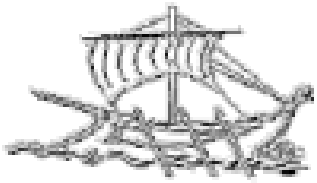


ΑΕΙ ΠΕΙΡΑΙΑ Τ.Τ.
ΣΧΟΛΗ ΤΕΧΝΟΛΟΓΙΚΩΝ ΕΦΑΡΜΟΓΩΝ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΗΛΕΚΤΡΟΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ Τ.Ε.

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Ανάλυση συναισθήματος σε ελληνικό κείμενο με χρήση αλγόριθμων μηχανικής μάθησης

Μιχαήλ Γ. Λιουδάκης
Ελευθέριος Γ. Αλεξανδράκης

Εισηγητής: Δρ Γεώργιος Πρεζεράκος, Καθηγητής

ΑΘΗΝΑ

ΣΕΠΤΕΜΒΡΗΣ 2017

(Κενό φύλλο)

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Ανάλυση συναισθήματος σε ελληνικό κείμενο με χρήση αλγόριθμων μηχανικής μάθησης

Μιχαήλ Γ. Λιουδάκης
Α.Μ. 42877

Ελευθέριος Γ. Αλεξανδράκης
Α.Μ. 43355

Εισηγητής:

Δρ Γεώργιος Πρεζεράκος, Καθηγητής

Εξεταστική Επιτροπή:

Ημερομηνία εξέτασης:

(Κενό φύλλο)

ΔΗΛΩΣΗ ΣΥΓΓΡΑΦΕΑ ΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ

Οι κάτωθι υπογεγραμμένοι **Λιουδάκης Μιχαήλ**, του **Γεωργίου**, με αριθμό μητρώου **42877** και **Αλεξανδράκης Ελευθέριος**, του **Γεωργίου**, με αριθμό μητρώου **43355** φοιτητές του Τμήματος Μηχανικών Η/Υ Συστημάτων Τ.Ε. του Α.Ε.Ι. Πειραιά Τ.Τ. πριν αναλάβουμε την εκπόνηση της Πτυχιακής Εργασίας μας, δηλώνουμε ότι ενημερωθήκαμε για τα παρακάτω:

«Η Πτυχιακή Εργασία (Π.Ε.) αποτελεί προϊόν πνευματικής ιδιοκτησίας τόσο του συγγραφέα, όσο και του Ιδρύματος και θα πρέπει να έχει μοναδικό χαρακτήρα και πρωτότυπο περιεχόμενο.

Απαγορεύεται αυστηρά οποιοδήποτε κομμάτι κειμένου της να εμφανίζεται αυτούσιο ή μεταφρασμένο από κάποια άλλη δημοσιευμένη πηγή. Κάθε τέτοια πράξη αποτελεί προϊόν λογοκλοπής και εγείρει θέμα Ηθικής Τάξης για τα πνευματικά δικαιώματα του άλλου συγγραφέα. Αποκλειστικός υπεύθυνος είναι ο συγγραφέας της Π.Ε., ο οποίος φέρει και την ευθύνη των συνεπειών, ποινικών και άλλων, αυτής της πράξης.

Πέραν των όποιων ποινικών ευθυνών του συγγραφέα σε περίπτωση που το Ίδρυμα του έχει απονείμει Πτυχίο, αυτό ανακαλείται με απόφαση της Συνέλευσης του Τμήματος. Η Συνέλευση του Τμήματος με νέα απόφασης της, μετά από αίτηση του ενδιαφερόμενου, του αναθέτει εκ νέου την εκπόνηση της Π.Ε. με άλλο θέμα και διαφορετικό επιβλέποντα καθηγητή. Η εκπόνηση της εν λόγω Π.Ε. πρέπει να ολοκληρωθεί εντός τουλάχιστον ενός ημερολογιακού 6μήνου από την ημερομηνία ανάθεσης της. Κατά τα λοιπά εφαρμόζονται τα προβλεπόμενα στο άρθρο 18, παρ. 5 του ισχύοντος Εσωτερικού Κανονισμού.»

(Κενό φύλλο)

ΕΥΧΑΡΙΣΤΙΕΣ

Η παρούσα πτυχιακή εργασία ολοκληρώθηκε μετά από επίμονες προσπάθειες, σε ένα ενδιαφέρον γνωστικό αντικείμενο, όπως αυτό της μηχανικής μάθησης. Την προσπάθειά μας αυτή υποστήριξε ο επιβλέπων καθηγητής μας, τον οποίο θα θέλαμε να ευχαριστήσουμε.

Ακόμα θα θέλαμε να ευχαριστήσουμε τον κ. Ιωάννη Τριαντάφυλλο για τις πολύτιμες συμβουλές του. Επίσης, οφείλουμε ένα ευχαριστώ στις οικογένειές μας και τους φίλους μας για την στήριξή τους όπου χρειάστηκε.

(Κενό φύλλο)

ΠΕΡΙΛΗΨΗ

Στόχος της πτυχιακής αυτής είναι η ανάλυση συναισθήματος σε ελληνικό κείμενο, το οποίο έχει εξαχθεί από κοινωνικά δίκτυα. Συγκεκριμένα, μελετήθηκαν διάφορες τεχνικές για την επίτευξη του στόχου αυτού με την χρήση δύο σετ δεδομένων – ένα κύριο και ένα βοηθητικό. Το βοηθητικό σετ δεδομένων, το οποίο περιείχε αγγλικά κείμενα, χρησιμοποιήθηκε για την εξαγωγή χρήσιμων συμπερασμάτων ως προς την συμπεριφορά του συστήματος αλλά και την επιλογή σημαντικών παραμέτρων. Το κύριο σετ δεδομένων χρησιμοποιήθηκε ως τελική αξιολόγηση του συστήματος, καθώς τα δεδομένα του ήταν στα ελληνικά. Όλα τα δεδομένα, εξήχθησαν από ένα κοινωνικό δίκτυο (Twitter) και βαθμονομήθηκαν ως προς το συναίσθημα που φέρουν με κάποια ετικέτα. Τα ετικετοποιημένα, πλέον, δεδομένα εισήχθησαν στο σύστημα με σκοπό την εκπαίδευσή του. Έπειτα, το σύστημα αξιολογήθηκε με συγκεκριμένες μετρικές ως προς την επίδοσή του να προβλέπει άγνωστα δεδομένα για το συναίσθημα που φέρουν. Τέλος, γίνεται μια προσπάθεια κατασκευής ενός δεύτερου συστήματος που παράγει διαφορετικά αποτελέσματα με μια τελείως διαφορετική προσέγγιση. Σκοπός είναι η μελλοντική συγχώνευσή του με το υπάρχον σύστημα για την περαιτέρω βελτίωση της απόδοσης.

ABSTRACT

The present thesis concerns the development of sentiment analysis in a Greek context, extracted from social media. For that purpose, different techniques were examined by using two different datasets, a main and a supporting one. The supporting dataset, which contained English texts, was used to extract useful outcomes regarding system's behavior and the choice of important parameters. The main dataset was used as a final evaluation of the system, as its data were consisted by Greek text. Every text, was extracted from social media (Twitter) and was labeled according to their sentiment conveyed. Labeled data were inserted to system so as to be trained. Afterwards, the system was evaluated with specific metrics for its performance to predict unlabeled data for the sentiment conveyed. Finally, there is an attempt for a construction of a second system, which outputs different outcomes with a totally different approach. The aim is the future merge with the current system for further boost of its performance.

ΕΠΙΣΤΗΜΟΝΙΚΗ ΠΕΡΙΟΧΗ: Μηχανική Μάθηση
ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: ανάλυση συναισθήματος, αλγόριθμοι ταξινόμησης, πολικότητα κειμένου, feature extraction, machine learning

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

ΚΑΤΑΛΟΓΟΣ ΔΙΑΓΡΑΜΜΑΤΩΝ

ΓΛΩΣΣΑΡΙ

NLP Natural Language Processing

TF-IDF Term Frequency – Inverse Document Frequency

BOW Bag Of Words

SVM Support Vector Machines

KNN K-Nearest Neighbors

SVC Support Vector Classification

ROC Receiver Operation Characteristic

Placeholder Σύμβολο υποκατάστασης

Hyperplane Υπερεπίπεδο

IR Information Retrieval

Classifier Ταξινομητής

Feature Χαρακτηριστικό

Label Ετικέτα

Accuracy Ακρίβεια

Stop words Συχνά χρησιμοποιούμενες λέξεις όπως προθέσεις, σύνδεσμοι, άρθρα κλπ.

Stemming Αφαίρεση κατάληξης λέξεων

(Κενό φύλλο)

ΚΕΦΑΛΑΙΟ 1

ΕΙΣΑΓΩΓΗ

Σε αυτό το κεφάλαιο αναλύονται οι βασικές έννοιες του αντικειμένου της πτυχιακής εργασίας και γίνεται μια εισαγωγή στο θεωρητικό υπόβαθρο.

1.1 Ανάλυση Συναισθήματος

Η ανάλυση συναισθήματος είναι ένα συνεχώς αναπτυσσόμενο πεδίο της μηχανικής μάθησης. Οι έρευνες που έχουν γίνει έχουν τεράστιο εύρος: ξεκινάνε από την ταξινόμηση ολόκληρων κειμένων και φτάνουν έως την ταξινόμηση λέξεων και φράσεων.

Συγκεκριμένα, στην ανάλυση συναισθήματος σε κοινωνικά δίκτυα όπως το Twitter, η ταξινόμηση ενός tweet μοιάζει πιο πολύ με ανάλυση συναισθήματος μιας πρότασης παρά κάποιου κειμένου. Αυτό συμβαίνει λόγω των περιορισμών που υπάρχουν στους χαρακτήρες ενός tweet. Επίσης, αν σε αυτό προστεθεί και η ανεπίσημη αλλά και ειδικευμένη, πολλές φορές, γλώσσα που χρησιμοποιείται, τότε ο στόχος γίνεται ακόμα δυσκολότερος.

Ο εντοπισμός συναισθήματος είναι ένα κλασικό πρόβλημα στον κλάδο της ταξινόμησης κειμένου. Σε αντίθεση με άλλες εργασίες του συγκεκριμένου κλάδου, ο στόχος δεν είναι να αναγνωριστούν τα θέματα, οι οντότητες ή οι συγγραφείς ενός κειμένου αλλά να αξιολογηθεί το εκφραζόμενο συναίσθημα σαν θετικό, αρνητικό ή ουδέτερο. Οι περισσότερες προσεγγίσεις που χρησιμοποιούνται για την επίτευξη αυτού του στόχου, συνήθως, περιλαμβάνουν μεθόδους από την μηχανική μάθηση, την υπολογιστική γλωσσολογία και την στατιστική.

Επειδή το Twitter είναι μια από τις πλουσιότερες πηγές άντλησης απόψεων, πολλές προσεγγίσεις έχουν βασιστεί στην ανάλυση tweets. Κάθε προσέγγιση χρησιμοποιεί διαφορετικά features: από καθορισμένες πολικότητες λέξεων ή φράσεων έως στην χρήση emoticons, πεζών/κεφαλαίων γραμμάτων, επιμήκυνση λέξεων ή φωνητικών χαρακτηριστικών. Ο στόχος, συνήθως, είναι ο εντοπισμός του συναισθήματος που εκφράζεται σε ένα tweet σαν σύνολο. Πολλές φορές, βέβαια, μπορεί να πρέπει να αναγνωριστεί το συναίσθημα σε ένα tweet με βάση ένα συγκεκριμένο θέμα. Η διαφορά των δύο πλαισίων είναι ότι ένα αρνητικό -με την

ευρεία έννοια του όρου- tweet θα μπορούσε να θεωρηθεί ουδέτερο εφόσον δεν αναφέρεται στο ως προς εξέταση θέμα.

1.2 Μηχανική Μάθηση

Η μηχανική μάθηση είναι ένας τομέας της επιστήμης της Πληροφορικής, που εξελίχθηκε από την μελέτη της αναγνώρισης προτύπων και την θεωρία υπολογιστικής μάθησης στην τεχνητή νοημοσύνη. Το 1959, ο Arthur Samuel όρισε την μηχανική μάθηση ως: “Τομέας μελέτης που δίνει στους υπολογιστές την δυνατότητα να μάθουν χωρίς να έχουν προγραμματιστεί λεπτομερώς”. Η μηχανική μάθηση ερευνά την μελέτη και κατασκευή αλγορίθμων που μπορούν να μάθουν από δεδομένα και να κάνουν πρόβλεψη σε αυτά. Τέτοιοι αλγόριθμοι λειτουργούν με την κατασκευή ενός μοντέλου από δείγματα εισόδου. Ο σκοπός είναι ο αλγόριθμος να κάνει μια πρόβλεψη ή να πάρει μια απόφαση στηριζόμενη σε δεδομένα, αντί να ακολουθήσει αυστηρά τις στατικές προγραμματιστικές οδηγίες.

Η μηχανική μάθηση φαίνεται να συσχετίζεται και ορισμένες φορές να επικαλύπτει την υπολογιστική στατιστική, μια μέθοδος που, επίσης, μέσω της χρήσης των υπολογιστών επικεντρώνεται στην δημιουργία προβλέψεων. Συνδέεται άρρηκτα με την μαθηματική βελτιστοποίηση, η οποία προσφέρει μεθόδους, θεωρία και πεδία εφαρμογών στον κλάδο. Η μηχανική μάθηση βρίσκει εφαρμογή σε ένα μεγάλο εύρος υπολογιστικών εργασιών, όπου ο λεπτομερής σχεδιασμός και προγραμματισμός είναι αδύνατο να επιτευχθεί. Ως παραδείγματα τέτοιων εφαρμογών μπορούμε να θέσουμε τα παρακάτω: φιλτράρισμα spam, οπτική αναγνώριση χαρακτήρα, μηχανές αναζήτησης, υπολογιστική όραση. Η μηχανική μάθηση μερικές φορές συγχέεται με την εξόρυξη δεδομένων, ωστόσο το συγκεκριμένο πεδίο επικεντρώνεται περισσότερο στην εξερευνητική ανάλυση Δεδομένων, γνωστή και ως μάθηση χωρίς επίβλεψη. Εντός του πεδίου της ανάλυσης δεδομένων, η μηχανική μάθηση είναι μια μέθοδος που χρησιμοποιείται για την σχεδίαση σύνθετων μοντέλων και αλγορίθμων με σκοπό την πρόβλεψη. Τα αναλυτικά αυτά μοντέλα, βασιζόμενα σε ιστορικές συσχετίσεις και τάσεις από τα εκάστοτε δεδομένα, επιτρέπουν σε ερευνητές, επιστήμονες δεδομένων, μηχανικούς και αναλυτές να παράγουν αξιόπιστες, επαναλαμβανόμενες αποφάσεις και αποτελέσματα, με σκοπό να εκμαιεύσουν “κρυφές γνώσεις”.

Θεωρητικά, η μηχανική μάθηση διακρίνεται σε τρεις μεγάλες κατηγορίες, με βάση την διαθέσιμη φύση της μάθησης, σε ένα σύστημα εκμάθησης. Οι κατηγορίες είναι οι παρακάτω:

1.2.1 Εποπτευόμενη μάθηση

Ως εποπτευόμενη μάθηση ορίζεται η διαδικασία κατά την οποία, ένας «δάσκαλος» παρέχει στον υπολογιστή κάποια παραδείγματα ως εισόδους και κάποιες επιθυμητές εξόδους. Ο στόχος είναι να μάθει έναν γενικό κανόνα, που να του επιτρέπει να χαρτογραφεί εισόδους σε εξόδους. Αναλυτικότερα, η εποπτευόμενη μάθηση είναι η εργασία της μηχανικής μάθησης, σύμφωνα με την οποία κατασκευάζεται μια συνάρτηση από κάποια δεδομένα προς εκπαίδευση. Τα δεδομένα αυτά, χαρακτηρίζονται από ετικέτες και αποτελούνται από ένα σετ παραδειγμάτων. Στην εποπτευόμενη μάθηση, κάθε παράδειγμα είναι ένα ζευγάρι που αποτελείται από ένα αντικείμενου εισόδου -συνήθως είναι διάνυσμα- και μια επιθυμητή τιμή εξόδου, γνωστή και ως “εποπτευόμενο σήμα”. Ένας αλγόριθμος εποπτευόμενης μάθησης αναλύει τα δεδομένα που έχουν εκπαιδευτεί και παράγει μια τεκμαιρόμενη συνάρτηση, η οποία μπορεί να χρησιμοποιηθεί για την χαρτογράφηση νέων παραδειγμάτων. Ιδανικά, ο αλγόριθμος δύναται να καθορίζει σωστά τις ετικέτες κλάσεων στις περιπτώσεις που τα παραδείγματα είναι άγνωστα. Για την επίτευξη του επιθυμητού αποτελέσματος, απαιτείται από τον αλγόριθμο εκμάθησης να γενικεύσει τα προς εκπαίδευση δεδομένα, σε άγνωστες καταστάσεις με κάποιον “λογικό” τρόπο.

1.2.2 Μη εποπτευόμενη μάθηση

Στον αλγόριθμο μάθησης δεν δίνονται ετικέτες, αφήνοντάς τον να βρει μόνος του την δομή της εισόδου, κάτι το οποίο μπορεί να είναι από μόνο του ένας στόχος (η ανακάλυψη κρυμμένων προτύπων σε δεδομένα) ή ένα μέσο προς έναν άλλο στόχο. Πιο συγκεκριμένα, η μη εποπτευόμενη μάθηση είναι η εργασία της μηχανικής μάθησης κατά την οποία κατασκευάζεται μια συνάρτηση για να περιγράψει μια κρυφή δομή από μη ετικετοποιημένα δεδομένα. Από την στιγμή που τα παραδείγματα που δίνονται στον μαθητευόμενο δεν έχουν ετικέτα, δεν υπάρχει σήμα σφάλματος ή επιβράβευσης για την αξιολόγηση μιας πιθανής κατάστασης.

Αυτό ξεχωρίζει την μη εποπτευόμενη μάθηση από την εποπτευόμενη και την ενισχυτική μάθηση.

1.2.3 Ενισχυτική μάθηση

Ένα πρόγραμμα αλληλοεπιδρά με ένα δυναμικό περιβάλλον, στο οποίο πρέπει να πετύχει έναν συγκεκριμένο στόχο (όπως η οδήγηση ενός οχήματος ή η εκμάθηση ενός παιχνιδιού παίζοντας με κάποιον αντίπαλο), χωρίς κάποιον δάσκαλο να του περιγράφει λεπτομερώς αν είναι κοντά στον στόχο του. Η διαφορά της ενισχυτικής μάθησης από την εποπτευόμενη έγκειται στο ζευγάρι σωστών εισόδων/εξόδων, τα οποία δεν παρουσιάζονται ποτέ, ούτε από υπο-βέλτιστες ενέργειες λεπτομερώς σωστές. Μεταξύ της εποπτευόμενης και μη εποπτευόμενης μάθησης υπάρχει η ημι-εποπτευόμενη μάθηση. Στην περίπτωση αυτή, ο δάσκαλος δίνει ένα ελλιπές εκπαιδευμένο σήμα: ένα εκπαιδευμένο σετ με μερικές (συνήθως πολλές) εξόδους να λείπουν. Πιο συγκεκριμένα, η ημι-εποπτευόμενη μάθηση είναι μια κλάση των εργασιών και τεχνικών εποπτευόμενης μάθησης, στην οποία γίνεται χρήση και μη ετικετοποιημένων δεδομένων για εκπαίδευση. Πολλές φορές, τα δεδομένα αποτελούνται σε μεγάλο ποσοστό τους από δεδομένα με ετικέτες και σε ένα μικρό ποσοστό χωρίς αυτές. Ερευνητές του χώρου της μηχανικής μάθησης κατέληξαν στο συμπέρασμα ότι τα δεδομένα χωρίς ετικέτες, όταν χρησιμοποιούνται σε συνδυασμό με ένα μικρό ποσοστό ετικετοποιημένων δεδομένων, μπορούν να παράγουν αξιοσημείωτη βελτίωση στην μαθησιακή ακρίβεια.

