons που είναι εικονίδια και χρησιμοποιούνται για να απεικονίσουν τη συναισθηματική κατάσταση του χρήστη.

3 Θεωρητικό Υπόβαθρο της Μεθοδολογίας

Στο κεφάλαιο αυτό θα γίνει μια θεωρητική ανάλυση των βασικών μεθόδων που χρησιμοποιούνται στη διαδικασία της ανάλυσης συναισθήματος, κάποιες από τις οποίες θα χρησιμοποιηθούν στην πρότασή μας. Αρχικά, θα γίνει μια σύντομη περιγραφή σχετικά με τη μηχανική μετάφραση και έπειτα, θα γίνει μια ανάλυση των μεθόδων για την προ-επεξεργασία των δεδομένων. Στη συνέχεια, περιγράφονται οι διάφορες τεχνικές για την εξαγωγή χαρακτηριστικών και τέλος, παρουσιάζονται κάποιοι βασικοί ταξινομητές για την ταξινόμηση συναισθήματος, καθώς και μέθοδοι για την αξιολόγηση της απόδοσης του μοντέλου.

3.1 Μηχανική μετάφραση

Η μηχανική μετάφραση είναι μια υπο-περιοχή της υπολογιστικής γλωσσολογίας και ερευνά τη χρήση λογισμικού για τη μετάφραση κειμένου ή λόγου από μία γλώσσα σε μια άλλη. Στη βασική της μορφή, η μηχανική μετάφραση λειτουργεί με απλή μετάφραση λέξη προς λέξη και αντικατάστασή της με εκείνη που της αντιστοιχεί. Ωστόσο, αυτή η πρακτική δεν παράγει καλά αποτελέσματα και συνεπώς, δε μπορεί να θεωρηθεί επαρκής, ειδικά αν σκεφτεί κανείς την ευρεία χρήση του διαδικτύου από την παγκόσμια κοινότητα και τις ανάγκες που το γεγονός αυτό δημιουργεί. Οι βασικές προσεγγίσεις είναι τα συστήματα βάσει κανόνων (rule-based), τα στατιστικά και τα νευρωνικά συστήματα. Η πρώτη προσέγγιση βασίζεται στη γραμματική και το συντακτικό μιας γλώσσας. Τα στατιστικά συστήματα παράγουν μετάφραση που βασίζεται σε στατιστικές μεθόδους που εφαρμόζονται σε δίγλωσσα παράλληλα κείμενα. Τα νευρωνικά συστήματα χρησιμοποιούν μεθόδους μηχανικής μάθησης βασισμένες στην λειτουργία νευρωνικών δικτύων, οι οποίες απαιτούν πολύ μεγάλες ποσότητες κειμένου. Στην τελευταία κατηγορία ανήκει και η υπηρεσία της Google, της οποίας το API[6] χρησιμοποιήθηκε για τις ανάγκες της εργασίας (Irfan, 2017).

[6] https://cloud.google.com/translate/

3.2 Pre-processing δεδομένων

Η εξόρυξη κειμένου (text mining) είναι η διαδικασία εξαγωγής πληροφορίας από ένα δομημένο ή μη σώμα κειμένου και ταυτόχρονα, η εύρεση μοτίβων (patterns) που χαρακτηρίζουν τα εκάστοτε δεδομένα. Οι τεχνικές εξόρυξης γνώμης χρησιμοποιούνται σε διάφορα πεδία, όπως είναι η επεξεργασία φυσικής γλώσσας, η ανάκτηση πληροφορίας, η ταξινόμηση και η συσταδοποίηση κειμένου (S. Vijayarani, 2015).

Με τον όρο προ-επεξεργασία (pre-processing) δεδομένων εννοούμε όλες εκείνες τις ενέργειες που απαιτούνται προκειμένου ένα κείμενο δεδομένων να είναι κατάλληλο για εξαγωγή χαρακτηριστικών και στη συνέχεια, εφαρμογή των αλγορίθμων μηχανικής μάθησης. Το pre-processing των δεδομένων είναι ένα σημαντικό κομμάτι και αποτελεί το πρώτο στάδιο της διαδικασίας για την εξόρυξη κειμένου. Υπάρχουν τρία βασικά βήματα προ-επεξεργασίας δεδομένων, τα οποία είναι: tokenization, stemming ή lemmatization και η απαλοιφή των stop words. Πέρα από τα παραπάνω, ανάλογα με τη δομή των δεδομένων και το είδος της ανάλυσης που σκοπεύει να κάνει κάποιος, εφαρμόζονται επιπλέον τεχνικές. Για παράδειγμα, στην περίπτωση του Twitter, πριν την προ-επεξεργασία θα πρέπει να γίνει διαχείριση των ειδικών tokens που αναφέραμε στο προηγούμενο κεφάλαιο, όπως τα usernames, τα hashtags κλπ.

3.2.1 Tokenization

Πρόκειται για την τεχνική κατάτμησης κατά την οποία μεγαλύτερα τμήματα κειμένου χωρίζονται σε μικρότερα κομμάτια (tokens). Για παράδειγμα, μεγάλα τμήματα κειμένου χωρίζουν σε προτάσεις, οι προτάσεις σε λέξεις κ.ο.κ. Στην πραγματικότητα, η εφαρμογή αυτής της τεχνικής δεν είναι πάντα τόσο απλή, καθώς συχνά υπάρχουν λάθη, με αποτέλεσμα να μην είναι σαφές πότε σηματοδοτείται το τέλος μιας λέξης ή μιας πρότασης. Ακόμα όμως και όταν δεν υπάρχουν λάθη στη σύνταξη, τα σημεία στίξης σε μια λέξη ή πρόταση δημιουργούν ασάφεια ως προς τον τρόπο που θα πρέπει να κατατμηθεί (Kaplan, 2005).

3.2.2 Stemming και Lemmatization

Οι τεχνικές stemming και lemmatization αποτελούν τεχνικές κανονικοποίησης του κειμένου (text normalization) και εφαρμόζονται για να προετοιμάσουν το κείμενο για περαιτέρω ανάλυση.

3.2.2.1 Stemming

Stemming είναι η διαδικασία μείωσης μιας λέξης στην απλούστερή της μορφή, δηλαδή η τεχνική με την οποία αφαιρείται η κατάληξη κάποιας λέξης προκειμένου να πάρουμε τη ρίζα της (stem). Για παράδειγμα:

introduction, introducing, introduces → introduc

gone, going, goes → go

Εφαρμόζεται προκειμένου να επιφέρει μείωση στο μέγεθος του λεξιλογίου προς επεξεργασία, ωστόσο κάποιες φορές η περικοπή της λέξης δεν ανταποκρίνεται εννοιολογικά στην αρχική της σημασία, ενώ κάποιες άλλες, η λέξη που επιστρέφεται δεν είναι πραγματική και δημιουργεί σύγχυση (Matthew J. Denny, 2017). Υπάρχουν τρεις κατηγορίες αλγορίθμων για stemming: μέθοδοι περικοπής (truncating), στατιστικές μέθοδοι και μικτές. Οι μέθοδοι περικοπής χρησιμοποιούνται προκειμένου να αφαιρέσουν το πρόθεμα ή την κατάληξη μιας λέξης. Οι στατιστικές μέθοδοι αφαιρούν τις καταλήξεις, αφού προηγουμένως εφαρμόσουνε κάποια στατιστική διαδικασία. Οι μικτές μέθοδοι περιλαμβάνουν μορφολογική ανάλυση ως προς τις συντακτικές διαφοροποιήσεις σχετικά με τις διάφορες κλίσεις, τον πληθυντικό αριθμό, τα γένη και ανάλυση ως προς τα παράγωγα που συνδέονται με τα μέρη του λόγου μιας πρότασης.

Ένας από τους πιο δημοφιλείς αλγόριθμους είναι ο Porter Stemmer, ο οποίος προτάθηκε το 1980. Ανήκει στην κατηγορία των truncating μεθόδων και βασίζεται στην ιδέα ότι οι καταλήξεις στην αγγλική γλώσσα αποτελούνται από άλλες μικρότερες και πιο απλές (Zahurul Islam, 2010). Είναι ιδιαίτερα αποτελεσματικός αλγόριθμος συγκριτικά με τους υπόλοιπους, ωστόσο οι λέξεις που επιστρέφει δεν είναι αληθινές (S. Vijayarani, 2015).

3.2.2.2 Lemmatization

Η λογική και η βασική λειτουργία της τεχνικής αυτής είναι σε γενικές γραμμές ίδια με αυτή του stemming. Ομοίως λοιπόν, η τεχνική lemmatization μειώνει τις διαφορετικές μορφές μιας λέξης και τις επαναφέρει στο λήμμα τους (lemma). Η διαφορά ανάμεσα στις δύο μεθόδους έγκειται στο γεγονός ότι στην περίπτωση που εφαρμόζεται lemmatization λαμβάνεται υπόψη η συντακτική λειτουργία και η έννοια της λέξης, κάτι που η τεχνική stemming αγνοεί. Επομένως, χρησιμοποιώντας lemmatization, ο αλγόριθμος καταφέρνει να επιστρέψει ως αποτέλεσμα τη γλωσσολογικά σωστή ρίζα της εκάστοτε λέξης και το αποτέλεσμα αυτό είναι υπαρκτή λέξη (Jivani, 2011). Για παράδειγμα:

introduction, introducing, introduces → introduce

gone, going, goes, went → go

3.2.3 Απαλοιφή των Stop words

Σε γενικές γραμμές, η διαδικασία της προ-επεξεργασίας των δεδομένων έχει ως στόχο την τροποποίηση του κειμένου με τρόπο τέτοιο ώστε να μπορεί να εφαρμοστεί η εξαγωγή χαρακτηριστικών. Πέρα όμως από αυτό, έχει ως στόχο την αποβολή της άχρηστης πληροφορίας. Τα stop words είναι τμήμα της φυσικής γλώσσας και πρόκειται για λέξεις που δεν προσφέρουν κάτι στο νόημα του κειμένου, όπως είναι τα άρθρα, οι προθέσεις, οι αντωνυμίες κ.ά. Αφαιρώντας τις λέξεις αυτές μειώνεται ταυτόχρονα και η πολυπλοκότητα του προβλήματος. Υπάρχουν τέσσερις μέθοδοι αφαίρεσης των stop words. Στην κλασική (Classic) μέθοδο αφαιρούνται stop words με βάση κάποια έτοιμη λίστα (Jivani, 2011). Στις μεθόδους Z, επιπλέον της έτοιμης λίστας, χρησιμοποιούνται τρεις μέθοδοι: αφαίρεση των πιο συχνά εμφανιζόμενων λέξεων, αφαίρεση των λέξεων που εμφανίζονται μόνο μία φορά και αφαίρεση των λέξεων που εμφανίζονται συχνά στο σύνολο των κειμένων. Στη μέθοδο κοινής πληροφορίας (Mutual Information) υπολογίζεται η κοινή πληροφορία μεταξύ μιας λέξης και μιας κλάσης και σε περίπτωση που είναι χαμηλή, αυτό σημαίνει ότι η λέξη δεν προσφέρει ιδιαίτερη πληροφορία για την κλάση και προτείνεται η αφαίρεσή της. Στη μέθοδο τυχαίας δειγματοληψίας βασισμένη σε όρους (Term Based Random Sampling) ταξινομούνται οι όροι με βάση το μέτρο απόκλισης Kullback-Leibler σε τυχαία επιλεγμένα διαδικτυακά έγγραφα και στη συνέχεια, κατασκευάζεται η λίστα των stop words (Jivani, 2011) (Sharma, 2012).

3.2.4 Part-Of-Speech Tagging

Με τον όρο part-of-speech εννοούμε τα μέρη του λόγου, όπως τα ουσιαστικά, τα ρήματα, τα επίθετα, τα επιρρήματα κ.ά. Part-of-speech tagging είναι η διαδικασία σημείωσης των λέξεων ενός κειμένου και η αντιστοίχισή τους με ένα συγκεκριμένο μέρος του λόγου με βάση τον ορισμό τους και τη σημασία τους στο κείμενο. Η διαδικασία γίνεται με τη βοήθεια αλγορίθμων που εμπίπτουν σε δύο κατηγορίες, τους αλγόριθμους βάσει κανόνων και τους πιθανοτικούς. Στην πρώτη περίπτωση, ο αλγόριθμος δέχεται σαν είσοδο μια αλληλουχία κατατμημένων λέξεων και ενός tagset και επιστρέφει ως αποτέλεσμα την αλληλουχία των λέξεων, όπου κάθε λέξη συνοδεύεται από την κατάλληλη ετικέτα. Στη δεύτερη περίπτωση, ο αλγόριθμος βασισμένος σε μοντέλα όπως το κρυφό μαρκοβιανό μοντέλο (HMM), έχοντας ένα σώμα κειμένου

υπολογίζει την πιθανοτική κατανομή των διαφόρων αλληλουχιών των ετικετών και επιλέγει την καλύτερη. Η απόδοση ετικέτας αποτελεί μια λειτουργία αποσαφήνισης, καθώς μία λέξη μπορεί να συνδέεται με περισσότερες από μία ετικέτες. Οπότε στόχος είναι επιλογή της σωστής ετικέτας ανάλογα με το νόημα του κειμένου. Η μέθοδος του part-of-speech tagging εξυπηρετεί την καλύτερη διερεύνηση ενός σώματος κειμένου από γλωσσολογικής πλευράς. Επίσης, πέρα από ότι προσφέρει πληροφορία για μια λέξη, προσφέρει και γνώση για τις γειτονικές της. Μία από τις πιο σημαντικές συλλογές ετικετών για τα αγγλικά είναι το Penn Treebank tagset, το οποίο έχει χρησιμοποιηθεί για την απόδοση ετικέτας σε πολλά σώματα κειμένου (Πίνακας 1) (Daniel Jurafsky, 2018).

Tag	Description	Tag	Description
CC	Coordinating conjunction	PRP$	Possessive pronoun
CD	Cardinal number	RB	Adverb
DT	Determiner	RBR	Adverb, comparative
EX	Existential *there*	RBS	Adverb, superlative
FW	Foreign word	RP	Particle
IN	Preposition or subordinating conjunction	SYM	Symbol
JJ	Adjective	TO	*to*
JJR	Adjective, comparative	UH	Interjection
JJS	Adjective, superlative	VB	Verb, base form
LS	List item marker	VBD	Verb, past tense
MD	Modal	VBG	Verb, gerund or present participle
NN	Noun, singular or mass	VBN	Verb, past participle
NNS	Noun, plural	VBP	Verb, non-3rd person singular present
NNP	Proper noun, singular	VBZ	Verb, 3rd person singular present
NNPS	Proper noun, plural	WDT	Wh-determiner
PDT	Predeterminer	WP	Wh-pronoun
POS	Possessive ending	WP$	Possessive wh-pronoun
PRP	Personal pronoun	WRB	Wh-adverb

Πίνακας 1 Penn Treebank Part-Of-Speech (POS) tags

(Liu, Sentiment Analysis and Opinion Mining, 2012)

3.3 Εξαγωγή χαρακτηριστικών

Τη διαδικασία της προ-επεξεργασίας των δεδομένων ακολουθεί αυτή της εξαγωγής χαρακτηριστικών. Η μέθοδος αυτή έχει ως στόχο να μετατρέψει τα δεδομένα σε μορφή κατάλληλη (διανυσματική) ώστε να χρησιμοποιηθούν από τους αλγόριθμους ταξινόμησης. Παράλληλα, με την εξαγωγή χαρακτηριστικών γίνεται προσπάθεια διατηρηθούν μόνο τα δεδομένα που αποτελούν χρήσιμη πληροφορία. Οι μέθοδοι

εξαγωγής και επιλογής χαρακτηριστικών χρησιμοποιούνται ξεχωριστά ή σε συνδυασμό προκειμένου να βελτιώσουν την αποτελεσματικότητα και την ακρίβεια του μοντέλου (Hiroshi Motoda, 2002). Παρακάτω, αναλύονται συνοπτικά μερικές από τις κυριότερες τεχνικές εξαγωγής χαρακτηριστικών.

