ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

-Πτυχιακή Εργασία-

ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ ΣΤΗΝ ΑΝΑΓΝΩΡΙΣΗ ΟΡΘΟΓΡΑΦΙΚΩΝ ΛΑΘΩΝ ΛΕΞΕΩΝ ΕΝΤΟΣ ΛΕΞΙΛΟΓΙΟΥ ΣΤΗΝ ΕΛΛΗΝΙΚΗ ΓΛΩΣΣΑ

Νίκη Καπετανοπούλου

ΑΜ: Π2013120

Επιβλέπουσα: Κάτια Κερμανίδου

Νοέμβριος 2017

Επιβλέπουσα

Κάτια Κερμανίδου, Επίκουρη Καθηγήτρια,

Τμήμα Πληροφορικής , Ιόνιο Πανεπιστήμιο

Τριμελής Επιτροπή

Κάτια Κερμανίδου, Επίκουρη Καθηγήτρια,

Τμήμα Πληροφορικής, Ιόνιο Πανεπιστήμιο

Φοίβος Μυλωνάς, Αναπληρωτής Καθηγητής,

Τμήμα Πληροφορικής, Ιόνιο Πανεπιστήμιο

Σπύρος Σιούτας, Επίκουρος Καθηγητής,

Τμήμα Πληροφορικής, Ιόνιο Πανεπιστήμιο

Ευχαριστίες

Για την εκπόνηση της παρούσας πτυχιακής θα ήθελα πρώτα να ευχαριστήσω την επιβλέπουσα μου, κ. Κάτια Κερμανίδου, επίκουρη καθηγήτρια του Ιονίου Πανεπιστημίου για την πολύτιμη βοήθεια, καθοδήγηση και κατανόηση που μου παρείχε κατά της διάρκεια της δουλειάς μου. Επιπλέον, θα ήθελα να ευχαριστήσω και τα υπόλοιπα μέλη της εξεταστικής επιτροπής, τον αναπληρωτή καθηγητή κ.Φοίβο Μυλωνά και τον επίκουρο καθηγητή κ.Σπύρο Σιούτα για την ανάγνωση της πτυχιακής μου εργασίας και την παρουσία τους. Επιπροσθέτως θα ήθελα να ευχαριστήσω την οικογένεια μου και συγκεκριμένα τους γονείς μου Πασχάλη και Βασιλική για την ολόψυχη αγάπη και ηθική υποστήριξη που μου παρείχαν καθώς και τον αδερφό μου Βασίλη.

Πρόλογος

Η μηχανική μάθηση είναι ένα ενδιαφέρον επιστημονικό υποπεδίο της Τεχνητής νοημοσύνης το οποίο έχει αναπτυχθεί πολύ κατά τα τελευταία χρόνια και συνδέεται άμεσα με την αναγνώριση προτύπων. Αφορά συστήματα τα οποία έχουν την ιδιότητα να δημιουργούν νέα γνώση ή μοντέλα και να αλληλεπιδρούν από ένα σύνολο διακριτών στοιχείων ή δεδομένων μέσω κάποιας προηγούμενης γνώσης αλλά και να κάνουν προβλέψεις σχετικά με αυτά. Η μηχανική μάθηση στην αναγνώριση ορθογραφικών λαθών λέξεων εντός λεξιλογίου στην Ελληνική γλώσσα έχει ως κύριο σκοπό την άρση της αμφισημίας λέξεων που γράφονται με περισσότερους από έναν τρόπους χρησιμοποιώντας τεχνικές μηχανικής μάθησης ώστε να αποσαφηνιστούν τα διάφορα χαρακτηριστικά των συμφραζόμενων λέξεων όπως τα μέρος του λόγου τους και ο αριθμός τους που έπονται κι προηγούνται της αμφίσημης λέξης. Μια ορθογραφικά αμφίσημη λέξη μπορεί να πάρει πολλές διαφορετικές έννοιες όντας ομόηχη ή έχοντας τυπογραφικά λάθη με αποτέλεσμα να είναι άκρως σημαντική η πλήρης κατανόηση των συμφραζομένων. Έτσι μέσα από πειραματισμούς με ποικίλους συνδυασμούς χαρακτηριστικών και αλγορίθμων οδηγούμαστε στον βέλτιστο συνδυασμό τους με απώτερο σκοπό την σωστή ορθογραφία των ζητούμενων λέξεων.

Περίληψη

Στην παρούσα διπλωματική εργασία μελετάται το πρόβλημα της αναγνώρισης ορθογραφικών λαθών σε κείμενα της ελληνικής γλώσσας με τη χρήση μεθόδων μηχανικής μάθησης. Ο αυτόματος ορθογραφικός έλεγχος των κειμένων είναι μία διαδικασία για την οποία έχουν αναπτυχθεί διάφορα παραδοσιακά εργαλεία, τα οποία, ωστόσο, τις περισσότερες φορές αποτυγχάνουν να εντοπίσουν λάθη που μπορεί να οφείλονται σε χρήση λανθασμένης γραμματικής ή στην επιλογή μη κατάλληλων λέξεων με βάση τα συμφραζόμενα.

Η μηχανική μάθηση έρχεται να προσφέρει πιο προηγμένες μεθόδους για την αυτόματη αναγνώριση ορθογραφικών λαθών, οι οποίες μαθαίνουν από τα δεδομένα και κάνουν προβλέψεις σχετικά με αυτά. Το πρόβλημα της αναγνώρισης ορθογραφικών λαθών υπάγεται στον κλάδο της Επεξεργασίας Φυσικής Γλώσσας (Natural Language Processing), όπου μέσω μίας πληθώρας τεχνικών μηχανικής μάθησης πραγματοποιείται η αυτοματοποιημένη ανάλυση του προφορικού και γραπτού λόγου.

Στην παρούσα, λοιπόν, διπλωματική εργασία, αφού οριστούν οι βασικές έννοιες που διέπουν το πρόβλημα, θα γίνει η ανάλυση διαφόρων τεχνικών μηχανικής μάθησης και η εφαρμογή αυτών σε δεδομένα που έχουν προέλθει από κείμενα ηλεκτρονικών εφημερίδων. Πιο συγκεκριμένα, θα μελετηθούν ομόηχα ρήματα τα οποία έχουν κατάληξη –ται και –τε. Τα πειράματα θα διεξαχθούν στην πλατφόρμα WEKA και η αξιολόγηση των διαφορετικών μοντέλων θα βασιστεί σε μετρικές που χρησιμοποιούνται ευρέως στα προβλήματα μηχανικής μάθησης.

Abstract

The problem of automatic spelling correction is a procedure for which many traditional tools have been developed, which, most of the times, fail to detect errors occurring due to the use of incorrect grammar or the selection of words that are not suitable according to the context of the text. The subject of the current thesis is the study of the spelling correction problem in Greek text by using a variety of machine learning techniques.

Machine learning, which is a subfield of computer science, provides solutions to the aforementioned problem, since it explores the study and construction of a variety of algorithms that can learn from data and make predictions on them. The problem of spelling correction belongs to the field of Natural Language Processing (NLP), which is broadly defined as the automated manipulation of natural language, namely speech and text.

In the current thesis, after defining the main concepts that describe the problem of spelling correction, I will analyze different machine learning techniques and apply them on data collected from online newspapers. More specifically, I will study verbs having exactly the same pronunciation, which differ only by their endings. The verbs I will focus on have ‘ται’ and ‘τε’ endings. Experiments will be conducted on WEKA platform and the models’ evaluation will be based on metrics widely used in machine learning problems.

Περιεχόμενα

Λίστα εικόνων

Λίστα πινάκων

Α΄ ΕΙΣΑΓΩΓΗ

Α΄.1 Αντικείμενο της Πτυχιακής Εργασίας

Το αντικείμενο της παρούσας πτυχιακής εργασίας είναι η μελέτη του προβλήματος της αναγνώρισης ορθογραφικών λαθών σε κείμενα της ελληνικής γλώσσας μέσω της χρήσης μεθόδων μηχανικής μάθησης. Η αναγνώριση ορθογραφικών λαθών υπάγεται στον ευρύτερο κλάδο της Επεξεργασίας Φυσικής Γλώσσας (Natural Language Processing), όπου μέσω μίας πληθώρας τεχνικών μηχανικής μάθησης, γίνεται η αυτοματοποιημένη ανάλυση του λόγου, προφορικού και γραπτού. Ο δομικός λίθος αυτών των τεχνικών είναι ο τεράστιος όγκος των διαθέσιμων δεδομένων, τα οποία χρησιμοποιούνται ώστε τα μοντέλα μηχανικής μάθησης να εκπαιδεύονται και να ‘μαθαίνουν’ να δίνουν απαντήσεις στα εκάστοτε προβλήματα.

Επιπλέον γίνεται πραγμάτευση των ορθογραφικών λαθών σε λέξεις που έχουν περισσότερες από μία έννοιες και είναι αδύνατο να αποσαφηνιστεί η σημασιολογία τους δίχως την χρήση τεχνικών μηχανικής μάθησης. Έτσι είναι απαραίτητη η γνώση των χαρακτηριστικών των συμφραζομένων λέξεων, δηλαδή των λέξεων που βρίσκονται πριν και μετά της αμφίσημης λέξης προκειμένου αποκαλυφθεί το νόημα της λέξης που προκαλεί αμφισημία. Με την γραμματική ανάλυση των συμφραζομένων όπως με την εύρεση του προσώπου και του αριθμού αυτών γίνεται ευκολότερη η διασαφήνιση της αμφίσημης λέξης. Είναι αρκετά συνηθισμένο φαινόμενο η εύρεση λέξεων οι οποίες παρουσιάζουν είτε τυπογραφικά λάθη είτε σημασιολογικά είτε και τα δύο. Υπάρχουν και τα γνωστικά λάθη κατά τα οποία ο συγγραφέας αδυνατεί να γνωρίζει τη σωστή ορθογραφία των λέξεων. Κάποιες λέξεις συνήθως “ακούγονται” το ίδιο αλλά έχουν διαφορετικό νόημα και κατάληξη. Αυτές οι λέξεις είναι γνωστές ως ομόηχες. Για παράδειγμα οι λέξεις “αυτοί” και “αυτή” οι οποίες αν και ακούγονται το ίδιο έχουν διαφορετικές καταλήξεις άρα και νόημα και ο τρόπος για να είναι δυνατή η εύρεση της σωστής λέξης είναι η ανάλυση των συμφραζομένων.

Άλλο ένα ενδιαφέρον παράδειγμα είναι οι λέξεις “πολύ” και “πολλοί” οι οποίες ακούγονται και αυτές το ίδιο αλλά έχουν διαφορετικό νόημα. Η πρώτη είναι ενικός αριθμός ενώ η δεύτερη πληθυντικός αριθμός. Με την συντακτική ανάλυση των συμφραζομένων επιτυγχάνεται η γνώση της σωστής λέξης. Επιπλέον, γίνεται η χρήση POS (part-of-speech) tags και συντακτικής ανάλυσης προκειμένου να γίνουν γνωστά τα χαρακτηριστικά των συμφραζομένων λέξεων και να γίνει επιτυχής η άρση της αμφισημίας.

Επιπροσθέτως θα γίνει χρήση της τεχνικής μειονοτικής υπερδειγματοληψείας (SMOTE) για το πρόβλημα της τάξης της ανισορροπίας. Έτσι, με την χρήση ποικίλων αλγόριθμων μηχανικής μάθησης και διαφορετικούς συνδυασμούς σετ χαρακτηριστικών στους αλγόριθμους αυτούς οδηγούμαστε στο πιο ολοκληρωμένο και άρτιο αποτέλεσμα για την σωστή ορθογραφία των ορθογραφικά αμφίσημων λέξεων. Ο όρος «αμφισημία» είναι μια περίπλοκη λέξη κατά την οποία μια λέξη μπορεί να σημαίνει κάτι το τελείως διαφορετικό απ’ ότι θεωρούσαμε εξαρχής. Δηλαδή, από την προφορά της μια λέξη να ακούγεται το ίδιο αλλά στον γραπτό λόγο να παρουσιάζει σημαντικές διαφορές. Για παράδειγμα οι

λέξεις “πολύ” και “πολλοί” οι οποίες “ακούγονται” το ίδιο αλλά γράφονται με διαφορετικές καταλήξεις. Επιπλέον οι λέξεις αυτές έχουν διαφορετικές σημασίες και κατά επέκταση διαφορετικές ερμηνείες. Η Ελληνική γλώσσα έχει πλούσια σε όρους μορφολογία και αμφισημία καθώς περιλαμβάνει στο λεξιλόγιο της ποικίλες λέξεις με διαφορετικά νοήματα. Παρακάτω βρίσκονται κάποια παραδείγματα συχνών ομόηχων λέξεων στις οποίες οι κατάληξη τους τελειώνει σε “τε” και “ται”. Το πρώτο (τε) αναφέρεται σε λέξεις οι οποίες βρίσκονται σε δεύτερο πρόσωπο πληθυντικό αριθμό ενεργητική φωνή. Το δεύτερο (ται) αναφέρεται σε λέξεις που βρίσκονται σε τρίτο πρόσωπο ενικό αριθμό παθητική φωνή. Η σωστή αναγνώριση της κάθε μιας ξεχωριστά μπορεί να επιτευχθεί από την γνώση των χαρακτηριστικών των συμφραζομένων λέξεων.

Οι λέξεις οι οποίες επιλέχθηκαν για την συγγραφή της παρούσας πτυχιακής εργασίας έχουν όλες κατάληξη σε –ται και –τε όντας ένα συχνό ορθογραφικό λάθος το οποίο λαμβάνει χώρα σε μεγάλο βαθμό στα ελληνικά κείμενα.

Παραδείγματα:

Εσείς γράφετε την ορθογραφία σωστά.

Αυτή η λέξη γράφεται με διπλό σίγμα.

Να καθαρίζετε συχνά το σπίτι.

Αυτό καθαρίζεται μόνο με απορρυπαντικό.

Στο κάθε παράδειγμα παραπάνω οι λέξεις διαβάζονται ακριβώς το ίδιο οπότε ο μόνος τρόπος για να γίνει γνωστό το πραγματικό νόημα της κάθε λέξης είναι η πλήρης κατανόηση ολόκληρων των προτάσεων. Μόνο με αυτόν τον τρόπο επιτυγχάνεται εξ'ολοκλήρου η άρση της αμφισημίας. Τα παραδοσιακά εργαλεία αυτόματης διόρθωσης σε κείμενα της ελληνικής γλώσσας βασίζονται στη διόρθωση λέξεων που δεν είναι συμβατές με λέξεις που υπάρχουν σε κάποιο λεξικό. Επομένως, η μέθοδος που προτείνεται έρχεται να δώσει λύση στην περίπτωση της διόρθωσης λέξεων οι οποίες έχουν λανθασμένη ορθογραφία με βάση τα συμφραζόμενα, ωστόσο αποτελούν έγκυρες λέξεις που μπορούν να βρεθούν σε οποιοδήποτε λεξικό. Στην παρούσα πτυχιακή εργασία μελετήθηκε μόνο ένα ζεύγος τέτοιων λέξεων, ωστόσο θα μπορούσαν να εξεταστούν μελλοντικά και άλλες περιπτώσεις ομόφωνων, δηλαδή λέξεων που προφέρονται ακριβώς το ίδιο αλλά έχουν διαφορετική κατάληξη.