1.3 Επεξεργασία Φυσικής Γλώσσας

Η επεξεργασία φυσικής γλώσσας είναι ένας τομέας της επιστήμης των υπολογιστών, της τεχνητής νοημοσύνης και της υπολογιστικής γλωσσολογίας και αφορά την αλληλεπίδραση μεταξύ υπολογιστών και της ανθρώπινης (φυσικής) γλώσσας. Κύριος σκοπός του συγκεκριμένου τομέα, είναι η κατανόηση της φυσικής γλώσσας, κάτι το οποίο θα δώσει την δυνατότητα στους υπολογιστές να αντλούν νόημα από μια είσοδο φυσικής γλώσσας. Ενώ, ένας άλλος σημαντικός σκοπός της επεξεργασίας φυσικής γλώσσας είναι η παραγωγή φυσικής γλώσσας.

1.3.1 Κατανόηση φυσικής γλώσσας

Η κατανόηση φυσικής γλώσσας είναι ένα υποπεδίο της επεξεργασίας φυσικής γλώσσας που ασχολείται με την κατανόηση ανάγνωσης, κάτι το οποίο θεωρείται από τα πιο δύσκολα προβλήματα της τεχνητής νοημοσύνης. Η διαδικασία της αποσυναρμολόγησης και της ανάλυσης των λέξεων είναι πολύ πιο σύνθετη και δύσκολη από την αντίστροφη διαδικασία συναρμολόγησης η οποία παράγει έξοδο σε φυσική γλώσσα, επειδή η ύπαρξη άγνωστων και απρόσμενων χαρακτηριστικών στην είσοδο και η ανάγκη να προσδιορίσει συντακτικά και σημασιολογικά σχεδιαγράμματα που να ταιριάζουνε σε αυτό, παράγοντες οι οποίοι είναι προ-προσδιορισμένοι στην παραγωγή γλώσσας. Μερικές εφαρμογές του υποπεδίου είναι: η συγκέντρωση ειδήσεων, η ταξινόμηση κειμένου, η ενεργοποίηση φωνής και η ανάλυση περιεχομένου μεγάλης κλίμακας.

1.3.2 Παραγωγή φυσικής γλώσσας

Παραγωγή φυσικής γλώσσας είναι η εργασία της επεξεργασίας φυσικής γλώσσας κατά την οποία παράγεται φυσική γλώσσα από ένα σύστημα αναπαράστασης μηχανής, γνωστό ως βάση γνώσεων ή λογική μορφή. Θα μπορούσαμε να πούμε ότι ένα σύστημα παραγωγής φυσικής γλώσσας είναι ένας μεταφραστής που μετατρέπει μια αναπαράσταση βασισμένη σε κάποιον υπολογιστή σε μια αναπαράσταση φυσικής γλώσσας. Οι μέθοδοι για την παραγωγή αυτή διαφέρουν κατά πολύ από εκείνες ενός μεταγλωττιστή λόγω των έμφυτης εκφραστικότητας στις φυσικές γλώσσες. Ένα παράδειγμα εφαρμογής είναι τα συστήματα που παράγουν επιστολές, τα οποία δεν περιλαμβάνουν γραμματικούς κανόνες αλλά μπορούν να παράξουν ένα γράμμα σε έναν πελάτη π.χ. ειδοποιώντας τον ότι έφτασε το όριο της πιστωτικής του κάρτας. Πιο σύνθετα συστήματα παραγωγής φυσικής γλώσσας δυναμικά δημιουργούν κείμενα ώστε να πετύχουν κάποιον επικοινωνιακό στόχο.

1.3.3 Επεξεργασία φυσικής γλώσσας με την χρήση μηχανικής μάθησης

Οι σύγχρονοι αλγόριθμοι επεξεργασίας φυσικής γλώσσας βασίζονται στη μηχανική μάθηση και ιδιαίτερα στην στατιστική μηχανική μάθηση. Η συγκεκριμένη προσέγγιση διαφέρει κατά πολύ από τις προηγούμενες προσπάθειες που έγιναν στον συγκεκριμένο τομέα, οι οποίες περιελάμβαναν πολλούς κανόνες γραμμένους με κώδικα από ανθρώπινο χέρι. Όμως, η πρόταση της μηχανικής μάθησης είναι να

μαθαίνονται αυτόματα αυτοί οι κανόνες μέσα από την ανάλυση μεγάλων συλλογών από δεδομένα τα οποία θα αποτελούνται από τυπικά παραδείγματα στον πραγματικό κόσμο και έχουν αξιολογηθεί με το χέρι με την σωστή τιμή ώστε να τα μάθει το σύστημα. Οι αλγόριθμοι μηχανικής μάθησης παίρνουν σαν είσοδο ένα μεγάλο σετ “χαρακτηριστικών” που παράγονται από τα δεδομένα. Μερικοί από τους πιο σύγχρονους αλγόριθμους παράγουν συστήματα περιπτώσεων (αν-τότε) και κανόνων τα οποία μπορούν κάλλιστα να συγκριθούν με τους κανόνες που θα γραφόντουσαν με το χέρι. Σημαντικές εργασίες στην επεξεργασία φυσικής γλώσσας:

1) Αυτόματη περίληψη
2) Αυτόματη μετάφραση (από μια φυσική γλώσσα σε άλλη)
3) Οπτική αναγνώριση χαρακτήρα
4) Ετικετοποίηση μερών του λόγου
5) Απάντηση ερωτήσεων
6) Ανάλυση συναισθήματος
7) Ανάκτηση πληροφορίας
8) Εξόρυξη πληροφορίας
9) Αναγνώριση φωνής
10) Αναγνώριση θέματος
11) Κατάτμηση λέξεων
12) Επεξεργασία φωνής

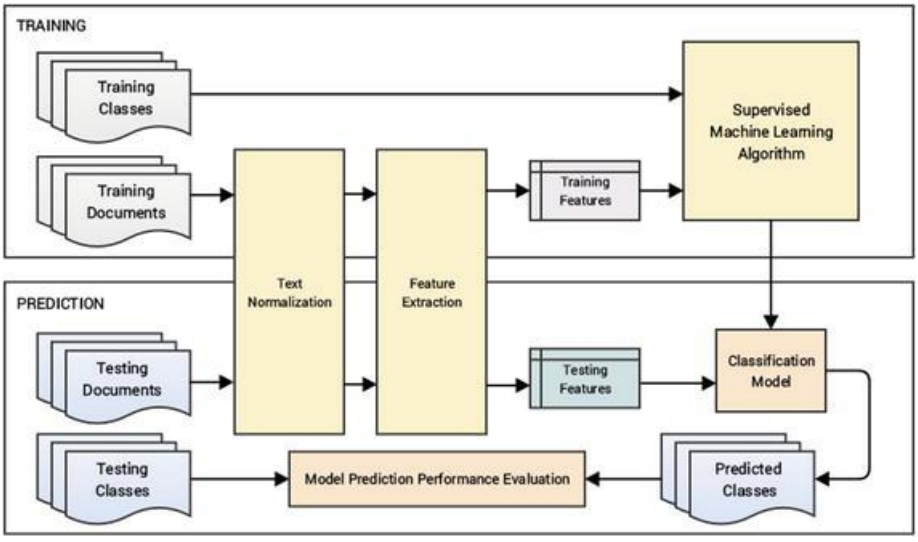
ΚΕΦΑΛΑΙΟ 2

ΑΡΧΙΤΕΚΤΟΝΙΚΗ ΣΥΣΤΗΜΑΤΟΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

Σε αυτό το κεφάλαιο αναλύονται τα βασικά μέρη των συστήματος μηχανικής μάθησης για την ταξινόμηση κειμένου.

2.1 Εισαγωγή

Για την κατασκευή ενός αυτοματοποιημένου συστήματος ταξινόμησης κειμένου πρέπει να ακολουθηθεί μια σειρά βημάτων για την εκπαίδευση αλλά και την δοκιμή του. Πολύ σημαντικός παράγοντας σε ολόκληρη την διαδικασία είναι η εξασφάλιση της πηγής των δεδομένων και η ανάκτησή τους, με σκοπό την τροφοδότηση του συστήματος. Τα κυριότερα βήματα για την ανάπτυξη ενός τέτοιου συστήματος, έχοντας ήδη εξασφαλίσει το σετ δεδομένων μας, είναι τα εξής:



Σχήμα 2.1: Διάγραμμα ροής ενός συστήματος ταξινόμησης

Το σχήμα 2.1 χωρίζεται σε δύο μέρη: στο μέρος της Εκπαίδευσης και στο μέρος της Πρόβλεψης, τα οποία είναι και οι κύριες διεργασίες για την κατασκευή ενός τέτοιου συστήματος. Σε γενικές γραμμές, ένα σετ δεδομένων χωρίζεται σε δύο ή τρία μέρη. Το κάθε μέρος χρησιμοποιείται για διαφορετική χρήση: το πρώτο χρησιμοποιείται για εκπαίδευση, το δεύτερο για επαλήθευση (προαιρετικό) και το τρίτο για την δοκιμή του συστήματος. Στο σχήμα φαίνεται μια επικάλυψη της

Κανονικοποίησης Κειμένου και της Εξαγωγής Features, η οποίο δηλώνει ότι ανεξάρτητα από το έγγραφο που θα κατηγοριοποιηθεί και προβλεφθεί η κλάση του, πρέπει να περάσει την ίδια σειρά μετασχηματισμών στις διεργασίες εκπαίδευσης και πρόβλεψης. Αρχικά, κάθε έγγραφο δέχεται μια προ-επεξεργασία και κανονικοποιείται, τότε εξάγονται συγκεκριμένα features σχετικά με το έγγραφο. Αυτές οι διεργασίες γίνονται πάντα και στα δύο μέρη για να εξασφαλιστεί πως το μοντέλο ταξινόμησης εκτελεί με συνέπεια τις προβλέψεις του. Στην διεργασία της Εκπαίδευσης, κάθε έγγραφο έχει τη δική του αντίστοιχη κλάση/κατηγορία που του δόθηκε χειροκίνητα ή επιμελήθηκαν προγενέστερα. Αυτά τα έγγραφα εκπαίδευσης επεξεργάζονται και κανονικοποιούνται στο βήμα Κανονικοποίησης Κειμένου, δίνοντας καθαρισμένο και τυποποιημένο κείμενο προς εκπαίδευση. Τα αποτελέσματα δίνονται στη φάση Εξαγωγής Features όπου διαφορετικά είδη τεχνικών χρησιμοποιούνται για την εξαγωγή features με νόημα από τα επεξεργασμένα κείμενα. Αυτά τα features είναι συνήθως αριθμητικοί πίνακες ή διανύσματα διότι οι κλασικοί αλγόριθμοι Μηχανικής Μάθησης δουλεύουν πάνω σε αριθμητικά διανύσματα. Από την στιγμή που θα παραχθούν αυτά τα features, επιλέγεται ένας αλγόριθμος Μηχανικής Μάθησης για την εκπαίδευση του μοντέλου.

Η εκπαίδευση του μοντέλου περιλαμβάνει την τροφοδότηση των διανυσμάτων με τα features των εγγράφων και τα αντίστοιχα labels, έτσι ώστε ο αλγόριθμος να μπορέσει να «μάθει» διάφορα μοτίβα που αντίστοιχα με την κάθε κλάση/κατηγορία και να μπορέσει ξαναχρησιμοποιήσει αυτήν την γνώση για να προβλέψει τις κλάσεις νέων εγγράφων. Αρκετά συχνά χρησιμοποιείται, προαιρετικά, ένα σετ δεδομένων για την αξιολόγηση της επίδοσης του αλγόριθμου ταξινόμησης για να εξασφαλιστεί ότι γενικεύει σε ικανοποιητικό βαθμό τα δεδομένα κατά την διάρκεια της εκπαίδευσης. Ο συνδυασμός αυτών των features και του αλγόριθμου Μηχανικής Μάθησης αποτελούν το Μοντέλο Ταξινόμησης, το οποίο είναι και το τελευταίο σκέλος του μέρους της Εκπαίδευσης. Πολλές φορές, αυτό το μοντέλο «κουρδίζεται» χρησιμοποιώντας διάφορες παραμέτρους του μοντέλου με την διαδικασία να ονομάζεται hyperparameter tuning, με σκοπό την επίτευξη της βέλτιστης επίδοσης.

Η διαδικασία της Πρόβλεψης περιλαμβάνει είτε την προσπάθεια πρόβλεψης της κλάσης νέων εγγράφων είτε την αξιολόγηση των προβλέψεων στο σετ δεδομένων προς δοκιμή. Το δοκιμαστικό αυτό σετ δεδομένων περνάει από την ίδια διαδικασία κανονικοποίησης και εξαγωγής features, με τα εξαγόμενα features να περνιούνται

στο εκπαιδευμένο Μοντέλο Ταξινόμησης, το οποίο προβλέπει την πιθανή κλάση για κάθε έγγραφο με βάση τα μοτίβα που «έμαθε» προηγουμένως. Αν τα έγγραφα έχουν χαρακτηριστεί με κάποια κλάση χειροκίνητα, τότε μπορεί να αξιολογηθεί η επίδοση του μοντέλου με την σύγκριση των αρχικών labels με τις προβλέψεις με διάφορες μετρικές όπως η ακρίβεια (accuracy). Αυτές οι μετρικές δίνουν μια εικόνα για το πόσο καλά μπορεί να προβλέψει το μοντέλο νέα έγγραφα.

2.2 Κανονικοποίηση Κειμένου - Προεπεξεργασία

Σε αυτήν την ενότητα, αναλύουμε το κομμάτι που θα κάνει μια προεπεξεργασία του κειμένου πριν αυτό περάσει στον classifier. Με αυτή την κανονικοποίηση του κειμένου πετυχαίνουμε μια ομοιομορφία στα δεδομένα που θα μπορέσει να βοηθήσει το μοντέλο στην εξαγωγή features. Επίσης, επιτυγχάνεται η αποφυγή τροφοδότησης περιττής πληροφορίας στον classifier ώστε το σύστημα να έχει μεγαλύτερη ταχύτητα αλλά και ακρίβεια στον εντοπισμό της χρήσιμης πληροφορίας. Αν και υπάρχουν πολλές και διαφορετικές τεχνικές, εμείς επιλέξαμε και δοκιμάσαμε διάφορους συνδυασμούς των παρακάτω:

- Αφαίρεση των hyperlinks
- Αφαίρεση αριθμών
- Αφαίρεση αρκετά χρησιμοποιούμενων λέξεων (stop words)
- Αφαίρεση σημείων στίξης μέσα στην πρόταση
- Αφαίρεση σημείων στίξης στο τέλος της πρότασης
- Αφαίρεση καταλήξεων (stemming)
- Μετατροπή όλων των γραμμάτων σε πεζά

2.3 Εξαγωγή Χαρακτηριστικών

Υπάρχουν πολλές τεχνικές για την εξαγωγή features που μπορούν να εφαρμοστούν σε δεδομένα που αποτελούνται από κείμενα, αλλά πρώτα ας αναφερθούμε στο τι εννοούμε με τον όρο features, γιατί μας χρειάζονται και ποια είναι η χρησιμότητά τους. Σε ένα σετ δεδομένων, συνήθως υπάρχουν πολλά σημεία δεδομένων. Τις περισσότερες φορές οι γραμμές του σετ δεδομένων και οι στήλες είναι διάφορα features ή ιδιότητες του σετ δεδομένων, με συγκεκριμένες τιμές για κάθε γραμμή ή παρατήρηση. Στην ορολογία της Μηχανικής Μάθησης, τα features είναι μοναδικά, μετρήσιμα στοιχεία ή ιδιότητες για κάθε παρατήρηση ή σημείο δεδομένων σε ένα σετ δεδομένων. Τα features είναι, συνήθως, αριθμητικά και μπορούν να είναι απόλυτες αριθμητικές τιμές ή κατηγορικά features που μπορούν να κωδικοποιηθούν ως δυαδικά για κάθε κατηγορία στη λίστα. Η διαδικασία εξαγωγής και επιλογής των features είναι αποτέλεσμα αναλυτικής στατιστικής ανάλυσης και επιτήρησης των δεδομένων, και ονομάζεται feature extraction ή feature engineering.

Τα εξαγόμενα features τροφοδοτούν τον αλγόριθμο Μηχανικής Μάθησης για την εκμάθηση μοτίβων που θα μπορούν να εφαρμοστούν σε μελλοντικά καινούργια σημεία δεδομένων για την απόκτηση πληροφοριών. Οι αλγόριθμοι αυτοί, περιμένουν features με την μορφή αριθμητικών διανυσμάτων επειδή κάθε αλγόριθμος είναι, κατά κύριο λόγο, μια μαθηματική λειτουργία βελτιστοποίησης και ελαχιστοποίησης απώλειας και λάθους όταν κάνει την προσπάθεια εκμάθησης μοτίβων από σημεία δεδομένων και παρατηρήσεων. Για αυτόν τον λόγο, στα δεδομένα κειμένων υπάρχει μια επιπλέον δυσκολία στον μετασχηματισμό των δεδομένων και στην εξαγωγή αριθμητικών features από αυτόν.

Παρακάτω θα αναλυθούν κάποιες τεχνικές εξαγωγής features που χρησιμοποιούνται, ειδικά, για δεδομένα κειμένων. Οι τεχνικές που χρησιμοποιήθηκαν και θα αναλυθούν είναι οι εξής:

- Το μοντέλο Vector Space
- Το μοντέλο Bag of Words
- Το μοντέλο TF-IDF
- Το μοντέλο NLP

Κάτι που πρέπει να σημειωθεί για την εξαγωγή features είναι ότι από την στιγμή που θα κατασκευαστεί ένα τέτοιο μοντέλο με την χρήση κάποιων μετασχηματισμών και μαθηματικών λειτουργιών, πρέπει να εξασφαλιστεί ότι θα επαναχρησιμοποιηθεί η ίδια διαδικασία όταν θα γίνεται η εξαγωγή features από τα νέα έγγραφα και δεν θα κατασκευαστεί εκ νέου όλος ο αλγόριθμος βασιζόμενος στα νέα έγγραφα αυτή την φορά.

2.3.1 Μοντέλο Vector Space

Το μοντέλο Vector Space είναι μια έννοια και ένα μοντέλο που είναι πολύ χρήσιμο στις περιπτώσεις που πρέπει να αντιμετωπιστούν δεδομένα κειμένων και είναι πολύ δημοφιλές στην ανάκτηση πληροφοριών και στην κατάταξη εγγράφων. Το μοντέλο Vector Space (ή Term Vector μοντέλο) ορίζεται ως ένα μαθηματικό και αλγεβρικό μοντέλο για τον μετασχηματισμό και την αναπαράσταση κειμένων ως αριθμητικά διανύσματα συγκεκριμένων όρων, τα οποία αποτελούν τις διαστάσεις του διανύσματος. Μαθηματικά αυτό ορίζεται ως εξής:

Έστω ότι υπάρχει ένα έγγραφο D σε έναν διανυσματικό χώρο εγγράφων VS. Ο αριθμός των διαστάσεων ή στηλών του κάθε εγγράφου θα είναι ο αριθμός του συνόλου των διακριτών όρων ή λέξεων όλων των εγγράφων στον διανυσματικό χώρο.

Άρα, ο διανυσματικός χώρος μπορεί να δηλωθεί ως:

$$VS = \{W_1, W_2, \dots, W_n\}$$

όπου υπάρχουν n διακριτές λέξεις σε όλα τα έγγραφα

Σε αυτόν τον διανυσματικό χώρο μπορεί να δηλωθεί το έγγραφο D ως:

$$D = \{W_{D1}, W_{D2}, \dots, W_{Dn}\}$$

όπου το WDn δηλώνει το βάρος της λέξης στο έγγραφο D

Αυτό το βάρος είναι μια αριθμητική τιμή και μπορεί να αναπαριστά οτιδήποτε, από την συχνότητα εμφάνισης της λέξης στο έγγραφο μέχρι το μέσο όρο της συχνότητας εμφάνισής της ή ακόμα και το βάρος του TF-IDF.