3.3.1 Bag of words

Ο απλούστερος τρόπος εξαγωγής χαρακτηριστικών είναι η τεχνική bag of words. Με αυτή την τεχνική δημιουργείται ένα σύνολο χαρακτηριστικών, το λεξικό και στη συνέχεια, κάθε πρόταση ή κείμενο αναπαριστάται ως ένα αραιό διάνυσμα με τιμές 0 για την απουσία και 1 για την παρουσία της προς εξέταση λέξης στο λεξικό. Έχει πάρει την ονομασία της από το γεγονός ότι δεν λαμβάνει υπόψη τη σειρά των λέξεων ή πόσες φορές μια λέξη εμφανίζεται στο κείμενο, αλλά εξετάζει μόνο αν η λέξη ανήκει στη λίστα με τα χαρακτηριστικά (term occurence) (Resham N. Waykole, 2018).

Για παράδειγμα:

A = *"John likes to watch movies. Mary likes movies too."*

B = *"John also likes to watch football games."*

Το λεξικό είναι:

"John", "likes", "to", "watch", "movies", "Mary", "too", "also", "football", "games"

Term occurrence αναπαράσταση:

A = [1,1,1,1,1,1,1,0,0,0]

B = [1,1,1,1,0,0,0,1,1,1]

3.3.1.1 Term Frequency

Πρόκειται για μια ελαφρώς διαφορετική εκδοχή της τεχνικής bag of words, κατά την οποία λαμβάνεται υπόψη η συχνότητα εμφάνισης των λέξεων σε ένα κείμενο. Επομένως, το κείμενο αναπαριστάται και πάλι ως ένα αραιό διάνυσμα, με τιμές όμως ακέραιους αριθμούς που αντιστοιχούν στην συχνότητα εμφάνισης των λέξεων στο κείμενο. Η τιμή του μέτρου term frequency υπολογίζεται ως εξής:

$$TF(t) = \frac{Number\ of\ times\ term\ t\ appears\ in\ a\ document}{Total\ number\ of\ terms\ in\ the\ document}$$

Το πρόβλημα με αυτή τη μέθοδο είναι ότι οι όροι που χρησιμοποιούνται συχνά γίνονται κυρίαρχοι, ενώ πιθανότατα δεν παρέχουν ιδιαίτερη πληροφόρηση για τη σωστή εκπαίδευση του μοντέλου. Ουσιαστικά, πρόκειται για stop words, που θα πρέπει να αφαιρούνται ή να αγνοούνται (Resham N. Waykole, 2018).

Για το παραπάνω παράδειγμα, η term frequency αναπαράσταση είναι:

A = [1,2,1,1,2,1,1,0,0,0]

B = [1,1,1,1,0,0,0,1,1,1]

Τη λύση στο παραπάνω πρόβλημα δίνει η τεχνική Tf-Idf, η οποία αναλύεται παρακάτω.

3.3.1.2 Tf-Idf

Η προσέγγιση Tf-Idf (Term frequency – Inverse document frequency) είναι μία παραλλαγή της μεθόδου term frequency και έχει ως στόχο να απεικονίσει πόσο σημαντική είναι μια λέξη σε ένα κείμενο που αποτελεί μέρος μιας συλλογής κειμένων. Η τιμή του μέτρου Tf-Idf αυξάνεται ανάλογα με τον αριθμό που μια λέξη εμφανίζεται σε ένα κείμενο και αντισταθμίζεται από τον αριθμό των κειμένων που περιέχουν τη λέξη. Η Tf-Idf είναι από τις πιο δημοφιλείς μεθόδους στην κατηγορία term-weighting (Daniel Jurafsky, 2018). Η τιμή του Idf υπολογίζεται ως εξής:

$$IDF(t) = log_e(\frac{Total\ number\ of\ documents}{Number\ of\ documents\ with\ term\ t\ in\ it})$$

Συνεπώς, η τιμή του όρου Idf είναι υψηλή για τις λέξεις που δεν εμφανίζονται συχνά σε μία συλλογή κειμένων και χαμηλή για εκείνες που εμφανίζονται συχνά. Επομένως, για το μέτρο Tf-Idf, έχουμε:

$$TF - IDF score = TF * IDF$$

3.3.2 Word2Vec

Το Word2Vec είναι ένα νευρωνικό μοντέλο και χρησιμοποιείται για την κατασκευή word embeddings. Το μοντέλο λαμβάνει σαν είσοδο ένα σώμα κειμένου μεγάλου όγκου και κατασκευάζει ένα διανυσματικό χώρο εκατοντάδων διαστάσεων, όπου σε κάθε μοναδική λέξη του κειμένου αντιστοιχίζεται ένα διάνυσμα στο χώρο. Το Word2Vec χρησιμοποιεί δύο υλοποιήσεις, είτε την continuous bag of words είτε την continuous skip gram. Στην πρώτη περίπτωση, το μοντέλο προβλέπει την τρέχουσα λέξη με βάση τις γειτονικές της λέξεις, χωρίς να επηρεάζεται από τη σειρά των λέξεων. Στη δεύτερη περίπτωση, το μοντέλο χρησιμοποιεί την τρέχουσα λέξη προκειμένου να προβλέψει τις γειτονικές της (Daniel Jurafsky, 2018).

3.4 Αλγόριθμοι ταξινόμησης

Στο προηγούμενο κεφάλαιο έγινε εκτενής αναφορά στη μηχανική μάθηση και τις κατηγορίες αλγορίθμων. Παρακάτω θα γίνει συνοπτική περιγραφή δύο βασικών αλγορίθμων επιβλεπόμενης μάθησης, του Naive Bayes και του Random Forest, που είναι και οι αλγόριθμοι που θα χρησιμοποιηθούν στην πρότασή μας.

3.4.1 Naive Bayes

Οι ταξινομητές Naive Bayes είναι μια οικογένεια πιθανοτικών αλγορίθμων οι οποίοι βασίζονται στο θεώρημα του Bayes, βασική αρχή του οποίου είναι η απλοϊκή υπόθεση ότι τα χαρακτηριστικά είναι πλήρως ανεξάρτητα μεταξύ τους. Ερευνώνται συστηματικά ήδη από τη δεκαετία του 1960 και παραμένουν μέχρι και σήμερα μια δημοφιλής μέθοδος ταξινόμησης κειμένου. Ο Naive Bayes είναι μια απλή μέθοδος κατασκευής ταξινομητών, όπου σε κάθε στιγμιότυπο του προβλήματος αναθέτεται μία ετικέτα κλάσης και αναπαριστάται ως διάνυσμα στο χώρο με βάση τα χαρακτηριστικά. Ουσιαστικά, μοντελοποιεί την κατανομή των κειμένων σε κλάσεις, χρησιμοποιώντας ένα πιθανοτικό μοντέλο με υπόθεση ανεξαρτησίας σε ό,τι αφορά στην κατανομή των διαφορετικών όρων (Charu C. Aggarwal, 2012). Παρά την απλοϊκή κατασκευή του και τις υπεραπλουστευτικές υποθέσεις στις οποίες βασίζεται, ο Naive Bayes επιτυγχάνει πολύ καλά αποτελέσματα σε διάφορα πολύπλοκα προβλήματα. Το πλεονέκτημά του έγκειται στο γεγονός ότι απαιτεί ένα μικρό σύνολο δεδομένων προς εκπαίδευση προκειμένου να κάνει την ταξινόμηση, το οποίο σημαίνει μικρό χρόνο εκπαίδευσης και

χαμηλές απαιτήσεις από πλευράς μνήμης. Επομένως, χρησιμοποιώντας το θεώρημα Bayes, η λειτουργία του μοντέλου συνοψίζεται ως εξής:

$$p(C_k \mid \mathbf{x}) = \frac{p(C_k)\, p(\mathbf{x} \mid C_k)}{p(\mathbf{x})}$$

Όπου:

- $p(C_k \mid x)$: η πιθανότητα να συμβεί το C_k με δεδομένο το x
- $p(C_k)$: η πιθανότητα να συμβεί το C_k
- $p(x \mid C_k)$: η πιθανότητα να συμβεί το x με δεδομένο το C_k
- $p(x)$: η πιθανότητα να συμβεί το x

ή πιο απλά:

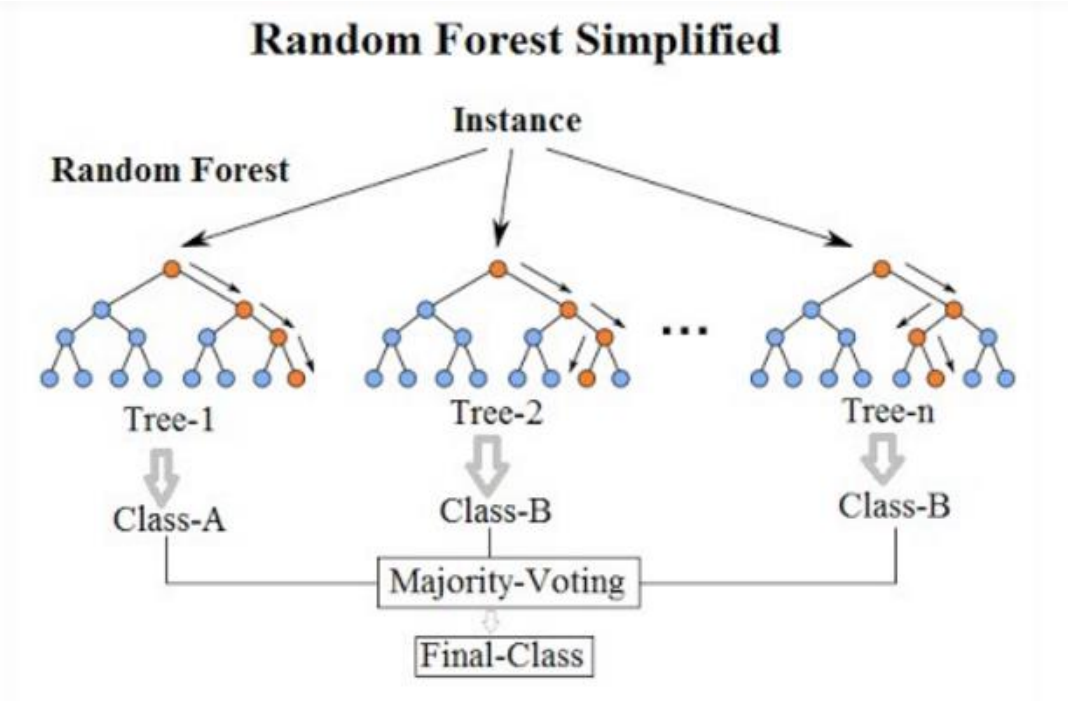
$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}$$

Τα μοντέλα που χρησιμοποιούνται συνήθως, για την ταξινόμηση με Naive Bayes είναι το μοντέλο Multivariate Bernoulli και το Multinomial μοντέλο. Και τα δύο μοντέλα υπολογίζουν την πιθανότητα μιας κλάσης με βάση την κατανομή των λέξεων στο κείμενο. Επίσης, αγνοούν την πραγματική θέση των λέξεων στο κείμενο και λειτουργούν με τη λογική bag of words. Η σημαντικότερη διαφορά μεταξύ των δύο είναι ότι στην περίπτωση του μοντέλου Bernoulli χρησιμοποιείται η εμφάνιση ή μη μιας λέξης προκειμένου να εξαχθούν χαρακτηριστικά, ενώ στο Multinomial μοντέλο λαμβάνεται υπόψη η συχνότητα εμφάνισης των λέξεων (Charu C. Aggarwal, 2012).

3.4.2 Random Forest

Ο αλγόριθμος Random Forest αποτελεί μια δημοφιλή μέθοδο επιβλεπόμενης μάθησης κατάλληλη για ταξινόμηση, αλλά και για παλινδρόμηση (regression). Πρόκειται για μια προσέγγιση που ανήκει στις ensemble μεθόδους μάθησης, βασική αρχή των οποίων είναι ότι μία ομάδα αδύναμων learners μπορεί να διαμορφώσει έναν ισχυρό learner. Αυτό σημαίνει ότι χρησιμοποιούνται πολλοί αλγόριθμοι μάθησης προκειμένου να επιτευχθεί καλύτερο αποτέλεσμα συγκριτικά με τη χρήση μεμονωμένα ενός από αυτούς (Mauro Castelli, 2019). Ο Random Forest λειτουργεί κατασκευάζοντας ένα μεγάλο αριθμό δένδρων απόφασης κατά τη διαδικασία της εκπαίδευσης και παράγει

ως έξοδο την κλάση που εμφανίζεται συχνότερα στα μεμονωμένα δένδρα (classification) ή τη μέση πρόβλεψη (regression). Ο συγκεκριμένος αλγόριθμος έχει καλή απόδοση συγκριτικά με άλλους, θεωρείται αρκετά σταθερός, αντιστέκεται στο over-fitting και αποδίδει καλά σε διάφορα domains και ανεξάρτητα από το γεγονός αν τα δεδομένα είναι κείμενου ή αριθμητικά. Στα αρνητικά, ο αλγόριθμος παρουσιάζει μεγάλη πολυπλοκότητα λόγω της παρουσίας πολλών δένδρων απόφασης, γεγονός που συνεπάγεται και περισσότερο χρόνο πρόβλεψης για νέα δεδομένα. Τέλος, πρέπει να αναφερθεί ότι ο Random Forest είναι ένα εργαλείο μοντελοποίησης προβλέψεων και όχι ένα περιγραφικό εργαλείο, οπότε η ερμηνεία και περιγραφή των σχέσεων που υπάρχουν στα δεδομένα είναι ανέφικτη, αν αναλογιστεί κανείς και την παρουσία ενός μεγάλου αριθμού δένδρων απόφασης στο τελικό μοντέλο (Louppe, 2014). Ο Random Forest μοντελοποιείται οπτικά ως εξής:



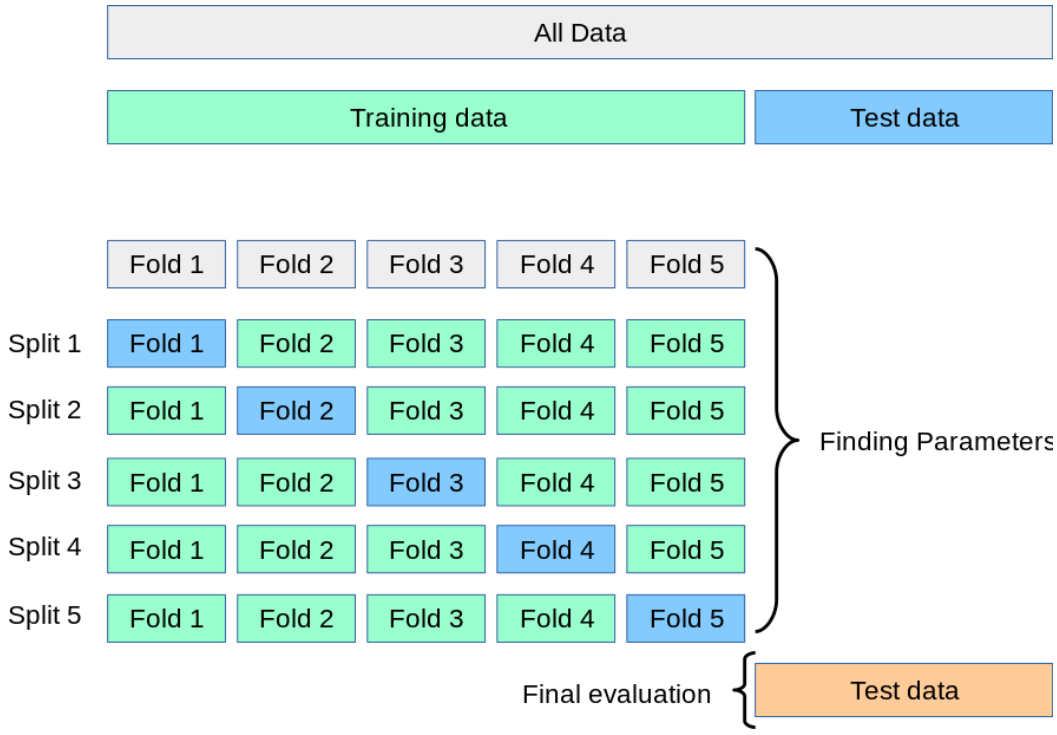
Εικόνα 4 Απεικόνιση και λειτουργία του αλγόριθμου Random Forest (Medium)

3.5 Αξιολόγηση μοντέλου

Στην ενότητα αυτή γίνεται μια συνοπτική περιγραφή των μεθόδων και των στατιστικών μέτρων που θα χρησιμοποιηθούν στην αξιολόγηση του ταξινομητή μας.