Α΄.2 Σκοπός της Πτυχιακής Εργασίας

Ο στόχος της παρούσας πτυχιακής εργασίας είναι η ανάλυση διάφορων κατηγοριών αλγορίθμων μηχανικής μάθησης, η εφαρμογή της σε δεδομένα που αφορούν κείμενα ελληνικής γλώσσας ψηφιακών εφημερίδων και η ανάδειξη του αλγορίθμου με την καλύτερη απόδοση, κατόπιν μιας σειράς πειραματισμών. Τα πειράματα θα διεξαχθούν στο πρόγραμμα Weka και θα δοθεί ιδιαίτερη έμφαση στην προεπεξεργασία των δεδομένων και στον τρόπο με τον υπολογίζονται τα διάνυσματα των χαρακτηριστικών που τροφοδοτούνται στα μοντέλα. Η αξιολόγηση των μοντέλων θα βασιστεί σε μετρικές αξιολόγησης που χρησιμοποιούνται ευρέως στα προβλήματα μηχανικής μάθησης.

Επιπροσθέτως, η μηχανική μάθηση στην αναγνώριση ορθογραφικών λαθών λέξεων στην Ελληνική γλώσσα έχει ως κύριο σκοπό την άρση της αμφισημίας λέξεων που γράφονται με ποικίλους και διαφορετικούς τρόπους, αναλύοντας τα συμφραζόμενα των λέξεων που προηγούνται και έπονται της συντακτικά αμφίσημης λέξης. Τα χαρακτηριστικά αυτά μπορεί να είναι ο αριθμός, το γένος, η πτώση ή και το μέρος του λόγου. Προκειμένου να αποσαφηνιστεί πλήρως η έννοια της αμφίσημης λέξης χρησιμοποιούνται τεχνικές μηχανικής μάθησης με συνδυασμούς χαρακτηριστικών (datasets) και αλγόριθμους. Στην παρούσα πτυχιακή εργασία θα γίνει χρήση των παρακάτω αλγορίθμων: NaiveBayes, RandomForest, kStar, Id3, DecisionStump, J48.

Α΄.2 Δομή της Πτυχιακής Εργασίας

Η παρούσα πτυχιακή εργασία οργανώνεται ως εξής :

Το *πρώτο* κεφάλαιο είναι η εισαγωγή η οποία περιέχει το αντικείμενο της πτυχιακής εργασίας, το σκοπό και την δομή της.

Το *δεύτερο* κεφάλαιο το οποίο είναι το θεωρητικό υπόβαθρο κατά το οποίο γίνεται λόγος για τον ορισμό της μηχανικής μάθησης, τα είδη της τα οποία είναι η επιβλεπόμενη μάθηση, η μη επιβλεπόμενη και η ενισχυτική μάθηση και για τις κατηγορίες προβλημάτων μηχανικής μάθησης οι οποίες είναι η κατηγοριοποίηση, η συσταδοποίηση, η παλινδρόμηση, η εκτίμηση πυκνότητας και η μείωση διαστατικότητας. Στη συνέχεια γίνεται ανάλυση των αλγορίθμων μηχανικής μάθησης καθώς και αξιολόγηση των μοντέλων μηχανικής μάθησης η οποία περιλαμβάνει τις μετρικές Accuracy, Precision, Recall και F-measure, Cross-Validation και το φαινόμενο της υπερεκπαίδευσης. Επιπροσθέτως, γίνεται λόγος για τις εφαρμογές μηχανικής μάθησης.

Το *τρίτο* κεφάλαιο περιλαμβάνει εμβάθυνση στην μηχανική μάθηση και στην αναγνώριση των ορθογραφικών λαθών. Αρχικά γίνεται λόγος για το βασικό πρόβλημα που υπάρχει στα ορθογραφικά λάθη τα οποία έχουν κατάληξη σε –τε και –ται. Εν ακολουθία, αναλύεται η έννοια της αυτόματης αναγνώρισης ορθογραφικών λαθών γενικότερα και, πιο ειδικά, στην περίπτωση της ελληνικής γλώσσας. Επίσης, γίνεται αναφορά σε μία σειρά τεχνικών που έχουν προταθεί στη βιβλιογραφία και αφορούν το υπό μελέτη πρόβλημα.

Το *τέταρτο* κεφάλαιο περιλαμβάνει την ανάλυση της μηχανικής μάθησης σε προβλήματα επεξεργασίας φυσικής γλώσσας. Παρακάτω, γίνεται λόγος για τις τεχνικές μηχανικής μάθησης και εν τέλει συμπεριλαμβάνεται μία επισκόπηση προσεγγίσεων που έχουν προταθεί στη βιβλιογραφία για την επίλυση του υπό μελέτη προβλήματος.

Στο *πέμπτο* κεφάλαιο, αφού δοθούν κάποιες βασικές πληροφορίες για το περιβάλλον WEKA όπου πραγματοποιήθηκαν οι πειραματισμοί, θα γίνει μία παρουσίαση των βημάτων που ακολουθήθηκαν κατά την εκτέλεση των πειραμάτων, καθώς και των αποτελεσμάτων που προέκυψαν.

Στο *έκτο* κεφάλαιο, εξάγονται ορισμένα συμπεράσματα και προτείνονται βελτιώσεις για την επέκταση της παρούσας εργασίας μελλοντικά.

Β΄ Θεωρητικό υπόβαθρο

Β΄.1 Ορισμός μηχανικής μάθησης

Η «μηχανική μάθηση» (machine learning) αποτελεί έναν κλάδο της τεχνητής νοημοσύνης που βασίζεται στην ιδέα ότι οι μηχανές πρέπει να είναι ικανές να μαθαίνουν και να προσαρμόζονται μέσω της εμπειρίας. Αξιοποιώντας τις δυνατότητες της μηχανικής μάθησης, οι υπολογιστές έχουν τη δυνατότητα να διαχειρίζονται νέες καταστάσεις μέσω της ανάλυσης, της αυτοεκπαίδευσης, της παρατήρησης και της εμπειρίας. Ένα γενικό ορισμό έρχεται να δώσει το 1997, ο Tom M. Mitchell, ένας πρωτοπόρος της μηχανικής μάθησης, ο οποίος προέβλεψε την εξέλιξη και τη συνέργεια ανθρώπου και μηχανών. Σύμφωνα, λοιπόν, με τον Mitchell [1] [2],

> *«Δεδομένης μιας εργασίας Τ, ενός μέτρου απόδοσης Ρ και μιας εμπειρίας Ε, μία μηχανή 'μαθαίνει' αν το σύστημα βελτιώνει την απόδοσή του Ρ στην εργασία Τ, ακολουθώντας την εμπειρία Ε».*Η μηχανική μάθηση γεννήθηκε από τη σύμπραξη της επιστήμης των υπολογιστών και της στατιστικής. Ενώ, λοιπόν, η επιστήμη των υπολογιστών απαντά στο ερώτημα 'Πώς μπορούμε να κατασκευάσουμε μηχανές που επιλύουν προβλήματα και, ποια προβλήματα είναι εγγενώς εύκολα/δύσκολα διαχειρίσιμα;', η στατιστική χαρακτηρίζεται από την ερώτηση 'Τι συμπέρασμα συνάγεται από τα δεδομένα σε συνδυασμό με υποθέσεις που αφορούν το μοντέλο και σε τι βαθμό αξιοπιστίας;'. Το καθοριστικό ερώτημα στο οποίο απαντά η μηχανική μάθηση επηρεάζεται από τα προβλήματα που καλούνται να επιλύσουν οι δύο προαναφερθείσες επιστήμες. Ενώ η επιστήμη των υπολογιστών έχει επικεντρωθεί, κυρίως, στο πώς να προγραμματίζονται χειροκίνητα οι υπολογιστές, η μηχανική μάθηση στοχεύει στον αυτοπρογραμματισμό των υπολογιστών μέσω εμπειρίας και κάποιας αρχικής δομής. Επίσης, ενώ η στατιστική ασχολείται κατά κύριο λόγο με τα συμπεράσματα που μπορούν να συναχθούν από τα δεδομένα, τη μηχανική μάθηση την απασχολούν ερωτήματα που αφορούν τις υπολογιστικές αρχιτεκτονικές και τους αλγορίθμους που μπορούν να χρησιμοποιηθούν, ώστε να αποθηκευθούν, να ανακτηθούν και να συγχωνευθούν τα δεδομένα.

Επίσης, απαντά σε ερωτήματα που σχετίζονται με τους τρόπους με τους οποίους διαφορετικές υποεργασίες μάθησης μπορούν να λειτουργήσουν αρμονικά σε ένα ευρύτερο σύστημα, καθώς και ερωτήματα που αφορούν τη διαχείριση των υπολογιστικών διεργασιών [2].

Αναδυόμενη από την αναγνώριση προτύπων και τη θεωρία ότι οι υπολογιστές μπορούν να μάθουν χωρίς να είναι προγραμματισμένοι για να εκτελούν συγκεκριμένες διεργασίες, η μηχανική μάθηση σχετίζεται με την ανάλυση και σχεδίαση αλγορίθμων που μαθαίνουν από τα δεδομένα.

Η διαθεσιμότητα του τεράστιου όγκου δεδομένων, των λεγόμενων ‘Big Data’, είχε σαν φυσικό επακόλουθο την έκθεση των μοντέλων μηχανικής μάθησης σε νέα δεδομένα στα οποία έχουν την ικανότητα να προσαρμόζονται με αυτόνομο τρόπο. Ο καταιγισμός με νέα δεδομένα σε συνδυασμό με

την τεράστια υπολογιστική ισχύ και τις διαθέσιμες αποθήκες δεδομένων οδήγησαν στη γρήγορη και αυτόματη δημιουργία μοντέλων που αναλύουν μεγάλους όγκους σύνθετων δεδομένων και παράγουν αξιόπιστα και ακριβή αποτελέσματα σε πολύ μεγάλη κλίμακα. Αυτοί είναι και οι λόγοι που οδήγησαν στη ραγδαία εξέλιξη και ηχηρή παρουσία της μηχανικής μάθησης στον ερευνητικό χώρο τα τελευταία χρόνια.

Η ύπαρξη του μεγάλου όγκου δεδομένων και η έκθεση των μοντέλων μηχανικής μάθησης σε αυτά, οδήγησε στην ανάπτυξη ενός νέου κλάδου, αυτόν της εξόρυξης δεδομένων (data mining). Η εξόρυξη δεδομένων ταυτίζεται με την υπολογιστική διαδικασία ανακάλυψης προτύπων σε τεράστια σύνολα δεδομένων, χρησιμοποιώντας τεχνικές από τους χώρους της μηχανικής μάθησης, της στατιστικής και των βάσεων δεδομένων [3]. Οι εφαρμογές της εξόρυξης δεδομένων συναντώνται σε διάφορους χώρους, όπως στην ιατρική, όπου γίνεται η διάγνωση του καρκίνου μέσω του εντοπισμού των καρκινικών κυττάρων έναντι των φυσιολογικών.

Στον τραπεζικό κλάδο, τα μοντέλα της εξόρυξης δεδομένων χρησιμοποιούνται για την ανίχνευση απάτης, ενώ στη βιομηχανία του εμπορίου γίνεται η ανάλυση παρελθοντικών δεδομένων και η διαχείριση αποθέματος. Τέλος, οι τεράστιοι όγκοι δεδομένων σε διάφορους επιστημονικούς κλάδους, όπως αυτούς της βιολογίας, της φυσικής και της αστρονομίας, αναλύονται και επεξεργάζονται αποτελεσματικά μόνο από τους υπολογιστές [4].

Ωστόσο, η μηχανική μάθηση συνδέεται άρρηκτα και με τον κλάδο της τεχνητής νοημοσύνης. Προκειμένου να αναπτύξει ευφυία, ένα σύστημα που βρίσκεται σε ένα περιβάλλον που μεταβάλλεται διαρκώς, θα πρέπει να αναπτύξει δεξιότητες που να του επιτρέπουν να μαθαίνει και να προσαρμόζεται σε αλλαγές, ώστε να μπορεί να ανταποκρίνεται σε όλα τα δυνατά σενάρια.

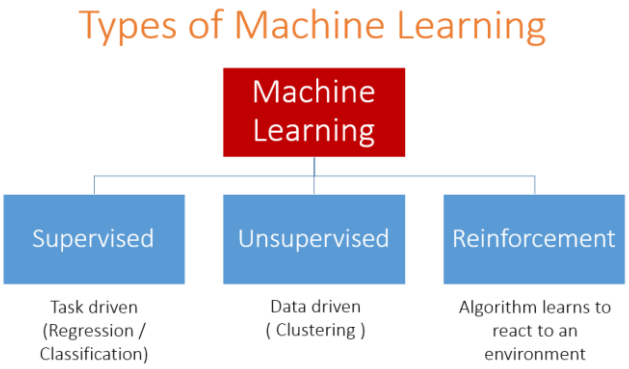
Συνδυάζοντας τις βασικές έννοιες που διέπουν την εξόρυξη δεδομένων και την τεχνητή νοημοσύνη, μπορούμε να αποφανθούμε ότι η μηχανική μάθηση ταυτίζεται με τη χρήση δεδομένων και παρελθοντικής εμπειρίας, της λεγόμενης μεταγνώσης, για τον προγραμματισμό των υπολογιστών, ώστε να μαθαίνουν να επιλύουν προβλήματα και να βελτιστοποιούν την απόδοσή τους. Το παραγόμενο μοντέλο πρέπει να καταλήγει στον καλύτερο συνδυασμό παραμέτρων μέσω της διαδικασίας της εκπαίδευσης και να μπορεί να χρησιμοποιείται είτε σαν προγνωστικό μοντέλο για την εκτέλεση προβλέψεων για μελλοντικές καταστάσεις, είτε σαν περιγραφικό μοντέλο για την απόκτηση χρήσιμης γνώσης από τα δεδομένα, είτε να συνδυάζει και τις δύο περιπτώσεις.

Οι δύο βασικές κατευθύνσεις στις οποίες πρέπει να δοθεί προσοχή κατά τη δημιουργία ενός μοντέλου μηχανικής μάθησης αφορούν τη διαδικασία εκπαίδευσής του καθώς και τη μετέπειτα αξιολόγησή του. Αρχικά, όσον αφορά τη διαδικασία εκπαίδευσης, θα πρέπει να χρησιμοποιηθούν οι κατάλληλες τεχνικές που θα επιλύσουν το πρόβλημα βελτιστοποίησης, καθώς και να γίνει η αποτελεσματική διαχείριση των διαθέσιμων δεδομένων. Επιπλέον, μετά την εκπαίδευση του μοντέλου, θα πρέπει να εξασφαλιστεί η υψηλή του απόδοση, καθώς και να ληφθούν υπόψη ζητήματα που αφορούν τη χωρική και χρονική πολυπλοκότητα του αλγορίθμου.