2.3.2 Μοντέλο Bag of Words

Το μοντέλο Bag of Words είναι, ίσως, ένα από τις πιο απλές αλλά και πιο ισχυρές τεχνικές για την εξαγωγή features από κείμενα. Η κύρια ιδέα του μοντέλου αυτού

είναι η μετατροπή των κειμένων σε διανύσματα ώστε κάθε διάνυσμα να αναπαριστά την συχνότητα όλων των διακριτών λέξεων που υπάρχουν στον διανυσματικό χώρο εγγράφων για το συγκεκριμένο έγγραφο. Έτσι, θεωρώντας το διάνυσμα από την προηγούμενη μαθηματική εξίσωση για το έγγραφο D, το βάρος της κάθε λέξης θα ισούται με την συχνότητα εμφάνισής της στο έγγραφο.

Επιπλέον, δίνεται η δυνατότητα να δημιουργηθεί ένα ίδιο μοντέλο για τις εμφανίσεις κάθε λέξης ή για τις εμφανίσεις n λέξεων (n-grams), κάτι που θα έκανε το μοντέλο n-gram Bag of Words ώστε η συχνότητα των διακριτών n-grams σε κάθε έγγραφο να λαμβάνεται υπόψη.

Το μοντέλο bag-of-words είναι μια απλοποιημένη αναπαράσταση που χρησιμοποιείται στην Επεξεργασία Φυσική Γλώσσας (Natural Language Processing – NLP) και στην Εξαγωγή Πληροφορίας (Information Retrieval – IR). Στο συγκεκριμένο μοντέλο, ένα κείμενο αναπαρίσταται ως η σακούλα (multiset) των λέξεων του, αγνοώντας την γραμματική και την σειρά των λέξεων αλλά κρατώντας την πολλαπλότητά τους.

Στα μαθηματικά, ένα multiset (ή σακούλα) είναι μια γενίκευση της έννοιας του set, στο οποίο επιτρέπονται οι πολλαπλές εμφανίσεις ενός στοιχείου. Για παράδειγμα, {a, a, b} και {a, b} είναι διαφορετικά multisets αν και είναι στο ίδιο set. Διαφέρουν στην πολλαπλότητα του στοιχείου a. Η πολλαπλότητα ενός στοιχείου είναι ο αριθμός των εμφανίσεων του συγκεκριμένου στοιχείου σε ένα multiset.

Το μοντέλο bag of words χρησιμοποιείται συχνά σε μεθόδους ταξινόμησης κειμένων όπου η εμφάνιση (ή και η συχνότητα εμφάνισης) κάθε λέξης χρησιμοποιείται ως feature στην εκπαίδευση ενός classifier.

Παράδειγμα

Έστω ότι έχουμε τις παρακάτω προτάσεις:

(1) John likes to watch movies. Mary likes movies too.

(2) John also likes to watch football games.

Από αυτές τις δυο προτάσεις, κατασκευάζεται η παρακάτω λίστα:

```
[ "John",
  "likes",
  "to",
  "watch",
```

"movies",
"also",
"football",
"games",
"Mary",
"too"]

Στην πράξη, το μοντέλο bag-of-words χρησιμοποιείται, κατά κύριο λόγο, ως εργαλείο για την παραγωγή features. Μετά τον μετασχηματισμό του κειμένου σε «σακούλα με λέξεις», μπορούν να υπολογιστούν διάφορα μέτρα για τον χαρακτηρισμό του κειμένου. Οι πιο συνηθισμένοι τύποι χαρακτηριστικών ή features που υπολογίζονται από το μοντέλο bag-of-words είναι: η συχνότητα ενός όρου, και συγκεκριμένα, ο αριθμός των φορών που ένας όρος εμφανίζεται στο κείμενο. Για τα παραπάνω παραδείγματα, κατασκευάζονται οι παρακάτω λίστες για την καταγραφή της συχνότητας των όρων όλων των μοναδικών λέξεων:

(1) [1, 2, 1, 1, 2, 0, 0, 0, 1, 1]

(2) [1, 1, 1, 1, 0, 1, 1, 1, 0, 0]

Κάθε καταχώρηση στις λίστες αναφέρεται στον αριθμό των εμφανίσεων της αντίστοιχης λέξης. Για παράδειγμα, η πρώτη καταχώρηση («1») δηλώνει ότι η λέξη «John» (η οποία είναι η πρώτη καταχώρηση στην σακούλα) εμφανίζεται μόνο μια φορά στην πρώτη πρόταση ενώ η δεύτερη καταχώρηση («2») ότι η λέξη «likes» εμφανίζεται δυο φορές.

Πιο αναλυτικά:

	John	**likes**	**to**	**watch**	**movies**	**also**	**football**	**games**	**Mary**	**too**
(1)	1	2	1	1	2	0	0	0	1	1
(2)	1	1	1	1	0	1	1	1	0	0

Κείμενο	**Αναπαράσταση BoW**
John likes to watch movies. Mary likes movies too.	[1, 2, 1, 1, 2, 0, 0, 0, 1, 1]
John also likes to watch football games.	[1, 1, 1, 1, 0, 1, 1, 1, 0, 0]

Πίνακας 2.1: Αναλυτική αναπαράσταση κειμένου στο μοντέλο bag of words

Παρ ’όλα αυτά, η συχνότητα όρων δεν είναι πάντα η καλύτερη αναπαράσταση ενός κειμένου. Αυτό συμβαίνει γιατί λέξεις όπως προθέσεις, άρθρα, σύνδεσμοι κλπ. έχουν, σχεδόν πάντα, την υψηλότερη συχνότητα σε ένα κείμενο. Για αυτόν τον λόγο, η μέτρηση των εμφανίσεων μιας λέξης δεν σημαίνει ότι η συγκεκριμένη λέξη είναι και σημαντική. Για να αντιμετωπιστεί αυτό το πρόβλημα, μια από τις πιο δημοφιλείς λύσεις «κανονικοποίησης» της συχνότητας όρων είναι η κατανομή βάρους σε έναν όρο με την αντίστροφη συχνότητα εγγράφων (Term Frequency – Inverse Document Frequency, TF-IDF).

2.3.3 Μοντέλο n-gram

Το bag of word μοντέλο είναι μια αναπαράσταση ενός εγγράφου χωρίς σειρά. Για παράδειγμα, στην παραπάνω πρόταση (1) η αναπαράσταση bag-of-words δεν επισημαίνει το γεγονός ότι το όνομα ενός ατόμου ακολουθείται πάντα από το ρήμα «likes» στο κείμενο. Ως εναλλακτική, το μοντέλο n-gram μπορεί να χρησιμοποιηθεί για την αποθήκευση αυτής της πληροφορίας που υπάρχει στο κείμενο.

Πιο συγκεκριμένα, ένα n-gram είναι μια συνεχόμενη ακολουθία n αντικειμένων μιας δοθείσας ακολουθίας κειμένου ή ομιλίας. Τα αντικείμενα μπορεί να είναι φωνήματα, συλλαβές, γράμματα ή λέξεις ανάλογα την εφαρμογή.

Εφαρμόζοντας, λοιπόν, ένα bigram μοντέλο στο παραπάνω παράδειγμα (1), το κείμενο θα αναλυθεί ως εξής:

["John likes",
"likes to",
"to watch",
"watch movies",
"Mary likes",
"likes movies",
"movies too"]

Ουσιαστικά, μπορούμε να δούμε το μοντέλο bag-of-words ως μια ειδική περίπτωση του n-gram μοντέλου όπου το n=1.

Το μοντέλο bag of words έχει ένα βασικό μειονέκτημα: παράγει αραιούς πίνακες (sparse matrices). Ένας πίνακας λέγεται αραιός αν ένα μεγάλο ποσοστό των στοιχείων του έχουν μηδενική τιμή. Σε αντίθεση, αν τα περισσότερα στοιχεία του

πίνακα είναι μη μηδενικά, τότε ο πίνακας θεωρείται πυκνός (dense). Δεν υπάρχει ακριβές ποσοστό σε σχέση με τον αριθμό των μηδενικών στοιχείων, επάνω από το οποίο ένας πίνακας χαρακτηρίζεται ως αραιός. Αρκεί όμως, για παράδειγμα, να πούμε ότι με περισσότερο από 80% μηδενικά ένας πίνακας χαρακτηρίζεται ως αραιός. Αραιοί πίνακες συναντώνται συχνά σε μεγάλα επιστημονικά προβλήματα (επίλυση εξισώσεων κ.λπ.). Το πρόβλημα με τη διαχείριση των αραιών πινάκων είναι ότι δαπανάται πολύ χώρος για την αποθήκευση μηδενικών.

Στο μοντέλο bag of words, οι αραιοί πίνακες είναι ένα συχνό φαινόμενο διότι κάθε έγγραφο αναπαρίσταται ως ένα διάνυσμα τιμών. Αυτό το διάνυσμα μπορεί να αποτελείται είτε από δυαδικές τιμές (οι οποίες δηλώνουν την παρουσία ή όχι μιας λέξης) είτε από απόλυτες τιμές (οι οποίες δηλώνουν την συχνότητα εμφάνισης) είτε από κανονικοποιημένες τιμές. Για αυτόν τον λόγο όταν τα διανύσματα των features είναι αραιά, τότε και ο πίνακας θα είναι αραιός. Το αν θα είναι αραιά τα διανύσματα των features εξαρτάται στο μέγεθος του λεξιλογίου, την έκταση και την ποικιλία των εγγράφων. Για παράδειγμα, αν έχουμε ένα dataset πολύ μικρών και παρόμοιων εγγράφων, θα μπορούσαμε να έχουμε έναν πυκνό πίνακα, κάτι το οποίο είναι πολύ σπάνιο στην πράξη.

Ας υποθέσουμε ότι έχουμε τις παρακάτω 3 προτάσεις:

1) Hello World, the sun is shining
2) Hello world, the weather is nice
3) Hello world, the wind is cold

Τότε το λεξιλόγιό μας (ας υποθέσουμε ότι είναι 1-gram και χωρίς αφαίρεση των stop words) θα ήταν κάπως έτσι:

[hello, world, the, wind, weather, sun, is, shining, nice, cold]

Τα δυαδικά διανύσματα των features θα ήταν:

1) [1, 1, 1, 0, 0, 0, 1, 1, 0, 0]
2) [1, 1, 1, 0, 0, 1, 0, 1, 1, 0]
3) [1, 1, 1, 1, 0, 0, 1, 0, 0, 1]

Και ο τελικός πίνακας θα ήταν:

[[1, 1, 1, 0, 0, 0, 1, 1, 0, 0]
[1, 1, 1, 0, 0, 1, 0, 1, 1, 0]
[1, 1, 1, 1, 0, 0, 1, 0, 0, 1]]

Αναλυτικότερα:

	hello	**world**	**the**	**wind**	**weather**	**sun**	**is**	**shining**	**nice**	**cold**
1)	1	1	1	0	0	0	1	1	0	0
2)	1	1	1	0	0	1	0	1	1	0
3)	1	1	1	1	0	0	1	0	0	1

Κείμενο	**Αναπαράσταση BoW**
Hello World, the sun is shining	[1, 1, 1, 0, 0, 0, 1, 1, 0, 0]
Hello world, the weather is nice	[1, 1, 1, 0, 0, 1, 0, 1, 1, 0]
Hello world, the wind is cold	[1, 1, 1, 1, 0, 0, 1, 0, 0, 1]
Τελικός Πίνακας	[[1, 1, 1, 0, 0, 0, 1, 1, 0, 0] [1, 1, 1, 0, 0, 1, 0, 1, 1, 0] [1, 1, 1, 1, 0, 0, 1, 0, 0, 1]]

Πίνακας 2.2: Παράδειγμα μετατροπής κειμένων σε διάνυσμα με τη χρήση του μοντέλου bag of words

Όπως φαίνεται, έχουμε 17x1 και 13x0: κάτι, που εξ ορισμού, δεν μπορεί να είναι αραιός πίνακας. Παρόλα αυτά, δεν είναι ένα πιθανό σενάριο σε μια πραγματική εφαρμογή.

2.3.4 Μοντέλο TF-IDF

Αν και το Bag of Words μοντέλο είναι καλό, τα διανύσματα είναι εξ ολοκλήρου βασισμένα στις απόλυτες συχνότητες εμφανίσεων των λέξεων. Αυτό δημιουργεί κάποια πιθανά προβλήματα όταν λέξεις που εμφανίζονται συχνά σε όλα τα έγγραφα, θα έχουν υψηλότερες συχνότητες και θα επισκιάζουν άλλες λέξεις που μπορεί να μην εμφανίζονται αλλά ίσως είναι πιο σημαντικές και αποτελεσματικές ως features για την αναγνώριση κάποιων κατηγοριών των εγγράφων. Τα προβλήματα αυτά έρχεται να λύσει ο TF-IDF (Term Frequency – Inverse Document Frequency).

Ο TF-IDF είναι ένα αριθμητικό στατιστικό που έχει ως σκοπό να αντικατοπτρίσει πόσο σημαντική είναι μια λέξη τόσο βάση του αριθμού εμφάνισής της στο κείμενο όσο και στο σύνολο των κειμένων. Η τιμή του TF-IDF αυξάνεται αναλογικά με τον αριθμό των φορών που μια λέξη εμφανίζεται σε ένα έγγραφο, αλλά συχνά

αντισταθμίζεται από την συχνότητα της λέξης στην συλλογή των εγγράφων, κάτι που βοηθάει στην προσαρμογή του γεγονότος ότι μερικές λέξεις εμφανίζονται πιο συχνά γενικότερα.

Ο TF-IDF είναι συνδυασμός δύο μετρικών: της συχνότητας όρου και της αντίστροφης συχνότητας εγγράφου. Αυτή η τεχνική είχε αναπτυχθεί αρχικά ως μια μετρική για την κατάταξη συναρτήσεων στην απεικόνιση αποτελεσμάτων των μηχανών αναζήτησης και, πλέον, έχει γίνει μέρος της ανάκτησης πληροφοριών και της εξαγωγής features σε κείμενα.

Μαθηματικά, ο TF-IDF είναι παράγωγο δύο μετρικών και μπορεί να γραφεί ως:

$$tfidf = tf \times idf$$

όπου tf (term frequency) και idf (inverse – document frequency) είναι οι δύο μετρικές

Η συχνότητα όρου (term frequency) είναι ο υπολογισμός που γίνεται και στο μοντέλο Bag of Words. Η συχνότητα όρου σε οποιοδήποτε διάνυσμα εγγράφου εκφράζεται από την τιμή της συχνότητας ενός συγκεκριμένου όρου σε ένα συγκεκριμένο έγγραφο.

Μαθηματικά μπορεί να αναπαρασταθεί ως:

$$tf(w, D) = f_{wD}$$

όπου fwD είναι η συχνότητα της λέξης w στο έγγραφο D

Υπάρχουν πολλές άλλες αναπαραστάσεις και υπολογισμοί της συχνότητας όρου, όπως η μετατροπή της συχνότητας σε δυαδικό feature όπου το 1 σημαίνει την ύπαρξη του όρου στο έγγραφο και το 0 την μη ύπαρξή του. Μερικές φορές μπορεί να κανονικοποιηθεί η αρχική τιμή της συχνότητας χρησιμοποιώντας λογάριθμους ή τον μέσο όρο της συχνότητας.

Η αντίστροφη συχνότητα εγγράφου (inverse – document frequency) υπολογίζεται με την διαίρεση του συνόλου των εγγράφων στο σετ δεδομένων προς τη συχνότητα του κάθε όρου στο έγγραφο και εφαρμόζοντας λογαριθμική κλιμάκωση στο αποτέλεσμα της διαίρεσης.

Μαθηματικά μπορεί να αναπαρασταθεί ως εξής:

$$idf(t) = \log \frac{C}{df(t)}$$

Όπου idf(t) είναι η αντίστροφη συχνότητα εγγράφου του όρου t, C το σύνολο των εγγράφων στο σετ δεδομένων και df(t) η συχνότητα του αριθμού των εγγράφων στα οποία εμφανίζεται ο όρος t.

Οπότε η τιμή του TF-IDF μπορεί να υπολογιστεί με τον πολλαπλασιασμό των δύο παραπάνω εξισώσεων (tf και idf).

2.3.5 Μοντέλο NLP

Το μοντέλο αυτό βασίζεται στην εξόρυξη πληροφορίας κατευθείαν από το κείμενο. Αυτό προϋποθέτει διαφορετική επεξεργασία στο ίδιο το κείμενο για την εύρεση features που θα βοηθήσουν στην ταξινόμηση κάθε καινούργιου κειμένου. Τέτοια features μπορεί να είναι:

- Το ποσοστό των κεφαλαίων γραμμάτων σε μια πρόταση
- Το ποσοστό χρήσης κάποιων μερών του λόγου
- Τα ίδια συνεχόμενα γράμματα
- Η εύρεση συγκεκριμένων λέξεων
- Η σωστή ορθογραφία των λέξεων

Συνεπώς, κρίσιμο κομμάτι για βελτίωση της απόδοσης είναι η επιλογή σωστών feature για την μεγιστοποίηση της απόδοσης του αλγορίθμου.

2.4 Αλγόριθμοι Ταξινόμησης

2.4.1 Εισαγωγή

Οι αλγόριθμοι ταξινόμησης είναι αλγόριθμοι εποπτευόμενης Μηχανικής Μάθησης που χρησιμοποιούνται για την ταξινόμηση, ταξινόμηση ή την προσθήκη ετικετών σε σημεία δεδομένων βασιζόμενοι σε παρατηρήσεις του παρελθόντος. Κάθε αλγόριθμος ταξινόμησης, όντας ένας αλγόριθμος εποπτευόμενης εκμάθησης, χρειάζεται δεδομένα για εκπαίδευση. Αυτά τα δεδομένα αποτελούνται από ένα σετ παρατηρήσεων εκπαίδευσης όπου κάθε παρατήρηση είναι ένα ζευγάρι που αποτελείται από ένα σημείο δεδομένων εισόδου (συνήθως διάνυσμα features) και το αντίστοιχο αποτέλεσμα εξόδου για αυτή την παρατήρηση εισόδου. Υπάρχουν τρεις βασικές διαδικασίες που περνάνε οι αλγόριθμοι ταξινόμησης:

- Η εκπαίδευση είναι η διαδικασία που ο αλγόριθμος εποπτευόμενης εκμάθησης αναλύει και προσπαθεί να ανακαλύψει μοτίβα από τα δεδομένα εκπαίδευσης ώστε να μπορεί να αναγνωρίσει ποια μοτίβα οδηγούν στο συγκεκριμένο αποτέλεσμα. Αυτά τα αποτελέσματα είναι γνωστά ως ετικέτες κλάσης/μεταβλητές κλάσης. Συνήθως εκτελείται η εξαγωγή features ή το features engineering για να παραχθούν features

με κάποια αξία από τα δεδομένα πριν την εκπαίδευση. Αυτά τα σετ features τροφοδοτούν τον αλγόριθμο επιλογής μας, ο οποίος προσπαθεί να αναγνωρίσει και να μάθει μοτίβα από αυτά και τα αντίστοιχα αποτελέσματά τους. Το αποτέλεσμα είναι μια συναγόμενη λειτουργία γνωστή ως μοντέλο ή μοντέλο ταξινόμησης. Αυτό το μοντέλο αναμένεται να έχει γενικευτεί αρκετά από τα μοτίβα που έμαθε από τα δεδομένα εκπαίδευσης ώστε να μπορεί να προβλέψει τις κλάσεις ή τα αποτελέσματα των νέων σημείων δεδομένων στο μέλλον.