3.5.1 Cross-validation

Η μέθοδος cross-validation είναι μια στατιστική μέθοδος η οποία χρησιμοποιείται για την αξιολόγηση ενός μοντέλου. Περιλαμβάνει την τυχαία κατάτμηση του συνόλου των δεδομένων σε k αμοιβαία αποκλειόμενα ίσα μέρη και για κάθε μέρος ο ταξινομητής εκπαιδεύεται με τα μέρη που απομένουν, δηλαδή για k-1 (Kohavi, 1995). Έπειτα, γίνονται προβλέψεις για τα εκάστοτε μέρη και στη συνέχεια, γίνεται σύγκριση με τις ετικέτες που έχουν αποδοθεί στα δεδομένα. Αυτό επιτρέπει να διαπιστώσει κανείς αν κάποια πρόβλεψη ήταν σωστή και πότε έγινε λάθος και στη συνέχεια, να κατασκευαστούν τα μέτρα υπολογισμού της απόδοσης του μοντέλου. Η μέθοδος αυτή είναι κατάλληλη για τις περιπτώσεις που υπάρχει περιορισμένος όγκος δεδομένων, καθώς δίνει τη δυνατότητα να συμμετάσχει στη διαδικασία της εκπαίδευσης το σύνολο των δεδομένων. Επίσης, παρέχει μεγαλύτερη αξιοπιστία, καθώς στη φάση της δοκιμής του ταξινομητή συμμετέχει διαδοχικά ολόκληρο το σύνολο δεδομένων. Η διαδικασία για k-fold cross-validation για k = 5, σχηματοποιείται παρακάτω:



Εικόνα 5 Διαδικασία 5-fold Cross-validation

(Scikit-learn)

3.5.2 Στατιστικά μέτρα αξιολόγησης

Μετά την κατασκευή ενός ταξινομητή, ακολουθεί η αξιολόγησή του ως προς την ακρίβεια των προβλέψεών του. Η σωστή αξιολόγηση ενός μοντέλου είναι πολύ σημαντική προκειμένου να μπορεί να χρησιμοποιηθεί περαιτέρω και έχει ως βάση τη σύγκριση των αποτελεσμάτων της κατηγοριοποίησης με τη χρήση του ταξινομητή και των ετικετών που έχουν αποδοθεί εξ αρχής (Liu, Web Data Mining – Exploring Hyperlinks, Contents and Usage Data, 2011).

3.5.2.1 Accuracy

Το κυριότερο μέτρο αξιολόγησης είναι η ορθότητα (accuracy), που υπολογίζεται ως ο αριθμός των ορθά ταξινομημένων δεδομένων διαιρεμένος με το σύνολο του συνόλου των δεδομένων δοκιμής. Κάποιες φορές χρησιμοποιείται επίσης, το ποσοστό σφάλματος (error rate), το οποίο είναι 1 – *accuracy*.

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

ή διαφορετικά:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Ωστόσο, το μέτρο του accuracy δεν είναι πάντα κατάλληλο, καθώς στις περιπτώσεις που μας ενδιαφέρουν περισσότερες από μία κλάσεις, δεν παρέχει πάντα αξιόπιστη πληροφορία. Για παράδειγμα, στην περίπτωση email ham / spam classification, αν υποθέσουμε ότι τα δεδομένα μας είναι κατά 99% ham, τότε στην περίπτωση που ο ταξινομητής προβλέπει ότι όλα τα email είναι ham, αυτό σημαίνει ότι θα έχουμε 99% accuracy, το οποίο όμως, δεν είναι αξιόπιστο, αφού θα υπάρχει απόλυτη αποτυχία στην πρόβλεψη των spam. Για το λόγο αυτό, πέρα από το μέτρο του accuracy, χρησιμοποιούνται επιπλέον μέτρα για τον υπολογισμό της απόδοσης ενός μοντέλου, όπως είναι τα μέτρα confusion matrix, precision, recall και f1-score.

3.5.2.2 Confusion matrix

Είναι ένα από τα κλασικά μέτρα απόδοσης και πρόκειται για ένα πίνακα που αναπαριστά την απόδοση ενός αλγόριθμου και επιτρέπει να διαπιστώσει κανείς εύκολα

αν υπάρχει σύγχυση στην κατηγοριοποίηση μεταξύ δύο κλάσεων. Οι γραμμές απεικονίζουν τις προβλέψεις, ενώ οι στήλες τις πραγματικές κλάσεις ή μπορεί να συμβαίνει και το αντίστροφο. Η μήτρα σύγχυσης έχει την παρακάτω μορφή:

	Predicted negative	Predicted positive
Actual negative	TN	FP
Actual positive	FN	TP

Πίνακας 2 Confusion Matrix

όπου:

TN: ο αριθμός των ορθών ταξινομήσεων στην αρνητική κλάση

FN: ο αριθμός των λανθασμένων ταξινομήσεων στην αρνητική κλάση

FP: ο αριθμός των λανθασμένων ταξινομήσεων στη θετική κλάση

TP: ο αριθμός των ορθών ταξινομήσεων στη θετική κλάση

3.5.2.3 Precision, recall, f1-score

Με βάση το μέτρο του confusion matrix και τις πληροφορίες που παρέχει, δημιουργούνται κάποια ακόμα μέτρα απόδοσης του μοντέλου, όπως είναι η ακρίβεια (precision), η ανάκληση (recall) και ο ισορροπημένος αρμονικός μέσος (f1-score).

Η ακρίβεια εστιάζει σε μια συγκεκριμένη κλάση και έχει να κάνει με το ποσοστό των θετικών (ή των αρνητικών) προβλέψεων που ήταν ορθές. Εκφράζεται με τον τύπο:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Αντίστοιχα, η ανάκληση εστιάζει στην κατηγοριοποίηση και υπολογίζει τι ποσοστό της θετικής (ή της αρνητικής) κλάσης προβλέπεται σωστά. Εκφράζεται με τον τύπο:

$$\text{Recall} = \frac{TP}{TP + FN}$$

Το μέτρο f1-score αποτελεί ένα ακόμα στατιστικό μέτρο και πρόκειται για τον αρμονικό μέσο των precision και recall:

$$f_1\text{-score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

4 Μεθοδολογία

Στο κεφάλαιο αυτό, γίνεται συγκριτική μελέτη τριών διαφορετικών τρόπων για την ανάλυση συναισθήματος ελληνικών tweets. Η βασική διαφοροποίηση έγκειται στη διαδικασία του pre-processing των δεδομένων, όπου χρησιμοποιούνται οι δύο πιο δημοφιλείς βιβλιοθήκες στο πεδίο της επεξεργασίας φυσικής γλώσσας (NLP). Συγκεκριμένα, στην πρώτη περίπτωση χρησιμοποιείται η κλασική βιβλιοθήκη επεξεργασίας γλώσσας NLTK[7] σε συνδυασμό με μηχανικά μεταφρασμένα ελληνικά δεδομένα στα αγγλικά. Στη δεύτερη περίπτωση, χρησιμοποιούνται και πάλι μηχανικά μεταφρασμένα δεδομένα τα οποία επεξεργάζονται κατάλληλα με χρήση της βιβλιοθήκης spaCy[8]. Στην τρίτη περίπτωση, χρησιμοποιείται η spaCy για απευθείας επεξεργασία ελληνικών δεδομένων, μιας και είναι η μόνη που προσφέρει αυτή τη δυνατότητα.

Παρακάτω, γίνεται αναλυτική περιγραφή της μεθοδολογίας για το σύνολο των περιπτώσεων, αναφέροντας τις κοινές τεχνικές που χρησιμοποιούν σε κάθε στάδιο της διαδικασίας, καθώς και τα σημεία στα οποία διαφοροποιούνται. Το σύνολο της διαδικασίας έγινε χρησιμοποιώντας τη γλώσσα Python[9], έκδοσης 3.7.

4.1 Συλλογή δεδομένων

Το σύνολο των δεδομένων αποτελείται από 1876 ελληνικά tweets και προκύπτει από δύο επιμέρους datasets. Το πρώτο dataset (“GRGE”) περιέχει θετικά, αρνητικά και ουδέτερα tweets πολιτικού περιεχομένου σχετιζόμενο με τις εκλογές του Ιανουαρίου 2015 στην Ελλάδα (Adam Tsakalidis, 2018). Από το παραπάνω dataset αφαιρέθηκε η ουδέτερη κλάση και χρησιμοποιήθηκαν τελικά 661 tweets, από τα οποία 582 αρνητικά και 79 θετικά. Το δεύτερο dataset αποτελείται από θετικά, αρνητικά και ουδέτερα ελληνικά και αγγλικά tweets που αφορούν στο 53ο Φεστιβάλ Ταινιών Θεσσαλονίκης, από το οποίο αφαιρέθηκαν τα αγγλικά tweets, καθώς και το σύνολο των ουδέτερων tweets. (E. Schinas, 2013). Άρα, το δεύτερο dataset αποτελείται από 1215, 895 θετικά και 320 αρνητικά. Κατά συνέπεια, η αναλογία θετικών και αρνητικών στο σύνολο των δεδομένων που χρησιμοποιήθηκε είναι 974 προς 902.

[7] https://www.nltk.org/

[8] https://spacy.io/

[9] www.python.org/

Topic	Number of Tweets	Positive	Negative
Greek Elections 2015	661	79	582
53th Thessaloniki International Film Festival (#tiff)	1215	895	320
Total	1876	974	902

Πίνακας 3 Περιγραφή Dataset

4.1.1 Προετοιμασία δεδομένων

Μετά τη συλλογή των δεδομένων και πριν τη διαδικασία του pre-processing, ακολουθεί η προετοιμασία, κατά την οποία γίνεται το καθάρισμα των δεδομένων το οποίο γίνεται με χρήση *regular expressions*. Συγκεκριμένα, εφόσον πρόκειται για δεδομένα του Twitter, αφαιρούνται οι χρήστες, τα ReTweets, τα URLs, τα Hashtags και τα περιττά κενά, όπως φαίνεται στο τμήμα του κώδικα παρακάτω (Εικόνα 6):

```
def CleanTweets(aList):
    cleanTweetsList = []
    for tweet in aList:
        myStr = ' '. join(tweet[:-1])
        removeUser = re.sub(r"@\S*", "", myStr)
        removeRTs = re.sub(r"RT\S*", "", removeUser)
        removeHTT = re.sub(r"htt\S*", "", removeRTs)
        removeURL = re.sub(r"\S*\.gr|\S*\.com", "", removeHTT)
        removeHashtag = re.sub(r"#\S*", "", removeURL)
        removePunctuation = re.sub(r"[^\w\s]", "", removeHashtag)
        removeDigits = re.sub(r"\d", "", removePunctuation)
        removeSpaces = re.sub(r"\s+", " ", removeDigits)
        removeEdgeSpaces = removeSpaces.strip()
        cleanTweetsList.append([removeEdgeSpaces, tweet[-1]])
    return(cleanTweetsList)
```

Εικόνα 6 Προετοιμασία δεδομένων

4.1.2 Μηχανική μετάφραση

Για τις δύο πρώτες περιπτώσεις που χρησιμοποιούνται τα μηχανικά μεταφρασμένα δεδομένα, πραγματοποιείται μηχανική μετάφραση με τη βοήθεια του Google Translate API[10]. Με την εκτέλεση της παρακάτω εντολής εγκαθίσταται το API:

pip install google-cloud-translate.

Παρακάτω, η Εικόνα 7 δείχνει το τμήμα του κώδικα που χρησιμοποιήθηκε για τη μετάφραση των tweets, για το οποίο στη συνέχεια, γίνεται αποκωδικοποίηση των

[10] https://cloud.google.com/translate/

ειδικών οντοτήτων της HTML που εμφανίζονται στα δεδομένα μετά την ολοκλήρωση της μετάφρασης, π.χ. ", #, κ.ά.:

```
def gr2enTranslation(aList):
    gr2en_translated = []
    for tweet in aList:
        tmp = []
        translate = translateClient.translate(tweet[0], target_language='en')
        tmp = [translate['translatedText'], tweet[-1]]
        myStr = ' '.join(tmp)
        # decode HTML special entities in Python string
        removeHTML = html.unescape(myStr)
        gr2en_translated.append(removeHTML)
    return(gr2en_translated)
```

Εικόνα 7 Μετάφραση δεδομένων

4.2 Pre-processing δεδομένων

Μετά το καθάρισμα των δεδομένων και τη μηχανική μετάφραση, προκύπτουν δύο αρχεία, ένα αρχείο με τα ελληνικά και ένα με τα μηχανικά μεταφρασμένα σε αγγλικά tweets. Επομένως, τα δεδομένα μας είναι πλέον έτοιμα για τη διαδικασία του pre-processing.

4.2.1 Tokenization

Το πρώτο στάδιο της διαδικασίας είναι η κατάτμηση των δεδομένων ανά λέξη. Για την πρώτη περίπτωση, χρησιμοποιείται η βιβλιοθήκη NLTK, η οποία εγκαθίσταται με την παρακάτω εντολή:

pip install nltk.

Προκειμένου να διαχωριστούν τα δεδομένα, χρησιμοποιείται το word_tokenize από το πακέτο nltk.tokenize, όπως φαίνεται παρακάτω.

```
def Tokenize(aList):
    tokensList = []
    for tweet in aList:
        tokens = word_tokenize(tweet)
        tokensList.append(tokens)
    return(tokensList)
```

Εικόνα 8 Tokenize στην NLTK

Στη δεύτερη περίπτωση, χρησιμοποιείται η βιβλιοθήκη spaCy, η εγκατάσταση της οποίας γίνεται με την παρακάτω εντολή:

pip install spacy.

Τα δεδομένα μας είναι μηχανικά μεταφρασμένα, όπως και στην παραπάνω περίπτωση, επομένως πρόκειται για αγγλικά. Ακολουθεί εγκατάσταση του αγγλικού μοντέλου με την εντολή:

python -m spacy download en_core_web_sm.

Η συνολική διαδικασία του pre-processing για τα αγγλικά γίνεται χρησιμοποιώντας το παραπάνω γλωσσικό μοντέλο ως εξής:

nlp = spacy.load('en_core_web_sm').