Η δυναμική που χαρακτηρίζει τη μηχανική μάθηση τα τελευταία χρόνια, έρχεται να δικαιολογήσει τις ποικίλες εφαρμογές της. Ενδεικτικά, αξίζει να αναφερθεί η εφαρμογή της μηχανικής μάθησης για την εξασφάλιση της ασφάλειας των δεδομένων και τον εντοπισμό εισβολέων σε ιδιωτικά δίκτυα. Επίσης, η μηχανική μάθηση χρησιμοποιείται για την ανίχνευση απάτης (fraud detection), το φιλτράρισμα ηλεκτρονικής αλληλογραφίας (spam filtering) και την πρόβλεψη χρηματιστηριακών δεικτών. Τέλος, είναι αξιοσημείωτη η χρήση της μηχανικής μάθησης στον κλάδο της υγείας, για τον πρώιμο εντοπισμό του καρκίνου, στον κλάδο του μάρκετινγκ για την οργάνωση στοχευμένων εκστρατειών καθώς και στην επεξεργασία φυσικής γλώσσας (natural language processing - NLP) [5]

Β΄.2 Είδη μηχανικής μάθησης

Κάθε μοντέλο μηχανικής μάθησης αντιλαμβάνεται την εμπειρία από προηγούμενες καταστάσεις ως δεδομένα, τα οποία περιγράφουν ένα πρόβλημα, και καλείται να προβλέψει μελλοντικές συμπεριφορές.



Εικόνα 1 Είδη μηχανικής μάθησης

Πηγή: http://en.proft.me/2015/12/24/types-machine-learning-algorithms/

Ανάλογα με τη φύση των σημάτων ή την ύπαρξη ανατροφοδότησης σε ένα σύστημα μάθησης, η μηχανική μάθηση διακρίνεται στα τρία παρακάτω είδη [6]:

Β΄.2.1 Επιβλεπόμενη μάθηση (supervised learning)

Όσον αφορά την επιβλεπόμενη μάθηση, χρησιμοποιούνται δεδομένα, τα οποία ονομάζονται δεδομένα εκπαίδευσης (training set), ώστε να γίνει η κατασκευή μίας συνάρτησης που θα αντιστοιχίζει δεδομένες εισόδους σε συγκεκριμένες εξόδους. Τα δεδομένα εκπαίδευσης αποτελούνται από ένα σύνολο εγγραφών, όπου κάθε εγγραφή χαρακτηρίζεται από ένα διάνυσμα εισόδου και την επιθυμητή τιμή της εξόδου. Ο αλγόριθμος επιβλεπόμενης μάθησης αναλύει και επεξεργάζεται τα δεδομένα εκπαίδευσης, ώστε να παράγει την κατάλληλη συνάρτηση, η οποία θα χρησιμοποιείται για τη μετέπειτα απεικόνιση νέων δεδομένων, όπου πλέον η έξοδος θα είναι άγνωστη. Να σημειωθεί ότι ο βασικός στόχος κατά τη διαδικασία της εκπαίδευσης είναι η γενίκευση του αλγορίθμου, ώστε η αντιστοίχιση των νέων εισόδων στις αντίστοιχες εξόδους να γίνεται με τη μέγιστη δυνατή ακρίβεια.

Παρακάτω δίνονται αναλυτικά τα βήματα που ακολουθούνται κατά τη δημιουργία ενός μοντέλου επιβλεπόμενης μηχανικής μάθησης [7]

1. Αρχικά, θα πρέπει να διευκρινιστεί το είδος των δεδομένων εκπαίδευσης. Για παράδειγμα, στην περίπτωση της ανάλυσης γραφικού χαρακτήρα, οι εγγραφές μπορεί να είναι ένας απλός χαρακτήρας γραμμένος στο χέρι, μία λέξη από τέτοιους χαρακτήρες ή μια γραμμή από λέξεις.
2. Στη συνέχεια, θα πρέπει να γίνει η συλλογή των δεδομένων εκπαίδευσης. Όπως προαναφέρθηκε, στην περίπτωση της επιβλεπόμενης μάθησης, τα δεδομένα εκπαίδευσης πρέπει να αποτελούνται τόσο από τα διανύσματα εισόδου όσο και από τις επιθυμητές εξόδους. Τα δεδομένα αυτά είναι δυνατό να συλλεχθούν από ειδικούς ή να είναι αποτέλεσμα πειραματικών μετρήσεων.

3. Στο βήμα αυτό θα πρέπει να αποφασιστεί ο τρόπος με τον οποίο θα αναπαρασταθούν τα χαρακτηριστικά εισόδου, τα οποία θα τροφοδοτηθούν στον αλγόριθμο μάθησης. Συνήθως, το διάνυσμα εισόδου μετατρέπεται σε ένα διάνυσμα χαρακτηριστικών, το οποίο περιγράφει το αντικείμενο. Ο αριθμός των χαρακτηριστικών δεν πρέπει να είναι ιδιαίτερα μεγάλος, λόγω του φαινομένου της κατάρας της διαστατικότητας (curse of dimensionality), ωστόσο, θα πρέπει τα χαρακτηριστικά που θα επιλεγούν να εμπεριέχουν αρκετή πληροφορία, ώστε να μπορεί να προβλεφθεί με ακρίβεια η έξοδος.
4. Η επιλογή του κατάλληλου αλγορίθμου ανήκει στο επόμενο στάδιο της διαδικασίας. Ο αλγόριθμος αυτός επιλέγεται από ένα ευρύ φάσμα τεχνικών μηχανικής μάθησης, όπως είναι τα δέντρα απόφασης, η μέθοδος Naïve Bayes, τα νευρωνικά δίκτυα, τα μοντέλα SVM, κτλ.
5. Αφού ολοκληρωθεί η σχεδίαση, στον αλγόριθμο τροφοδοτούνται τα δεδομένα εκπαίδευσης. Αρκετοί αλγόριθμοι επιβλεπόμενης μάθησης απαιτούν από το χρήστη να ορίσει ορισμένες παραμέτρους ελέγχου, τις λεγόμενες υπερπαραμέτρους. Οι παράμετροι αυτές μπορούν να αναπροσαρμόζονται, μέχρις ότου βελτιστοποιηθεί η απόδοση του αλγορίθμου σε ένα υποσύνολο των δεδομένων εκπαίδευσης, το οποίο λέγεται σύνολο επαλήθευσης (validation set), ή μέσω της διαδικασίας cross-validation.
6. Τέλος, θα πρέπει να ελεγχθεί πόσο ακριβή είναι τα αποτελέσματα που παράγει ο αλγόριθμος. Αφού γίνει η αναπροσαρμογή των παραμέτρων και ολοκληρωθεί η διαδικασία της εκπαίδευσης, θα πρέπει να μετρηθεί η απόδοση του αλγορίθμου μάθησης στο σύνολο ελέγχου (test set).

Αφού ολοκληρωθούν όλα τα παραπάνω βήματα, το τελικό μοντέλο είναι έτοιμο να εφαρμοστεί σε νέα δεδομένα, με άγνωστες εξόδους. Η διαδικασία που ακολουθείται έχει ως εξής: το μοντέλο τροφοδοτείται με τις νέες εισόδους, εξάγεται το διάνυσμα των χαρακτηριστικών και γίνεται η πρόβλεψη, οπότε, τελικά, επιστρέφονται οι προβλέψεις των κλάσεων της μεταβλητής εξόδου.

Β΄.2.2 Μη επιβλεπόμενη μάθηση (unsupervised learning)

Στην περίπτωση της μη επιβλεπόμενης μάθησης ο αλγόριθμος προσπαθεί να ανακαλύψει την κρυμμένη δομή των δεδομένων, καθώς δεν είναι γνωστές οι κλάσεις της εξόδου. Σε αντίθεση με την επιβλεπόμενη μάθηση, ο στόχος της μη επιβλεπόμενης μάθησης είναι πιο υποκειμενικός, καθώς δεν υπάρχει συγκεκριμένος στόχος για την ανάλυση, όπως είναι η πρόβλεψη της εξόδου. Η μη επιβλεπόμενη μάθηση χρησιμοποιείται στη στατιστική για την εκτίμηση πυκνότητας, καθώς και για να εντοπίσει τα σημαντικότερα χαρακτηριστικά των δεδομένων (Principal Component Analysis - PCA).

Β΄.2.3 Ενισχυτική μάθηση (reinforcement learning)

Η ενισχυτική μάθηση αποτελεί ένα είδος δυναμικού προγραμματισμού που εκπαιδεύει αλγορίθμους χρησιμοποιώντας ένα σύστημα επιβραβεύσεων και ποινών. Αυτό το είδος μάθησης βρίσκει εφαρμογή σε διάφορους τομείς, όπως στη θεωρία πληροφοριών, στη θεωρία παιγνίων, στη νοημοσύνη σμήνους και στους γενετικούς αλγορίθμους.

Ένας αλγόριθμος ή πράκτορας ενισχυτικής μάθησης μαθαίνει αλληλεπιδρώντας με το περιβάλλον του, χωρίς ανθρώπινη παρέμβαση, όπου ανάλογα με την απόδοσή του, δέχεται μία σειρά επιβραβεύσεων ή ποινών.

Πιο συγκεκριμένα, σε αυτό το είδος μάθησης συναντώνται οι έννοιες της εξερεύνησης (exploration) και της εκμετάλλευσης (exploitation). Η εξερεύνηση αναφέρεται στην ανακάλυψη άγνωστου χώρου, ενώ η εκμετάλλευση στην αξιοποίηση της ήδη υπάρχουσας γνώσης. Επειδή οι αλγόριθμοι ενισχυτικής μάθησης χρησιμοποιούν τεχνικές δυναμικού προγραμματισμού, το περιβάλλον δομείται ως μία διαδικασία απόφασης Markov (Markov Decision Process) [8].

Β΄.3 Κατηγορίες προβλημάτων μηχανικής μάθησης

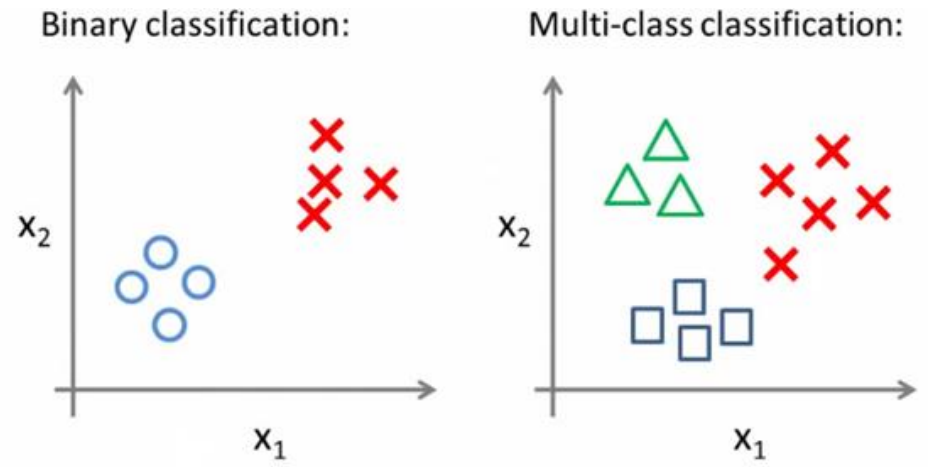
Στην ενότητα αυτή θα γίνει αναφορά στα βασικά προβλήματα που καλούνται να επιλύσουν οι αλγόριθμοι μηχανικής μάθησης.

Β΄.3.1 Κατηγοριοποίηση (classification)

Η κατηγοριοποίηση, η οποία ανήκει στην επιβλεπόμενη μάθηση, έχει σαν στόχο τον προσδιορισμό της κλάσης εξόδου κάθε νέας παρατήρησης. Η εκπαίδευση ενός μοντέλου κατηγοριοποίησης γίνεται σε ένα σύνολο δεδομένων το οποίο αποτελείται από εγγραφές, όπου κάθε εγγραφή (**x**, y) χαρακτηρίζεται από το διάνυσμα **x** των χαρακτηριστικών και την κλάση εξόδου y. Ένας γενικός ορισμός της κατηγοριοποίησης είναι ο εξής [9]:

> *Ως κατηγοριοποίηση ορίζεται η διαδικασία εκμάθησης μίας συνάρτησης στόχου f, μέσω της οποίας κάθε σύνολο χαρακτηριστικών* ***x*** *αντιστοιχίζεται σε μία από τις προκαθορισμένες κλάσεις της εξόδου y.*

Η κατηγοριοποίηση διακρίνεται σε δύο κατηγορίες ανάλογα με τον αριθμό των τιμών που λαμβάνει η έξοδος. Στην περίπτωση που οι τιμές της εξόδου είναι μόνο δύο έχουμε δυαδική κατηγοριοποίηση (binary classification), ενώ αν οι κλάσεις της εξόδου είναι περισσότερες από δύο, τότε μιλάμε για κατηγοριοποίηση πολλών κλάσεων (multi-class classification). Στην παρακάτω εικόνα δίνεται ένα παράδειγμα δυαδικής και πολλών κλάσεων κατηγοριοποίησης αντίστοιχα.



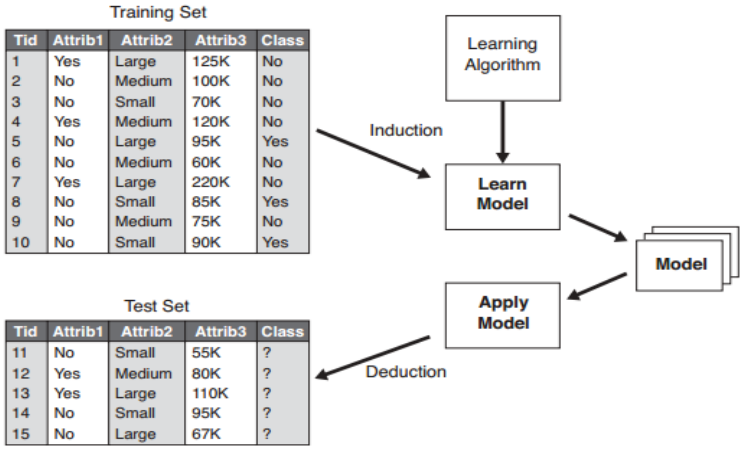
Εικόνα 2 Παράδειγμα κατηγοριοποίησης

Πηγή: www.holehouse.org

Ανάλογα με το σκοπό για τον οποίο προορίζονται, τα μοντέλα κατηγοριοποίησης διακρίνονται στα περιγραφικά και τα προγνωστικά. Στην περίπτωση των περιγραφικών μοντέλων, γίνεται μία συνοπτική παρουσίαση των δεδομένων και η διάκριση των κλάσεων βάσει των χαρακτηριστικών τους. Τα προγνωστικά μοντέλα, από την άλλη μεριά, στοχεύουν στην πρόβλεψη της κλάσης εξόδου δεδομένων που τροφοδοτούνται στο σύστημα. Μπορούμε να φανταστούμε το μοντέλο σαν ένα μαύρο κουτί το οποίο δέχεται τα δεδομένα εισόδου και εξάγει την προβλεπόμενη κλάση εξόδου.

Μερικά παραδείγματα διαδεδομένων τεχνικών κατηγοριοποίησης είναι τα δέντρα απόφασης, τα νευρωνικά δίκτυα, οι μηχανές SVM και οι τεχνικές Naïve Bayes. Σε κάθε τεχνική εφαρμόζεται ένας αλγόριθμος εκμάθησης μέσω του οποίου αποτυπώνεται με το βέλτιστο τρόπο η σχέση μεταξύ του συνόλου χαρακτηριστικών και της κλάσης εξόδου των δεδομένων εισόδου. Το παραγόμενο μοντέλο πρέπει να συσχετίζει αποτελεσματικά τα άγνωστα δεδομένα και να προβλέπει σωστά τις κλάσεις εξόδου αυτών. Επομένως, ο βασικός στόχος του αλγορίθμου εκμάθησης είναι να κατασκευάζει μοντέλα με καλή ικανότητα γενίκευσης, δηλαδή μοντέλα που προβλέπουν με υψηλή ακρίβεια τις κλάσεις εξόδου των άγνωστων παρατηρήσεων.