- Η αξιολόγηση περιλαμβάνει την προσπάθεια δοκιμής της επίδοσης των προβλέψεων του μοντέλου για την βαθμολόγηση της εκπαίδευσης και της εκμάθησης από το σετ δεδομένων προς εκπαίδευση. Για αυτό χρησιμοποιείται ένα σετ δεδομένων επαλήθευσης (validation dataset), ώστε να δοκιμαστεί η επίδοση του μοντέλου κάνοντας προβλέψεις πάνω στο συγκεκριμένο σετ δεδομένων και επαληθεύοντας αυτές τις προβλέψεις με τις αληθινές κλάσεις τους. Επίσης, αρκετά συχνά χρησιμοποιείται και η τεχνική της διασταυρωμένης επαλήθευσης (cross-validation), κατά την οποία τα δεδομένα διαιρούνται σε υποσύνολα (folds) των οποίων ένα μεγάλο κομμάτι χρησιμοποιείται για εκπαίδευση, ενώ τα υπολειπόμενα δεδομένα χρησιμοποιούνται για την επαλήθευση του εκπαιδευόμενου μοντέλου. Να σημειωθεί πως, κάθε φορά, το μοντέλο συντονίζεται με βάση τα αποτελέσματα της επαλήθευσης για την επίτευξη της βέλτιστης σύνθεσης που θα αποφέρει τη μέγιστη ακρίβεια και το ελάχιστο ποσοστό λάθους. Επίσης, το μοντέλο αξιολογείται με ένα σετ δεδομένων προς δοκιμή (test dataset), με τη διαφορά ότι δεν θα συντονιστεί με τα αποτελέσματα αυτής της αξιολόγησης καθώς είναι πιθανό να το καταστήσει μεροληπτικό ή να επιφέρει υπερφόρτωση (overfit) σε πολύ συγκεκριμένα features του σετ δεδομένων. Αυτό το σετ δεδομένων προς δοκιμή είναι ένα αντιπροσωπευτικό δείγμα από τα νέα, πραγματικά δεδομένα στα οποία το μοντέλο θα κάνει προβλέψεις και τι επίδοση θα έχει σε αυτά τα νέα δεδομένα.
- Ο συντονισμός, γνωστός και ως hyperparameter tuning, είναι η διαδικασία η οποία εστιάζει στην προσπάθεια βελτιστοποίησης του μοντέλου. Η βελτιστοποίηση περιλαμβάνει την μεγιστοποίηση της

ικανότητας πρόβλεψης και την ελαχιστοποίηση των λαθών. Κάθε μοντέλο είναι μια μαθηματική συνάρτηση με πολλές παραμέτρους που καθορίζουν την πολυπλοκότητα, την μαθησιακή ικανότητα και άλλα χαρακτηριστικά του μοντέλου. Αυτές οι παράμετροι ονομάζονται hyperparameters επειδή δεν μπορούν να μαθευτούν απευθείας από τα δεδομένα και πρέπει να δηλωθούν πριν την λειτουργία και την εκπαίδευση του μοντέλου. Ως εκ τούτου, η διαδικασία επιλογής των βέλτιστων hyperparameters του μοντέλου για την επίτευξη υψηλής ακρίβειας στις προβλέψεις ονομάζεται συντονισμός του μοντέλου. Ο συντονισμός επιτυγχάνεται με διάφορους τρόπου και τεχνικές όπως της τυχαιοποιημένης αναζήτησης (randomized search) και της αναζήτησης πλέγματος (grid search).

Υπάρχουν πολλοί τύποι αλγόριθμων ταξινόμησης, εμείς όμως θα εστιάσουμε σε εκείνους που χρησιμοποιούνται κατά κύριο λόγο στην ταξινόμηση κειμένων και εκείνους που χρησιμοποιήσαμε στο σύστημά μας. Οι αλγόριθμοι αυτοί είναι:

1) Naive Bayes
2) Support Vector Machines
3) Decision Trees
4) Random Forest
5) K Nearest Neighbors
6) Logistic Regression
7) Neural Networks
8) Gradient Boosting

Επίσης, υπάρχουν συνδυαστικές τεχνικές που ονομάζονται ensemble, οι οποίες χρησιμοποιούν τον συνδυασμό δύο ή περισσότερων μοντέλων για την εκμάθηση και πρόβλεψη αποτελεσμάτων. Οι συνδυαστικές τεχνικές είναι δύσκολες στην επίτευξη καλών αποδόσεων γιατί είναι αρκετά επιρρεπείς στην υπερφόρτωση.

2.4.2 Naive Bayes

Στην πραγματικότητα, ο naive Bayes δεν είναι ένας μοναδικός αλγόριθμος για την εκπαίδευση ενός classifier. Είναι μια οικογένεια αλγορίθμων που βασίζονται σε μια κοινή αρχή. Ο naive Bayes είναι ένας αλγόριθμος εποπτευόμενης μάθησης που

χρησιμοποιεί το διάσημο θεώρημα του Bayes, κάνοντας μια «αφελή» υπόθεση πως κάθε feature είναι ανεξάρτητο από τα άλλα. Στην θεωρία των πιθανοτήτων, ανεξάρτητες ονομάζονται δυο τιμές όταν η παρουσία της μιας δεν επηρεάζει την πιθανότητα εμφάνισης της άλλης.

Μαθηματικά, το θεώρημα του Bayes γράφεται ως εξής:

Δεδομένης μιας κλάσης y και ενός σετ n features σε μορφή διανύσματος {x1, x2, …, xn), χρησιμοποιώντας το θεώρημα του Bayes η πιθανότητα εμφάνισης της κλάσης y με τα δεδομένα features θα είναι:

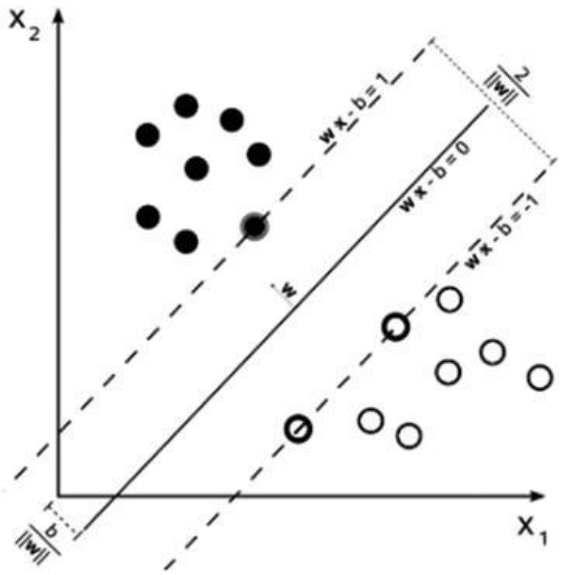
$$P(y|x_1, x_2, \dots, x_n) = \frac{P(y) \times P(x_1, x_2, \dots, x_n|y)}{P(x_1, x_2, \dots, x_n)}$$

Αυτός ο classifier πολλές φορές χαρακτηρίζεται ως απλός, όπως φαίνεται και από το όνομά του, αλλά και από τις αρκετές υποθέσεις που γίνονται στα δεδομένα και στα features που ίσως να μην ισχύουν σε πραγματικές συνθήκες. Παρ' όλα αυτά, ο αλγόριθμος αυτός δουλεύει αξιοσημείωτα καλά σε περιπτώσεις που σχετίζονται με ταξινόμηση, περιλαμβάνοντας την ταξινόμηση εγγράφων με πολλές κλάσεις, τα φίλτρα spam κλπ. Ένα μεγάλο πλεονέκτημα των μοντέλων naive Bayes είναι ότι χρειάζονται έναν μικρό αριθμό δεδομένων για εκπαίδευση για να κάνουν κάποια εκτίμηση των παραμέτρων που είναι απαραίτητες για την ταξινόμηση. Ως αποτέλεσμα, μπορεί να εκπαιδευτεί πολύ γρήγορα σε σύγκριση με άλλους classifiers και να δουλέψει αρκετά καλά χωρίς να έχουμε αρκετά δεδομένα για εκπαίδευση. Τα μοντέλα δεν αποδίδουν καλά όταν έχουν πολλά features και αυτό το φαινόμενο είναι γνωστό ως η κατάρα των διαστάσεων. Ο naive Bayes αντιμετωπίζει αυτής της φύσεως τα προβλήματα με τον διαχωρισμό των features σχετιζόμενων με την μεταβλητή κλάση και υπολογίζοντας κάθε υπόθεση ανεξάρτητα σαν να ήταν μοναδική. Ο πολυωνυμικός naive Bayes είναι μια επέκταση του naive Bayes όπου ο αριθμός των κλάσεων είναι παραπάνω από δύο. Σε αυτήν την περίπτωση τα διανύσματα των features υποτίθεται ότι είναι μετρητές λέξεων από το Bag of Words μοντέλο, αλλά και τα βάρη του TF-IDF μπορούν να δουλέψουν κάλλιστα. Ένας περιορισμός που υπάρχει είναι ότι αρνητικά features δεν μπορούν να δοθούν στον αλγόριθμο.

2.4.3 Support Vector Machines

Στην μηχανική μάθηση, τα Support Vector Machines (SVM) είναι αλγόριθμοι εποπτευόμενης μάθησης που χρησιμοποιούνται για ταξινόμηση (classification),

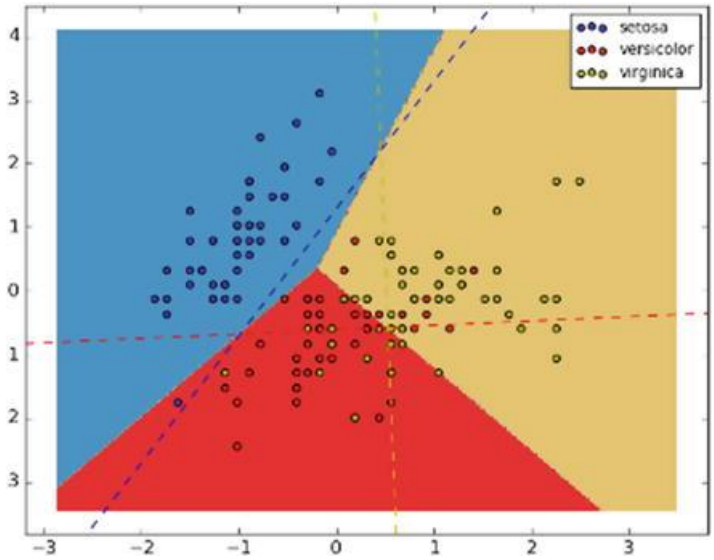
παλινδρόμηση (regression), καινοτομία (novelty) και εντοπισμό ανωμαλιών ή εξωστρέφειας (anomaly/outlier detection). Θεωρώντας ένα δυαδικό πρόβλημα ταξινόμησης, εάν υπάρχουν δεδομένα προς εκπαίδευση τέτοια ώστε κάθε σημείο δεδομένων ή παρατήρησης να ανήκει σε μια συγκεκριμένη κλάση, ο αλγόριθμος SVM μπορεί να εκπαιδευτεί με βάση αυτά τα δεδομένα με σκοπό να μπορεί να αναθέσει μελλοντικά σημεία δεδομένων σε μια από τις δύο κλάσεις. Αυτός ο αλγόριθμος αναπαριστά τα προς εκπαίδευση δεδομένα ως σημεία στον χώρο με τέτοιο τρόπο ώστε τα σημεία που ανήκουν σε κάποια κλάση να μπορούν να διαχωριστούν από ένα ευρύ κενό ανάμεσά τους, που ονομάζεται hyperplane. Με αυτόν τον διαχωρισμό επιτυγχάνεται η πρόβλεψη των καινούργιων σημείων δεδομένων με γνώμονα την πλευρά του hyperplane που έχουν τοποθετηθεί. Αυτή η διαδικασία ακολουθείται για μια κλασική περίπτωση γραμμικής ταξινόμησης. Παρ' όλα αυτά, ο SVM μπορεί, επίσης, να εκτελέσει περιπτώσεις μη γραμμικής ταξινόμησης με μια πολύ ενδιαφέρουσα προσέγγιση που ονομάζεται μέθοδος πυρήνα, όπου συναρτήσεις του πυρήνα χρησιμοποιούνται για την λειτουργία σε υψηλών διαστάσεων χώρους από features που είναι μη γραμμικά διαχωριζόμενοι. Ο αλγόριθμος SVM παίρνει ως είσοδο ένα σετ δεδομένων προς εκπαίδευση και κατασκευάζει ένα hyperplane με μια συλλογή από hyperplanes για έναν χώρο υψηλών διαστάσεων από features. Όσο μεγαλύτερα είναι τα περιθώρια του hyperplane, τόσο καλύτερος είναι και ο διαχωρισμός ώστε να οδηγεί σε μικρότερα ποσοστά λάθους στην γενίκευση του classifier.



Σχήμα 2.2: Χώρος διανυσμάτων με hyperplane

Στο σχήμα 2.2 φαίνεται ο SVM δύο κλάσεων που απεικονίζει το hyperplane και τα βοηθητικά διανύσματα. Υπάρχουν δύο βασικοί τύποι περιθωρίων που βοηθούν στον διαχωρισμό των σημείων δεδομένων που ανήκουν σε διαφορετικές κλάσεις. Όταν τα δεδομένα είναι γραμμικά διαχωρίσιμα, όπως στο παραπάνω σχήμα, υπάρχουν δύο «σκληρά» περιθώρια που απεικονίζονται από τα δύο παράλληλα hyperplanes με τις διακεκομμένες γραμμές, τα οποία βοηθούν στον διαχωρισμό των σημείων δεδομένων που ανήκουν στις δύο διαφορετικές κλάσεις. Αυτό επιτυγχάνεται λαμβάνοντας υπόψη ότι η απόσταση μεταξύ τους είναι όσο το δυνατόν μεγαλύτερη. Η περιοχή που περικλείεται από αυτά τα δύο hyperplanes αποτελεί το περιθώριο με το hyperplane του μέγιστου περιθωρίου να βρίσκεται στη μέση.

Για ένα πρόβλημα ταξινόμησης με πολλαπλές κλάσεις, αν υπάρχουν n κλάσεις, για κάθε κλάση εκπαιδεύεται ένας classifier που βοηθάει στον διαχωρισμό μεταξύ κάθε κλάσης και των άλλων n-1 κλάσεων. Κατά τη διάρκεια της πρόβλεψης, οι βαθμολογίες (αποστάσεις από τα hyperplanes) για κάθε classifier υπολογίζονται, και η μεγαλύτερη βαθμολογία επιλέγεται δίνοντας την αντίστοιχη κλάση ως πρόβλεψη.



Σχήμα 2.3: Εκπαίδευση τριών classifiers σε ένα πρόβλημα SVM τριών κλάσεων στο διάσημο σετ δεδομένων iris

Στο σχήμα 2.3 φαίνεται πως ένα σύνολο τριών SVM classifiers έχει εκπαιδευτεί για κάθε μία εκ των τριών κλάσεων και συνδυάζονται για τις τελικές προβλέψεις ώστε τα σημεία δεδομένων που ανήκουν σε κάθε κλάση να παίρνουν τα σωστά labels.

2.4.4 Decision Tree

Ένα decision tree είναι ένα εργαλείο στήριξης για αποφάσεις το οποίο χρησιμοποιεί έναν δενδροειδή γράφο ή μοντέλο των αποφάσεων και των πιθανών συνέπειών τους, συμπεριλαμβάνοντας τυχαία γεγονότα, κόστος πόρων και χρησιμότητας. Κάθε decision tree μοιάζει στη δομή με flowchart όπου ο κάθε κόμβος αντιπροσωπεύει μια δοκιμή με ένα χαρακτηριστικό (π.χ. αν ένα κέρμα θα έρθει κορώνα ή γράμματα), κάθε κλαδί αντιπροσωπεύει το αποτέλεσμα της δοκιμής και κάθε φύλλο αντιπροσωπεύει μια κλάση ετικέτας (δηλαδή μια απόφαση που πάρθηκε μετά από τον υπολογισμό όλων των χαρακτηριστικών). Τα μονοπάτια από την ρίζα στο φύλλο αντιπροσωπεύουν τους κανόνες ταξινόμησης.

2.4.5 Random Forest

Τα random forests ή random decision forests είναι μια συνδυαστική μέθοδος μάθησης για ταξινόμηση, παλινδρόμηση και άλλες εργασίες, τα οποία λειτουργούν με την δημιουργία ενός πλήθους decision trees κατά την φάση της εκπαίδευσης και παράγουν την κλάση που εμφανίζεται πιο συχνά (ταξινόμηση) ή την μέση τιμή των προβλέψεων (παλινδρόμηση) του κάθε δέντρου. Τα random forests αντιμετωπίζουν και διορθώνουν σε μεγάλο βαθμό το πρόβλημα των decision trees κατά το οποίο υπάρχει overfitting του σετ εκπαίδευσης.

2.4.6 K-Nearest-Neighbors

Ο αλγόριθμος k-nearest-neighbors είναι μια μη παραμετρική μέθοδος που χρησιμοποιείται για ταξινόμηση και παλινδρόμηση. Και στις δύο περιπτώσεις, οι είσοδοι αποτελούνται από τα k κοντινότερα παραδείγματα στο χώρο. Η έξοδος, ωστόσο, εξαρτάται από την χρήση του k-NN, αν γίνεται για ταξινόμηση ή παλινδρόμηση:

- Στην ταξινόμηση, η έξοδος του αλγόριθμου είναι μέλος μιας κατηγορίας. Ένα αντικείμενο κατηγοριοποιείται από την πλειοψηφία των ψήφων των

γειτόνων του, με το αντικείμενο να ανατίθεται στην κλάση που είναι περισσότερο κοινή στους k κοντινότερους γείτονες (με το k να είναι ένας θετικός ακέραιος, συνήθως μικρός αριθμός). Αν k = 1, τότε το αντικείμενο ανατίθεται, απλά, στην κλάση του πιο κοντινού γείτονα.

- Στην παλινδρόμηση, η έξοδος είναι αντικειμενική τιμή του αντικειμένου. Αυτή η τιμή είναι ο μέσος όρος των τιμών των k κοντινότερων γειτόνων.

Ο k-NN είναι ένας τύπος μάθησης που βασίζεται στα παραδείγματα, ή αλλιώς lazy learning, όπου η συνάρτηση προσεγγίζεται μόνο σε τοπικό επίπεδο και όλοι οι υπολογισμοί αναβάλλονται μέχρι την ταξινόμηση. Ο αλγόριθμος k-NN είναι από τους απλούστερους αλγόριθμους μηχανικής μάθησης.

Τόσο για την ταξινόμηση όσο και για την παλινδρόμηση, μπορεί να είναι χρήσιμος με την ανάθεση βαρών στις συνεισφορές των γειτόνων, έτσι ώστε οι κοντινότεροι γείτονες να συνεισφέρουν περισσότερο στο μέσο όρο από τους πιο μακρινούς. Για παράδειγμα, μια κοινή κατανομή βαρών αποτελείται από την ανάθεση του βάρους 1/d στον κάθε γείτονα, όπου το d είναι η απόσταση του γείτονα.

2.4.7 Logistic Regression

Στην στατιστική, logistic regression είναι ένα μοντέλο παλινδρόμησης στο οποίο η εξαρτώμενη μεταβλητή είναι κατηγορική. Με τον όρο κατηγορική εννοείται, ότι η μεταβλητή μπορεί να πάρει μια τιμή από έναν περιορισμένο και συνήθως συγκεκριμένο αριθμό πιθανών τιμών. Μια από τις πιθανές μορφές που μπορεί να λάβει το μοντέλο αυτό, είναι με μια δυαδική εξαρτώμενη μεταβλητή, η οποία μπορεί να πάρει μόνο δύο τιμές: 0 και 1. Οι παραπάνω τιμές, αναπαριστούν αποτελέσματα όπως επιτυχία/αποτυχία, νίκη/ήττα, ζωντανός/νεκρός ή υγιής/άρρωστος. Πιο συγκεκριμένα, μορφές στις οποίες η εξαρτώμενη μεταβλητή έχει περισσότερες από δύο κατηγορίες αποτελεσμάτων, αναλύονται σε πολυωνυμικό logistic regression. Αν οι πολλαπλές κατηγορίες είναι διατεταγμένες, τότε αναλύονται σε διατεταγμένο logistic regression. Το δυαδικό μοντέλο χρησιμοποιείται για την εκτίμηση της πιθανότητας ενός δυαδικού αποτελέσματος που βασίζεται σε μια ή περισσότερες προγνωστικές (ή ανεξάρτητες) μεταβλητές (features). Έτσι, μας δίνεται η δυνατότητα να πούμε ότι η παρουσία ενός παράγοντα κινδύνου (risk factor) αυξάνει την πιθανότητα ενός δεδομένου αποτελέσματος σε συγκεκριμένο ποσοστό.