Παρόμοια, στην τρίτη περίπτωση η επεξεργασία των ελληνικών δεδομένων γίνεται εγκαθιστώντας το αντίστοιχο γλωσσικό μοντέλο με την εντολή:

python -m spacy download el_core_news_sm

και χρησιμοποιώντας το, με τη βοήθεια της εντολής:

nlp = spacy.load('el_core_news_sm').

Η κατάτμηση των δεδομένων και για τις δύο περιπτώσεις γίνεται χρησιμοποιώντας το παρακάτω τμήμα του κώδικα:

```
def Tokenize(aList):
    tokensList = []
    for tweet in aList:
        removeEdgeSpaces = tweet.strip()
        rev = nlp(removeEdgeSpaces)
        tmp = [word.text for word in rev[:-1]]
        tokensList.append(tmp + [str(rev[-1])])
    return(tokensList)
```

Εικόνα 9 Tokenize στη spaCy

4.2.2 Lowercase

Στο στάδιο αυτό πραγματοποιείται lowercase για το σύνολο των δεδομένων και αφαιρούνται οι λέξεις εκείνες που έχουν μήκος μικρότερο των 2 χαρακτήρων. Η διαδικασία είναι ίδια και για τις τρεις περιπτώσεις και συνοψίζεται στο παρακάτω τμήμα του κώδικα:

```
def Lowercase(aList):
    lowerList = []
    for tweet in aList:
        tmp = []
        for word in tweet[:-1]:
            if len(word) > 2:
                tmp.append(word.lower())
        lowerList.append(tmp + [tweet[-1]])
    return(lowerList)
```

Εικόνα 10 Lowercase

4.2.3 Αφαίρεση Stop words

Σε αυτό το βήμα της διαδικασίας αφαιρούνται λέξεις οι οποίες θεωρητικά, δεν προσφέρουν πληροφορία στην απόφαση του αλγόριθμου για κατηγοριοποίηση. Στην πρώτη περίπτωση, η αφαίρεση γίνεται με χρήση της συνάρτησης stopwords του πακέτου nltk.corpus της NLTK, αφαιρώντας από τα δεδομένα το σύνολο των stop words της αγγλικής γλώσσας, όπως φαίνεται παρακάτω:

```
def RemoveStopwords(aList):
    myStopwords = set(stopwords.words('english'))
    stopwordsFreeList = []
    for tweet in aList:
        tmpList = []
        for word in tweet[:-1]:
            if word not in myStopwords:
                tmpList.append(word)
        stopwordsFreeList.append(tmpList + [tweet[-1]])
    return(stopwordsFreeList)
```

Εικόνα 11 Απαλοιφή Stop words στην NLTK

Στη δεύτερη περίπτωση όπου χρησιμοποιείται η spaCy για τα αγγλικά δεδομένα αφαιρούνται τα αγγλικά stopwords εισάγοντας το module spacy.lang.en με τον τρόπο που φαίνεται παρακάτω (Εικόνα 12).

```
def RemoveStopWords(aList):
    StopWords = spacy.lang.en.stop_words.STOP_WORDS
    stopWordsFree = []
    for tweet in aList:
        tmp = []
        for word in tweet[:-1]:
            if word not in StopWords:
                tmp.append(word)
        stopWordsFree.append(tmp + [tweet[-1]])
    return(stopWordsFree)
```

Εικόνα 12 Απαλοιφή Stop words στη spaCy

Για την τρίτη περίπτωση ο κώδικας είναι πανομοιότυπος με αυτό της Εικόνας 12, με τη διαφορά ότι αφαιρούνται οι αντίστοιχες ελληνικές λέξεις χρησιμοποιώντας το module spacy.lang.el.

4.2.4 Lemmatization

Όπως, αναλύθηκε σε προηγούμενο κεφάλαιο, στο στάδιο αυτό οι λέξεις που διαφοροποιούνται ως προς την κλίση, τα γένη κλπ. επαναφέρονται στην πιο απλή μορφή τους. Η διαδικασία για την περίπτωση της NLTK απεικονίζεται παρακάτω, όπου καλείται η κλάση WordNetLemmatizer() του πακέτου nltk.stem και δημιουργείται το αντίστοιχο αντικείμενο:

```
def Lemmatizing(aList):
    lm = WordNetLemmatizer()
    lemmatizedList = []
    for tweet in aList:
            tmpList = []
            for word in tweet[:-1]:
                tmpList.append(lm.lemmatize(word))
            lemmatizedList.append(tmpList + [tweet[-1]])
    return(lemmatizedList)
```

Εικόνα 13 Lemmatization στην NLTK

Για τις περιπτώσεις των αγγλικών και ελληνικών δεδομένων, ο παρακάτω κώδικας απεικονίζει τη διαδικασία, χρησιμοποιώντας βέβαια, το αντίστοιχο γλωσσικό μοντέλο:

```
def Lemmatize(aList):
    lemmaList = []
    for tweet in aList:
        rev = nlp(tweet)
        tmp = [word.lemma_ for word in rev[:-1]]
        lemmaList.append(tmp + [str(rev[-1])])
    return(lemmaList)
```

Εικόνα 14 Lemmatization στην spaCy

4.2.5 Part-of-speech tagging

Ακολουθεί η διαδικασία για απόδοση ετικέτας με τα μέρη του λόγου. Για την περίπτωση της NLTK, η απόδοση ετικέτας γίνεται χρησιμοποιώντας τον tagger nltk.pos_tag, όπως φαίνεται στην Εικόνα 15.

```
def POSTagging(aList):
    taggedList = []
    for tweet in aList:
        tagged = nltk.pos_tag(tweet[:-1])
        new_tagged = []
        for word in tagged:
            new_tagged.append(word[0] + '_' + word[1])
        taggedList.append(new_tagged + [tweet[-1]])
    return(taggedList)
```

Εικόνα 15 POS tagging στην NLTK

Για τις περιπτώσεις της spaCy, ο κώδικας είναι και πάλι κοινός, χρησιμοποιώντας για την πρώτη το αγγλικό και για τη δεύτερη το ελληνικό μοντέλο.

```
def PosTag(aList):
    posList = []
    nlp = spacy.load('en_core_web_sm')
    for tweet in aList:
        rev = nlp(tweet)
        tmp = [(word.text, word.pos_) for word in rev[:-1]]
        #merge word and pos in a single string
        temp = [(word[0] + '_' + word[1]) for word in tmp]
        posList.append(temp + [str(rev[-1])])
    return(posList)
```

Εικόνα 16 POS tagging στην spaCy

4.3 Εξαγωγή χαρακτηριστικών

Μετά το pre-processing, ακολουθεί η εξίσου σημαντική διαδικασία της εξαγωγής χαρακτηριστικών, από την οποία εξάγονται οι λέξεις που θεωρείται ότι συμβάλλουν στη διαδικασία κατηγοριοποίησης. Στη μεθοδολογία μας χρησιμοποιούνται δύο κλασικές και δημοφιλείς τεχνικές, η Count Vectorizer και η TF-IDF. Συνοπτικά, η πρώτη βασίζεται στη λογική του term frequency, ενώ η δεύτερη ισοσκελίζει το θόρυβο που ενδεχομένως, προκαλούν οι όροι που εμφανίζονται πολύ συχνά στο σύνολο των δεδομένων. Στη συνέχεια, χρησιμοποιείται 5-fold cross-validation, προκειμένου να έχουμε όσο το δυνατόν πιο αξιόπιστα αποτελέσματα, αξιοποιώντας το σύνολο των δεδομένων, τόσο στη διαδικασία της εκπαίδευσης (training), όσο και στη διαδικασία της δοκιμής (testing). Η διαδικασία και ο κώδικας είναι κοινός και για τις τρεις περιπτώσεις.

4.3.1 Count Vectorizer

Συγκεκριμένα, για την Count Vectorizer αρχικά δημιουργείται το αντικείμενο της Count Vectorizer και στη συνέχεια, δημιουργεί το σύνολο των χαρακτηριστικών με βάση το dataset, μετατρέπει τα δεδομένα και επιστρέφει ένα πίνακα με τις μετρήσεις των χαρακτηριστικών στο σύνολο των δεδομένων (Εικόνα 17).

```
def CountVect(aList):
    cv = CountVectorizer(lowercase=False)
    X = cv.fit_transform(aList).toarray()
    # Sort, inverse and get indices of idf weights from more to less important
    X_sum = X.sum(axis=0)
    indices = np.argsort(X_sum)[::-1]
    # Get features
    feature_names = cv.get_feature_names()
    # Number of features
    top_n = 100
    # Save to top_features top_n features
    top_features = [feature_names[i] for i in indices[:top_n]]
    return(top_features, X)
```

Εικόνα 17 Count Vectorizer

4.3.2 Tf-Idf

Παρόμοια, δημιουργείται το αντικείμενο της TfIdf Vectorizer και έπειτα, δημιουργείται το σύνολο των χαρακτηριστικών με βάση το dataset και επιστρέφει ένα πίνακα με τα βάρη των χαρακτηριστικών στο σύνολο των δεδομένων (Εικόνα 18).

```
def TfIdf(aList):
    tf = TfidfVectorizer(lowercase=False)
    X = tf.fit_transform(aList).toarray()
    # Sort, inverse and get indices of idf weights from more to less important
    indices = np.argsort(tf.idf_)[::-1]
    # Get features
    feature_names = tf.get_feature_names()
    # Number of features
    top_n = 100
    # Save to list top_n features
    top_features = [feature_names[i] for i in indices[:top_n]]
    return(top_features, X)
```

Εικόνα 18 Tf-Idf

4.3.3 5-fold Cross-validation

Για τη διαδικασία k-fold cross-validation όπως φαίνεται στην Εικόνα 19, χρησιμοποιούνται 5-folds. Ανακατεύουμε τα δεδομένα και στη συνέχεια, το σύνολο των δεδομένων χωρίζεται σε πέντε διαδοχικά ίσα μέρη, τέσσερα από τα οποία χρησιμοποιούνται για την εκπαίδευση των δεδομένων και ένα για τη δοκιμή. Η διαδικασία επαναλαμβάνεται πέντε φορές και με αυτό τον τρόπο όλα τα μέρη συμμετέχουν τόσο στην εκπαίδευση όσο και στη δοκιμή.

```
def TrainTestSet(aList, X):
    Lbls = np.array(aList)
    kfold = KFold(n_splits=5, shuffle=True, random_state=1)
    for train_set,test_set in kfold.split(X):
        X_train, X_test = X[train_set], X[test_set]
        y_train, y_test = Lbls[train_set], Lbls[test_set]
    return(kfold, X_train, X_test, y_train, y_test)
```

Εικόνα 19 5-fold Cross-validation

4.4 Ταξινομητές

Στο στάδιο αυτό χρησιμοποιούνται οι αλγόριθμοι ταξινόμησης προκειμένου να γίνει η εκπαίδευση των δεδομένων και η δημιουργία του μοντέλου πρόβλεψης.

4.4.1 Random Forest

Παρακάτω παρουσιάζεται ο κώδικας για τον ταξινομητή Random Forest (Εικόνα 20). Συγκεκριμένα, καλείται ο αλγόριθμος, ο οποίος χρησιμοποιεί 200 δέντρα απόφασης και στη συνέχεια, γίνεται η εκπαίδευση του μοντέλου με βάση τα tweets και τις αντίστοιχες ετικέτες τους. Τέλος, γίνεται η πρόβλεψη για τα δεδομένα δοκιμής από το μοντέλο που εκπαιδεύσαμε.

```
def RFClassifier(X_train, y_train, X_test):
    RF_classifier = RandomForestClassifier(n_estimators=200, random_state=0)
    # Train the algorithm
    RF_classifier.fit(X_train, y_train)
    # Predict sentiment
    RF_predictions = RF_classifier.predict(X_test)
    return(RF_classifier, RF_predictions)
```

Εικόνα 20 Εκπαίδευση και πρόβλεψη με Random Forest

4.4.2 Multinomial Naive Bayes

Παρόμοια είναι και η διαδικασία που ακολουθείται για τον Naive Bayes, όπου χρησιμοποιείται ο αλγόριθμος με την παράμετρο εξομάλυνσης alpha με τιμή 0.1, γίνεται η εκπαίδευση και τέλος, η πρόβλεψη των δεδομένων (Εικόνα 21).

```
def MNBClassifier(X_train, y_train, X_test):
    MNB_classifier = MultinomialNB(alpha=0.1)
    # Train the algorithm
    MNB_classifier.fit(X_train, y_train)
    # Predict sentiment
    MNB_predictions = MNB_classifier.predict(X_test)
    return(MNB_classifier, MNB_predictions)
```

Εικόνα 21 Εκπαίδευση και πρόβλεψη με Multinomial Naive Bayes

5 Αποτελέσματα

Στο σύνολο της μεθοδολογίας χρησιμοποιείται 5-fold cross-validation. Αυτό σημαίνει ότι παράγονται για κάθε πείραμα πέντε confusion matrices και πέντε classification reports. Για λόγους πρακτικότητας και καλύτερης παρουσίασης των αποτελεσμάτων, υπολογίζεται ένας μέσος πίνακας σύγχυσης με βάση τους πέντε πίνακες που εξάγονται για κάθε πείραμα. Από το μέσο confusion matrix, υπολογίζεται στη συνέχεια το classification report. Τα αποτελέσματα παρουσιάζονται παρακάτω, σε τρεις ενότητες: 5.1 NLTK με μηχανική μετάφραση, 5.2 SpaCy με μηχανική μετάφραση και 5.3 SpaCy με ελληνικά. Κάθε μία από τις ενότητες περιέχει οκτώ πειράματα, πάνω στα οποία τελικά γίνεται η τελική σύγκριση.

5.1 NLTK με μηχανική μετάφραση

Παρακάτω, απεικονίζονται τα αποτελέσματα για οκτώ πειράματα, χρησιμοποιώντας την πρώτη μέθοδο (NLTK με μηχανική μετάφραση). Για κάθε ένα από τα πειράματα παρουσιάζονται δύο πίνακες με τα αποτελέσματα. Ο πρώτος πίνακας αποτελεί το classification report, ενώ ο δεύτερος περιέχει τις τιμές του accuracy για κάθε ένα από τα πέντε folds, τη μέση τιμή του accuracy, την τυπική απόκλιση και τον αριθμό των χαρακτηριστικών.

5.1.1 TF-IDF, RF και απαλοιφή StopWords

Στους παρακάτω πίνακες, στην αρνητική κλάση παρατηρούμε σχετικά χαμηλό precision και υψηλό recall. Αυτό σημαίνει ότι 77% των αρνητικών προβλέψεων ήταν σωστές, ενώ το 88% της αρνητικής κλάσης κατηγοριοποιήθηκε σωστά. Για τη θετική κλάση, 87% των προβλέψεων ήταν σωστές, ενώ 75% της κλάσης κατηγοριοποιήθηκε σωστά. Η τυπική απόκλιση μένει σε χαμηλά ποσοστά (2%), που σημαίνει ότι δεν υπάρχουν μεγάλες διακυμάνσεις ανά fold και άρα, το μοντέλο μας μπορεί να θεωρηθεί ότι παράγει αξιόπιστα αποτελέσματα.