Στην εικόνα που ακολουθεί παρουσιάζεται η διαδικασία που ακολουθείται για την εκπαίδευση ενός μοντέλου κατηγοριοποίησης.



Εικόνα 3 Διαδικασία κατασκευής ενός μοντέλου κατηγοριοποίησης

Πηγή: [4]

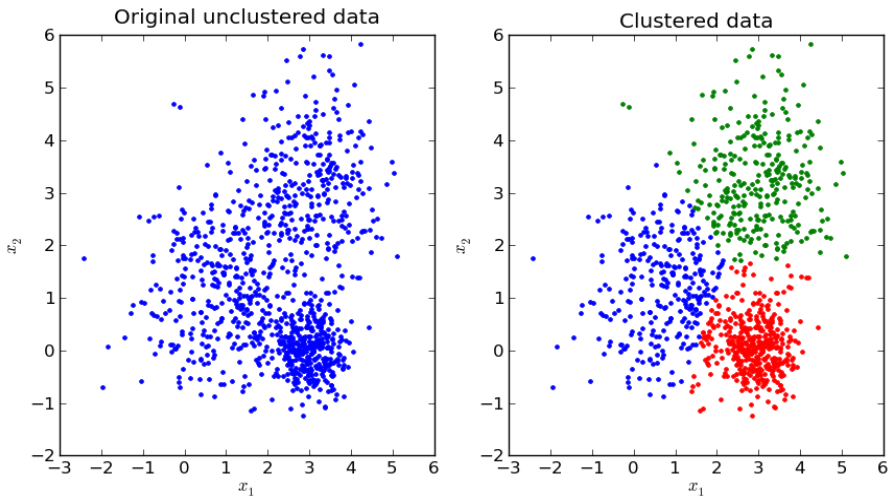
Παρατηρούμε στην παραπάνω εικόνα ότι τα δεδομένα εκπαίδευσης εισάγονται στο μοντέλο το οποίο γενικεύεται από τον αλγόριθμο εκμάθησης. Εφόσον πρόκειται για κατηγοριοποίηση, τα δεδομένα εκπαίδευσης περιλαμβάνουν και την πληροφορία για τις κλάσεις εξόδου, οι οποίες στη συγκεκριμένη περίπτωση λαμβάνουν τις τιμές ‘Yes’ και ‘No’. Αφού βρεθεί το μοντέλο με τη μέγιστη δυνατή ακρίβεια πρόβλεψης, μέσω της αναπροσαρμογής των παραμέτρων, αυτό εφαρμόζεται στα δεδομένα ελέγχου, όπου, πλέον, οι άγνωστες κλάσεις εξόδου προβλέπονται από το παραγόμενο μοντέλο.

Β΄.3.2 Συσταδοποίηση (clustering)

Η συσταδοποίηση ανήκει στη μη επιβλεπόμενη μάθηση και αφορά το διαχωρισμό μίας συλλογής δειγμάτων σε ομάδες, βάσει κάποιας ομοιότητας [10].

Τα δείγματα αυτά, συνήθως, αναπαρίστανται σαν ένα διάνυσμα μετρήσεων ή ένα σημείο στον πολυδιάστατο χώρο.

Ουσιαστικά, μπορούμε να πούμε ότι τα δείγματα που ανήκουν σε μία ομάδα παρουσιάζουν περισσότερες ομοιότητες σε σχέση με δείγματα που ανήκουν σε διαφορετικές ομάδες. Ένα παράδειγμα συσταδοποίησης φαίνεται στην εικόνα που ακολουθεί, όπου αριστερά δίνεται το σύνολο των αρχικών δεδομένων και δεξιά ο διαχωρισμός των σημείων σε 3 διαφορετικές ομάδες (διαφορετικά χρώματα) μέσω του αλγορίθμου k-means.



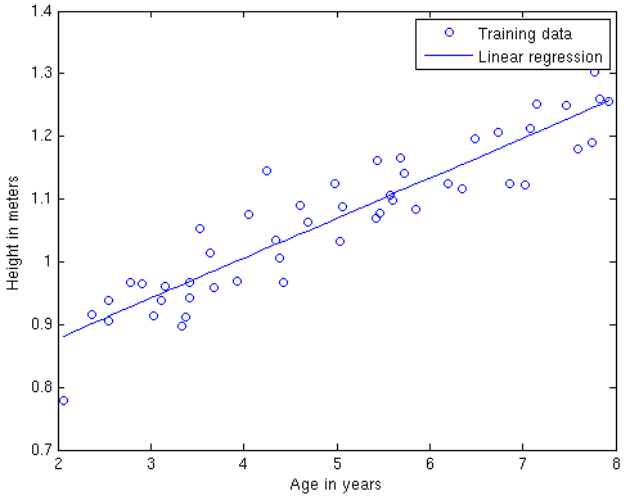
Εικόνα 4 Παράδειγμα συσταδοποίησης

Η ποικιλία μεταξύ των διαφορετικών τεχνικών για την αναπαράσταση και οργάνωση των δεδομένων καθώς και για τον υπολογισμό των μετρικών ομοιότητας μεταξύ των διαφορετικών δειγμάτων, είχε σαν αποτέλεσμα την ανάπτυξη πολλών διαφορετικών μεθόδων συσταδοποίησης, με την k-means να εμφανίζεται σαν την πιο διαδεδομένη.

Σε αντιδιαστολή με την κατηγοριοποίηση, στη συσταδοποίηση δεν είναι γνωστή η επιθυμητή έξοδος, οπότε ο αλγόριθμος πρέπει να ανακαλύψει κρυμμένα πρότυπα και συσχετίσεις μεταξύ των δεδομένων, ώστε να τα οργανώσει σε διαφορετικές ομάδες (clusters).Η συσταδοποίηση συναντάται σε πολλά προβλήματα μηχανικής μάθησης, όπως είναι η ανάκτηση κειμένου, η τμηματοποίηση εικόνας και η ανάλυση προτύπων. Η παρουσία της σε αρκετά από αυτά τα προβλήματα ενισχύεται λόγω της ύπαρξης ελάχιστης παρελθοντικής πληροφορίας για τα δεδομένα, γεγονός που οδηγεί στην αναζήτηση συσχετίσεων μεταξύ των δειγμάτων και στην ανακάλυψη κρυφών δομών.

Β΄.3.3 Παλινδρόμηση (regression)

Η παλινδρόμηση ή αλλιώς γραμμική παλινδρόμηση, όπως και η κατηγοριοποίηση, ανήκει στην επιβλεπόμενη μάθηση, ωστόσο διαφοροποιείται από αυτήν, καθώς οι έξοδοι λαμβάνουν συνεχείς και όχι διακριτές τιμές. Αποτελεί μία στατιστική διαδικασία για την εκτίμηση της σχέσης μεταξύ μιας εξαρτημένης μεταβλητής και μιας ή περισσότερων ανεξάρτητων μεταβλητών. Πιο συγκεκριμένα, μέσω της διαδικασίας της παλινδρόμησης μελετάται η μεταβολή της τιμής της εξαρτημένης μεταβλητής όταν μεταβάλλεται οποιαδήποτε από τις ανεξάρτητες μεταβλητές. Παρακάτω δίνεται ένα παράδειγμα παλινδρόμησης, όπου φαίνεται η γραμμική εξάρτηση του ύψους των παιδιών από την ηλικία.



Εικόνα 5 Παράδειγμα παλινδρόμησης

Πηγή: http://openclassroom.stanford.edu

Ένα μοντέλο παλινδρόμησης μπορεί να περιγραφεί με βάση τις παρακάτω παραμέτρους και μεταβλητές:

- Τις άγνωστες παραμέτρους $\boldsymbol{\beta}$, που αναπαριστούν ένα διάνυσμα.
- Τις ανεξάρτητες μεταβλητές $\boldsymbol{X}$.
- Την εξαρτημένη μεταβλητή $\boldsymbol{Y}$.

Επομένως, μπορούμε να πούμε ότι ένα μοντέλο παλινδρόμησης συσχετίζει την $\boldsymbol{Y}$ με μία συνάρτηση των $\boldsymbol{X}$ και $\boldsymbol{\beta}$

$$\boldsymbol{Y} \approx f(\boldsymbol{X}, \boldsymbol{\beta})$$

Εκτός από τη γραμμική παλινδρόμηση, υπάρχει και η λογιστική παλινδρόμηση, η οποία αποτελεί ένα είδος πιθανοτικού στατιστικού μοντέλου κατηγοριοποίησης.

Χρησιμοποιείται για να παράγει μία δυαδική πρόβλεψη μιας κατηγορικής μεταβλητής (π.χ. 0 ή 1, επιτυχία ή αποτυχία, υγιής ή ασθενής), η οποία εξαρτάται από μία ή περισσότερες μεταβλητές πρόβλεψης. Πέρα από τη δυαδική περίπτωση, η εξαρτημένη μεταβλητή μπορεί να έχει περισσότερες από δύο κατηγορίες, οπότε σε αυτήν την περίπτωση μιλάμε για πολλαπλή λογιστική παλινδρόμηση.

Β΄.3.4 Εκτίμηση πυκνότητας (density estimation)

Η εκτιμηση πυκνότητας ανήκει στη μη επιβλεπόμενη μάθηση και αφορά τη μελέτη παρατηρήσιμων δεδομένων για την κατασκευή της εκτίμησης μιας μη παρατηρήσιμης, κρυφής συνάρτησης πυκνότητας πιθανότητας. Η μη παρατηρήσιμη συνάρτηση πυκνότητας θεωρείται σαν την πυκνότητα σύμφωνα με την οποία κατανέμεται ένας μεγάλος πληθυσμός, όπου τα δεδομένα θεωρούνται ένα τυχαίο δείγμα αυτού.

Μία πολύ συχνή χρήση της εκτίμησης πυκνότητας εντοπίζεται στην εξερεύνηση των ιδιοτήτων ενός δοθέντος συνόλου δεδομένων, από όπου μπορεί να ανακαλυφθεί πολύτιμη γνώση, όπως για παράδειγμα δεδομένα που δεν ακολουθούν ομοιόμορφη (uniform) ή πολυτροπική (multimodal) κατανομή. Μία σημαντική πτυχή της στατιστικής, που συχνά παραγκωνίζεται, είναι η παρουσίαση των δεδομένων πίσω στον πελάτη, έτσι ώστε να παρέχεται επεξήγηση των συμπερασμάτων που είναι πιθανό να προέκυψαν με διαφορετικούς τρόπους. Οι εκτιμήσεις πυκνότητας είναι ιδανικές για αυτό το σκοπό, καθώς μπορούν να γίνουν εύκολα κατανοητές και από άλλα άτομα, πέρα των μαθηματικών.

Β΄.3.5 Μείωση διαστατικότητας (dimensionality reduction)

Μία ακόμη εφαρμογή της μη επιβλεπόμενης μάθησης είναι η μείωση διαστατικότητας, η οποία αναφέρεται στη διαδικασία μείωσης του αριθμού των ανεξάρτητων μεταβλητών, μέσω της εύρεσης ενός αριθμού κύριων συνιστωσών. Δηλαδή, αν φανταστούμε ότι το διάνυσμα χαρακτηριστικών τοποθετείται σε ένα N-διάστατο χώρο, με τη μείωση της διαστατικότητας οι διαστάσεις του χώρου αυτού μειώνονται, οπότε ελαττώνεται και η πολυπλοκότητα του προβλήματος. Η μείωση διαστατικότητας διακρίνεται στις διαδικασίες της επιλογής και εξαγωγής χαρακτηριστικών (feature selection – feature extraction).

Οι μέθοδοι επιλογής χαρακτηριστικών προσπαθούν να βρουν ένα υποσύνολο των αρχικών μεταβλητών με τρόπο ώστε να περιγράφεται το μεγαλύτερο ποσοστό της αρχικής πληροφορίας. Υπάρχουν τρεις στρατηγικές επιλογής χαρακτηριστικών: η στρατηγική *filter*, όπως για παράδειγμα το κέρδος πληροφορίας, η στρατηγική *wrapper*, όπου η αναζήτηση καθοδηγείται από την ακρίβεια και η στρατηγική *embedded*, κατά την οποία τα χαρακτηριστικά επιλέγονται για να προστεθούν ή να αφαιρεθούν κατά τη δημιουργία του μοντέλου, ανάλογα με τα σφάλματα πρόβλεψης. Σε ορισμένες περιπτώσεις, η ανάλυση δεδομένων και η κατηγοριοποίηση μπορούν να γίνουν στο μειωμένο χώρο με μεγαλύτερη ακρίβεια από ό,τι στον αρχικό χώρο.

Η εξαγωγή χαρακτηριστικών μετασχηματίζει τα δεδομένα από το χώρο των πολλών διαστάσεων σε ένα χώρο με λιγότερες διαστάσεις. Ο μετασχηματισμός των δεδομένων μπορεί να είναι γραμμικός, όπως στην περίπτωση της ανάλυσης κύριων συνιστωσών (principal component analysis), ωστόσο υπάρχουν αρκετές μέθοδοι που εκτελούν μείωση διαστατικότητας με μη γραμμικό τρόπο.

Β΄.4 Αξιολόγηση μοντέλων μηχανικής μάθησης

Προκειμένου να αξιολογηθεί η απόδοση των μοντέλων μηχανικής μάθησης, χρησιμοποιούνται οι μετρικές accuracy, precision, recall και F-measure, οι οποίες υπολογίζονται με βάση τον πίνακα σύγχυσης (confusion matrix) [11] που ακολουθεί:

		Κλάση πρόβλεψης	
		Απώλεια	**Μη απώλεια**
Πραγματική κλάση	**Απώλεια**	TP	FN
	Μη απώλεια	FP	TN

Πίνακας 1 Πίνακας σύγχυσης για την αξιολόγηση ταξινομητών

Ο πίνακας σύγχυσης βοηθά στην οπτικοποίηση της απόδοσης των αλγορίθμων επιβλεπόμενης μάθησης, όπου κάθε στήλη του αναπαριστά τα δείγματα της κλάσης πρόβλεψης, ενώ κάθε γραμμή του αντιστοιχεί στα δείγματα της πραγματικής κλάσης. Οι συμβολισμοί TP (True Positive) και FN (False Negative) αντιστοιχούν στις περιπτώσεις όπου τα θετικά παραδείγματα ταξινομούνται σωστά και τα αρνητικά παραδείγματα ταξινομούνται λάθος αντίστοιχα. Οι δύο επόμενοι συμβολισμοί, FP (False Positive) και TN (True Negative), δηλώνουν τα αρνητικά παραδείγματα που ταξινομούνται σωστά και τα θετικά παραδείγματα που ταξινομούνται λάθος αντίστοιχα.