2.4.8 Neural Networks

Τα νευρωνικά δίκτυα είναι μια υπολογιστική προσέγγιση που χρησιμοποιείται στον κλάδο της Πληροφορικής και σε άλλους ερευνητικούς κλάδους, βασιζόμενα σε μια μεγάλη αλληλουχία νευρωνικών μονάδων (ή αλλιώς τεχνητών νευρώνων). Τα νευρωνικά δίκτυα έχουν ως στόχο τους να μιμηθούν τον τρόπο που ένας πραγματικός εγκέφαλος λύνει προβλήματα με μεγάλες ομάδες βιολογικών νευρώνων που συνδέονται από νευράξονες. Κάθε νευρωνική μονάδα συνδέεται με πολλές άλλες. Οι σύνδεσμοι αυτοί μπορούν να είναι enforcing or inhibitory στην επίδρασή του στην κατάσταση ενεργοποίησης των συνδεδεμένων νευρωνικών μονάδων. Κάθε νευρωνική μονάδα μπορεί να έχει μια αθροιστική συνάρτηση, η οποία συνδυάζει τις τιμές όλων των εισόδων. Επίσης, μπορεί να υπάρξει μια συνάρτηση κατωφλίου ή ορίου σε κάθε σύνδεση ακόμα και στην ίδια τη μονάδα, τέτοια ώστε να εξασφαλίζεται, ότι το σήμα ξεπερνά κάποιο όριο πριν διαδοθεί σε άλλους νευρώνες. Τέτοια συστήματα είναι αυτό-εκπαιδευόμενα, αντί για ρητά προγραμματιζόμενα, και υπερέχουν σε πεδία που η λύση ή ο εντοπισμός χαρακτηριστικών (features) είναι δύσκολο να εκφραστούν σε ένα παραδοσιακό πρόγραμμα υπολογιστή.

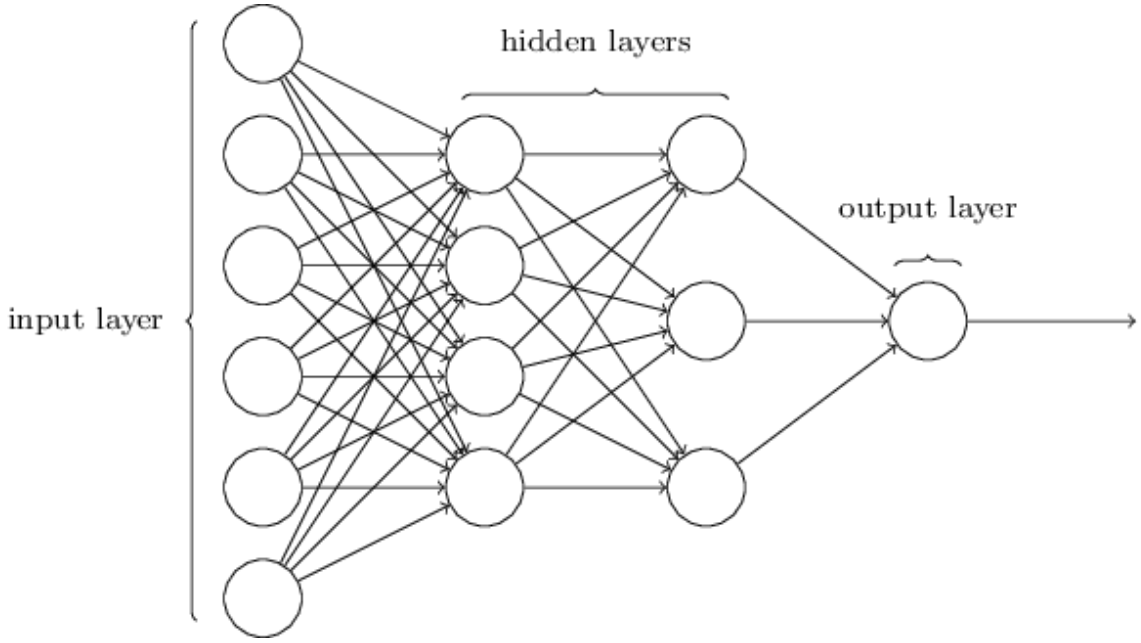
Συνήθως, τα νευρωνικά δίκτυα συντίθενται από πολλαπλά στρώματα ή από έναν σχεδιασμό κύβου, όπου το σήμα διασχίζει ένα μονοπάτι από το μπροστινό μέρος προς το πίσω. Η προς τα πίσω διάδοση είναι χρήση της μπροστινής διέγερσης για την επαναφορά των βαρών στις «μπροστινές» νευρωνικές μονάδες. Το συγκεκριμένο επιτυγχάνεται πολλές φορές σε συνδυασμό με την εκπαίδευση όπου το σωστό αποτέλεσμα είναι γνωστό. Περισσότερο καινοτόμα δίκτυα έχουν λίγο πιο ελεύθερη ροή σε όρους διέγερσης και αναχαίτισης με τις συνδέσεις που αλληλεπιδρούν να είναι πολύ περισσότερο σύνθετες και χαοτικές. Τα δυναμικά νευρωνικά δίκτυα είναι και τα πιο προχωρημένα, σε σημείο που μπορούν να σχηματίζουν καινούργιες συνδέσεις αλλά και νευρωνικές μονάδες ενώ παράλληλα, να απενεργοποιούν άλλες. Όλα αυτά επιτυγχάνονται μέσω κανόνων.

Ο στόχος των νευρωνικών δικτύων είναι να λύσουν προβλήματα με τον ίδιο τρόπο κατά τον οποίο θα λειτουργούσε ένας ανθρώπινος εγκέφαλος, παρά το γεγονός ότι αρκετά νευρωνικά δίκτυα είναι πιο αφηρημένα. Τα σύγχρονα νευρωνικά δίκτυα που χρησιμοποιούνται σε διάφορα έργα, συνήθως αποτελούνται από μερικές χιλιάδες έως και εκατομμύρια νευρωνικές μονάδες και συνδέσεις. Κάτι το

οποίο είναι ακόμα, αρκετές τάξεις μεγέθους πιο κάτω από την πολυπλοκότητα του ανθρώπινου εγκεφάλου και το καθιστά πιο κοντά στην υπολογιστική δύναμη ενός σκουληκιού.

Νέες έρευνες για τον ανθρώπινο εγκέφαλο, συχνά γεννούν καινούργιες ιδέες και στα νευρωνικά δίκτυα. Για παράδειγμα, μια νέα προσέγγιση χρησιμοποιεί συνδέσεις που εκτείνονται πολύ πιο μακριά συνδέοντας περισσότερα στρώματα επεξεργασίας, σε αντίθεση με το να μένουν σε παρακείμενους νευρώνες. Μια άλλη προσέγγιση που μελετάται, η «Βαθιά Μάθηση» (Deep Learning), εξερευνά τους διάφορους τύπους σήματος στο πέρασμα του χρόνου που διαδίδουν οι νευράξονες εισάγοντας, κατ' αυτόν τον τρόπο, μεγαλύτερη πολυπλοκότητα από ένα σετ δυαδικών μεταβλητών που απλά είναι on/off.

Αναλυτικότερα, η Βαθιά Μάθηση είναι ένας κλάδος της μηχανικής μάθησης που βασίζεται σε ένα σετ αλγορίθμων που προσπαθούν να μοντελοποιήσουν υψηλού επιπέδου αφηρημένες σχέσεις μεταξύ δεδομένων. Σε μια απλή περίπτωση, υπάρχουν δύο είδη νευρώνων: αυτοί που δέχονται το σήμα εισόδου και αυτοί που στέλνουν το σήμα εξόδου. Όταν το στρώμα εισόδου δέχεται μια είσοδο, την μεταβιβάζει σε μια τροποποιημένη μορφή στο επόμενο επίπεδο. Σε ένα βαθύ δίκτυο, υπάρχουν πολλά στρώματα μεταξύ της εισόδου και της εξόδου επιτρέποντας με αυτόν τον τρόπο στον αλγόριθμο να χρησιμοποιεί πολλαπλά στρώματα επεξεργασίας, τα οποία αποτελούνται από πολλαπλούς γραμμικούς και μη γραμμικούς μετασχηματισμούς.



Σχήμα 2.4: Ένα νευρωνικό δίκτυο με 4 στρώματα (1 εισόδου, 2 κρυφά και 1 εξόδου)

2.4.9 Gradient boosting

Το gradient boosting είναι μια τεχνική μηχανικής μάθησης για προβλήματα regression και classification, η οποία παράγει ένα μοντέλο πρόβλεψης στη μορφή συνόλου από αδύναμα μοντέλα πρόβλεψης, συνήθως decision tress. Το μοντέλο κατασκευάζεται ανά στάδια, όπως και άλλες μέθοδοι boosting, τα οποία γενικεύονται με την εφαρμογή βελτιστοποίησης σε μια αυθαίρετη διαφορίσιμη συνάρτηση απώλειας.

2.5 Αξιολόγηση Μοντέλων Ταξινόμησης

2.5.1 Εισαγωγή

Η εκπαίδευση, ο συντονισμός και η κατασκευή των μοντέλων είναι σημαντικά μέρη όλου του κύκλου ζωής, αλλά ακόμα πιο σημαντικό είναι να γνωρίζουμε το πόσο καλά αποδίδουν αυτά τα μοντέλα. Η απόδοση των μοντέλων ταξινόμησης συνήθως βασίζεται στην ευστοχία που έχουν οι προβλέψεις τους στα νέα σημεία δεδομένων. Τις περισσότερες φορές, γίνεται μέτρηση της απόδοσης σε ένα σετ δεδομένων που αποτελείται από σημεία δεδομένων τα οποία δεν χρησιμοποιήθηκαν για να επηρεάσουν ή να εκπαιδεύσουν τον classifier με κανέναν τρόπο. Αυτό το σετ δεδομένων αποτελείται από αρκετές παρατηρήσεις και τις αντίστοιχες ετικέτες τους. Αφού εξαχθούν τα features με τον ίδιο τρόπο που γινόταν στην εκπαίδευση του μοντέλου, τα features δίνονται στο ήδη εκπαιδευμένο μοντέλο και παρατηρούνται οι προβλέψεις του για κάθε σημείο δεδομένων. Αυτές οι προβλέψεις συγκρίνονται με τις πραγματικές ετικέτες για την εκτίμηση της ικανότητας και της ακρίβειας των προβλέψεων του μοντέλου.

Αρκετές μετρικές μπορούν να καθορίσουν την απόδοση των προβλέψεων ενός μοντέλου, με κυριότερες τις εξής:

- Accuracy
- Precision
- Recall
- F1 score

2.5.2 Confusion Matrix

Το confusion matrix είναι ένας εξαιρετικός τρόπος για την μέτρηση της απόδοσης του συστήματος για δύο κλάσεις ή παραπάνω. Το confusion matrix είναι μια

πινακοειδής δομή που βοηθάει στην οπτικοποίηση της απόδοσης των classifiers. Κάθε στήλη στον πίνακα αντιπροσωπεύει ταξινομημένες περιπτώσεις βασιζόμενες σε προβλέψεις και κάθε γραμμή του πίνακα εκπροσωπεί ταξινομημένες περιπτώσεις με βάση την πραγματική κλάση που ανήκουν. Συνήθως, επιλέγεται η ετικέτα μιας κλάσης ως η θετική κλάση, η οποία, τις περισσότερες φορές, είναι η κλάση ενδιαφέροντος. Το σχήμα που ακολουθεί δείχνει ένα κλασικό παράδειγμα confusion matrix δύο κλάσεων, όπου το p δηλώνει την θετική κλάση και το n την αρνητική κλάση.

	p' (Predicted)	n' (Predicted)
p (Actual)	True Positive	False Negative
n (Actual)	False Positive	True Negative

Σχήμα 2.5: Confusion Matrix

Ο χαρακτηρισμός True Positive (TP) υποδηλώνει τον αριθμό των σωστών προβλέψεων για την θετική κλάση. Ενώ, ο χαρακτηρισμός False Negative (FN) υποδεικνύει τον αριθμό των περιπτώσεων αυτής της κλάσης που προβλέφθηκαν λανθασμένα ως η αρνητική κλάση. Το False Positive (FP) είναι ο αριθμός των περιπτώσεων που προβλέφθηκαν λανθασμένα ως η θετική κλάση ενώ, στην πραγματικότητα, δεν ήταν. True Negative (TN) είναι ο αριθμός των περιπτώσεων που σωστά προβλέφθηκαν ως η αρνητική κλάση.

2.5.2 Accuracy

Η μετρική accuracy ορίζεται ως η συνολική ακρίβεια ή το ποσοστό των σωστών προβλέψεων του μοντέλου, η οποία περιγράφεται από την παρακάτω εξίσωση:

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

Όπου στον αριθμητή έχουμε όλες τις σωστές προβλέψεις που διαιρούνται από όλα τα αποτελέσματα που βρίσκονται στον παρονομαστή.

2.5.3 Precision

Η μετρική precision ορίζεται ως ο αριθμός των προβλέψεων που έγιναν και είναι, πραγματικά, σωστές ή σχετικές από όλες τις προβλέψεις που βασίζονται στην θετική κλάση. Για αυτόν τον λόγο είναι γνωστή και ως η θετική προβλεπόμενη τιμή και μπορεί να περιγραφεί από την παρακάτω φόρμουλα:

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

Όπου έχουμε τις σωστές προβλέψεις για την θετική κλάση στον αριθμητή να διαιρούνται από όλες τις προβλέψεις που έγιναν για την θετική κλάση, συμπεριλαμβάνοντας και τις λάθος.

2.5.4 Recall

Η μετρική recall ορίζεται ως ο αριθμός των προβλέψεων που έγιναν σωστά για την θετική κλάση. Η μετρική αυτή είναι γνωστή και ως hit rate, coverage ή sensitivity και δίνεται από την παρακάτω σχέση:

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

Όπου έχουμε τις σωστές προβλέψεις για την θετική κλάση στον αριθμητή και στον παρονομαστή τις σωστές και άστοχες προβλέψεις για την θετική κλάση, η διαίρεση αυτή δίνει και το hit rate.

2.5.5 F1 score

Το F1 score είναι μια άλλη μετρική ακρίβειας που υπολογίζεται από τον αρμονικό μέσο όρο του precision και του recall και αναπαρίσταται ως εξής:

$$F1\ score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

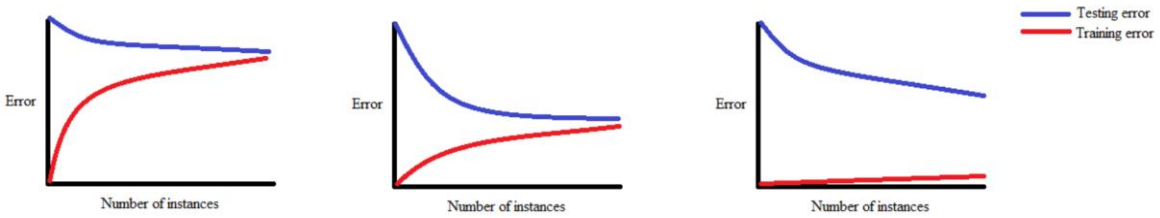
2.5.6 Καμπύλη Receiver Operating Characteristic (ROC Curve)

Η καμπύλη ROC είναι μια γραφική παράσταση που απεικονίζει τη διαγνωστική ικανότητα ενός δυαδικού συστήματος ταξινόμησης καθώς το όριο διάκρισής του ποικίλει. Η καμπύλη δημιουργείται με την απεικόνιση του true positive rate προς το false positive rate σε διάφορες τιμές κατωφλίων.

2.5.7 Καμπύλη Μάθησης (Learning Curve)

Η καμπύλη μάθησης είναι ένα γράφημα το οποίο συγκρίνει την απόδοση ενός μοντέλου κατά την εκπαίδευση και την δοκιμή δεδομένων ενώ τα παραδείγματα εκπαίδευσης μεταβάλλονται. Ο γενικός κανόνας λέει ότι η απόδοση θα έπρεπε να αυξάνεται όσο αυξάνεται και ο αριθμός των παραδειγμάτων. Η καμπύλη αυτή, μπορεί να μας δείξει πότε ένα μοντέλο έχει μάθει στο μέγιστο από τα δεδομένα που δέχεται. Υπάρχουν τρεις περιπτώσεις καμπυλών μάθησης:

1. ***Υψηλής προτίμησης:*** όταν τα σφάλματα σε εκπαίδευση και δοκιμή συγκλίνουν και είναι υψηλά. Τότε, ανεξάρτητα από τον αριθμό των δεδομένων που θα τροφοδοτήσουμε το μοντέλο, εκείνο δεν μπορεί να αναπαραστήσει την υποκείμενη σχέση και έχει υψηλά συστημικά σφάλματα.
2. ***Υψηλής διακύμανσης:*** όταν υπάρχει μεγάλο κενό μεταξύ των σφαλμάτων. Τότε, χρειάζονται δεδομένα για την βελτίωση του μοντέλου ή απλοποίησή του με λιγότερα ή λιγότερο σύνθετα features.
3. ***Ιδανική:*** όταν οι καμπύλες εκπαίδευσης και δοκιμής συγκλίνουν σε παρόμοιες τιμές. Τότε, το μοντέλο γενικεύει σε νέα δεδομένα. Όσο μικρότερο το κενό, τόσο καλύτερα γίνεται η γενίκευση αυτή.



Σχήμα 2.6: Οι τρεις τύποι καμπυλών μάθησης: (α) υψηλής προτίμησης, (β) ιδανική, (γ) υψηλής διακύμανσης

(Κενό φύλλο)

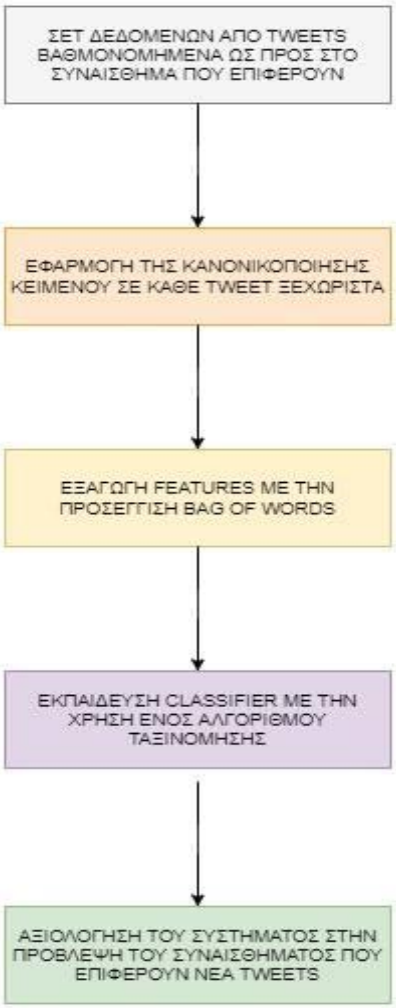
ΚΕΦΑΛΑΙΟ 3

ΥΛΟΠΟΙΗΣΗ ΚΑΙ ΑΠΟΤΕΛΕΣΜΑΤΑ

Σε αυτό το κεφάλαιο αναλύονται οι τεχνικές που χρησιμοποιήσαμε για την υλοποίηση του συστήματός μας καθώς και τα αποτελέσματα των δοκιμών με τις παρατηρήσεις που εξάγαμε.