	Precision	**Recall**	**F1-score**	**Support**
Negative	0.77	0.88	0.82	181
Positive	0.87	0.75	0.81	195
Avg/Total	0.82	0.82	0.81	376

Πίνακας 4 Classification Report για TF-IDF, RF και απαλοιφή Stop words

	k-fold (k=1)	**k-fold (k=2)**	**k-fold (k=3)**	**k-fold (k=4)**	**k-fold (k=5)**
Accuracy	0.84	0.81	0.82	0.81	0.79
Summary	**5-fold mean accuracy**: 0.81 **Standard deviation**: 0.02 **Number of features**: 4698				

Πίνακας 5 Accuracies, summary για TF-IDF, RF και απαλοιφή Stop words

5.1.2 TF-IDF, MNB και απαλοιφή StopWords

Στους παρακάτω πίνακες, 85% των αρνητικών προβλέψεων ήταν σωστές, ενώ από 79% της αρνητικής κλάσης κατηγοριοποιήθηκε σωστά. Αντίστοιχα, το 82% των θετικών προβλέψεων ήταν σωστές, ενώ το 87% της θετικής κλάσης κατηγοριοποιήθηκε σωστά. Η τυπική απόκλιση κυμαίνεται σε ποσοστό μόλις 1%.

	Precision	**Recall**	**F1-score**	**Support**
Negative	0.85	0.79	0.82	180
Positive	0.82	0.87	0.84	195
Avg/Total	0.83	0.83	0.83	375

Πίνακας 6 Classification Report για TF-IDF, MNB και απαλοιφή Stop words

	k-fold (k=1)	**k-fold (k=2)**	**k-fold (k=3)**	**k-fold (k=4)**	**k-fold (k=5)**
Accuracy	0.84	0.85	0.83	0.83	0.81
Summary	**5-fold mean accuracy**: 0.83 **Standard deviation**: 0.01 **Number of features**: 4698				

Πίνακας 7 Accuracies, summary για TF-IDF, MNB και απαλοιφή Stop words

5.1.3 Count Vectorizer, RF και απαλοιφή StopWords

Στους παρακάτω πίνακες, 80% των αρνητικών προβλέψεων ήταν σωστές, ενώ το 84% της αρνητικής κλάσης κατηγοριοποιήθηκε σωστά. Αντίστοιχα, το 84% των θετικών προβλέψεων ήταν σωστές, ενώ 81% της θετικής κλάσης κατηγοριοποιήθηκε σωστά. Η τυπική απόκλιση κυμαίνεται σε ποσοστό μόλις 1%.

	Precision	Recall	F1-score	Support
Negative	0.80	0.84	0.82	180
Positive	0.84	0.81	0.83	194
Avg/Total	0.82	0.82	0.82	374

Πίνακας 8 Classification Report για CV, RF και απαλοιφή Stop words

	k-fold (k=1)	k-fold (k=2)	k-fold (k=3)	k-fold (k=4)	k-fold (k=5)
Accuracy	0.83	0.83	0.83	0.82	0.80
Summary	**5-fold mean accuracy**: 0.82 **Standard deviation**: 0.01 **Number of features**: 4698				

Πίνακας 9 Accuracies, summary για CV, RF και απαλοιφή Stop words

5.1.4 Count Vectorizer, MNB και απαλοιφή StopWords

Στους παρακάτω πίνακες, το 84% των αρνητικών προβλέψεων ήταν σωστές, ενώ το 82% της αρνητικής κλάσης κατηγοριοποιήθηκε σωστά. Αντίστοιχα, το 84% των θετικών προβλέψεων ήταν σωστές, ενώ το 86% της θετικής κλάσης κατηγοριοποιήθηκε σωστά. Η τυπική απόκλιση κυμαίνεται σε ποσοστό μόλις 1%.

	Precision	Recall	F1-score	Support
Negative	0.84	0.82	0.83	180
Positive	0.84	0.86	0.85	195
Avg/Total	0.84	0.84	0.84	375

Πίνακας 10 Classification Report για CV, MNB και απαλοιφή Stop words

	k-fold (k=1)	k-fold (k=2)	k-fold (k=3)	k-fold (k=4)	k-fold (k=5)
Accuracy	0.85	0.85	0.84	0.84	0.81
Summary	**5-fold mean accuracy**: 0.84 **Standard deviation**: 0.01 **Number of features**: 4698				

Πίνακας 11 Accuracies, summary για CV, MNB και απαλοιφή Stop words

5.1.5 TF-IDF, RF και μη απαλοιφή StopWords

Στους παρακάτω πίνακες, το 81% των αρνητικών προβλέψεων ήταν σωστές, ενώ το 80% της αρνητικής κλάσης κατηγοριοποιήθηκε σωστά. Αντίστοιχα, το 81% των θετικών προβλέψεων ήταν σωστές, ενώ το 82% της θετικής κλάσης κατηγοριοποιήθηκε σωστά. Η τυπική απόκλιση κυμαίνεται σε ποσοστό 3%.

	Precision	**Recall**	**F1-score**	**Support**
Negative	0.81	0.80	0.80	181
Positive	0.81	0.82	0.82	194
Avg/Total	0.81	0.81	0.81	375

Πίνακας 12 Classification Report για TF-IDF, RF και μη απαλοιφή Stop words

	k-fold (k=1)	**k-fold (k=2)**	**k-fold (k=3)**	**k-fold (k=4)**	**k-fold (k=5)**
Accuracy	0.84	0.78	0.84	0.78	0.81
Summary	**5-fold mean accuracy**: 0.81 **Standard deviation**: 0.03 **Number of features**: 4889				

Πίνακας 13 Accuracies, summary για TF-IDF, RF και μη απαλοιφή Stop words

5.1.6 TF-IDF, MNB και μη απαλοιφή StopWords

Στους παρακάτω πίνακες, το 84% των αρνητικών προβλέψεων ήταν σωστές, ενώ το 82% της αρνητικής κλάσης κατηγοριοποιήθηκε σωστά. Αντίστοιχα, το 84% των θετικών προβλέψεων ήταν σωστές, ενώ το 86% της θετικής κλάσης κατηγοριοποιήθηκε σωστά. Η τυπική απόκλιση κυμαίνεται σε ποσοστό μόλις 1%.

	Precision	**Recall**	**F1-score**	**Support**
Negative	0.84	0.82	0.83	180
Positive	0.84	0.86	0.85	195
Avg/Total	0.84	0.84	0.84	375

Πίνακας 14 Classification Report για TF-IDF, MNB και μη απαλοιφή Stop words

	k-fold (k=1)	**k-fold (k=2)**	**k-fold (k=3)**	**k-fold (k=4)**	**k-fold (k=5)**
Accuracy	0.83	0.85	0.85	0.84	0.82
Summary	**5-fold mean accuracy**: 0.84				

	Standard deviation: 0.01 **Number of features**: 4889

Πίνακας 15 Accuracies, summary για TF-IDF, MNB και μη απαλοιφή Stop words

5.1.7 Count Vectorizer, RF και μη απαλοιφή StopWords

Στους παρακάτω πίνακες, το 85% των αρνητικών προβλέψεων ήταν σωστές, ενώ το 73% της αρνητικής κλάσης κατηγοριοποιήθηκε σωστά. Αντίστοιχα, το 78% των θετικών προβλέψεων ήταν σωστές, ενώ το 88% της θετικής κλάσης κατηγοριοποιήθηκε σωστά. Η τυπική απόκλιση κυμαίνεται σε ποσοστό 2%.

	Precision	**Recall**	**F1-score**	**Support**
Negative	0.85	0.73	0.79	181
Positive	0.78	0.88	0.83	195
Avg/Total	0.81	0.81	0.81	376

Πίνακας 16 Classification Report για CV, RF και μη απαλοιφή Stop words

	k-fold (k=1)	**k-fold (k=2)**	**k-fold (k=3)**	**k-fold (k=4)**	**k-fold (k=5)**
Accuracy	0.83	0.78	0.84	0.79	0.81
Summary	**5-fold mean accuracy**: 0.81 **Standard deviation**: 0.02 **Number of features**: 4889				

Πίνακας 17 Accuracies, summary για CV, RF, και μη απαλοιφή Stop words

5.1.8 Count Vectorizer, MNB και μη απαλοιφή StopWords

Στους παρακάτω πίνακες, το 84% των αρνητικών προβλέψεων ήταν σωστές, ενώ το 83% της αρνητικής κλάσης κατηγοριοποιήθηκε σωστά. Αντίστοιχα, το 84% των θετικών προβλέψεων ήταν σωστές, ενώ το 85% της θετικής κλάσης κατηγοριοποιήθηκε σωστά. Η τυπική απόκλιση κυμαίνεται σε ποσοστό μόλις 1%.

	Precision	**Recall**	**F1-score**	**Support**
Negative	0.84	0.83	0.83	180
Positive	0.84	0.85	0.85	195
Avg/Total	0.84	0.84	0.84	375

Πίνακας 18 Classification Report για CV, MNB και μη απαλοιφή Stop words

	k-fold (k=1)	**k-fold (k=2)**	**k-fold (k=3)**	**k-fold (k=4)**	**k-fold (k=5)**
Accuracy	0.84	0.85	0.85	0.84	0.82
Summary	**5-fold mean accuracy**: 0.84 **Standard deviation**: 0.01 **Number of features**: 4889				

Πίνακας 19 Accuracies, summary για CV, MNB και μη απαλοιφή Stop words

5.2 SpaCy με μηχανική μετάφραση

Παρόμοια με την πρώτη ενότητα, στη δεύτερη απεικονίζονται τα αποτελέσματα για τα ίδια οκτώ πειράματα, χρησιμοποιώντας τη δεύτερη μέθοδο αυτή τη φορά (spaCy με μηχανική μετάφραση). Για κάθε ένα από τα πειράματα παρουσιάζονται δύο πίνακες με τα αποτελέσματα. Ο πρώτος πίνακας αποτελεί το classification report, ενώ ο δεύτερος περιέχει τις τιμές του accuracy για κάθε ένα από τα πέντε folds, τη μέση τιμή του accuracy, την τυπική απόκλιση και τον αριθμό των χαρακτηριστικών.

5.2.1 TF-IDF, RF και απαλοιφή StopWords

Στους παρακάτω πίνακες, το 85% των αρνητικών προβλέψεων ήταν σωστές, ενώ το 74% της αρνητικής κλάσης κατηγοριοποιήθηκε σωστά. Αντίστοιχα, το 78% των θετικών προβλέψεων ήταν σωστές, ενώ το 88% της θετικής κλάσης κατηγοριοποιήθηκε σωστά. Η τυπική απόκλιση κυμαίνεται σε ποσοστό 2%.

	Precision	**Recall**	**F1-score**	**Support**
Negative	0.85	0.74	0.79	181
Positive	0.78	0.88	0.83	194
Avg/Total	0.82	0.81	0.81	375

Πίνακας 20 Classification Report για TF-IDF, RF και απαλοιφή Stop words

	k-fold (k=1)	**k-fold (k=2)**	**k-fold (k=3)**	**k-fold (k=4)**	**k-fold (k=5)**
Accuracy	0.80	0.82	0.84	0.78	0.81
Summary	**5-fold mean accuracy**: 0.81				

	Standard deviation: 0.02 **Number of features**: 3993

Πίνακας 21 Accuracies, summary για TF-IDF, RF και απαλοιφή Stop words

5.2.2 TF-IDF, MNB και απαλοιφή StopWords

Στους παρακάτω πίνακες, το 83% των αρνητικών προβλέψεων ήταν σωστές, ενώ το 79% της αρνητικής κλάσης κατηγοριοποιήθηκε σωστά. Αντίστοιχα, το 82% των θετικών προβλέψεων ήταν σωστές, ενώ το 85% της θετικής κλάσης κατηγοριοποιήθηκε σωστά. Η τυπική απόκλιση κυμαίνεται σε ποσοστό μόλις 1%.

	Precision	**Recall**	**F1-score**	**Support**
Negative	0.83	0.79	0.81	180
Positive	0.82	0.85	0.83	195
Avg/Total	0.82	0.82	0.82	375

Πίνακας 22 Classification Report για TF-IDF, MNB και απαλοιφή Stop words

	k-fold (k=1)	**k-fold (k=2)**	**k-fold (k=3)**	**k-fold (k=4)**	**k-fold (k=5)**
Accuracy	0.82	0.84	0.83	0.82	0.81
Summary	**5-fold mean accuracy**: 0.83 **Standard deviation**: 0.01 **Number of features**: 3993				

Πίνακας 23 Accuracies, summary για TF-IDF, MNB και απαλοιφή Stop words

5.2.3 Count Vectorizer, RF και απαλοιφή StopWords

Στους παρακάτω πίνακες, το 85% των αρνητικών προβλέψεων ήταν σωστές, ενώ το 71% της αρνητικής κλάσης κατηγοριοποιήθηκε σωστά. Αντίστοιχα, το 77% των θετικών προβλέψεων ήταν σωστές, ενώ το 89% της θετικής κλάσης κατηγοριοποιήθηκε σωστά. Η τυπική απόκλιση κυμαίνεται σε ποσοστό 3%.

	Precision	**Recall**	**F1-score**	**Support**
Negative	0.85	0.71	0.77	181
Positive	0.77	0.89	0.82	195
Avg/Total	0.81	0.80	0.80	376

Πίνακας 24 Classification Report για CV, RF και απαλοιφή Stop words

	k-fold (k=1)	**k-fold (k=2)**	**k-fold (k=3)**	**k-fold (k=4)**	**k-fold (k=5)**
Accuracy	0.79	0.84	0.82	0.75	0.81
Summary	**5-fold mean accuracy**: 0.80 **Standard deviation**: 0.03 **Number of features**: 3993				

Πίνακας 25 Accuracies, summary για CV, RF και απαλοιφή Stop words

5.2.4 Count Vectorizer, MNB και απαλοιφή StopWords

Στους παρακάτω πίνακες, το 83% των αρνητικών προβλέψεων ήταν σωστές, ενώ το 81% της αρνητικής κλάσης κατηγοριοποιήθηκε σωστά. Αντίστοιχα, το 83% των θετικών προβλέψεων ήταν σωστές, ενώ το 85% της θετικής κλάσης κατηγοριοποιήθηκε σωστά. Η τυπική απόκλιση κυμαίνεται σε ποσοστό μόλις 1%.

	Precision	**Recall**	**F1-score**	**Support**
Negative	0.83	0.81	0.82	181
Positive	0.83	0.85	0.84	194
Avg/Total	0.83	0.83	0.83	375

Πίνακας 26 Classification Report για CV, MNB και απαλοιφή Stop words

	k-fold (k=1)	**k-fold (k=2)**	**k-fold (k=3)**	**k-fold (k=4)**	**k-fold (k=5)**
Accuracy	0.84	0.83	0.83	0.82	0.82
Summary	**5-fold mean accuracy**: 0.83 **Standard deviation**: 0.01 **Number of features**: 3993				

Πίνακας 27 Accuracies, summary για CV, MNB και απαλοιφή Stop words

5.2.5 TF-IDF, RF και μη απαλοιφή StopWords

Στους παρακάτω πίνακες, το 82% των αρνητικών προβλέψεων ήταν σωστές, ενώ το 78% της αρνητικής κλάσης κατηγοριοποιήθηκε σωστά. Αντίστοιχα, το 81% των θετικών προβλέψεων ήταν σωστές, ενώ το 84% της θετικής κλάσης κατηγοριοποιήθηκε σωστά. Η τυπική απόκλιση κυμαίνεται σε ποσοστό μόλις 1%.