Σε γενικές γραμμές, μπορούμε να πούμε ότι ο πίνακας σύγχυσης δείχνει τους τρόπους κατά τους οποίους το μοντέλο συγχέει τις τιμές που δίνει στην έξοδο όταν κάνει προβλέψεις. Το πιο σημαντικό είναι ότι δε δίνει μόνο τη γνώση για τα λάθη που έκανε το μοντέλο μάθησης, αλλά και για τους τύπους των λαθών. Παρακάτω δίνεται η περιγραφή των τεσσάρων μετρικών αξιολόγησης.

Β΄.4.1 Μετρική Accuracy

Η μετρική accuracy, η οποία χρησιμοποιείται πολύ συχνά για την αξιολόγηση των μοντέλων μηχανικής μάθησης, δηλώνει το ποσοστό του συνολικού αριθμού των σωστών προβλέψεων και ο αντίστοιχος τύπος της είναι:

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

(1)

Έτσι, για παράδειγμα, αν ένα μοντέλο κατηγοριοποίησης ή όπως συχνά αναφέρεται ως ταξινομητής (classifier), μπορεί να προβλέψει σωστά τις κλάσεις των μισών εγγραφών, τότε λέμε ότι επιτυγχάνεται ακρίβεια ίση με 50%.

Β΄.4.2 Μετρική Precision

Η μετρική precision δηλώνει το ποσοστό των θετικών παραδειγμάτων που ταξινομήθηκαν σωστά και υπολογίζεται με βάση τον τύπο:

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

(2)

Για παράδειγμα, κατά την αναζήτηση κειμένου σε ένα σύνολο εγγράφων, σαν precision ορίζεται ο αριθμός των σωστών αποτελεσμάτων, αν αυτά διαιρεθούν με το σύνολο των επιστρεφόμενων αποτελεσμάτων. Προφανώς, όσο αυξάνεται το precision, τόσο μειώνεται ο αριθμός των FN, δηλαδή των αρνητικών παραδειγμάτων που δεν ταξινομούνται σωστά.

Β΄.4.3 Μετρική Recall

Η μετρική recall είναι το ποσοστό των θετικών παραδειγμάτων που αναγνωρίστηκαν σωστά από τον ταξινομητή και δίνεται από τον παρακάτω τύπο:

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

(3)

Κατά την κατηγοριοποίηση, μετρική recall με τιμή 1 για μία κλάση Α σημαίνει ότι κάθε στοιχείο από την κλάση Α αντιστοιχήθηκε ορθώς σε αυτή, ωστόσο απουσιάζει η πληροφορία για το ποσοστό των παραδειγμάτων που θεωρήθηκε λανθασμένα ότι ανήκουν επίσης στην κλάση Α. Για τη μετρική recall ισχύει ότι όσο μεγαλύτερη είναι η τιμή της τόσο λιγότερα θετικά παραδείγματα έχουν ταξινομηθεί λάθος.

Β΄.4.4 Μετρική F-Measure

Οι μετρικές precision και recall δεν μπορούν από μόνες τους να περιγράψουν την αποτελεσματικότητα ενός ταξινομητή, καθώς αν η μία μαρτυρά καλή απόδοση δε σημαίνει ότι θα συμβαίνει το ίδιο και με την άλλη. Για να ξεπεραστεί αυτό το πρόβλημα, χρησιμοποιείται μία τέταρτη μετρική, η F-measure, η οποία ορίζεται ως ο αρμονικός μέσος των μετρικών precision και recall, οπότε ο αντίστοιχος τύπος της είναι:

$$F - measure = \frac{2\ x\ Precision\ x\ Recall}{Precision+Recall}$$

(4)

Η μετρική F-measure προσεγγίζει το μέσο όρο των precision και recall όταν οι τιμές αυτών πλησιάζουν, ενώ, σε γενικές γραμμές θεωρείται ο αρμονικός μέσος τους, ο οποίος, στην περίπτωση δύο αριθμών, ταυτίζεται με το τετράγωνο του γεωμετρικού μέσου διαιρούμενο με τον αριθμητικό μέσο. Όταν η τιμή της μετρικής F-measure προσεγγίζει τη μονάδα, τότε έχει επιτευχθεί ένας καλός συνδυασμός των precision και recall.

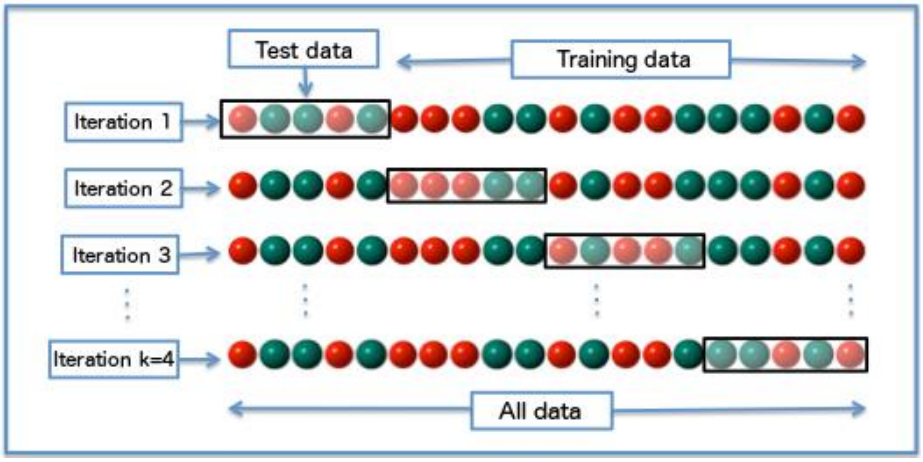
Β΄.5 Cross-validation

Το cross-validation είναι μία μέθοδος αξιολόγησης μοντέλων μηχανικής μάθησης, όπου εκτιμάται η ικανότητα γενίκευσης των αποτελεσμάτων μίας στατιστικής ανάλυσης σε ένα ανεξάρτητο σύνολο δεδομένων. Χρησιμοποιείται, κυρίως, σε προβλήματα πρόβλεψης, όπου ο στόχος είναι να εκτιμηθεί ο βαθμός ακρίβειας που θα έχει το μοντέλο πρόβλεψης στην πράξη. Σε ένα πρόβλημα πρόβλεψης, λοιπόν, το μοντέλο τροφοδοτείται με δεδομένα όπου είναι γνωστή η κλάση εξόδου, τα λεγόμενα δεδομένα εκπαίδευσης, μέσω των οποίων γίνεται η εκπαίδευσή του, και, στη συνέχεια, του παρέχονται άγνωστα δεδομένα, τα λεγόμενα δεδομένα αξιολόγησης ή ελέγχου, για να γίνει η αξιολόγησή του. Ο στόχος του cross-validation είναι να ορίσει ένα σύνολο δεδομένων στο οποίο θα ελεγχθεί το μοντέλο κατά τη φάση της εκπαίδευσης, έτσι ώστε να αποφευχθούν προβλήματα όπως αυτό της υπερεκπαίδευσης. Επίσης, μέσω του cross-validation γίνεται η επιλογή των βέλτιστων παραμέτρων του εκάστοτε αλγορίθμου.

Κάθε γύρος του cross-validation περιλαμβάνει το διαχωρισμό του συνόλου δεδομένων σε συμπληρωματικά υποσύνολα, όπου η ανάλυση διεξάγεται στο ένα υποσύνολο (σύνολο εκπαίδευσης) και η αξιολόγηση γίνεται στο άλλο υποσύνολο (σύνολο αξιολόγησης ή ελέγχου). Για να μειωθεί η μεταβλητότητα, εκτελούνται πολλαπλοί κύκλοι cross-validation, όπου κάθε φορά τα δεδομένα διαχωρίζονται με διαφορετικό τρόπο, οπότε, τελικά, λαμβάνεται ο μέσος όρος των αποτελεσμάτων αξιολόγησης που έχουν ληφθεί από όλους τους γύρους εκτέλεσης του cross-validation.

Το cross-validation είναι αρκετές φορές προτιμότερο να χρησιμοποιείται έναντι της συμβατικής διαδικασίας ελέγχου (πχ όταν το σύνολο δεδομένων διασπάται σε δύο σύνολα, 70% για την εκπαίδευση και 30% για τον έλεγχο του μοντέλου), καθώς συχνά ο αριθμός των δεδομένων δεν είναι αρκετά μεγάλος, ώστε σε περίπτωση διαχωρισμού του αρχικού συνόλου να μη χαθεί πολύτιμη πληροφορία. Σε γενικές γραμμές, η μέθοδος cross-validation συνδυάζει το σφάλμα πρόβλεψης, ώστε να εξάγει μία πιο ακριβή εκτίμηση της προβλεπτικής απόδοσης του μοντέλου.

Μία πολύ γνωστή κατηγορία του cross-validation είναι το k-fold cross-validation. Στην περίπτωση του k-fold cross-validation το αρχικό δείγμα χωρίζεται τυχαία σε k ισάριθμα υποσύνολα. Από αυτά τα υποσύνολα, ένα χρησιμοποιείται σαν το σύνολο αξιολόγησης για τον έλεγχο του μοντέλου και τα υπόλοιπα k-1 αποτελούν τα δεδομένα εκπαίδευσης. Η διαδικασία επαναλαμβάνεται k φορές, όπου καθένα από τα k υποσύνολα επιτελεί μόνο μία φορά το ρόλο του συνόλου αξιολόγησης. Από τα k αποτελέσματα που έχουν προκύψει από τους διαφορετικούς γύρους, λαμβάνεται ο μέσος όρος ώστε να προκύψει η τελική αξιολόγηση. Το πλεονέκτημα αυτής της μεθόδου είναι ότι όλες οι παρατηρήσεις χρησιμοποιούνται τόσο για την εκπαίδευση όσο και για την αξιολόγηση και, παράλληλα, κάθε παρατήρηση χρησιμοποιείται μόνο μία φορά για την αξιολόγηση. Πολύ συχνά, σε προβλήματα ταξινόμησης χρησιμοποιείται η μέθοδος 10-fold cross-validation.



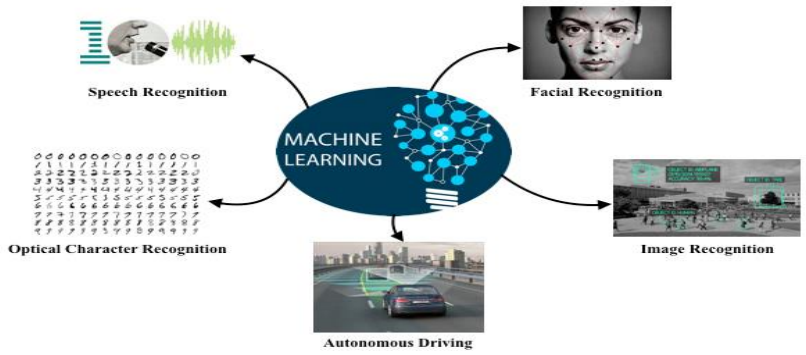
Εικόνα 6 Παράδειγμα k-fold cross-validation με k=4
Πηγή: https://en.wikipedia.org/wiki/Cross-validation_(statistics)

Β΄.6 Το φαινόμενο της Υπερεκπαίδευσης

Ένα από τα σημαντικότερα προβλήματα που καλούνται να ξεπεράσουν οι τεχνικές μηχανικής μάθησης είναι αυτό της υπερεκπαίδευσης. Η υπερεκπαίδευση συμβαίνει όταν ένα μοντέλο είναι υπερβολικά σύνθετο, όπως, για παράδειγμα, όταν έχει πάρα πολλές παραμέτρους σε σχέση με τον αριθμό των παρατηρήσεων. Στην περίπτωση αυτή το μοντέλο που παράγεται έχει πολύ μικρή προβλεπτική ικανότητα, καθώς αντιδρά με υπερβολικό τρόπο σε μικρές διακυμάνσεις των δεδομένων εκπαίδευσης. Η υπερεκπαίδευση γίνεται αντιληπτή μέσω της παρατήρησης των σφαλμάτων, όπου το σφάλμα κατά την εκπαίδευση του μοντέλου είναι πολύ μικρό, αλλά όταν το μοντέλο εφαρμόζεται σε νέα δεδομένα κατά τη διαδικασία ελέγχου, η τιμή του σφάλματος λαμβάνει υπερβολικά μεγάλη τιμή.

Β΄.7 Εφαρμογές μηχανικής μάθησης

Η ραγδαία εξέλιξη που χαρακτηρίζει τη μηχανική μάθηση τα τελευταία χρόνια, δικαιολογεί και την εφαρμογή της σε μία πληθώρα προβλημάτων που συναντώνται σε διαφορετικούς επιστημονικούς κλάδους. Ενδεικτικά, αναφέρονται παρακάτω ορισμένες από τις πιο σημαντικές εφαρμογές της μηχανικής μάθησης [2].



Εικόνα 7 Εφαρμογές της μηχανικής μάθησης

Πηγή: https://www1.in.tum.de/lehrstuhl_1/people/people-archive/74-research/research-topics/896-machine-learning-applications

- *Αναγνώριση φωνής*. Τα σύγχρονα συστήματα αναγνώρισης φωνής έχουν μεταβεί από το χειροκίνητο προγραμματισμό τους στην αυτοματοποιημένη διαχείρισή τους, καθώς η ακρίβεια που επιτυγχάνεται στη δεύτερη περίπτωση είναι πολύ υψηλότερη συγκριτικά με τις παραδοσιακές μεθόδους.
- *Υπολογιστική όραση*. Πολλά σύγχρονα συστήματα υπολογιστικής όρασης, από την αναγνώριση προσώπου μέχρι την ταξινόμηση κυτταρικών εικόνων από μικροσκόπια, βασίζονται στη μηχανική μάθηση, γιατί, όπως και πριν, τα συστήματα αυτά είναι πιο ακριβή από τα παραδοσιακά συστήματα, τα οποία προγραμματίζονται χειροκίνητα. Σαν παράδειγμα τέτοιας εφαρμογής αξίζει να αναφερθεί η ταξινόμηση των γραμμάτων στα ταχυδρομεία των ΗΠΑ, όπου η ταξινόμηση γίνεται με βάση τις γραμμένες στο χέρι διευθύνσεις.
- *Βιολογία*. Η μηχανική μάθηση πλέον χρησιμοποιείται για την ανίχνευση πιθανών εξάρσεων ασθενειών. Αυτό επιτυγχάνεται μέσω της εκμάθησης ιδιοτήτων τυπικών δειγμάτων, ώστε να μπορούν να ανιχνευθούν πιθανές ανωμαλίες, οι οποίες αντιστοιχούν σε κρούσματα.
- *Έλεγχος ρομπότ*. Μέσω των τεχνικών μηχανικής μάθησης τα ρομπότ αποκτούν καινούριες δεξιότητες ή προσαρμόζονται στο εκάστοτε περιβάλλον. Λόγω των πραγματικών συνθηκών στις οποίες εκτίθενται τα ρομπότ, ανακύπτουν διάφορες δυσκολίες, όπως οι πολλές διαστάσεις ή εμπόδια πραγματικού χρόνου για τη συλλογή δεδομένων και τη μάθηση. Η μηχανική μάθηση λύνει τα προβλήματα της χαρτογράφησης του χώρου, της τοποθέτησης του ρομπότ στο χώρο καθώς και της αποφυγής εμποδίων.
- *Εφαρμογή σε εμπειρικές επιστήμες*. Πολλές επιστήμες που βασίζονται σε αναλύσεις τεράστιου όγκου δεδομένων καταφεύγουν στις τεχνικές της μηχανικής μάθησης. Κάποια παραδείγματα είναι η εκμάθηση μοντέλων για την έκφραση γονιδίων στα κύτταρα, ο εντοπισμός ασυνήθιστων αστρονομικών αντικειμένων και ο χαρακτηρισμός των σύνθετων προτύπων της εγκεφαλικής δραστηριότητας μέσα από τη χρήση απεικονίσεων από μαγνητικούς τομογράφους.