3.1 Επισκόπηση Συστήματος

Η δομή του συστήματος που υλοποιήσαμε καθώς και τα κύρια μέρη του φαίνονται στο παρακάτω διάγραμμα ροής:



Διάγραμμα 4.1: Διάγραμμα ροής του συστήματος

3.1.1 Εργαλεία που χρησιμοποιήσαμε

Το κύριο εργαλείο που χρησιμοποιήσαμε ήταν η γλώσσα προγραμματισμού με την οποία κάναμε την υλοποίηση και αυτή ήταν η γλώσσα Python. Επιλέξαμε την Python, και συγκεκριμένα το περιβάλλον ανάπτυξης Spyder, γιατί έχει πολλές και χρήσιμες βιβλιοθήκες που θα μας βοηθούσανε στην υλοποίησή μας. Τέτοιες βιβλιοθήκες, τις οποίες και χρησιμοποιήσαμε, ήταν:

- ***Sklearn***: Η βιβλιοθήκη sklearn ήταν εκείνη που χρησιμοποιήσαμε περισσότερο, καθώς μπορεί να υλοποιήσει τα περισσότερα μέρη ενός συστήματος μηχανικής μάθησης. Οι δυνατότητες που μας παρείχε ήταν από την εξαγωγή features μέχρι την υλοποίηση των αλγόριθμων ταξινόμησης αλλά και την αξιολόγηση του συστήματος με πολύ απλό και κατανοητό τρόπο.
- ***Numpy***: Η βιβλιοθήκη numpy, αν και έχει πάρα πολλές δυνατότητες και παρέχει πολλές ευκολίες σε ότι έχει να κάνει με επιστημονικούς υπολογισμούς και διάφορους μετασχηματισμούς, εμείς την χρησιμοποιήσαμε για κάποιους υπολογισμούς και θα λέγαμε ότι δεν είδαμε τις δυνατότητές της σε βάθος.
- ***NLTK***: Η βιβλιοθήκη NLTK (Natural Language ToolKit) είναι μια βιβλιοθήκη που υλοποιεί διάφορες τεχνικές επεξεργασίας φυσικής γλώσσας. Εμείς την χρησιμοποιήσαμε, κυρίως, στην κανονικοποίηση του κειμένου. Για να είμαστε ακριβείς, στην κανονικοποίηση κειμένου του αγγλικού σετ δεδομένων.

3.1.2 Δυσκολίες που συναντήσαμε

Η βασική δυσκολία που συναντήσαμε στο όλο εγχείρημά μας ήταν η έλλειψη βιβλιογραφίας και παρόμοιων συστημάτων για ελληνικό κείμενο. Μπορεί στο εξωτερικό, και όταν έχουμε να κάνουμε για αγγλικό κείμενο, ο τομέας της ταξινόμησης κειμένου και της μηχανικής μάθησης να έχει προχωρήσει αρκετά, αλλά στην Ελλάδα είναι ακόμα στις αρχές του. Για αυτόν τον λόγο, συγκεκριμένα, το υποσύστημα του NLP που είχαμε σαν σκοπό να υλοποιήσουμε σαν παράλληλο υποσύστημα μαζί με εκείνο του bag of words, μας καθυστέρησε ιδιαίτερα στην υλοποίηση. Πιστεύουμε, όμως, ότι κάναμε μια αρκετά καλή προσπάθεια σε κάτι τόσο καινούργιο και ενδιαφέρον όπως ο τομέας ταξινόμησης ελληνικού κειμένου.

3.2 Δεδομένα

Για την υλοποίηση του συστήματος χρησιμοποιήσαμε δύο διαφορετικά σετ δεδομένων, ένα με αγγλικά κείμενα και ένα με ελληνικά. Αυτό συνέβη διότι μας ήταν πιο εύκολο, αρχικά, να χρησιμοποιήσουμε ένα σετ δεδομένων με αγγλικές λέξεις και φράσεις ώστε να ακολουθήσουμε κάποιες πεπατημένες τεχνικές και υλοποιήσεις, προκειμένου να εξοικειωθούμε με το αντικείμενο. Δηλαδή, υπήρχαν έτοιμες βιβλιοθήκες και εργαλεία που μας βοήθησαν σε αυτόν τον στόχο, πριν ξεκινήσουμε να υλοποιήσουμε κάτι μόνοι μας. Από την στιγμή που εξοικειωθήκαμε και καταλήξαμε σε κάποια συμπεράσματα ως προς την συμπεριφορά του συστήματός μας, χρησιμοποιήσαμε το ελληνικό σετ δεδομένων ώστε να καταλήξουμε σε κάποια τελικά συμπεράσματα ως προς την συμπεριφορά του συστήματος. Άλλωστε, στην προσέγγιση του bag of words που επιλέξαμε για την υλοποίηση, η γλώσσα του κειμένου δεν έχει ιδιαίτερη σημασία. Αντίθετα, στην προσέγγιση του NLP που ξεκινήσαμε να υλοποιούμε, η γλώσσα του κειμένου είναι καθοριστικός παράγοντας για την υλοποίησή της.

3.2.1 Αγγλικό σετ δεδομένων

Τα δεδομένα με αγγλικό κείμενο που χρησιμοποιήθηκαν αποτελούνταν από κείμενα μικρού μήκους, τα οποία εξωρύχθηκαν από το γνωστό κοινωνικό δίκτυο Twitter. Τα tweets ήταν στο σύνολό τους 2005, εκ των οποίων τα 1402 εξέφραζαν κάποιο αρνητικό συναίσθημα ενώ τα υπόλοιπα (632) κάποιο θετικό. Σαφώς, το στάδιο εκπαίδευσης ενός αλγορίθμου είναι κρίσιμο κομμάτι με αποτέλεσμα τα δεδομένα να έχουν σημαντικό ρόλο στην απόδοση του συστήματος. Συνεπώς, μπορούμε να πούμε ότι το ιδανικό θα ήταν να έχουμε μεγάλο πλήθος δεδομένων και πιο ισομοιρασμένα νούμερα στις δύο κατηγορίες.

Το αρχείο που χρησιμοποιήθηκε ήταν ένα αρχείο σε μορφή csv που βρήκαμε από την βιβλιογραφία. Η πληροφορία που κρατήθηκε για το κάθε tweet ήταν:

- ID του tweet
- Συναίσθημα που επιφέρει
- Κείμενο

Όπως ήταν αναμενόμενο, η γλώσσα που χρησιμοποιούνταν στα περισσότερα tweets ήταν τελείως «ανεπίσημη». Αυτό σημαίνει ότι περιείχε πολλά συνεχόμενα

σημεία στίξης, αρκετές συντομογραφίες, πολλά ορθογραφικά λάθη αλλά και τεράστια χρήση των emoticons.

Παράδειγμα tweet με θετικό συναίσθημα

Not long till LONDON BABY!!! and then the EMIRATES ON SUNDAY!!!! YEAH!! cant wait...

Παράδειγμα tweet με αρνητικό συναίσθημα

i really feel bad bout eating a cheeseburger and a donut for dinner ugh! i so need to burn this off tomorrow! :| darn McDonalds!!!!

3.2.2 Ελληνικό σετ δεδομένων

Το ελληνικό σετ δεδομένων που χρησιμοποιήσαμε αποτελούνταν από tweets, τα οποία, κατά κύριο λόγο, εξέφραζαν άποψη για ένα από τα παρακάτω αντικείμενα:

1. Φαγητά
2. Τράπεζες
3. Τηλεπικοινωνίες

Για αυτόν τον λόγο, διαιρέσαμε τα δεδομένα στις τρεις κατηγορίες που αναφέρθηκαν. Έπειτα, υπήρχε για το κάθε tweet μια ανάλυση του κειμένου του και τι χαρακτηριστικά έχει αυτό. Τέτοια χαρακτηριστικά θεωρήθηκαν τα παρακάτω:

- ***Αντικειμενικότητα:*** Αυτό το χαρακτηριστικό είναι ένας δείκτης για το αν το κείμενο που αναλύουμε επιφέρει ή όχι κάποιο συναίσθημα. Σπάνια, ένα αντικειμενικό κείμενο επιφέρει κάποιο συναίσθημα και αυτό διότι είναι κατά κύριο λόγο κάποιο γεγονός ή είδηση.
- ***Συναίσθημα:*** Τα συναισθήματα που είχαν τα δεδομένα μας ήταν τρία: θετικό, αρνητικό και ουδέτερο. Παρά την πληροφορία που είχαμε για τα ουδέτερα tweets, επιλέξαμε να μην την χρησιμοποιήσουμε και να υλοποιήσουμε ένα σύστημα ταξινόμησης δυο κατηγοριών.

Μια γενική εικόνα του σετ δεδομένων φαίνεται στον παρακάτω πίνακα:

	Θετικά	**Αρνητικά**	**Ουδέτερα**	**Σύνολο**
Φαγητά	407 (4.7%)	1012 (11.8%)	1220 (14.2%)	2639 (30.7%)
Τράπεζες	261 (3%)	936 (10.9%)	780 (9.1%)	1977 (23%)
Τηλεπικοινωνίες	752 (8.8)	654 (7.6%)	2560 (29.8%)	3966 (46.2%)
Σύνολο	1420 (16.5%)	2602 (30.3%)	4560 (53.1%)	8582

Πίνακας 4.1: Στατιστική εικόνα του ελληνικού σετ δεδομένων

3.3 Κανονικοποίηση Κειμένου

Η κανονικοποίηση κειμένου σε ένα σύστημα ταξινόμησης κειμένου, είναι ίσως η πιο σημαντική διαδικασία. Αυτό συμβαίνει γιατί σε ένα (σωστά) επεξεργασμένο κείμενο είναι πολύ πιο εύκολο να εξωρυχθεί κάποιο σημαντικό feature ή πληροφορία, παρά σε ένα κείμενο που περιέχει αρκετό θόρυβο και άρα αρκετή περιττή πληροφορία. Βέβαια, το «σωστά» επεξεργασμένο κείμενο ποικίλλει και δεν υπάρχει κάποια χρυσή τομή για να βρεθεί. Αυτό βρίσκεται μόνο από πολλές δοκιμές και συνδυασμούς μέχρι να καταλήξουμε στον καλύτερο συνδυασμό που μας επιφέρει τα καλύτερα αποτελέσματα.

3.3.1 Τεχνικές που υλοποιήθηκαν

Οι τεχνικές κανονικοποίησης κειμένου που υλοποιήσαμε και οι λόγοι για τους οποίους επιλέχθηκαν, αναλύονται παρακάτω:

- Αφαίρεση των hyperlinks: σε ένα tweet που υπάρχει ένα hyperlink (π.χ. www.google.gr) αφαιρείται τελείως ή αντικαθίσταται από μια λέξη placeholder (π.χ. @link). Μετά από δοκιμές καταλήξαμε στην αφαίρεση των links διότι το μοντέλο μας μπορεί να θεωρήσει ως feature την λέξη που χρησιμοποιούμε ως placeholder, καθώς υπάρχει η πιθανότητα να εμφανίζεται αρκετές φορές.

  Κείμενο πριν: Ευχαριστώ το www.google.com για τις πληροφορίες!

  Κείμενο μετά: Ευχαριστώ το για τις πληροφορίες!

- Αφαίρεση αριθμών: οι αριθμοί τις περισσότερες φορές δεν φέρουν κάποιο συναίσθημα σε μια πρόταση και δεν καθορίζουν κάτι στο νόημά της. Για αυτόν τον λόγο, δοκιμάσαμε να τους αφαιρούμε από τα tweet και να δούμε τι αποτέλεσμα θα έχουμε.

  Κείμενο πριν: Σήμερα γίνομαι 15 χρονών!!!!!

  Κείμενο μετά: Σήμερα γίνομαι χρονών!!!!!
- Αφαίρεση αρκετά χρησιμοποιούμενων λέξεων (stop words): η αφαίρεση προθέσεων, άρθρων, συνδέσμων, αντωνυμιών κλπ. είναι μια συνηθισμένη τεχνική στην προ-επεξεργασία κειμένων. Αυτό συμβαίνει γιατί μια πρόταση είναι σπάνιο να καθορίζεται ως προς το συναίσθημα ή το νόημα που επιφέρει από τέτοιες λέξεις. Επίσης, ο classifier θα τροφοδοτηθεί με περιττές λέξεις στις οποίες θα δώσει μεγάλη βαρύτητα λόγω της συχνής εμφάνισής τους.

  Κείμενο πριν: Θέλω να πάω διακοπές πριν τον Αύγουστο.

  Κείμενο μετά: Θέλω πάω διακοπές Αύγουστο.
- Αφαίρεση σημείων στίξης μέσα στην πρόταση: η ιδέα πίσω από αυτό το είδος καθαρισμού είναι ότι τα σημεία στίξης μέσα στην πρόταση δεν έχουν άλλη χρήση από το να διαχωρίζουν τις προτάσεις. Σπάνια καθορίζουν κάποιο συναίσθημα του συνολικού tweet.

  Κείμενο πριν: Επιτέλους, πήρα καινούργιο κινητό. Ευχαριστώ μπαμπά!

  Κείμενο μετά: Επιτέλους πήρα καινούργιο κινητό Ευχαριστώ μπαμπά!
- Αφαίρεση σημείων στίξης στο τέλος της πρότασης: διαφοροποιήσαμε την αφαίρεση των σημείων στίξης στο τέλος της πρότασης γιατί πολύ συχνά η τελευταία πρόταση φέρει το νόημα ενός κειμένου και ως εκ τούτου το σημείο στίξης που βρίσκεται στο τέλος αυτής ίσως κρύβει κάποια σημασία.

  Κείμενο πριν: Δεν μπορώ άλλο διάβασμα…..

  Κείμενο μετά: Δεν μπορώ άλλο διάβασμα
- Αφαίρεση καταλήξεων (stemming): η αφαίρεση των καταλήξεων έγινε για την διευκόλυνση του classifier να εξάγει ως κοινά features λέξεις (ίσως και φράσεις) που διαφοροποιούνται στο ελάχιστο (π.χ. είμαι καλά – είμαστε καλά). Στόχος με αυτή την τεχνική είναι ο εντοπισμός των ριζών των λέξεων.

Κείμενο πριν: Έχω κουραστεί να καθαρίζω.

Κείμενο μετά: Έχ κουραστ να καθαρ.

- Μετατροπή όλων των γραμμάτων σε πεζά: η μετατροπή αυτή γίνεται για την καλύτερη ομοιομορφία των λέξεων ώστε ο classifier να μην εξάγει ως διαφορετικά features μια λέξη που ξεκινάει από κεφαλαίο γράμμα και μια λέξη που γράφτηκε με όλα τα γράμματα πεζά.

  Κείμενο πριν: ΤΕΡΑΣΤΙΑ ΝΙΚΗ ΓΙΑ ΤΗΝ ΕΘΝΙΚΗ!!!

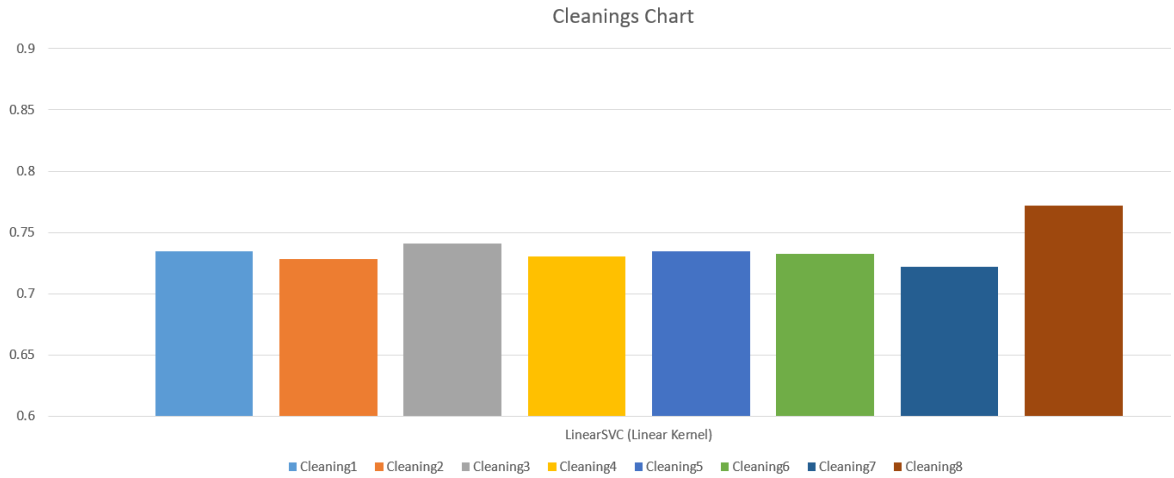
  Κείμενο μετά: τεραστια νικη για την εθνικη!!!

Παράδειγμα

Αρχικό tweet	Δεν μπορώ άλλη ζέστη! 42 βαθμούς το Σάββατο σύμφωνα με το www.meteo.gr!
Αφαίρεση hyperlink	Δεν μπορώ άλλη ζέστη! 42 βαθμούς το Σάββατο σύμφωνα με το www.meteo.gr!
Αφαίρεση αριθμών	Δεν μπορώ άλλη ζέστη! 42 βαθμούς το Σάββατο σύμφωνα με το !
Αφαίρεση stop words	Δεν μπορώ άλλη ζέστη! βαθμούς το Σάββατο σύμφωνα με το !
Αφαίρεση σημείων στίξης μέσα στην πρόταση	μπορώ άλλη ζέστη! βαθμούς Σάββατο σύμφωνα!
Αφαίρεση σημείων στίξης στο τέλος της πρότασης	μπορώ άλλη ζέστη βαθμούς Σάββατο σύμφωνα!
Αφαίρεση καταλήξεων	μπορώ άλλη ζέστη βαθμούς Σάββατο σύμφωνα
Μετατροπή όλων των γραμμάτων σε πεζά	μπορ άλλ ζέστ βαθμ Σάββατ σύμφων
Τελικό tweet	μπορ άλλ ζέστ βαθμ σάββατ σύμφων

Πίνακας 4.2: Παράδειγμα καθαρισμού ενός tweet

3.3.2 Απότελέσματα δοκιμών κανονικοποίησης κειμένου



Διάγραμμα 4.2: Αποτελέσματα δοκιμών διαφορετικών καθαρισμών κειμένου με την χρήση του αλγόριθμου LinearSVC και της μετρικής F1 score

Cleaning1: Καθαρισμός links, καθαρισμός stop words, μετατροπή όλων των γραμμάτων σε πεζά, καθαρισμός σημείων στίξεων στο τέλος του κειμένου, καθαρισμός σημείων στίξεων μέσα στο κείμενο, καθαρισμός καταλήξεων.

Cleaning2: Καθαρισμός links, καθαρισμός stop words, μετατροπή όλων των γραμμάτων σε πεζά, καθαρισμός σημείων στίξεων στο τέλος του κειμένου, καθαρισμός σημείων στίξεων μέσα στο κείμενο.

Cleaning3: Καθαρισμός links, μετατροπή όλων των γραμμάτων σε πεζά, καθαρισμός σημείων στίξεων στο τέλος του κειμένου, καθαρισμός σημείων στίξεων μέσα στο κείμενο.

Cleaning4: Καθαρισμός links, μετατροπή όλων των γραμμάτων σε πεζά.

Cleaning5: Καθαρισμός links.

Cleaning6: Κανένας καθαρισμός, χρήση αυτούσιου κειμένου.

Cleaning7: Καθαρισμός links, καθαρισμός αριθμών, καθαρισμός stop words, μετατροπή όλων των γραμμάτων σε πεζά, καθαρισμός σημείων στίξεων στο τέλος του κειμένου, καθαρισμός σημείων στίξεων μέσα στο κείμενο, καθαρισμός καταλήξεων.

Cleaning8: Καθαρισμός links, μετατροπή όλων των γραμμάτων σε πεζά, καθαρισμός σημείων στίξεων στο τέλος του κειμένου, καθαρισμός σημείων στίξεων μέσα στο κείμενο, καθαρισμός καταλήξεων.

3.3.3 Παρατηρήσεις

Από τον πίνακα φαίνεται ότι καλύτερος συνδυασμός καθαρισμού είναι ο καθαρισμός νούμερο 8. Ο συγκεκριμένος καθαρισμός διαφοροποιείται από τους άλλους στο ότι δεν χρησιμοποιεί την αφαίρεση των stop words κάτι που φαίνεται να επηρεάζει την επίδοση του συστήματος. Ο λόγος φαίνεται να είναι ότι, παρά την αφαίρεση περιττών (και ίσως άχρηστων λέξεων), χάνονται και λέξεις που καθορίζουν το νόημα και το συναίσθημα που επιφέρει η πρόταση. Χαρακτηριστικό παράδειγμα είναι αυτό που δόθηκε παραπάνω. Η φράση «δεν μπορώ» μετά από έναν τέτοιο καθαρισμό θα γίνει «μπορώ», κάτι το οποίο αλλάζει άρδην την σημασία της φράσης και κατά συνέπεια της πρότασης. Για αυτόν τον λόγο στα επόμενα πειράματα του συστήματός μας θα αποφύγουμε την αφαίρεση των stop words και θα χρησιμοποιούμε τον καθαρισμό 8.