	Precision	Recall	F1-score	Support
Negative	0.82	0.78	0.80	180
Positive	0.81	0.84	0.82	194
Avg/Total	0.81	0.81	0.81	374

Πίνακας 28 Classification Report για TF-IDF, RF και μη απαλοιφή Stop words

	k-fold (k=1)	k-fold (k=2)	k-fold (k=3)	k-fold (k=4)	k-fold (k=5)
Accuracy	0.81	0.79	0.84	0.81	0.81
Summary	**5-fold mean accuracy**: 0.81 **Standard deviation**: 0.01 **Number of features**: 4262				

Πίνακας 29 Accuracies, summary για TF-IDF, RF και μη απαλοιφή Stop words

5.2.6 TF-IDF, MNB και μη απαλοιφή StopWords

Στους παρακάτω πίνακες, το 84% των αρνητικών προβλέψεων ήταν σωστές, ενώ το 82% της αρνητικής κλάσης κατηγοριοποιήθηκε σωστά. Αντίστοιχα, το 83% των θετικών προβλέψεων ήταν σωστές, το 85% της θετικής κλάσης κατηγοριοποιήθηκε σωστά. Η τυπική απόκλιση κυμαίνεται σε ποσοστό μόλις 2%.

	Precision	Recall	F1-score	Support
Negative	0.84	0.82	0.83	180
Positive	0.83	0.85	0.84	195
Avg/Total	0.83	0.83	0.83	375

Πίνακας 30 Classification Report για TF-IDF, MNB και μη απαλοιφή Stop words

	k-fold (k=1)	k-fold (k=2)	k-fold (k=3)	k-fold (k=4)	k-fold (k=5)
Accuracy	0.84	0.81	0.87	0.84	0.82
Summary	**5-fold mean accuracy**: 0.83 **Standard deviation**: 0.02 **Number of features**: 4262				

Πίνακας 31 Accuracies, summary για TF-IDF, MNB και μη απαλοιφή Stop words

5.2.7 Count Vectorizer, RF και μη απαλοιφή StopWords

Στους παρακάτω πίνακες, το 84% των αρνητικών προβλέψεων ήταν σωστές, ενώ το 77% της αρνητικής κλάσης κατηγοριοποιήθηκε σωστά. Αντίστοιχα, το 80% των θετικών προβλέψεων ήταν σωστές, ενώ το 87% της θετικής κλάσης κατηγοριοποιήθηκε σωστά. Η τυπική απόκλιση κυμαίνεται σε ποσοστό 2%.

	Precision	**Recall**	**F1-score**	**Support**
Negative	0.84	0.77	0.80	181
Positive	0.80	0.87	0.83	195
Avg/Total	0.82	0.82	0.82	376

Πίνακας 32 Classification Report για CV, RF και μη απαλοιφή Stop words

	k-fold (k=1)	**k-fold (k=2)**	**k-fold (k=3)**	**k-fold (k=4)**	**k-fold (k=5)**
Accuracy	0.82	0.80	0.85	0.81	0.83
Summary	**5-fold mean accuracy**: 0.82 **Standard deviation**: 0.02 **Number of features**: 4262				

Πίνακας 33 Accuracies, summary για CV, RF και μη απαλοιφή Stop words

5.2.8 Count Vectorizer, MNB και μη απαλοιφή StopWords

Στους παρακάτω πίνακες, το 83% των αρνητικών προβλέψεων ήταν σωστές, ενώ το 82% της αρνητικής κλάσης κατηγοριοποιήθηκε σωστά. Αντίστοιχα, το 84% των θετικών προβλέψεων ήταν σωστές, ενώ το 85% της θετικής κλάσης κατηγοριοποιήθηκε σωστά. Η τυπική απόκλιση κυμαίνεται σε ποσοστό 2%.

	Precision	**Recall**	**F1-score**	**Support**
Negative	0.83	0.82	0.83	180
Positive	0.84	0.85	0.84	195
Avg/Total	0.83	0.83	0.83	375

Πίνακας 34 Classification Report για CV, MNB και μη απαλοιφή Stop words

	k-fold (k=1)	**k-fold (k=2)**	**k-fold (k=3)**	**k-fold (k=4)**	**k-fold (k=5)**
Accuracy	0.84	0.82	0.87	0.83	0.82
Summary	**5-fold mean accuracy**: 0.83				

	Standard deviation: 0.02 **Number of features**: 4262

Πίνακας 35 Accuracies, summary για CV, MNB και μη απαλοιφή Stop words

5.3 SpaCy για ελληνικά tweets

Στην ενότητα αυτή, παρόμοια με τις προηγούμενες δύο, απεικονίζονται τα αποτελέσματα για τα οκτώ πειράματα, χρησιμοποιώντας την τρίτη μέθοδο (spaCy με ελληνικά). Για κάθε ένα από τα πειράματα παρουσιάζονται δύο πίνακες με τα αποτελέσματα. Ο πρώτος πίνακας αποτελεί το classification report, ενώ ο δεύτερος περιέχει τις τιμές του accuracy για κάθε ένα από τα πέντε folds, τη μέση τιμή του accuracy, την τυπική απόκλιση και τον αριθμό των χαρακτηριστικών.

5.3.1 TF-IDF, RF και απαλοιφή StopWords

Στους παρακάτω πίνακες, το 79% των αρνητικών προβλέψεων ήταν σωστές, ενώ το 84% της αρνητικής κλάσης κατηγοριοποιήθηκε σωστά. Αντίστοιχα, το 85% των θετικών προβλέψεων ήταν σωστές, ενώ το 79% της θετικής κλάσης κατηγοριοποιήθηκε σωστά. Η τυπική απόκλιση κυμαίνεται σε ποσοστό 3%.

	Precision	**Recall**	**F1-score**	**Support**
Negative	0.79	0.84	0.82	180
Positive	0.85	0.79	0.82	195
Avg/Total	0.82	0.82	0.82	375

Πίνακας 36 Classification Report για TF-IDF, RF και απαλοιφή Stop words

	k-fold (k=1)	**k-fold (k=2)**	**k-fold (k=3)**	**k-fold (k=4)**	**k-fold (k=5)**
Accuracy	0.84	0.81	0.85	0.81	0.77
Summary	**5-fold mean accuracy**: 0.82 **Standard deviation**: 0.03 **Number of features**: 6732				

Πίνακας 37 Accuracies, summary για TF-IDF, RF και απαλοιφή Stop words

5.3.2 TF-IDF, MNB και απαλοιφή StopWords

Στους παρακάτω πίνακες, το 84% των αρνητικών προβλέψεων ήταν σωστές, ενώ το 81% της αρνητικής κλάσης κατηγοριοποιήθηκε σωστά. Αντίστοιχα, το 83% των θετικών προβλέψεων ήταν σωστές, ενώ το 86% της θετικής κλάσης κατηγοριοποιήθηκε σωστά. Η τυπική απόκλιση κυμαίνεται σε ποσοστό 2%.

	Precision	**Recall**	**F1-score**	**Support**
Negative	0.84	0.81	0.82	180
Positive	0.83	0.86	0.84	194
Avg/Total	0.83	0.83	0.83	374

Πίνακας 38 Classification Report για TF-IDF, MNB και απαλοιφή Stop words

	k-fold (k=1)	**k-fold (k=2)**	**k-fold (k=3)**	**k-fold (k=4)**	**k-fold (k=5)**
Accuracy	0.84	0.80	0.86	0.83	0.83
Summary	**5-fold mean accuracy**: 0.83 **Standard deviation**: 0.02 **Number of features**: 6732				

Πίνακας 39 Accuracies, summary για TF-IDF, MNB και απαλοιφή Stop words

5.3.3 Count Vectorizer, RF και απαλοιφή StopWords

Στους παρακάτω πίνακες, το 81% των αρνητικών προβλέψεων ήταν σωστές, ενώ το 78% της αρνητικής κλάσης κατηγοριοποιήθηκε σωστά. Αντίστοιχα, το 80% των θετικών προβλέψεων ήταν σωστές, ενώ το 84% της θετικής κλάσης κατηγοριοποιήθηκε σωστά. Η τυπική απόκλιση κυμαίνεται σε ποσοστό 3%.

	Precision	**Recall**	**F1-score**	**Support**
Negative	0.81	0.78	0.80	180
Positive	0.80	0.84	0.82	195
Avg/Total	0.81	0.81	0.81	375

Πίνακας 40 Classification Report για CV, RF και απαλοιφή Stop words

	k-fold (k=1)	**k-fold (k=2)**	**k-fold (k=3)**	**k-fold (k=4)**	**k-fold (k=5)**
Accuracy	0.86	0.77	0.83	0.79	0.78
Summary	**5-fold mean accuracy**: 0.81				

	Standard deviation: 0.03 **Number of features**: 6732

Πίνακας 41 Accuracies, summary για CV, RF και απαλοιφή Stop words

5.3.4 Count Vectorizer, MNB και απαλοιφή StopWords

Στους παρακάτω πίνακες, το 84% των αρνητικών προβλέψεων ήταν σωστές, ενώ το 81% της αρνητικής κλάσης κατηγοριοποιήθηκε σωστά. Αντίστοιχα, το 83% των θετικών προβλέψεων ήταν σωστές, ενώ το 86% της θετικής κλάσης κατηγοριοποιήθηκε σωστά. Η τυπική απόκλιση κυμαίνεται σε ποσοστό 2%.

	Precision	**Recall**	**F1-score**	**Support**
Negative	0.84	0.81	0.83	181
Positive	0.83	0.86	0.85	195
Avg/Total	0.84	0.84	0.84	376

Πίνακας 42 Classification Report για CV, MNB και απαλοιφή Stop words

	k-fold (k=1)	**k-fold (k=2)**	**k-fold (k=3)**	**k-fold (k=4)**	**k-fold (k=5)**
Accuracy	0.85	0.81	0.86	0.84	0.84
Summary	**5-fold mean accuracy**: 0.84 **Standard deviation**: 0.02 **Number of features**: 6732				

Πίνακας 43 Accuracies, summary για CV, MNB και απαλοιφή Stop words

5.3.5 TF-IDF, RF και μη απαλοιφή StopWords

Στους παρακάτω πίνακες, το 79% των αρνητικών προβλέψεων ήταν σωστές, ενώ το 77% της αρνητικής κλάσης κατηγοριοποιήθηκε σωστά. Αντίστοιχα, το 80% των θετικών προβλέψεων ήταν σωστές, ενώ το 82% της θετικής κλάσης κατηγοριοποιήθηκε σωστά. Η τυπική απόκλιση κυμαίνεται σε ποσοστό μόλις 1%.

	Precision	**Recall**	**F1-score**	**Support**
Negative	0.79	0.77	0.78	180
Positive	0.80	0.82	0.81	195
Avg/Total	0.79	0.79	0.79	375

Πίνακας 44 Classification Report για TF-IDF, RF και μη απαλοιφή Stop words

	k-fold (k=1)	**k-fold (k=2)**	**k-fold (k=3)**	**k-fold (k=4)**	**k-fold (k=5)**
Accuracy	0.79	0.81	0.81	0.77	0.79
Summary	**5-fold mean accuracy**: 0.79 **Standard deviation**: 0.01 **Number of features**: 7001				

Πίνακας 45 Accuracies, summary για TF-IDF, RF και μη απαλοιφή Stop words

5.3.6 TF-IDF, MNB και μη απαλοιφή StopWords

Στους παρακάτω πίνακες, το 85% των αρνητικών προβλέψεων ήταν σωστές, ενώ το 82% της αρνητικής κλάσης κατηγοριοποιήθηκε σωστά. Αντίστοιχα, το 84% των θετικών προβλέψεων ήταν σωστές, ενώ το 86% της θετικής κλάσης κατηγοριοποιήθηκε σωστά. Η τυπική απόκλιση κυμαίνεται σε ποσοστό μόλις 1%.

	Precision	**Recall**	**F1-score**	**Support**
Negative	0.85	0.82	0.83	181
Positive	0.84	0.86	0.85	195
Avg/Total	0.84	0.84	0.84	376

Πίνακας 46 Classification Report για TF-IDF, MNB και μη απαλοιφή Stop words

	k-fold (k=1)	**k-fold (k=2)**	**k-fold (k=3)**	**k-fold (k=4)**	**k-fold (k=5)**
Accuracy	0.84	0.83	0.86	0.84	0.85
Summary	**5-fold mean accuracy**: 0.84 **Standard deviation**: 0.01 **Number of features**: 7001				

Πίνακας 47 Accuracies, summary για TF-IDF, MNB και μη απαλοιφή Stop words

5.3.7 Count Vectorizer, RF και μη απαλοιφή StopWords

Στους παρακάτω πίνακες, το 85% των αρνητικών προβλέψεων ήταν σωστές, ενώ το 70% της αρνητικής κλάσης κατηγοριοποιήθηκε σωστά. Αντίστοιχα, το 76% των θετικών προβλέψεων ήταν σωστές, ενώ το 88% της θετικής κλάσης κατηγοριοποιήθηκε σωστά. Η τυπική απόκλιση κυμαίνεται σε ποσοστό 2%.

	Precision	Recall	F1-score	Support
Negative	0.85	0.70	0.76	181
Positive	0.76	0.88	0.82	195
Avg/Total	0.80	0.79	0.79	376

Πίνακας 48 Classification Report για CV, RF και μη απαλοιφή Stop words

	k-fold (k=1)	k-fold (k=2)	k-fold (k=3)	k-fold (k=4)	k-fold (k=5)
Accuracy	0.77	0.79	0.82	0.78	0.80
Summary	**5-fold mean accuracy**: 0.79 **Standard deviation**: 0.02 **Number of features**: 7001				

Πίνακας 49 Accuracies, summary για CV, RF και μη απαλοιφή Stop words

5.3.8 Count Vectorizer, MNB και μη απαλοιφή StopWords

Στους παρακάτω πίνακες, το 84% των αρνητικών προβλέψεων ήταν σωστές, ενώ το 84% της αρνητικής κλάσης κατηγοριοποιήθηκε σωστά. Αντίστοιχα, το 85% των θετικών προβλέψεων ήταν σωστές, ενώ το 85% της θετικής κλάσης κατηγοριοποιήθηκε σωστά. Η τυπική απόκλιση κυμαίνεται σε ποσοστό μόλις 1%.

	Precision	Recall	F1-score	Support
Negative	0.84	0.84	0.84	180
Positive	0.85	0.85	0.85	194
Avg/Total	0.84	0.84	0.84	374

Πίνακας 50 Classification Report για CV, MNB και μη απαλοιφή Stop words

	k-fold (k=1)	k-fold (k=2)	k-fold (k=3)	k-fold (k=4)	k-fold (k=5)
Accuracy	0.85	0.83	0.85	0.84	0.85
Summary	**5-fold mean accuracy**: 0.84 **Standard deviation**: 0.01 **Number of features**: 7001				

Πίνακας 51 Accuracies, summary για CV, MNB και μη απαλοιφή Stop words

5.4 Συγκεντρωτικά αποτελέσματα – Διαγράμματα

5.4.1 Mean accuracy για το σύνολο των πειραμάτων

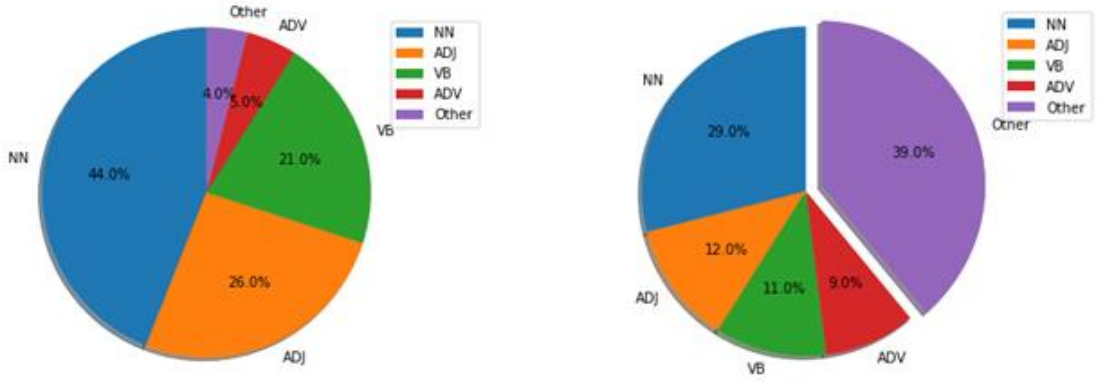
Από τον παρακάτω πίνακα αντιλαμβανόμαστε ότι υψηλότερα αποτελέσματα επιτυγχάνει η περίπτωση της μη απαλοιφής stop words και ο αλγόριθμος Multinomial Naive Bayes. Ο Random Forest συμπεριφέρεται περίπου με τον ίδιο τρόπο για τις περιπτώσεις απαλοιφής και μη απαλοιφής stop words. Σε ό,τι αφορά τις μεθόδους, παρατηρούμε ότι ελαφρώς καλύτερα αποτελέσματα επιτυγχάνει η spaCy για τα ελληνικά δεδομένα, ακολουθεί η NLTK και τέλος, η spaCy για τα μεταφρασμένα δεδομένα. Παρακάτω, θα επικεντρωθούμε στις διάφορες περιπτώσεις μη απαλοιφής stop words και θα διερευνήσουμε σε μεγαλύτερο βάθος τα χαρακτηριστικά τους.