Όλες οι παραπάνω εφαρμογές καθώς και άλλες πολλές που δεν αναφέρθηκαν, αναδεικνύουν τη σημασία της μηχανικής μάθησης και τη συμβολή της στην εξέλιξη πολλών επιστημονικών κλάδων και στην επίλυση προβλημάτων, τα οποία ήταν αδύνατο να επιλυθούν με αυστηρές αλγοριθμικές και στατιστικές μεθόδους. Η αξιοποίηση και ανάλυση του μεγάλου όγκου των διαθέσιμων δεδομένων οδηγεί στη συνεχή εξέλιξη της μηχανικής μάθησης και στην ανάπτυξη ολοένα και πιο προηγμένων τεχνικών.

Γ΄ Αυτόματη Αναγνώριση Ορθογραφικών Λαθών

Γ΄.1 Βασικό πρόβλημα αναγνώρισης ορθογραφικών λαθών

Τα ορθογραφικά λάθη τα οποία εντοπίζονται σε κείμενα της Ελληνικής γλώσσας μπορούν να είναι ομόηχα, όπου κατά αυτή την εκδοχή «ακούγονται» το ίδιο αλλά διαφέρουν τελείως στον γραπτό λόγο. Για παράδειγμα οι λέξεις ‘θεωρείτε’ και ‘θεωρείται’ οι οποίες έχουν διαφορετική κατάληξη και εν τέλει διαφορετική ερμηνεία. Η πρώτη απευθύνεται σε λέξεις οι οποίες βρίσκονται στο δεύτερο πρόσωπο πληθυντικό αριθμό ενεργητική φωνή και η δεύτερη λέξη απευθύνεται σε λέξεις οι οποίες βρίσκονται σε τρίτο πρόσωπο ενικό αριθμό παθητική φωνή. Από την αναγνώριση των συμφραζομένων και μόνο μπορεί να επιτευχθεί η παραγωγή της σωστής λέξης. Ωστόσο μπορεί να υπάρξει και η περίπτωση ορθογραφικών λαθών κατά την οποία μια λέξη είναι λανθασμένα γραμμένη σε συντακτικό επίπεδο. Ένα τέτοιο παράδειγμα αποτελεί η λέξη ‘ειρωνεία’ όπου κατά αυτή πολλές φορές την συναντάμε λανθασμένα γραμμένη ως ‘ειρωνία’. Επιπλέον μπορεί να υπάρξουν λάθη σε τονικό επίπεδο, σε μορφολογικό και σε στίξης. Στην παρούσα πτυχιακή εργασία θα υπάρξει ενασχόληση μόνο με τα ομόηχα ορθογραφικά λάθη και συγκεκριμένα με λέξεις που έχουν κατάληξη σε –τε και –ται.

Γ΄2 Εισαγωγικά στοιχεία

Η ύπαρξη ορθογραφικών λαθών αποτελεί ένα φαινόμενο άρρηκτα συνδεδεμένο με το γραπτό λόγο και, πλέον, αντιμετωπίζεται με προηγμένες τεχνικές που παρέχει ο κλάδος της πληροφορικής.

Τα ορθογραφικά λάθη εντοπίζονται σε κάθε είδους κείμενο, συμπεριλαμβανομένων και των μαθητικών και φοιτητικών δοκιμίων. Σε μία έρευνα του 2008, βρέθηκε ότι τα ορθογραφικά λάθη αποτελούσαν το 6.5% όλων των τύπων των λαθών που εντοπίστηκαν σε κείμενα των φοιτητών σε κολλέγια των ΗΠΑ, παρά το γεγονός ότι οι συγγραφείς είχαν πρόσβαση σε εργαλεία ορθογραφικού ελέγχου [12]. Τα αποτελέσματα μίας παρόμοιας έρευνας κατέταξαν τα ορθογραφικά λάθη μεταξύ των πέντε πιο συχνών λαθών σε κείμενα πρωτοετών φοιτητών των ΗΠΑ [13].

Όπως είναι αναμενόμενο, ορθογραφικά λάθη εντοπίζονται και σε κείμενα που συντάσσονται από άτομα που η γλώσσα του κειμένου δεν αποτελεί τη μητρική τους γλώσσα και τα λάθη αυτά είναι διαφορετικά από εκείνα που συναντώνται σε κείμενα που γράφονται σε γλώσσα όμοια με τη μητρική [14].

Στον κλάδο της αυτόματης αξιολόγησης του γραπτού λόγου, η ανίχνευση των ορθογραφικών λαθών χρησιμοποιείται σε εφαρμογές εκμάθησης γλωσσών καθώς και σε συστήματα αυτοματοποιημένης βαθμολόγησης, ιδίως όταν υπάρχει και ανατροφοδότηση από το χρήστη [15]. Ωστόσο, τα ορθογραφικά λάθη ασκούν μία βαθύτερη επιρροή στην αυτόματη αξιολόγηση κειμένου. Η μη βέλτιστη αυτόματη ανίχνευση γραμματικών και ορθογραφικών λαθών μπορεί να οφείλεται σε κακή απόδοση των NLP (Natural Language Processing) εργαλείων, όταν αυτά τροφοδοτούνται με θορυβώδη δεδομένα [16].

Στην περίπτωση της αυτόματης διόρθωσης λαθών σε προθέσεις και προσδιορισμούς στην αγγλική γλώσσα, παρατηρήθηκε ότι η διαδικασία διαταράσσεται συχνά από την ύπαρξη ορθογραφικών λαθών [17]. Μάλιστα, η αυτόματη διόρθωση των ορθογραφικών λαθών βελτιώνει την αυτόματη διόρθωση των λαθών προθέσεων και προσδιορισμών που γίνονται από άτομα που γράφουν σε γλώσσα διαφορετική από τη μητρική τους [18].

Τα ορθογραφικά λάθη δημιουργούν, επίσης, προβλήματα στα συστήματα αυτόματης βαθμολόγησης, όπου η προσοχή επικεντρώνεται στην αξιολόγηση της ορθότητας των απαντήσεων και όχι στην ποιότητα της γλώσσας [19].

Μία πτυχή του προβλήματος των ορθογραφικών λαθών είναι αυτή που αφορά τη διόρθωση ορθογραφικών λαθών μη-λέξεων (non-words. Οι μη-λέξεις αποτελούν αλληλουχία χαρακτήρων που δεν περιέχονται στο λεξικό. Οι δύο κλασικές μέθοδοι για τη διόρθωση λαθών μη-λέξεων είναι η ενσωμάτωση μονάδων για τον υπολογισμό της μετρικής 'edit distance', μέσω της οποίας ποσοτικοποιείται η ανομοιότητα μεταξύ δύο αλληλουχιών χαρακτήρων, καθώς και ο υπολογισμός της φωνητικής ομοιότητας. Μία από τις πρώτες προσεγγίσεις για τη διόρθωση λαθών μη-λέξεων περιελάμβανε τη χρήση συμφραζόμενων, ενώ μία πιο πρόσφατη προσέγγιση εκμεταλλεύεται τη συχνότητα των λέξεων σε ένα κείμενο σαν ένα επιπλέον χαρακτηριστικό για τη βαθμολόγηση κάθε υποψηφίου [15]. Η χρήση του συνόλου κειμένων Google Web1T n-gram για τη διόρθωση ορθογραφικών λαθών βασισμένη στα συμφραζόμενα περιγράφεται στο [20].

Το n-gram είναι μία έννοια που χρησιμοποιείται στην υπολογιστική γλωσσολογία και στις πιθανότητες και αποτελεί μία αλληλουχία από n τμήματα χαρακτήρων που βρίσκονται σε ένα δείγμα από κείμενο ή ομιλία. Τα τμήματα χαρακτήρων μπορεί να είναι λέξεις, γράμματα, συλλαβές ή φωνήματα, ανάλογα με τις ανάγκες της εκάστοτε εφαρμογής.

Γ΄.3 Ορθογραφικά λάθη στην ελληνική γλώσσα

Ένα συχνό φαινόμενο που συναντάται κατά τον ορθογραφικό έλεγχο των λέξεων είναι η ανίχνευση ορθογραφικών λαθών σε λέξεις που είναι σωστά γραμμένες βάσει του λεξικού, ωστόσο δεν είναι οι κατάλληλες λέξεις για να χρησιμοποιηθούν μέσα στην πρόταση.

Αντιθέτως, πρόκειται για λέξεις που φαίνονται ή ακούγονται όμοιες με τη σωστή λέξη. Για παράδειγμα, οι λέξεις ‘peace’ και ‘piece’ της αγγλικής γλώσσας προφέρονται με τον ίδιο τρόπο και είναι και οι δύο έγκυρες λέξεις, καθώς περιέχονται σε οποιοδήποτε λεξικό. Παρόλα αυτά, αν χρησιμοποιηθεί λανθασμένα η μία λέξη αντί της άλλης, το σφάλμα που θα υπάρχει μέσα στην πρόταση, λόγω του διαφορετικού νοήματος των δύο λέξεων, δε θα είναι δυνατό να ανιχνευθεί από ένα κοινό εργαλείο ορθογραφικού ελέγχου. Τέτοιου είδους σφάλματα αποτελούν το 25% με 50% των ορθογραφικών λαθών που παρατηρούνται σε κείμενα της αγγλικής γλώσσας [21].

Στην περίπτωση της ελληνικής γλώσσας, λόγω του λεξιλογικού της πλούτου, υπάρχουν περισσότεροι από ένας τρόποι για την προφορά ενός φωνήεντος. Για παράδειγμα, τα φωνήεντα ‘ο’ και ‘ω’ ακούγονται ακριβώς το ίδιο. Επίσης, τα φωνήεντα ‘ι’, ‘η’, ‘υ’ και οι δίφθογγοι ‘ει’, ‘οι’, ‘υι’ προφέρονται με όμοιο τρόπο, όπως αντίστοιχα και το φωνήεν ‘ε’ και ο δίφθογγος ‘αι’.

Όλοι οι προαναφερθέντες λόγοι δημιουργούν ασάφειες, οι οποίες, αν δεν έχουμε επαρκείς πληροφορίες για τα συμφραζόμενα της πρότασης ή ολόκληρου του κειμένου, είναι δύσκολο να ξεπεραστούν. Στον παρακάτω πίνακα παρουσιάζονται κάποια παραδείγματα τέτοιων περιπτώσεων, που στην ελληνική γραμματική χαρακτηρίζονται ως ομόφωνα. Τα ομόφωνα, δηλαδή, είναι λέξεις που διαφοροποιούνται ως προς τη γραφή και το νόημα, ωστόσο προφέρονται ακριβώς με τον ίδιο τρόπο. Στην παρούσα διπλωματική εργασία, θα γίνει ενασχόληση με την περίπτωση αναγνώρισης ορθογραφικών λαθών σε ομόφωνα, και πιο συγκεκριμένα θα μελετηθεί η περίπτωση των λέξεων με κατάληξη –τε και –ται.

Λέξη 1	*Νόημα*	*Λέξη 2*	*Νόημα*
λύπη	συναίσθημα (ουσιαστικό)	**λείπει**	απουσιάζει (ρήμα)
καλή	καλή (ουσιαστικό, ενικός αριθμός, θηλυκού γένους)	**καλοί**	καλοί (ουσιαστικό, πληθυντικός αριθμός, αρσενικού γένους)
πίνετε	πίνετε (ρήμα, ενεργητική φωνή, πληθυντικός αριθμός, 2ο πρόσωπο)	**πίνεται**	πίνεται (ρήμα, παθητική φωνή, ενικός αριθμός, 3ο πρόσωπο)

Πίνακας 2 Ομόφωνα στην ελληνική γλώσσα

Όπως είναι φυσικό, τα ορθογραφικά λάθη στην περίπτωση των ομοφώνων είναι ένα συχνό φαινόμενο κατά τη συγγραφή κειμένων στην ελληνική γλώσσα. Ένα μεγάλο ποσοστό των επιθέτων, αριθμητικών και αντωνυμιών έχουν κατάληξη ‘η’ για το θηλυκό γένος στον ενικό αριθμό και ‘οι’ για το αρσενικό γένος στον πληθυντικό αριθμό. Επίσης, ένα μεγάλο μέρος των ελληνικών ρημάτων καταλήγουν σε ‘αι’ στο 3ο πρόσωπο ενικού αριθμού παθητικής φωνής και σε ‘ε’ στο 2ο πρόσωπο πληθυντικού αριθμού ενεργητικής φωνής. Πέρα από την κατάληξη, το υπόλοιπο μέρος της λέξης είναι το ίδιο και στις δύο περιπτώσεις, δημιουργώντας με αυτόν τον τρόπο ένα σύνολο ομοφώνων που συναντώνται μέσα στα κείμενα με λανθασμένη ορθογραφία.

Καθώς το πρόβλημα που αναφέρθηκε επικεντρώνεται σε έγκυρες λέξεις, η μέθοδος για τη διόρθωση των ορθογραφικών λαθών θα πρέπει να επικεντρώνεται σε πληροφορίες που θα συλλέγονται από τις υπόλοιπες λέξεις μέσα στην πρόταση και το νόημα αυτών. Αυτό μας οδηγεί στο συμπέρασμα ότι οποιαδήποτε μεθοδολογία για τον ορθογραφικό έλεγχο σε ομόφωνα θα πρέπει να υιοθετήσει μία προσέγγιση βασισμένη στα συμφραζόμενα. Στην επόμενη ενότητα θα γίνει εκτεταμένη αναφορά στις τεχνικές αναγνώρισης ορθογραφικών λαθών που βασίζονται στα συμφραζόμενα.

Γ΄.4 Τεχνικές αυτόματης αναγνώρισης ορθογραφικών λαθών

Στην ενότητα αυτή, θα γίνει μία βιβλιογραφική αναφορά των τεχνικών που έχουν προταθεί για την αναγνώριση των ορθογραφικών λαθών, εστιάζοντας την προσοχή σε αυτές που βασίζονται στα συμφραζόμενα. Το πρόβλημα της αυτόματης αναγνώρισης ορθογραφικών λαθών τοποθετείται στο ευρύτερο πλαίσιο του αντικειμένου της Επεξεργασίας Φυσικής Γλώσσας (Natural Language Processing), το οποίο θα αναλυθεί στο επόμενο κεφάλαιο. Επομένως, ορισμένες έννοιες που μπορεί να χρησιμοποιηθούν σε αυτήν την ενότητα θα γίνουν περισσότερο κατανοητές στη συνέχεια.

Τα παραδοσιακά εργαλεία διόρθωσης ορθογραφικών λαθών βασίζονται σε λεξικά προκειμένου να εντοπίσουν αστοχίες στην ορθογραφία των λέξεων. Ωστόσο, όπως αναφέρθηκε και στην προηγούμενη ενότητα, στην περίπτωση των ομόφωνων λέξεων, δηλαδή των λέξεων που προφέρονται με τον ίδιο ακριβώς τρόπο, αλλά διαφοροποιούνται ως προς την ορθογραφία, ένα λεξικό δε θα μπορούσε να εντοπίσει τα ορθογραφικά λάθη.