3.4 Εξαγωγή Χαρακτηριστικών

Σε αυτό το κομμάτι του συστήματος, υλοποίησαμε δύο διαφορετικές προσεγγίσεις για την εξαγωγή features. Η πρώτη υλοποίηση βασίστηκε στο κλασικό μοντέλο bag of words, ενώ η δεύτερη είχε ως βάση την επεξεργασία φυσικής γλώσσας.

3.4.1 Προσέγγιση BoW

Η προσέγγιση bag of words είναι αρκετά απλή. Αφού όλα τα δεδομένα μας περάσουν από την κανονικοποίηση κειμένου και έχουν την μορφή που επιθυμούμε, δίνονται ως είσοδοι στο μοντέλο TF-IDF. Ο TF-IDF επιλέγει, με την σειρά του, τα κυριότερα features και τροφοδοτεί τον αλγόριθμο ταξινόμησης, ο οποίος θα εκπαιδευτετεί με αυτά. Μετά την εκπαίδευση του αλγόριθμου ταξινόμησης της επιλογής μας, υπάρχει και μια αξιολόγηση του συστήματος με ένα μέρος των δεδομένων που δεν χρησιμοποιήθηκαν για την εκπαίδευση. Αυτό γίνεται για να αποφευχθεί η πρόβλεψη ενός tweet που είναι ήδη γνωστή η κατηγορία ταξινόμησής του και έχουμε λανθασμένα αποτελέσματα στην αξιολόγηση. Επίσης, για την αποφυγή λανθασμένων αποτελεσμάτων, πολύ σημαντικό είναι να αφαιρεθούν τυχόν διπλότυπες εγγραφές από το σετ δεδομένων για τον λόγο που προαναφέρθηκε.

3.4.2 Προσέγγιση NLP

Η λογική για την επιλογή των feature ξεκίνησε με γνώμονα τον σκοπό του συστήματος που είναι η πολικότητα ενός κειμένου. Αυτό σημαίνει ότι επιλέχτηκαν features τα οποία καθορίζουν σε μεγάλο βαθμό το συναίσθημα ενός κειμένου όπως: τα emoticons (τα οποία χρησιμοποιούνται σε μεγάλο βαθμό στα κοινωνικά δίκτυα), το ποσοστό των κεφαλαίων και η ύπαρξη θετικού ή αρνητικού στοιχείου όπως λέξεις ή φράσεις. Επίσης, μπορεί να υπάρξει κατάταξη μεταξύ των χαρακτηριστικών για την απόφαση του αλγορίθμου. Δηλαδή, το χαρακτηριστικό της ύπαρξης και του πλήθους των θετικών/αρνητικών λέξεων έχει μεγαλύτερη βαρύτητα από τα emoticons. Επιπλέον, υπάρχει και ο συνδυασμός κάποιων feature. Για παράδειγμα, αν σε ένα κείμενο το ποσοστό των κεφαλαίων είναι έντονο και περιέχει αρνητικές λέξεις τότε μπορεί να θεωρηθεί ότι υπάρχει ενόχληση ή εκνευρισμός. Σε τέτοιες περιπτώσεις μπορεί να θεωρηθεί το κείμενο ως αρνητικό, αλλά χωρίς να είναι αρκετό το κριτήριο του συνδυασμού των χαρακτηριστικών.

Δομή και βήματα προσέγγισης

Ξεκινώντας με την δομή του κομματιού αυτού να επισημάνουμε ότι η διαδικασία της προεπεξεργασίας διαφέρει από το bag of words κομμάτι. Σε αντίθεση με το bag of words μοντέλο, εδώ είναι σημαντικό το σημείο εκτέλεσης του καθαρισμού διότι κάτι που μπορεί να θεωρηθεί ως θόρυβος στο bag of words εδώ αποτελεί feature. Ένα τέτοιο feature, για παράδειγμα, είναι το ποσοστό των κεφαλαίων. Αφού υπολογισθεί, τότε μπορούν τα κεφαλαία γράμματα να μετατραπούν σε πεζά.

Αρχικά, το πρόγραμμα ορίζει ένα μοντέλο καθαριστή που αφαιρεί θορύβους, κάποιοι από αυτούς είναι κοινοί με το bag of words όπως: σημεία στίξης, ηλεκτρονικές διευθύνσεις, email, stop words. Πριν εκτελεσθεί η διαδικασία του καθαρισμού, το αρχικό κείμενο χωρίζεται σε υποκείμενα αν υπάρξουν σημεία στίξης μεταξύ των προτάσεων. Αυτό γίνεται διότι μπορεί να υπάρχει διαφορετικό συναίσθημα σε κάθε πρόταση. Άρα, ο αλγόριθμος θα μπορούσε να βγάλει λάθος αποτέλεσμα αν υπολόγιζε όλες τις προτάσεις μαζί παρά μεμονωμένα. Στη συνέχεια, αφού υπολογίσει τα χαρακτηριστικά που έχουν οριστεί επιστρέφει τα στατιστικά αυτών από την κάθε πρόταση που χωρίστηκε το αρχικό κείμενο. Με τα στατιστικά αυτά, αφού κανονικοποιηθούν ή μετατραπούν σε κατανοητή μορφή για τον υπολογιστή γίνονται είσοδοι για τον αλγόριθμο ταξινόμησης, όπως ακριβώς και στο

bag of words. Τέτοιου είδους μοντέλα, βελτιώνονται με το πλήθος και την ποιότητα των features καθώς και με τον τρόπο που τα δεδομένα διαχειρίζονται.

Χαρακτηριστικά που χρησιμοποιήθηκαν

Το εύρος των χαρακτηριστικών που θα μπορούσε να χρησιμοποιηθεί για ένα τέτοιο σύστημα είναι από το ποσοστό των κεφαλαίων που συνηθέστερα υποδηλώνουν εκνευρισμό μέχρι την γραμματική ανάλυση σε ένα κείμενο που θα υποδήλωνε την αξιοπιστία αυτού.

- Στην αρχή, όπως αναφέρθηκε, υπολογίστηκε το ποσοστό των κεφαλαίων λέξεων σε ένα κείμενο και αν το ποσοστό αυτό είναι πάνω από μία τιμή κατωφλίου (80% στο σύστημά μας) τότε η ύπαρξη αρνητικού συναισθήματος γίνεται πιο έντονη.
- Στην ιδιαίτερη διάλεκτο των κοινωνικών δικτύων η ύπαρξη παρατεταμένων γραμμάτων σε μία λέξη εκφράζει θετικό συναίσθημα. Για παράδειγμα η φράση: «Είμαι τοσο χαρούμενοοοος που πέρασα στο πανεπιστήμιοοοο!!!» υποδήλωνει θετικό συναίσθημα χωρίς αμφιβολία. Άλλο ένα χαρακτηριστικό που μπορεί να εντοπισθεί σε ένα κείμενο κοινωνικού δικτύου, και ίσως σε πιο έντονο βαθμό, είναι τα emoticons. Τα emoticons μπορεί να θεωρηθούν χαρακτηριστικά με βαρύτητα σε συνδυασμό με την ύπαρξη κάποιου άλλου.
- Στο παραπάνω παράδειγμα για το θετικό αποτέλεσμα του αλγορίθμου συμβάλει και η ύπαρξη θετικών λέξεων σε αυτή. Δύο τέτοια χαρακτηριστικά είναι ο αριθμός των θετικών και αρνητικών λέξεων. Λέξεις που εκφράζουν απόλυτα θετικό ή αρνητικό συναίσθημα όπως: «αγαπώ» ή «μισώ». Έτσι, κάθε λέξη ελέγχεται ως προς το συναίσθημα που εκφράζει. Η ύπαρξη θετικών λέξεων σε συνδυασμό με χαρούμενα emoticons θα αποτελούν χαρακτηριστικό βαρύτητας για το αποτέλεσμα του αλγορίθμου.

Χαρακτηριστικά προς υλοποίηση

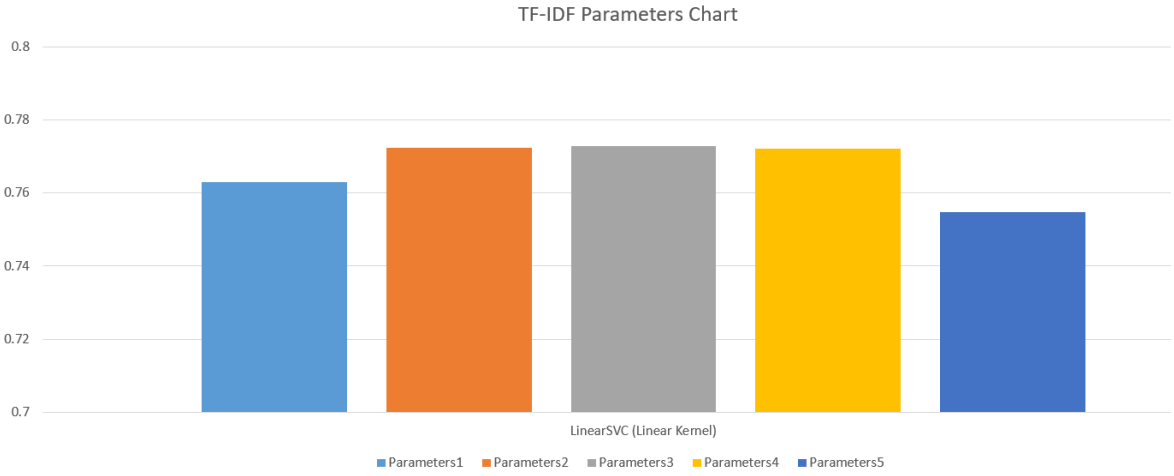
Ερευνήθηκαν χαρακτηριστικά που θα μπορούσε να βοηθήσουν στην απόδοση και την ταχύτητα του αλγορίθμου. Το POS (part of speech) tagging είναι ένα από αυτά. Το pos tagging χαρακτηριστικό χωρίζει τις λέξεις σε μέρη του λόγου (π.χ. αντικείμενα, ουσιαστικά, ρήματα κλπ.) σε μία πρόταση. Επειδή οι λέξεις που εκφράζουν συναίσθημα είναι επίθετα, ρήματα και ουσιαστικά με την βοήθεια του

pos tagging θα μπορούσε να ελέγχονται συγκεκριμένες κατηγορίες λέξεων. Επίσης, το pos tagging είναι αναγκαίο για την ανάλυση γραμματικών κανόνων σε μία πρόταση. Επίσης σαν χαρακτηριστικό αποτελούν λέξεις-φράσεις. Για τον λόγο ότι κάποιες λέξεις μπορεί να εκφράζουν θετικό ή αρνητικό συναίσθημα ή ακόμα και τίποτα, ο συνδυασμός τους με άλλες λέξεις μπορεί να υποδηλώνει κάτι. Για παράδειγμα, οι λέξεις «μηχανική» και «μάθηση» σημαίνουν δύο διαφορετικά πράγματα ενώ ο συνδυασμός «μηχανική μάθηση» αναφέρεται σε κάτι πολύ συγκεκριμένο. Τέλος, ένα πρόβλημα που προκύπτει στα χαρακτηριστικά που αφορούν τις λέξεις, μπορεί να θεωρηθεί η ορθογραφία. Έγινε μια προσπάθεια βελτίωσης στο πρόβλημα αυτό με έναν διορθωτή λέξεων, διορθώνοντας κάποιες περιπτώσεις αλλά έχοντας ακόμα περιθώρια εξέλιξης για μεγαλύτερη επιτυχία και καλύτερη ποιότητα κειμένου.

3.4.3 Αποτελέσματα δοκιμών εξαγωγής features

Τα αποτελέσματα και οι δοκιμές έγιναν επιλέγοντας την πρώτη προσέγγιση, εκείνη του bag of words. Σε αυτό το κομμάτι του συστήματος δοκιμάστηκαν κυρίως δύο πράγματα:

- Ο αριθμός των n-grams
- Ο αριθμός των features που τροφοδοτούμε τον TF-IDF



Διάγραμμα 4.3: Αποτελέσματα δοκιμών διαφορετικών παραμέτρων του μοντέλου TF-IDF με την χρήση του αλγόριθμου LinearSVC και της μετρικής F1 score

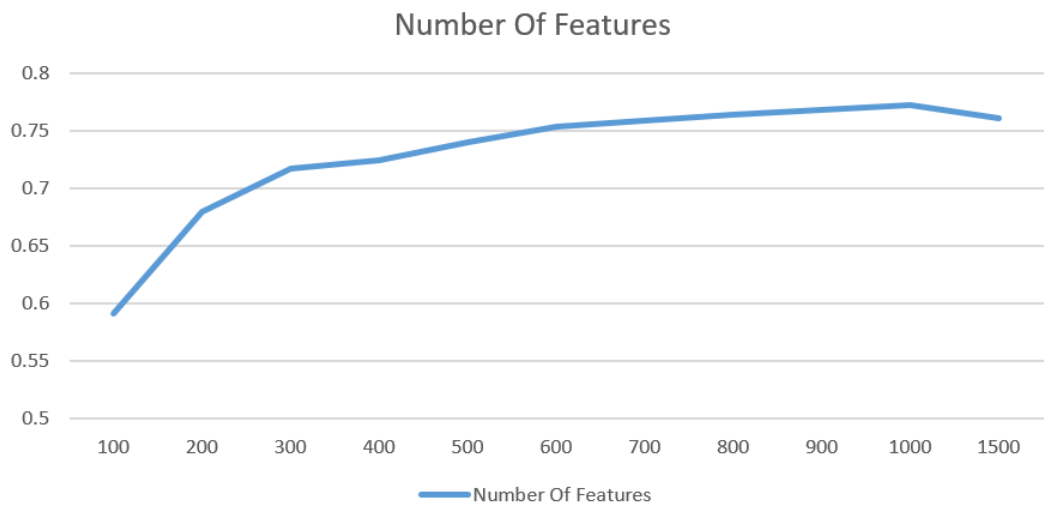
Parameters1: εύρος n-grams (1-2)

Parameters2: εύρος n-grams (1-3)

Parameters3: εύρος n-grams (1-4)

Parameters4: εύρος n-grams (1-5)

Parameters5: εύρος n-grams (1-6)



Διάγραμμα 4.4: Αποτελέσματα δοκιμών με διαφορετικό αριθμό features με την χρήση του αλγόριθμου LinearSVC και της μετρικής F1 score

3.4.4 Παρατηρήσεις

Όσον αφορά τον πρώτο πίνακα, το σύστημά μας φαίνεται να έχει την καλύτερη επίδοση με την χρήση μεσαίου εύρους n-grams, αφού με το μικρότερο εύρος (1 έως 2) και με το μεγαλύτερο (1-6) φαίνεται η επίδοση να μειώνεται.

Ο κύριος λόγος που μπορεί να εξηγηθεί αυτό το αποτέλεσμα είναι γιατί μια άρνηση (π.χ. δεν) επισυνάπτεται σε μια λέξη που προηγείται εκείνης ή την ακολουθεί. Αυτή η διαδικασία βελτιώνει την ακρίβεια της ταξινόμησης καθώς η άρνηση παίζει σημαντικό ρόλο σε μια γνώμη ή σε μια έκφραση συναισθήματος.

Επίσης, η χρήση της «αργκό» και η ανάγκη έκφρασης σε λίγες λέξεις στα κοινωνικά δίκτυα είναι ένας σημαντικός παράγοντας. Στο Twitter πολλές λέξεις γίνονται συντομογραφίες και οι φράσεις συρρικνώνονται. Αυτό έχει ως αποτέλεσμα τα μεσαίου μεγέθους n-grams να μπορούν να εντοπίσουν το νόημα, άρα και το συναίσθημα που επιφέρουν, αποτελεσματικότερα από τα μικρού ή μεγάλου μεγέθους n-grams. Αφού, τα μικρά n-grams χάνουν σημαντικές λέξεις ενώ τα μεγάλα n-grams χάνουν το νόημα από τις περιττές λέξεις.

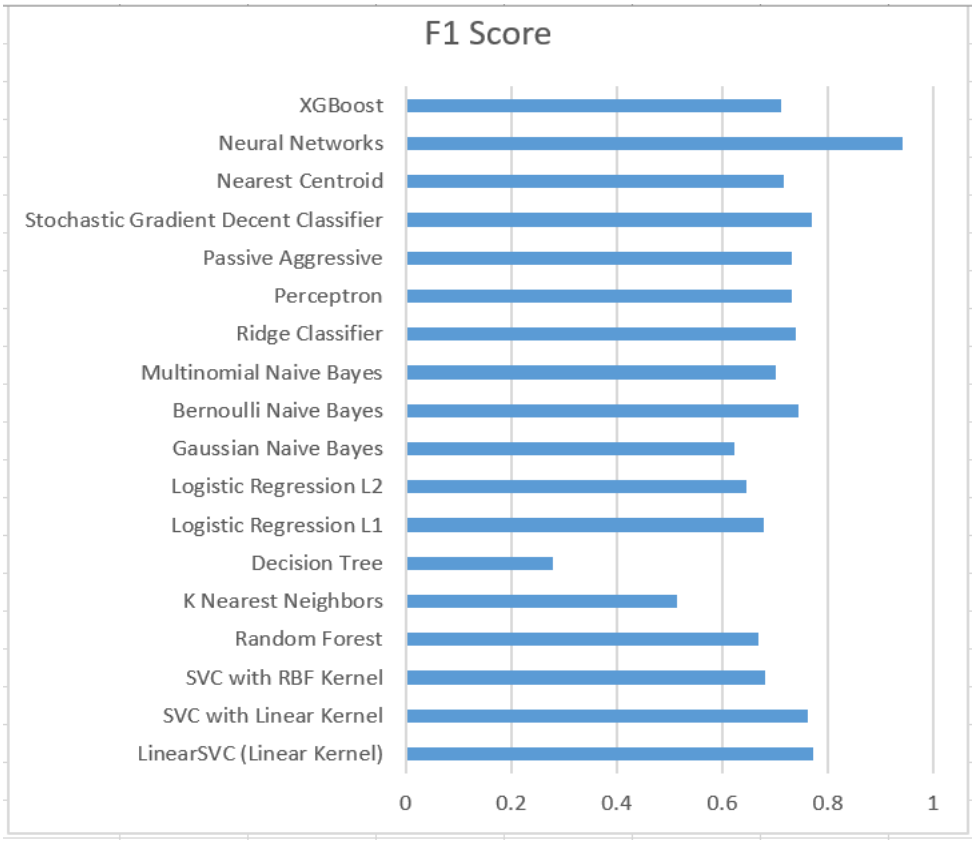
Από την καμπύλη του δεύτερου γραφήματος, μπορούμε να διακρίνουμε τα εξής: Αρχικά, με την αύξηση των features φαίνεται να αυξάνεται και η επίδοση του συστήματος σε μεγάλο βαθμό, καθώς από τα 100 στα 200 features έχουμε βελτίωση της μετρικής μας από 0.59 σε 0.68. Στην συνέχεια, η επίδοση δείχνει να σταθεροποιείται, αφού από την αύξηση των features από 500 σε 1000 έχουμε βελτίωση της απόδοσης ελάχιστα (0.03). Τελικά, με την αύξηση των features στον αριθμό 1500 ξεκινάει μια πτωτική πορεία της επίδοσης, κάτι που οφείλεται στο φαινόμενο του «overfitting». Το φαινόμενο αυτό, κάνει την εμφάνισή του σε συστήματα τα οποία έχουν «υπερτροφοδοτηθεί» από features, κάτι το οποίο μετατρέπει το σύστημα σε μη λειτουργικό. Ο μεγάλος αριθμός από features περισσότερο δείχνει να «συγχύζει» τον αλγόριθμο στην επιλογή μιας απόφασης, καθώς λαμβάνει υπόψιν του και πολλά ασήμαντα features, παρά τον βοηθάει στην λήψη της σωστής απόφασης.

Έχοντας στο μυαλό μας τις παραπάνω παρατηρήσεις, καταλαβαίνουμε ότι κάθε σύστημα πρέπει να δοκιμαστεί με πολλές και διάφορες παραμέτρους προκειμένου να βρεθεί η βέλτιστη επίδοσή του. Όπως είδαμε, δεν υπάρχει μια ιδανική πρακτική (π.χ. χρησιμοποιούμε τα περισσότερα ή λιγότερα n-grams ή features) για την επίτευξη αυτού του στόχου, αλλά επιτυγχάνεται με την συνεχή δοκιμή και καταγραφή των επιδόσεων του συστήματος.