	Algorithm	**NLTK/TF-IDF (mt)**	**NLTK/ CV (mt)**	**spaCy/TF-IDF (mt)**	**spaCy/ CV (mt)**	**spaCy/TF-IDF (gr)**	**spaCy/ CV (gr)**
Remove Stop Words	**RF**	0.8145	0.8225	0.8108	0.8012	0.815	0.807
	MNB	0.8316	0.8358	0.8252	0.8294	0.8326	0.839
No Stop Word removal	**RF**	0.8108	0.8102	0.8124	0.8204	0.7937	0.7921
	MNB	**0.8396**	**0.8401**	**0.8348**	**0.8342**	**0.8444**	**0.8443**

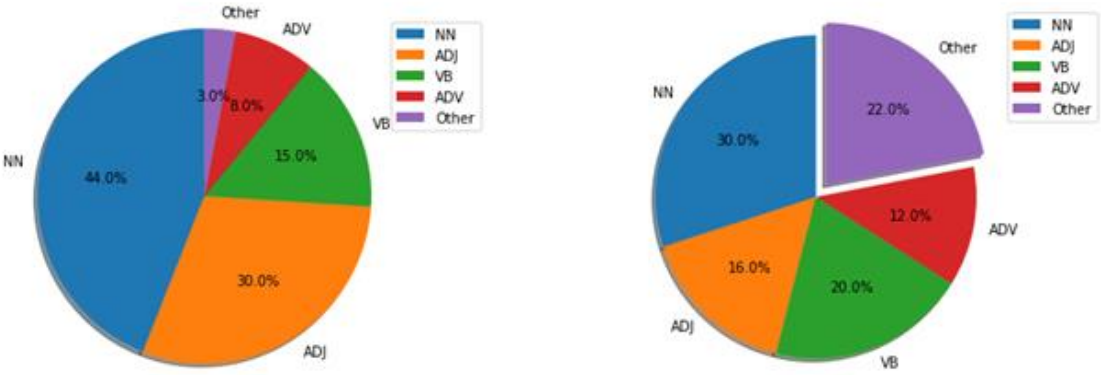
Πίνακας 52 Mean accuracy για NLTK (mt), spaCy (mt) και spaCy (gr)

5.4.2 Κατανομή των POS tags στα top 100 features

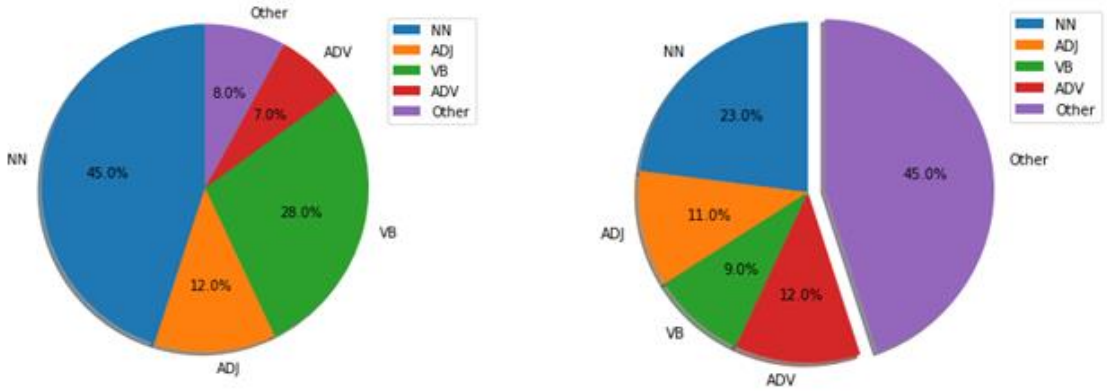
Παρακάτω, απεικονίζεται η ποσοστιαία συμμετοχή των POS tags στα εκατό πιο σημαντικά χαρακτηριστικά για τα υψηλότερα accuracies ανά περίπτωση χωρίς απαλοιφή των stop words. Συγκεκριμένα, εξετάζονται τα πιο βασικά μέρη του λόγου, δηλαδή, NN: ουσιαστικά, ADJ: επίθετα, VB: ρήματα και ADV: επιρρήματα. Όπως παρατηρείται, η συμμετοχή των ουσιαστικών (NN) είναι σε όλες τις περιπτώσεις εκείνη που συγκεντρώνει το μεγαλύτερο ποσοστό έναντι των υπολοίπων. Άξιο αναφοράς είναι επίσης, το γεγονός ότι στις περιπτώσεις που η εξαγωγή χαρακτηριστικών γίνεται με βάση τη συχνότητα των όρων (Count Vectorizer) έχουμε μεγαλύτερη διασπορά των ποσοστών των σημαντικότερων χαρακτηριστικών στα διάφορα POS tags, γι’ αυτό και το ποσοστό των υπόλοιπων (Other) είναι αρκετά υψηλό.



Εικόνα 22 Κατανομή POS tags για NLTK/IDF(mt) (αριστερά) και NLTK/CV(mt)



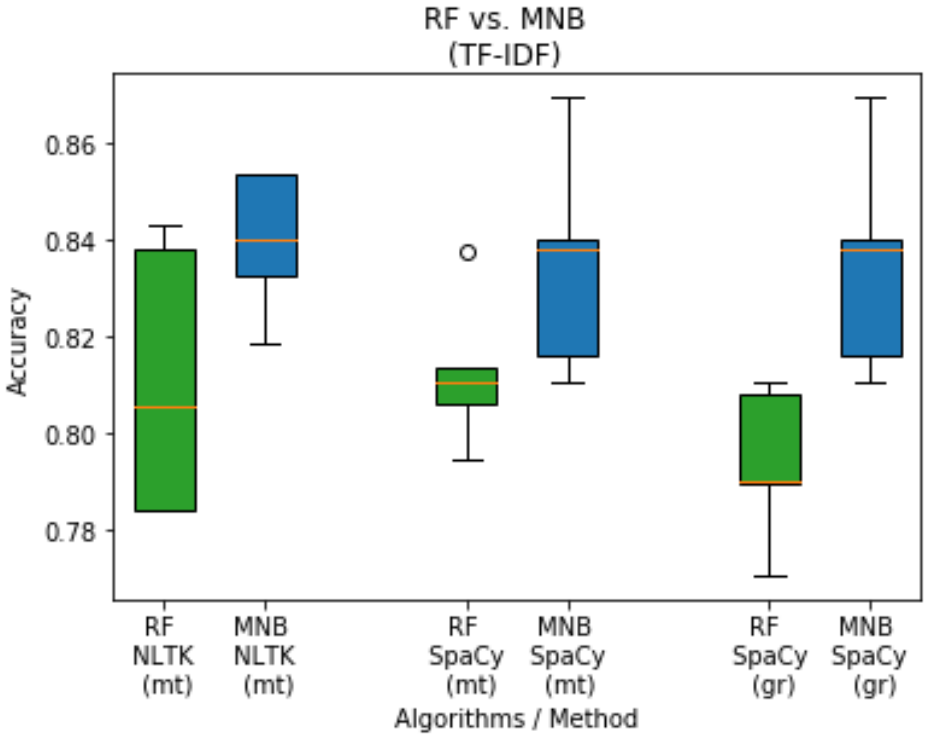
Εικόνα 23 Κατανομή POS tags για spaCy/IDF(mt) (αριστερά) και spaCy/CV(mt)



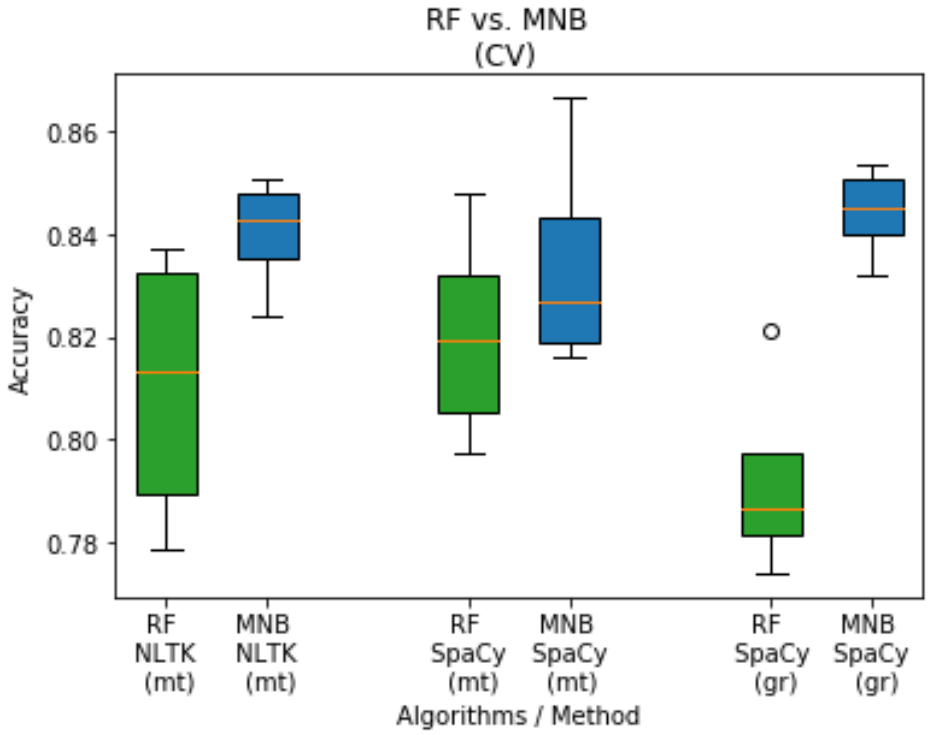
Εικόνα 24 Κατανομή POS tags για spaCy/IDF(gr) (αριστερά) και spaCy/CV(gr)

5.4.3 Σύγκριση RF και MNB accuracies

Παρακάτω ακολουθούν δύο διαγράμματα στα οποία συγκρίνουμε τα accuracies για το σύνολο των 5 folds για τους αλγόριθμους Random Forest και Multinomial Naive Bayes για την περίπτωση μη απαλοιφής stop words και ανά τεχνική εξαγωγής χαρακτηριστικών. Παρατηρούμε ότι σε όλες τις περιπτώσεις ο MNB υπερτερεί του RF και επίσης, εμφανίζει μικρότερη απόκλιση στα αποτελέσματα, γεγονός που δείχνει ότι το μοντέλο είναι περισσότερο αξιόπιστο. Επίσης, αξιοσημείωτο είναι ότι ο RF εμφανίζει το χαμηλότερο accuracy στην περίπτωση spaCy με ελληνικά και CV, εκεί που ο MNB έχει το υψηλότερο.



Εικόνα 25 Σύγκριση των RF και MNB accuracies ανά μέθοδο με χρήση TF-IDF



Εικόνα 26 Σύγκριση των RF και MNB accuracies ανά μέθοδο με χρήση CV

5.4.4 Σύγκριση RF και MNB accuracies με feature pruning

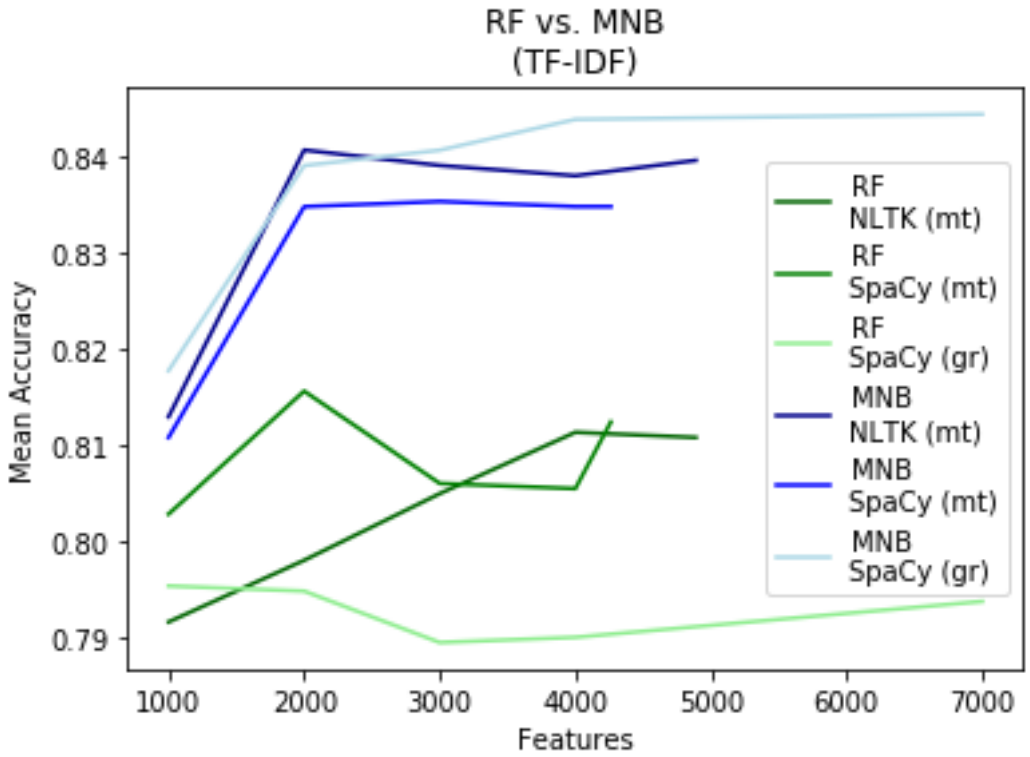
Ο παρακάτω πίνακας απεικονίζει τα μέσα accuracies για το σύνολο των πειραμάτων με feature pruning και χωρίς να αφαιρούμε τα stop words. Γενικά, παρατηρούμε ότι μόνο στην περίπτωση της spaCy με ελληνικά επιτυγχάνονται τα υψηλότερα αποτελέσματα με χρήση όλων των διαθέσιμων χαρακτηριστικών. Αντίθετα, στην περίπτωση των μεταφρασμένων δεδομένων, τόσο η NLTK όσο και η spaCy, επιτυγχάνουν τα υψηλότερα accuracies για 2000 και 3000 χαρακτηριστικά. Επίσης, αξίζει να σημειωθεί ότι η NLTK με CV επιτυγχάνει το καλύτερο accuracy για features=3000 για το σύνολο των αποτελεσμάτων των πειραμάτων μας.