Αυτό οφείλεται στο γεγονός ότι πρόκειται για έγκυρες λέξεις οι οποίες, ωστόσο, χρησιμοποιούνται με λάθος τρόπο μέσα στην πρόταση σύμφωνα με τους κανόνες της γραμματικής. Τη λύση σε αυτή την κατηγορία προβλημάτων έρχονται να δώσουν διάφορες τεχνικές που βασίζονται στα συμφραζόμενα και έχουν προταθεί στη βιβλιογραφία, οι οποίες χρησιμοποιούν τις αρχές της στατιστικής ή της μηχανικής μάθησης, αξιοποιώντας διάφορες ιδιότητες των λέξεων που προηγούνται ή έπονται της λέξης, η ορθογραφία της οποίας πρέπει να ελεγχθεί.

Οι τεχνικές που βασίζονται στη στατιστική αποτελούν πιο παραδοσιακές μεθόδους, καθώς τα τελευταία χρόνια οι επιστήμονες έχουν εστιάσει την προσοχή τους σε προσεγγίσεις που βασίζονται στη μηχανική μάθηση. Ένα χαρακτηριστικό παράδειγμα μεθόδου αναγνώρισης ορθογραφικών λαθών περιγράφεται στο [22], όπου ακολουθείται μία διαδικασία αναζήτησης του κοντινότερου γείτονα για την αυτόματη αναγνώριση ορθογραφικών λαθών.

Πιο συγκεκριμένα, η διαδικασία περιλαμβάνει την αντικατάσταση της λέξης με τη λανθασμένη ορθογραφία με τη σωστή λέξη από ένα λεξικό, βάσει μίας μετρικής ομοιότητας. Η μετρική αυτή υπολογίζεται με βάση τον αριθμό των 'trigrams' που είναι όμοια και στις δύο λέξεις. Το trigram είναι μία ειδική περίπτωση του n-gram, όπου το n=3. Δηλαδή, σε αυτήν την περίπτωση, λαμβάνονται υπόψη οι τρεις λέξεις πριν και μετά της αμφίσημης λέξης αντίστοιχα, έτσι ώστε να υπολογιστεί η μετρική ομοιότητας.

Ακόμη μία τεχνική που βασίζεται στις αρχές της στατιστικής περιγράφηκε στο [23], όπου προτάθηκε ένα πιθανοτικό μοντέλο για την αυτόματη αναγνώριση ορθογραφικών λαθών σε κείμενα της αγγλικής γλώσσας. Η βασική ιδέα της συγκεκριμένης τεχνικής είναι η χρήση ενός πιθανοτικού μοντέλου για την παραγωγή της λανθασμένης λέξης από την αντίστοιχη σωστή.

Για αυτό το λόγο, χρησιμοποιείται η δεσμευμένη πιθανότητα $P(Y|X)$, δηλαδή η πιθανότητα να παραχθεί μία λανθασμένη ορθογραφικά λέξη Y από μία ορθή λέξη X. Με άλλα λόγια, εξάγεται μία προσέγγιση της σωστής λέξης από τη λανθασμένη, με τέτοιο τρόπο ώστε να μειωθεί η μέση πιθανότητα σφάλματος κατά τη λήψη της απόφασης.

Οι συγγραφείς στο [24], μελετούν την αποτελεσματικότητα της εφαρμογής ενός στατιστικού μοντέλου στα πλαίσια της αναγνώρισης φωνής. Σε αυτό το μοντέλο, η συντακτική, σημασιολογική και πραγματική γνώση ενσωματώνεται σε δεσμευμένες πιθανότητες που υπολογίζονται με βάση trigram λέξεων.

Όπως προαναφέρθηκε, το πρόβλημα της αυτόματης αναγνώρισης ορθογραφικών λαθών προσεγγίζεται πλέον με μεθόδους μηχανικής μάθησης, καθώς επιτυγχάνεται αρκετά υψηλή ακρίβεια στην πρόβλεψη της σωστής λέξης. Προκειμένου να εκπαιδευτούν τα μοντέλα μηχανικής μάθησης, χρησιμοποιούνται σύνολα αμφίσημων λέξεων, τα οποία αποτελούν το σύνολο δεδομένων, και χαρακτηριστικά που βασίζονται στην ύπαρξη συγκεκριμένων λέξεων και part-of-speech (PoS) tags που προκύπτουν από τα συμφραζόμενα της λέξης [21].

Τα PoS tags αποτελούν ετικέτες που αποδίδονται στις λέξεις και φανερώνουν τι μέρος του λόγου είναι κάθε λέξη, δηλαδή αν είναι ρήμα, ουσιαστικό, επίρρημα, κτλ.

Μία μέθοδος που βασίζεται στα συμφραζόμενα προτείνεται στο [21], όπου γίνεται προσπάθεια αναγνώρισης των αμφίσημων λέξεων χρησιμοποιώντας τις καταλήξεις των λέξεων που περιβάλλουν τις λέξεις-στόχους. Αυτό σημαίνει ότι δεν απαιτούνται υψηλού επιπέδου γλωσσολογικές πληροφορίες, όπως PoS tags, συντακτική ανάλυση και μορφολογική επεξεργασία. Επειδή στην ελληνική γλώσσα υπάρχουν διάφοροι τύποι αμφίσημων λέξεων, το πρώτο σετ λέξεων που μελετούν οι συγγραφείς στο [21] είναι τα ρήματα ενεργητικής και παθητικής φωνής, τα οποία προφέρονται ακριβώς με τον ίδιο τρόπο, αλλά διαφοροποιούνται ως προς την κατάληξη, πχ:

- Εσείς **πίνετε** κρασί.
- Αυτό το νερό δεν **πίνεται**.

Παρατηρούμε, δηλαδή, ότι ο μοναδικός τρόπος για να διαπιστωθεί ποια λέξη πρέπει να χρησιμοποιηθεί είναι η κατανόηση του νοήματος ολόκληρης της πρότασης. Το επόμενο σετ αμφίσημων λέξεων που χρησιμοποιείται αφορά λέξεις θηλυκού και αρσενικού γένους που βρίσκονται στον ενικό και πληθυντικό αριθμό αντίστοιχα, όπως για παράδειγμα:

- Η δασκάλα μας είναι πολύ **καλή**.
- Οι μαθητές είναι **καλοί**.

Στο [25], οι συγγραφείς παρουσιάζουν μία μέθοδο που διορθώνει ταυτόχρονα λάθη μη-λέξεων και πραγματικών λέξεων.

Για να το επιτύχουν αυτό, ενσωματώνουν στη μέθοδό τους χαρακτηριστικά από το επίπεδο των χαρακτήρων καθώς επίσης και από το φωνητικό, λεξικολογικό, συντακτικό και σημασιολογικό επίπεδο.

Για το συντακτικό επίπεδο, τα χαρακτηριστικά αποτελούνται από τρία PoS tags, τα οποία τροφοδοτούνται στο μοντέλο για να ξεκινήσει η εκπαίδευσή του στα 14000 πιο συχνά λήμματα του λεξικού τους.

Οι συγγραφείς στο [26] χρησιμοποιούν τα σύνολα εκπαίδευσης μαζί με μία βάση δεδομένων που αποτελείται από n-grams για μεγέθη από 1 μέχρι m, όπου χρησιμοποιούνται το πολύ m λέξεις αριστερά και δεξιά από τη λέξη-στόχο αντίστοιχα. Ο αλγόριθμος επιλέγει τη λέξη με τη μεγαλύτερη πιθανότητα, ώστε να αντικαταστήσει τη λανθασμένη. Στην περίπτωση που δε βρεθεί κάποια υποψήφια λέξη, το m μειώνεται μέχρις ότου λάβει την τιμή m = 1, όπου το αποτέλεσμα αντιστοιχεί στην πιο συχνή λέξη.

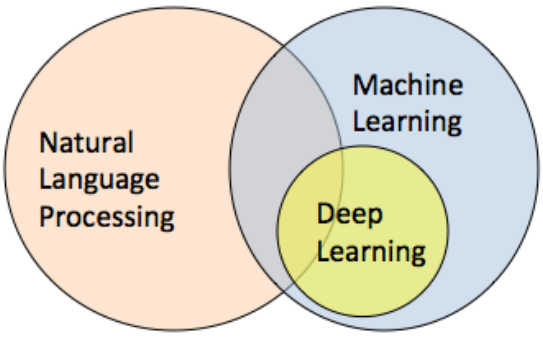
Στο [27] παρουσιάζεται ένας Window-based αλγόριθμος, όπου χρησιμοποιούνται δύο τύποι χαρακτηριστικών: τα συμφραζόμενα και οι συμφράσεις. Τα χαρακτηριστικά που προέρχονται από τα συμφραζόμενα ελέγχουν για την παρουσία μίας συγκεκριμένης λέξης μεταξύ των ±k λέξεων μεταξύ της λέξης-στόχου, ενώ αντίθετα οι συμφράσεις αναζητούν ένα πρότυπο που αποτελείται από m το πολύ γειτονικές λέξεις και/ή PoS tags γύρω από τη λέξη-στόχο.

Δ΄ Μηχανική Μάθηση και Αναγνώριση Ορθογραφικών Λαθών Λέξεων

Δ΄.1 Εφαρμογή της μηχανικής μάθησης σε προβλήματα επεξεργασίας φυσικής γλώσσας

Το πρόβλημα της αυτόματης αναγνώρισης ορθογραφικών λαθών εντάσσεται στο ευρύτερο αντικείμενο της Επεξεργασίας Φυσικής Γλώσσας (ΕΦΓ). Γι' αυτό το λόγο, κρίνεται σκόπιμο, σε αυτήν την ενότητα, να γίνει μία σύντομη αναφορά στον ορισμό, τα βασικά χαρακτηριστικά και τις εφαρμογές της ΕΦΓ.

Η Επεξεργασία Φυσικής Γλώσσας (Natural Language Processing - NLP) αναφέρεται σε ένα εύρος θεωρητικά υποκινούμενων υπολογιστικών τεχνικών για την αυτόματη ανάλυση και αναπαράσταση της ανθρώπινης γλώσσας, έτσι ώστε να διευκολύνεται η επικοινωνία μεταξύ ανθρώπων και μηχανών, να βελτιώνεται η επικοινωνία μεταξύ των ανθρώπων ή απλά να εκτελείται η επεξεργασία του προφορικού ή γραπτού λόγου. Η ΕΦΓ σχετίζεται με τη δημιουργία συστημάτων που μπορούν να καταλάβουν τη φυσική γλώσσα, ενώ η μηχανική μάθηση παράγει συστήματα που μπορούν να μάθουν από προηγούμενη εμπειρία. Όταν αυτές οι δύο έννοιες συναντώνται, τότε κατασκευάζονται συστήματα που μπορούν να μάθουν πώς να κατανοούν τη φυσική γλώσσα.



Εικόνα 8 Επεξεργασία Φυσικής Γλώσσας και Μηχανική Μάθηση

Πηγή: http://rutumulkar.com/blog/2016/NLP-ML

Οι ερευνητές της ΕΦΓ αποσκοπούν στην απόκτηση γνώσης σχετικά με το πώς οι άνθρωποι κατανοούν και χρησιμοποιούν τη γλώσσα, πληροφορίες οι οποίες είναι απαραίτητες για την εύρεση τεχνικών και ανάπτυξη εργαλείων που θα παρέχουν στους υπολογιστές τη δυνατότητα να χειρίζονται τις φυσικές (ανθρώπινες) γλώσσες.

Η έρευνα στον κλάδο της ΕΦΓ έχει παρουσιάσει ραγδαία εξέλιξη, καθώς από την επεξεργασία μιας πρότασης, που μπορούσε να πάρει έως και 7 λεπτά, έχουμε μεταβεί στην εποχή της Google, όπου εκατομμύρια ιστοσελίδων επεξεργάζονται σε λιγότερο από ένα δευτερόλεπτο.

Τα θεμέλια της Επεξεργασίας Φυσικής Γλώσσας εντοπίζονται σε διάφορους τομείς, όπως στην επιστήμη των υπολογιστών και πληροφοριών, τη γλωσσολογία, τα μαθηματικά, την τεχνητή νοημοσύνη, τη ρομποτική, την ψυχολογία, κτλ. Οι εφαρμογές της ΕΦΓ είναι ποικίλες και περικλείουν μελέτες διάφορων κλάδων, όπως είναι η μεταγλώττιση μηχανής, η επεξεργασία και περίληψη κειμένου φυσικής γλώσσας, οι διεπαφές χρηστών, η ανάκτηση πληροφοριών, η αναγνώριση φωνής και η τεχνητή νοημοσύνη [28].

Προς το παρόν, η ανάκτηση, ενσωμάτωση και επεξεργασία της διαδικτυακής πληροφορίας βασίζεται, κυρίως, σε αλγορίθμους που αφορούν την κειμενική αναπαράσταση των ιστοσελίδων. Οι συγκεκριμένοι αλγόριθμοι είναι αποτελεσματικοί στην ανάκτηση κειμένων, το διαχωρισμό τους σε επιμέρους μέρη, τον έλεγχο της ορθογραφίας και τον υπολογισμό του αριθμού των λέξεων. Ωστόσο, στην περίπτωση της ερμηνείας των προτάσεων και της εξαγωγής χρήσιμης πληροφορίας, οι δυνατότητες των προαναφερθέντων αλγορίθμων κρίνονται περιορισμένες. Η Επεξεργασία Φυσικής Γλώσσας απαιτεί υψηλού επιπέδου συμβολικές δυνατότητες οι οποίες περιλαμβάνουν [29]:

- Τη δημιουργία και διάδοση δυναμικών δεσμών.
- Τη χρήση αναδρομικών, συστατικών δομών.
- Την απόκτηση και πρόσβαση σε λεξιλογικές, σημασιολογικές και σποραδικές πληροφορίες.

- Τον έλεγχο πολλαπλών μονάδων μάθησης και τη δρομολόγηση της πληροφορίας μεταξύ αυτών των μονάδων.
- Την αναπαράσταση αφηρημένων εννοιών.

Όπως είναι γνωστό, η γλωσσολογική πληροφορία χαρακτηρίζεται από αμφισημία, η οποία είναι δυνατόν να αντιμετωπιστεί μέσω των μοντέλων μηχανικής μάθησης. Οι ΕΦΓ αλγόριθμοι που προτείνονται στη βιβλιογραφία βασίζονται σε στατιστική μηχανική μάθηση ή σε μοντέλα επιβλεπόμενης μηχανικής μάθησης. Πριν την εμφάνιση των τεχνικών μηχανικής μάθησης, όλες οι διεργασίες της Επεξεργασίας Φυσικής Γλώσσας διεκπεραιώνονταν με τη χρήση προσεγγίσεων που βασίζονταν στη σύνταξη μεγάλων συνόλων από κανόνες. Ωστόσο, οι τεχνικές αυτές έχουν πλέον ξεπεραστεί, καθώς η μηχανική μάθηση έχει καταστήσει ηχηρή την παρουσία της.