3.5 Αλγόριθμοι Ταξινόμησης

Για πολλούς, η πιο σημαντική επιλογή σε ένα σύστημα Μηχανικής Μάθησης είναι η επιλογή του αλγόριθμου που θα κάνει την ταξινόμηση. Όπως φαίνεται στο παρακάτω διάγραμμα, δοκιμάσαμε αρκετούς αλγόριθμους με σκοπό την επίτευξη της καλύτερης επίδοσης του συστήματός μας.

3.5.1 Αποτελέσματα δοκιμών αλγόριθμων ταξινόμησης



Διάγραμμα 4.5: Αποτελέσματα αλγορίθμων με την μετρική F1 score

3.5.2 Παρατηρήσεις

Σύμφωνα με τις δοκιμές μας, η μέγιστη επίδοση, στο σετ δεδομένων που χρησιμοποιήσαμε, επιτυγχάνεται από τα Νευρωνικά Δίκτυα. Τα Νευρωνικά Δίκτυα, όπως αναφέραμε και παραπάνω, έχουν την δυνατότητα να εκπαιδεύονται ξανά και ξανά μέσα από την ανατροφοδότηση. Όπως είναι λογικό, όσο αυξάναμε τις δύο βασικές παραμέτρους (ρυθμό εκμάθησης και αριθμό επαναλήψεων), αυξανόταν και η απόδοση. Αυτό όμως σήμαινε και την μείωση της απόκρισης και ταχύτητας του συστήματος. Τελικά, καταλήξαμε σε μια ενδιάμεση λύση, κατά την οποία το σκορ του συστήματός μας είναι αρκετά υψηλό φτάνοντας το 94% και παράλληλα ο χρόνος απόκρισης δεν είναι απαγορευτικός. Το αποτέλεσμα αυτό επετεύχθη χρησιμοποιώντας τις παρακάτω παραμέτρους:

- δύο στρώματα νευρώνων (Rectifier και Softmax)
- ρυθμό εκμάθησης (learning rate): 0.05
- αριθμό επαναλήψεων (iterations): 15

3.6 Meta-Classifier

Αυτό το κομμάτι των δοκιμών ήταν, ίσως, το πιο ενδιαφέρον αλλά και το πιο δύσκολο γιατί δεν υπήρχε αρκετή βιβλιογραφία και υλοποιήσεις παρόμοιων συστημάτων πάνω στις οποίες θα μπορούσαμε να στηριχθούμε. Αυτό συμβαίνει διότι η εκπαίδευση ενός classifier με τα αποτελέσματα πολλών και διαφορετικών μεταξύ τους classifier είναι μια καινούργια τεχνική που δοκιμάζεται στο πεδίο της μηχανικής μάθησης. Αυτός, ο μεταγενέστερος classifier, ο οποίος εκπαιδεύεται από τα αποτελέσματα άλλων classifier, ονομάζεται meta-classifier.

3.6.1 Αποτελέσματα δοκιμών meta-classifier

Αλγόριθμοι αρχικών classifier	**Αλγόριθμος meta-classifier**	**Accuracy**	**Precision**	**Recall**	**F1 score**
Linear SVC, SVC with Linear Kernel, SGD, XGB	Bernoulli Naive Bayes	0.870	0.802	0.775	0.788
K-Nearest Neighbors, Random Forest, Gaussian Naive Bayes	Logistic Regression L1	0.862	0.783	0.771	0.777
Linear SVC, SVC with Linear Kernel, SGD	Logistic Regression L1	0.855	0.760	0.794	0.771
Linear SVC, Random Forest, Gaussian Naive Bayes	Neural Networks	0.850	0.766	0.752	0.756

Πίνακας 4.3: Αποτελέσματα διάφορων classifier σε συνδυασμό με έναν meta-classifier

3.6.2 Παρατηρήσεις

Στον παραπάνω πίνακα βλέπουμε τις 4 καλύτερες επιδόσεις που πετύχαμε με την εκπαίδευσης ενός meta-classifier. Παρά το γεγονός ότι ευελπιστούσαμε να πετύχουμε καλύτερα αποτελέσματα από τα αποτελέσματα που είχαμε από την εκπαίδευση ενός και μόνο classifier, δεν σημαίνει πως δεν μείναμε ευχαριστημένοι. Καθώς, σχεδόν όλοι οι συνδυασμοί που δοκιμάσαμε πέτυχαν ποσοστά F1 score

μεγαλύτερα του 75%. Αυτό σημαίνει πως με κάποια παραπάνω έρευνα στους αλγόριθμους ταξινόμησης αλλά και στον συνδυασμό αυτών, θα καταφέρουμε να πετύχουμε αρκετά υψηλά ποσοστά απόδοσης του συστήματος.

3.7 Οριστικοποίηση Συστήματος

Από τα αποτελέσματα των δοκιμών και τις παρατηρήσεις που εξάγαμε καταλήξαμε στις εξής παραμέτρους κάθε κομματιού του συστήματός μας για την βέλτιστη απόδοση:

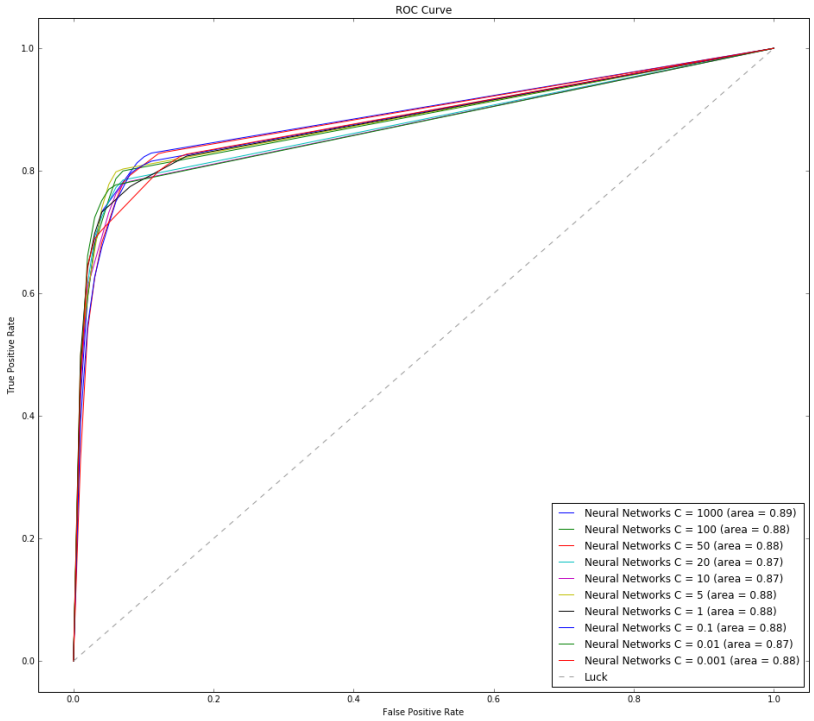
- ***Καθαρισμός κειμένου:*** Καθαρισμός links, μετατροπή όλων των γραμμάτων σε πεζά, καθαρισμός σημείων στίξεων στο τέλος του κειμένου, καθαρισμός σημείων στίξεων μέσα στο κείμενο, καθαρισμός καταλήξεων
- ***Αριθμός n-grams:*** 1-5
- ***Αριθμός features:*** 1000
- ***Αλγόριθμος ταξινόμησης:*** Νευρωνικά Δίκτυα δύο επιπέδων με παραμέτρους: learning rate = 0.05 και iterations = 15.

Με αυτές τις παραμέτρους θα δοκιμάσουμε να εκπαιδεύσουμε το σύστημά μας με την χρήση του ελληνικού σετ δεδομένων. Κάτι το οποίο αποτελεί για εμάς πρόκληση, καθώς υπάρχουν ελάχιστα συστήματα που να κάνουν ταξινόμηση κειμένου σε ελληνικό κείμενο.

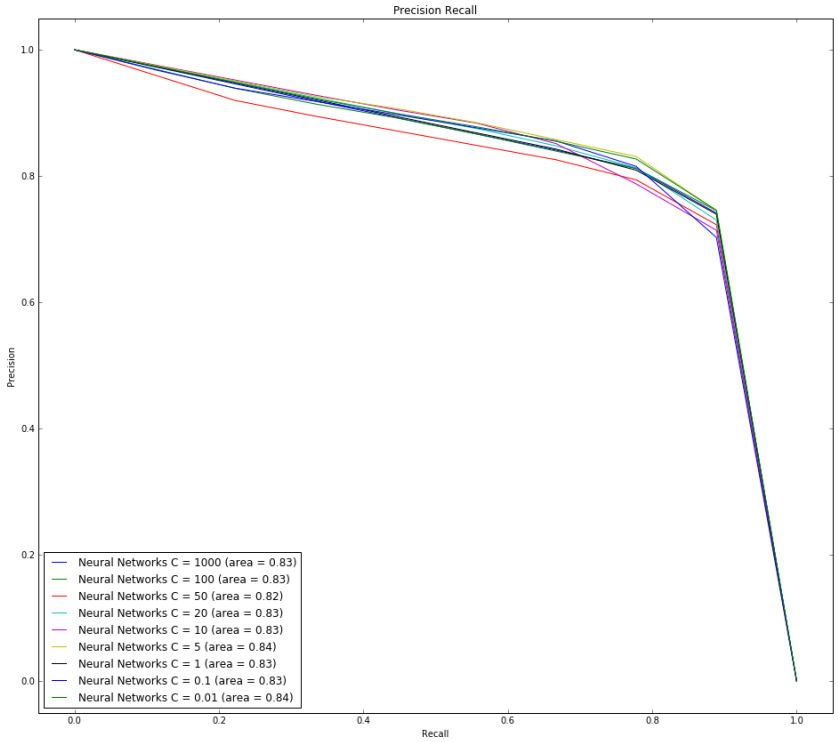
3.7.1 Αποτελέσματα

Accuracy	Precision	Recall	F1
0.930	0.893	0.797	0.833

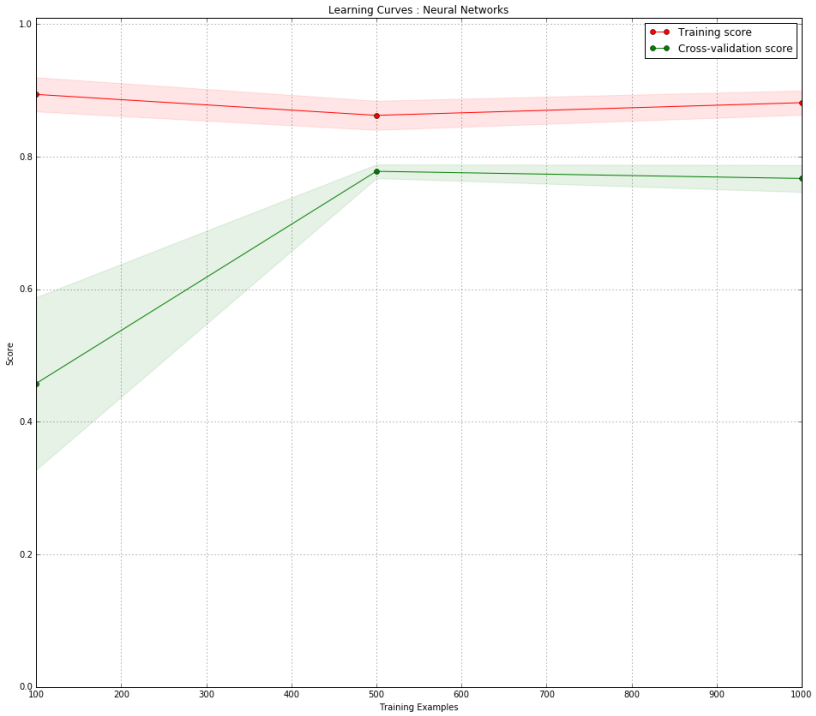
Πίνακας 4.4: Αποτελέσματα όλων των μετρικών του συστήματος με την χρήση του ελληνικού σετ δεδομένων



Διάγραμμα 4.6: Καμπύλη ROC του συστήματος με χρήση του ελληνικού σετ δεδομένων



Διάγραμμα 4.7: Καμπύλη Precision-Recall του συστήματος με χρήση του ελληνικού σετ δεδομένων



Διάγραμμα 4.8: Καμπύλη Μάθησης του συστήματος με χρήση του ελληνικού σετ δεδομένων

3.7.2 Συμπεράσματα

Τα αποτελέσματα των μετρικών αλλά και των καμπυλών φαίνονται να είναι αρκετά ικανοποιητικά. Συγκεκριμένα, τα ποσοστά των μετρικών accuracy, precision, recall και F1 score ξεπερνάνε το 80% (με εξαίρεση το recall που είναι οριακά στο 0.8), κάτι που δείχνει ότι το σύστημα παραμένει σταθερό στα αποτελέσματα που παράγει παρά την διαφοροποίηση του σετ δεδομένων και της γλώσσας από την οποία αποτελείται.

Όσον αφορά τις καμπύλες, τα συμπεράσματα που μπορούμε να εξάγουμε από τις δύο πρώτες (διαγράμματα 4.6 και 4.7) είναι πως το σύστημα έχει αρκετά εύστοχες προβλέψεις ακόμα και όταν οι πιθανότητες που δίνει για την επικρατούσα κλάση είναι κοντά στο 0.5. Αυτό έχει ως αποτέλεσμα να μην χρειάζεται να ανεβάσουμε το κατώτατο κατώφλι πιθανοτήτων ώστε να θεωρήσουμε σίγουρη μια πρόβλεψη του συστήματος σε κάποια κλάση για να πετύχουμε υψηλά ποσοστά ευστοχίας.

Τέλος, μερικές πολύ σημαντικές πληροφορίες μάς παρέχει η καμπύλη μάθησης (διάγραμμα 4.8). Αυτά που μας δείχνει η συγκεκριμένη καμπύλη είναι πως το σύστημά μας έχει ήδη μάθει από τα 500 παραδείγματα, κάτι το οποίο σημαίνει ότι

έχει μάθει ικανοποιητικά γρήγορα. Όμως, μάς δείχνει και το γεγονός ότι από εκείνο το σημείο και μετά παύει να μαθαίνει. Αυτό σημαίνει πως χρειάζεται παραπάνω δουλειά στο κομμάτι της εξαγωγής των features, καθώς η προσέγγιση bag of words έχει κάποια όρια σαν αυτό. Σε αυτό το πρόβλημα θα έρθει να δώσει λύση η προσέγγιση του NLP που έχουμε ως σκοπό να προσθέσουμε στο ήδη υπάρχον σύστημα.

3.8 Μελλοντικές Βελτιώσεις

Παρά τα καλά αποτελέσματα που φαίνεται να πετυχαίνει το σύστημά μας, πάντα υπάρχουν περιθώρια βελτίωσης. Μια βελτίωση που μπορεί, φυσικά, να έρθει στο σύστημα είναι η βελτίωση των ποσοστών στις διάφορες μετρικές που χρησιμοποιήσαμε αλλά και η συνεχής καλυτέρευση των καμπυλών που δείχνουν την συμπεριφορά του συστήματος.

Επίσης, κύριο μέλημά μας είναι η περαιτέρω ανάπτυξη του συστήματος εξαγωγής features με την προσέγγιση NLP. Πιστεύουμε πως όταν αυτή η προσέγγιση συνδυαστεί με εκείνη του bag of words σε έναν meta-classifier θα καταφέρουμε ένα πολύ καλό αποτέλεσμα με έναν τρόπο ο οποίος είναι ιδιαίτερα καινοτόμος και δεν έχει αναπτυχθεί από άλλα συστήματα.

Τέλος, η χρήση του ουδέτερου συναισθήματος θα μπορούσε να είναι μια πολύ ενδιαφέρουσα προσθήκη στο σύστημά μας. Δηλαδή, η δυνατότητα του συστήματός μας να ταξινομεί σε τρεις διαφορετικές κατηγορίες θα ήταν μια ιδιαίτερη πρόκληση για εμάς.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Samuel, Arthur (1959). "Some Studies in Machine Learning Using the Game of Checkers". IBM Journal of Research and Development. 3.

[2] S. Z. Li, Markov Random Field Modeling in Computer Vision, Springer-Verlag, 1995Ron Kohavi; Foster Provost (1998). "Glossary of terms". Machine Learning. 30: 271–274.

[3] Alpaydin, Ethem (2010). Introduction to Machine Learning. London: The MIT Press. ISBN 978-0-262-01243-0. Retrieved 4 February 2017.

[4] Mehryar Mohri, Afshin Rostamizadeh, Ameet Talwalkar (2012) Foundations of Machine Learning, The MIT Press ISBN 9780262018258.

[5] Jordan, Michael I.; Bishop, Christopher M. (2004). "Neural Networks". In Allen B. Tucker. Computer Science Handbook, Second Edition (Section VII: Intelligent Systems). Boca Raton, FL: Chapman & Hall/CRC Press LLC. ISBN 1-58488-360-X.

[6] Hastie,Trevor,Robert Tibshirani, Friedman,Jerome (2009). The Elements of Statistical Learning: Data mining,Inference,and Prediction. New York: Springer. pp. 485–586. ISBN 978-0-387-84857-0.

[7] Ijsmi, Editor (2017-08-27). "Natural Language Processing concepts and methods revisited". International Journal of Statistics and Medical Informatics. 4

[8] Harris, Zellig (1954). "Distributional Structure". Word. 10 (2/3): 146–62.

[9] Youngjoong Ko (2012). "A study of term weighting schemes using class information for text classification". SIGIR'12. ACM.

[10] Robertson, S. (2004). "Understanding inverse document frequency: On theoretical arguments for IDF". Journal of Documentation. 60 (5): 503–520.

[11] Pissanetzky, Sergio (1984). Sparse Matrix Technology. Academic Press.

[12] S. Z. Li, Markov Random Field Modeling in Computer Vision, Springer-Verlag, 1995.

[13] Cortes, C.; Vapnik, V. (1995). "Support-vector networks". Machine Learning. 20 (3): 273–297.

[14] Ben-Hur, Asa, Horn, David, Siegelmann, Hava, and Vapnik, Vladimir; "Support vector clustering" (2001) Journal of Machine Learning Research, 2: 125–137.

[15] Press, William H.; Teukolsky, Saul A.; Vetterling, William T.; Flannery, B. P. (2007). "Section 16.5. Support Vector Machines". Numerical Recipes: The Art of Scientific Computing (3rd ed.). New York: Cambridge University Press. ISBN 978-0-521-88068-8.

[16] Shawe-Taylor, J.; Cristianini, N. (2004). Kernel Methods for Pattern Analysis. Cambridge University Press.

[17] Quinlan, J. R. (1987). "Simplifying decision trees". International Journal of Man-Machine Studies. 27 (3): 221.

[18] Ho, Tin Kam (1995). Random Decision Forests (PDF). Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282.

[19] Ho, Tin Kam (1998). "The Random Subspace Method for Constructing Decision Forests" (PDF). IEEE Transactions on Pattern Analysis and Machine Intelligence. 20 (8): 832–844.

[20] Altman, N. S. (1992). "An introduction to kernel and nearest-neighbor nonparametric regression". The American Statistician. 46 (3): 175–185.

[21] David A. Freedman (2009). Statistical Models: Theory and Practice. Cambridge University Press. p. 128.

[22] Rennie, J.; Shih, L.; Teevan, J.; Karger, D. (2003). Tackling the poor assumptions of Naive Bayes classifiers (PDF). ICML.

[23] Fawcett, Tom (2006); An introduction to ROC analysis, Pattern Recognition Letters, 27, 861–874.

[24] Kouloumpis, Efthymios, Theresa Wilson, and Johanna Moore. Twitter Sentiment Analysis: The Good the Bad and the OMG! in Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM-2011). 2011.

[25] A. Pak, P. Paroubek, Twitter based system: using Twitter for disambiguating sentiment ambiguous adjectives, Proceedings of the 5th International Workshop on Semantic Evaluation, Association for Computational Linguistics, 2010, pp. 436–439.