Features # / Algorithm		**NLTK/TF-IDF (mt)**	**NLTK/CV (mt)**	**spaCy/TF-IDF (mt)**	**spaCy/CV (mt)**	**spaCy/TF-IDF (gr)**	**spaCy/CV (gr)**
1000	**RF**	0.7916	0.7974	0.8028	0.8220	0.7953	0.7878
	MNB	0.8129	0.8182	0.8108	0.8188	0.8177	0.8177
2000	**RF**	0.7980	0.8044	0.8156	0.8134	0.7948	0.7932
	MNB	**0.8406**	0.8449	0.8347	**0.8358**	0.8390	0.8353
3000	**RF**	0.8049	0.8033	0.8060	0.8113	0.7894	0.7942
	MNB	0.8390	**0.8465**	**0.8353**	0.8337	0.8406	0.8422
4000	**RF**	0.8113	0.8028	0.8054	0.8188	0.7900	0.7910

	MNB	0.8380	0.8406	0.8348	0.8342	0.8438	**0.8443**
All-features	**RF**	0.8108	0.8102	0.8124	0.8204	0.7937	0.7921
	MNB	0.8396	0.8401	0.8348	0.8342	**0.8444**	**0.8443**

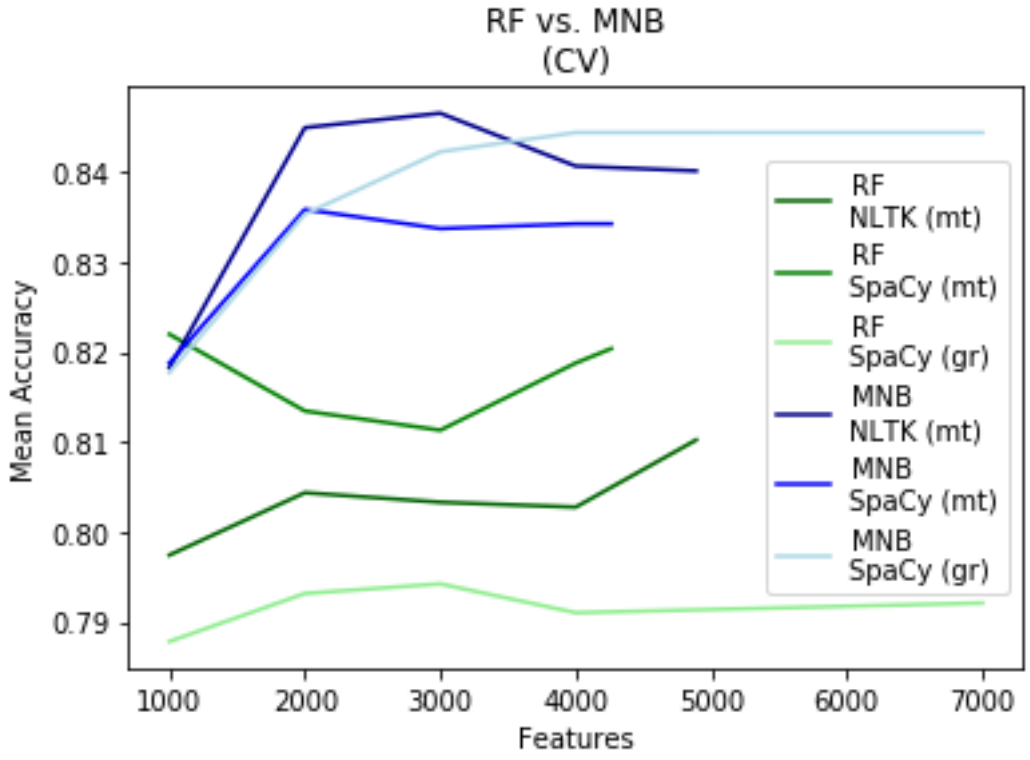
Πίνακας 53 Mean accuracy με feature pruning

Στο παρακάτω διάγραμμα συγκρίνουμε τους Random Forest και Multinomial Naive Bayes για τις περιπτώσεις μη απαλοιφής των stop words και για την τεχνική TF-IDF με σταδιακή αύξηση των χαρακτηριστικών που λαμβάνουν μέρος στην εκπαίδευση του μοντέλου. Γενικά, παρατηρείται αυξητική τάση όσο αυξάνεται ο αριθμός των χαρακτηριστικών και για τους δύο αλγόριθμους. Σε ό,τι αφορά στον MNB, υπάρχει μεγαλύτερη συμφωνία στη συμπεριφορά του και στις τρεις μεθόδους. Επίσης, περίπου στα 2000 features φαίνεται να σταματά η αυξητική του διάθεση σε ό,τι αφορά τις δύο πρώτες μεθόδους. Ο RF παρουσιάζει μια περισσότερο άναρχη και ξεχωριστή πορεία ανά μέθοδο. Ειδικά, στην περίπτωση του spaCy με αγγλικά δεδομένα εμφανίζει μεγάλες διακυμάνσεις καθώς τα χαρακτηριστικά αυξάνονται.



Εικόνα 27 Σύγκριση των RF και MNB accuracies για αύξηση των features με χρήση TF-IDF

Στο παρακάτω διάγραμμα, γίνεται η αντίστοιχη σύγκριση των RF και MNB για την περίπτωση εξαγωγής χαρακτηριστικών βάσει συχνότητας των όρων (Count Vectorizer). Παρατηρούμε ότι ο MNB και πάλι στα 2000 features φτάνει στην καλύτερη τιμή του για τις μεθόδους με αγγλικά δεδομένα και παρουσιάζει μια σταθερότητα ή τάση καθόδου στο accuracy περίπου από τα 3000 features και μετά. Σχετικά με τον RF, αξίζει να σημειωθεί ότι στα 1000 features με τη μέθοδο spaCy για αγγλικά καταφέρνει ελαφρώς υψηλότερα ποσοστά από οποιαδήποτε μέθοδο με MNB.



Εικόνα 28 Σύγκριση των RF και MNB accuracies για αύξηση των features με χρήση CV

6 Επίλογος

Η παρούσα εργασία πραγματεύεται τη διαγλωσσική ανάλυση συναισθήματος για ελληνικά δεδομένα του Twitter. Για το σκοπό αυτό και με βάση το accuracy της πρόβλεψης, συγκρίνονται τρεις μέθοδοι, από τις οποίες οι δύο μέθοδοι έχουν να κάνουν με ανάλυση συναισθήματος μηχανικά μεταφρασμένων δεδομένων στα αγγλικά, ενώ η τρίτη πραγματοποιεί ανάλυση συναισθήματος απευθείας σε ελληνικά δεδομένα. Η διαδικασία συνοψίζεται στη συλλογή και προ-επεξεργασία των δεδομένων, σε επόμενο στάδιο γίνεται η εξαγωγή χαρακτηριστικών και στη συνέχεια, εκπαιδεύεται το μοντέλο μας με τη βοήθεια δύο αλγόριθμων ταξινόμησης. Τέλος, ακολουθούν τα αποτελέσματα και η αξιολόγησή τους για μια σειρά πειραμάτων που διενεργήθηκαν για το σύνολο των μεθόδων.

6.1 Σύνοψη και συμπεράσματα

Για το σύνολο των τριών μεθόδων παρατηρούμε σε ό,τι αφορά τα αρχικά πειράματα ότι ο Naive Bayes συμπεριφέρεται καλύτερα από τον Random Forest, είτε πρόκειται για εξαγωγή χαρακτηριστικών με τη μέθοδο TF-IDF είτε με τη μέθοδο Count Vectorizer. Επίσης, παρατηρούμε ότι έχουμε καλύτερα αποτελέσματα στις περιπτώσεις που δεν υπάρχει αφαίρεση των stop words. Τέλος, σχετικά με τις μεθόδους που χρησιμοποιήθηκαν, καλύτερη κρίνεται εκείνη που χρησιμοποιεί τη βιβλιοθήκη spaCy με ελληνικά δεδομένα, ενώ με μικρή διαφορά ακολουθεί η NLTK και στη συνέχεια, επίσης με μικρή διαφορά η spaCy για μηχανικά μεταφρασμένα δεδομένα.

Στη συνέχεια, με βάση τα παραπάνω πειράματα, επικεντρωνόμαστε αποκλειστικά σε εκείνα που αφορούν στη μη απαλοιφή των stop words. Για αυτές τις περιπτώσεις, πραγματοποιούμε μια σειρά νέων πειραμάτων που εστιάζουν στο κλάδεμα (pruning) των χαρακτηριστικών. Για τα μεταφρασμένα δεδομένα τόσο με τη χρήση της NLTK όσο και με τη χρήση της spaCy, παρατηρούμε ότι τα καλύτερα αποτελέσματα επιτυγχάνονται με pruning, δηλαδή η αύξηση των χαρακτηριστικών δεν προσθέτει στο τελικό accuracy. Το παραπάνω, ωστόσο δεν ισχύει για τη μέθοδο με spaCy με ελληνικά δεδομένα, όπου επιτυγχάνει το καλύτερο accuracy με χρήση όλων των χαρακτηριστικών που έχει στη διάθεσή της.

Τα πολύ καλά αποτελέσματα και για τις τρεις μεθόδους, επιβεβαιώνουν ότι η ανάλυση συναισθήματος για ελληνικά δεδομένα, μπορεί να επιτευχθεί τόσο υιοθετώντας πρακτικές διαγλωσσικής ανάλυσης, δηλαδή με χρήση μηχανικής μετάφρασης και των υπαρχόντων εργαλείων για την αγγλική γλώσσα, όσο και με χρήση εργαλείων που έχουν αναπτυχθεί αποκλειστικά για τα ελληνικά.

6.2 Μελλοντικές Επεκτάσεις

Στο πλαίσιο αυτής της διπλωματικής εργασίας, ερευνήθηκαν τρεις διαφορετικές μέθοδοι για την ανάλυση συναισθήματος από ελληνικά δεδομένα του Twitter. Δεδομένου ότι τα tweets που χρησιμοποιήθηκαν αφορούν δύο συγκεκριμένες θεματικές, θα μπορούσε να ερευνηθεί σε επόμενη φάση κατά πόσο το μοντέλο μας θα μπορούσε να παραγάγει καλά αποτελέσματα σε tweets με μεγαλύτερη θεματική ποικιλία.

Ένας ακόμη τομέας που θα μπορούσε να εξεταστεί εκτενέστερα είναι ποια από τα μέρη του λόγου συμβάλουν περισσότερο στη διαδικασία ταξινόμησης και τι αποτελέσματα παράγουν αν τα χαρακτηριστικά εξάγονται αποκλειστικά με βάση μια συγκεκριμένη κατηγορία (π.χ. ουσιαστικά) ή συνδυασμούς αυτών (π.χ. ουσιαστικά, επίθετα).

7 Βιβλιογραφία

Abhijit Mishra, D. K. (2016, August). Leveraging Cognitive Features for Sentiment Analysis. *Association for Computational Linguistics* .

Adam Tsakalidis, S. P. (2018, July). Building and evaluating resources for sentiment analysis in the Greek language.

Adel Al-Shabi, A. A.-M. (2017, January). Cross-Lingual Sentiment Classification from English to Arabic using Machine Translation. *International Journal of Advanced Computer Science and Applications 8(12)* .

Alec Go, R. B. (2009). Twitter Sentiment Classification using Distant Supervision.

Alexander Pak, P. P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. *LREC* .

Alexandra Balahur, M. T. (2012). Comparative Experiments for Multilingual SentimentAnalysis Using Machine Translation. *SDAD@ECML/PKDD* .

Charu C. Aggarwal, C. Z. (2012). *Mining Text Data.* Kluwer Academic Publishers.

Daniel Jurafsky, J. H. (2018). *Speech and Language Processing.*

E. Schinas, S. P. (2013). "EventSense: Capturing the Pulse of Large-scale Events by Mining Social Media Streams". *Proceedings of the 17th Panhellenic Conference on Informatics* .

Hiroshi Motoda, H. L. (2002, May). Feature selection, extraction and construction. *Communication of IICM* .

Honglei Guo, H. Z. (2010, October). OpinionIt: a text mining system for cross-lingual opinion analysis. *CIKM'10, Proceedings of the 19th ACM international conference on Information and knowledge management* .

Irfan, M. (2017, October). Machine Translation.

Jivani, A. G. (2011, November). A Comparative Study of Stemming Algorithms. *IJCTA* .

Judith Hurwitz, D. K. (2018). *Machine Learning for Dummies.* John Wiley & Sons, Inc.

Kaplan, R. M. (2005, January). A Method for Tokenizing Text.

Koehrsen, W. (n.d.). Random Forest Simple Explanation. *https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d* .

Kohavi, R. (1995, August). A study of cross-validation and bootstrap for accuracy estimation and model selection. *International Joint Conference on Artificial Intelligence* .

Liu, B. (2012). *Sentiment Analysis and Opinion Mining.* Morgan & Claypool Publishers.

Liu, B. (2011). *Web Data Mining – Exploring Hyperlinks, Contents and Usage Data.*

Louppe, G. (2014, July). Understanding Random Forest.

Maite Taboada, J. B. (2011, June). Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics* .

Manning, J. (2014, January). Definition and Classes of Social Media. *Encyclopedia of social media and politics* .

Matthew J. Denny, A. S. (2017, September). Text Preprocessing For Unsupervised Learning: Why ItMatters, When It Misleads, And What To Do About It. *Political Analysis* .

Mauro Castelli, L. V. (2019). Supervised Learning: Classification.

Medium. (n.d.). Random Forest Simple Explanation. *https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d* .

Mohamed Abdalla, G. H. (2017). Cross-Lingual Sentiment Analysis Without (Good) Translation. *Proceedings of the The 8th International Joint Conference on Natural Language Processing* .

Mosab Faqeeh, N. A.-A. (2014). Cross-lingual Short-Text Document Classificationfor Facebook Comments. *International Conference on Future Internet of Things and Cloud* .

Najma Sultana, P. K. (2019, April). SENTIMENT ANALYSIS FOR PRODUCT REVIEW. *ICTACT Journal on Soft Computing* .

Ozgur, A. (2004, January). Supervised and Unsupervised Machine Learning Techniques for Text Document Categorization.

Pollyanna Gonçalves, M. A. (2013). Comparing and Combining Sentiment Analysis Methods. *Proceedings of the first ACM conference on Online social networks* .

Resham N. Waykole, A. D. (2018, April). A REVIEW OF FEATURE EXTRACTION METHODS FOR TEXT CLASSIFICATION . *International Journal of Advance Engineering and Research Development* .

S. Vijayarani, J. I. (2015). Preprocessing Techniques for Text Mining-An Overview Dr. *International Journal of Computer Science & Communication Networks* .

Scikit-learn. (n.d.). Cross-validation: evaluating estimator performance. *https://scikit-learn.org/stable/modules/cross_validation.html* .

Sharma, D. (2012, September). Stemming Algorithms: A Comparative Study and their Analysis. *International Journal of Applied Information Systems (IJAIS)* .

Statista. (n.d.). Number of internet users worldwide from 2005 to 2018 (in millions). *https://www.statista.com/statistics/273018/number-of-internet-users-worldwide/* .

Statista. (n.d.). Number of social media users worldwide from 2010 to 2021 (in billions). *https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/* .

Statista. (n.d.). Only 34% of All Tweets Are in English. *https://www.statista.com/chart/1726/languages-used-on-twitter/* .

Turney, P. D. (2002). Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* .

Vasileios Hatzivassiloglou, M. (1997). Predicting the Semantic Orientation of Adjectives. *Proceedings of the 8th Conference on European Chapter of the Association for Computational Linguistics* .

Wan, X. (2009, August). Co-Training for Cross-Lingual Sentiment Classification. *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP* .

Zahurul Islam, M. N. (2010, September). A Light Weight Stemmer for Bengali and Its Use in Spelling Checker.