Κάποιοι αλγόριθμοι, μάλιστα, όπως τα Δέντρα Απόφασης (Decision Trees) παράγουν συστήματα από ‘εάν-τότε’ κανόνες, όμοιους με τους χειρόγραφους κανόνες που χρησιμοποιούνταν παλιότερα. Οι αλγόριθμοι μηχανικής μάθησης, αρχικά, εκπαιδεύονται σε ένα μεγάλο σύνολο δεδομένων με χαρακτηρισμένες εξόδους, ώστε να γίνει η γενίκευση του μοντέλου και, στη συνέχεια, το μοντέλο, αφού αξιολογηθεί κατά τη φάση ελέγχου, χρησιμοποιείται για την πρόβλεψη της εξόδου άγνωστων πλέον δεδομένων.

Ωστόσο, σταδιακά, η έρευνα φαίνεται να επικεντρώνεται περισσότερο σε στοχαστικά μοντέλα μηχανικής μάθησης. Σε αυτά τα μοντέλα, σε κάθε διάνυσμα εισόδου ανατίθεται και μία πραγματική τιμή βάρους, έτσι ώστε να γενικεύονται πιθανοτικές αποφάσεις. Το πλεονέκτημα αυτών των μοντέλων είναι ότι έχουν την ικανότητα να εκφράζουν τη σχετική βεβαιότητα πολλών διαφορετικών πιθανών απαντήσεων και όχι μόνο μίας, παράγοντας με αυτόν τον τρόπο πιο αξιόπιστα αποτελέσματα όταν το μοντέλο ενσωματώνεται σε μεγαλύτερα συστήματα.

Τα συστήματα Επεξεργασίας Φυσικής Γλώσσας που βασίζονται σε τεχνικές μηχανικής μάθησης παρουσιάζουν πολλά πλεονεκτήματα έναντι των παραδοσιακών τεχνικών με τους χειρόγραφους κανόνες [30]:

- Οι διαδικασίες μάθησης που χρησιμοποιούνται κατά τη μηχανική μάθηση εστιάζουν αυτόματα στις πιο συνηθισμένες περιπτώσεις, ενώ κατά τη συγγραφή κανόνων με το χέρι δεν είναι πάντα προφανές προς τα πού θα πρέπει να κατευθυνθεί η προσπάθεια.
- Οι αυτόματες διαδικασίες μάθησης μπορούν να κάνουν τη χρήση αλγορίθμων συμπερασματικής στατιστικής (statistical inference) έτσι ώστε να παράγουν μοντέλα που είναι ανθεκτικά σε άγνωστες (για παράδειγμα στην περίπτωση λέξεων ή δομών που δεν έχει ξανασυναντήσει το μοντέλο) ή εσφαλμένες εισόδους (για παράδειγμα στην περίπτωση λέξεων με ορθογραφικά λάθη). Όπως μπορεί να αντιληφθεί κανείς, η διαχείριση τέτοιων εισόδων με τη χρήση χειρόγραφων κανόνων είναι εξαιρετικά δύσκολη, επιρρεπής σε λάθη και χρονοβόρα.
- Τα συστήματα που βασίζονται στην αυτόματη εκμάθηση των κανόνων μπορούν να γίνουν πιο ακριβή μέσω της τροφοδότησής τους με περισσότερα δεδομένα εισόδου. Από την άλλη μεριά, τα παραδοσιακά συστήματα των κανόνων μπορούν να επιτύχουν μεγαλύτερη ακρίβεια μόνο αυξάνοντας την πολυπλοκότητα των κανόνων. Μάλιστα, υπάρχει ένα όριο για το πόσο πολύπλοκο μπορεί να είναι ένα σύστημα με χειρόγραφους κανόνες, πέρα του οποίου το σύστημα γίνεται μη διαχειρίσιμο.

  Ωστόσο, η δημιουργία περισσότερων δεδομένων για την προώθησή τους στα συστήματα μηχανικής μάθησης απαιτεί περισσότερες ανθρωποώρες, χωρίς να απαιτείται αύξηση της πολυπλοκότητας.

Παρακάτω θα δοθούν ορισμένα από τα συχνότερα πεδία έρευνας στην Επεξεργασία Φυσικής Γλώσσας. Ορισμένα από αυτά έχουν εφαρμογές στον πραγματικό κόσμο, ενώ κάποια άλλα συναντώνται σαν επιμέρους πεδία που χρησιμοποιούνται για την επίλυση μεγαλύτερων ζητημάτων.

- **Εύρεση θέματος (Stemming)**

  Η εύρεση θέματος είναι η διαδικασία ελάττωσης μιας λέξης, έτσι ώστε να μείνει το μέρος το οποίο παράγεται από τη ρίζα. Το θέμα μιας λέξης, το οποίο δε μεταβάλλεται κατά την κλίση, προκύπτει αν αφαιρεθεί η κατάληξη από την αρχική λέξη. Όπως είναι αναμενόμενο, συγγενικές λέξεις έχουν το ίδιο θέμα, γι' αυτό το λόγο, αρκετές μηχανές αναζήτησης μεταχειρίζονται τέτοιες λέξεις σαν συνώνυμες, σύμφωνα με μία διαδικασία εξάπλωσης των λέξεων κλειδιών της αναζήτησης.

- **Εύρεση λήμματος (Lemmatization)**

  Στην υπολογιστική γλωσσολογία, η εύρεση λήμματος αποτελεί την αλγοριθμική διαδικασία προσδιορισμού του λήμματος μιας λέξης βάσει του νοήματός της.

  Σε αντίθεση με την εύρεση θέματος (stemming), η εύρεση λήμματος βασίζεται στην ορθή αναγνώριση του μέρους του λόγου και του νοήματος μιας λέξης, όχι μόνο μέσα στην πρόταση, αλλά και στα συμφραζόμενα, τα οποία μπορεί να αφορούν γειτονικές προτάσεις ή ακόμη και ολόκληρο το έγγραφο.

❖ **Αναγνώριση μερών του λόγου (Part-of-speech tagging)**

Δεδομένης μιας πρότασης, η διαδικασία αυτή αφορά την αναγνώριση του μέρους του λόγου κάθε λέξης. Αν και στην αγγλική γλώσσα, υπάρχουν πολλές λέξεις που αντιστοιχούν σε περισσότερα από ένα μέρη του λόγου, πχ book = βιβλίο (ουσιαστικό) ή κάνω κράτηση (ρήμα), στην ελληνική γλώσσα υπάρχουν ομόηχες λέξεις που διαφοροποιούνται, ωστόσο, ως προς την ορθογραφία τους, πχ αυτί / αυτή / αυτοί.

❖ **Διαχωρισμός λέξεων (Word segmentation)**

Το πεδίο αυτό της Επεξεργασίας Φυσικής Γλώσσας αφορά το διαχωρισμό ενός συνεχόμενου τμήματος κειμένου σε ξεχωριστές λέξεις. Για ορισμένες γλώσσες, όπως τα Αγγλικά και τα Ελληνικά, αυτή η διαδικασία είναι πολύ απλή, καθώς οι λέξεις διαχωρίζονται με κενά στις προτάσεις. Ωστόσο, σε ορισμένες γλώσσες, όπως τα Κινέζικα, τα όρια των λέξεων δηλώνονται με διαφορετικό τρόπο, με αποτέλεσμα ο διαχωρισμός των λέξεων να είναι μία σημαντική διαδικασία, όπου απαιτείται γνώση σχετικά με το λεξιλόγιο και τη μορφολογία των λέξεων της εκάστοτε γλώσσας.

❖ **Οπτική αναγνώριση χαρακτήρων (OCR)**

Η διαδικασία αυτή περιλαμβάνει την εύρεση του κειμένου που αναπαριστάται από μία εικόνα.

❖ **Ανάλυση συναισθήματος (Sentiment analysis)**

Η ανάλυση συναισθήματος περιλαμβάνει την εξαγωγή πληροφορίας από ένα σύνολο εγγράφων, έτσι ώστε να προσδιορίσει την ‘πολικότητα’ του κειμένου. Με άλλα λόγια, είναι δυνατόν να αποφανθεί αν το συναίσθημα που περικλείεται στο κείμενο είναι αρνητικό ή θετικό (στην περίπτωση της δυαδικής ταξινόμησης).

Η ανάλυση συναισθήματος είναι ιδιαίτερα χρήσιμη για τις εκστρατείες μάρκετινγκ, καθώς αναγνωρίζονται οι τάσεις των καταναλωτών μέσα από τα κοινωνικά δίκτυα.

- **Αναγνώριση ομιλίας (Speech recognition)**

Η αναγνώριση ομιλίας σχετίζεται με τη γραπτή αναπαράσταση του προφορικού λόγου, μέσα από την επεξεργασία μίας ηχητικής αναπαράστασης αυτού. Πρόκειται για ένα από τα δυσκολότερα προβλήματα του ΕΦΓ. Στο φυσικό λόγο δεν υπάρχουν ευδιάκριτες παύσεις μεταξύ των λέξεων, οπότε ο διαχωρισμός λέξεων (word segmentation) που αναφέρθηκε παραπάνω αποτελεί μία υποδιεργασία κατά την αναγνώριση ομιλίας.

Επίσης, στις περισσότερες γλώσσες, οι ήχοι των λέξεων συγχέονται μεταξύ τους λόγω της συνεχόμενης ροής, γεγονός που καθιστά αρκετά πολύπλοκη διαδικασία τη μετατροπή του αναλογικού σήματος σε διακριτούς χαρακτήρες.

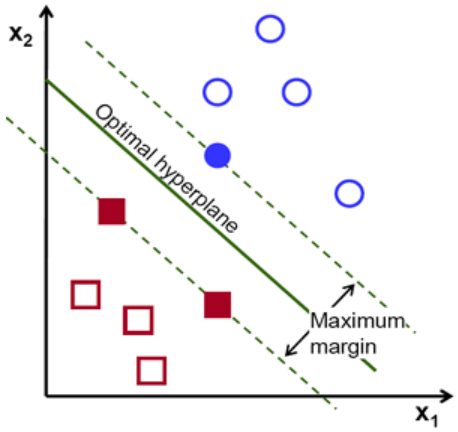
Τα πιο διαδομένα μοντέλα μηχανικής μάθησης που χρησιμοποιούνται στην Επεξεργασία Φυσικής Γλώσσας είναι τα Hidden Markov models (HMM), τα Random Forests (RF), τα Maximum Entropy (MaxEnt), τα Support Vector Machines (SVM), τα Decision Trees (DT), τα μοντέλα Naïve Bayes και τα νευρωνικά δίκτυα [31]. Σε επόμενη ενότητα θα γίνει η ανάλυση των βασικών χαρακτηριστικών κάθε μοντέλου.

Δ΄.2 Τεχνικές μηχανικής μάθησης για την αναγνώριση ορθογραφικών λαθών

Στην ενότητα αυτή θα αναλυθούν οι πιο βασικοί αλγόριθμοι που συναντώνται στη βιβλιογραφία και εφαρμόζονται στην Επεξεργασία Φυσικής Γλώσσας και, ειδικότερα, στην αναγνώριση ορθογραφικών λαθών.

Δ΄.2.1 Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines – SVM)

Τα μοντέλα SVM έχουν σαν βασικό στόχο να βρουν το βέλτιστο υπερεπίπεδο διαχωρισμού, το οποίο ταξινομεί όσο το δυνατόν με μεγαλύτερη ακρίβεια τα σημεία των δεδομένων, διαχωρίζοντάς τα σε δύο κλάσεις [32]. Τα σημεία εκπαίδευσης που είναι πιο κοντά στο βέλτιστο υπερεπίπεδο ονομάζονται διανύσματα υποστήριξης και είναι αυτά που χρησιμοποιούνται για να αποφασιστεί η κλάση της εξόδου. Προκειμένου να διαχειριστούν προβλήματα μη γραμμικότητας, τα μοντέλα SVM προβάλλουν, αρχικά, τα δεδομένα σε ένα χώρο υψηλότερων διαστάσεων, μέσω μιας συνάρτησης kernel, όπου, πλέον, προσπαθούν να βρουν το γραμμικό περιθώριο (margin) που θα διαχωρίζει τα δεδομένα.



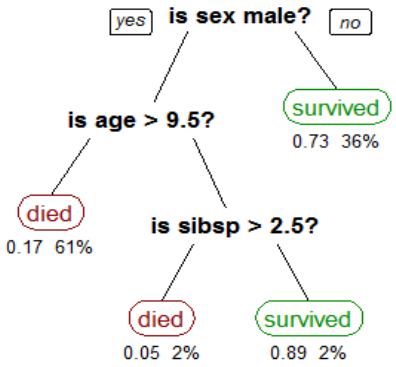
Εικόνα 9 Παράδειγμα μοντέλου SVM

Πηγή: http://opencv-python-tutroals.readthedocs.io/en/latest/

Δ΄.2.2 Δέντρα Απόφασης (Decision Trees - DT)

Τα δέντρα απόφασης βασίζονται στην ιδέα του ‘διαίρει και βασίλευε’ και χρησιμοποιούνται διεξοδικά σε προβλήματα αναγνώρισης ορθογραφικών λαθών. Η μέθοδος ξεκινά να ψάχνει το ‘καλύτερο’ χαρακτηριστικό στον κόμβο κορυφής και διασπά το δέντρο σε υποδέντρα. Αντίστοιχα, μέσω μιας αναδρομικής διαδικασίας, το υποδέντρο διαχωρίζεται με τη σειρά του, ακολουθώντας τον ίδιο κανόνα.

Ο διαχωρισμός σταματά όταν έχουμε φτάσει στα φύλλα που περιέχουν τις κλάσεις εξόδου ή όταν η μετρική που δίνει το καλύτερο μοντέλο είναι ίση με μηδέν. Όταν κατασκευαστεί το δέντρο, μπορούν να ληφθούν οι κανόνες μέσω της διάσχισης κάθε κλαδιού [33].



Εικόνα 10 Παράδειγμα δέντρου απόφασης

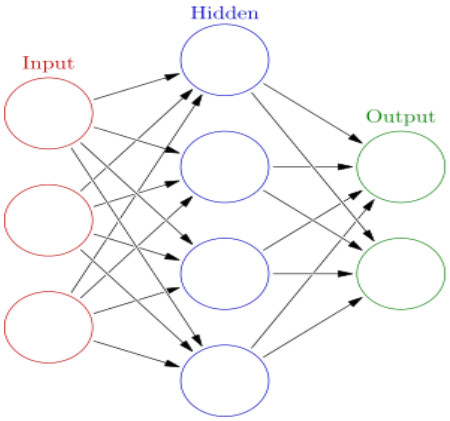
Πηγή: https://en.wikipedia.org/wiki/Decision_tree_learning

Η εύρεση του 'καλύτερου χαρακτηριστικού' γίνεται μέσω της χρήσης μετρικών, οι οποίες εξετάζουν, σε γενικές γραμμές την ομοιογένεια που παρουσιάζει η μεταβλητή εξόδου μέσα στα υποσύνολα. Οι βασικότερες μετρικές είναι δύο, ο δείκτης gini και το κέρδος πληροφορίας. Ο δείκτης gini μετρά πόσες φορές ένα τυχαία επιλεγόμενο στοιχείο από το σύνολο θα ταξινομηθεί λανθασμένα, αν η ταξινόμησή του γίνεται τυχαία, σύμφωνα με την κατανομή των κλάσεων στο υποσύνολο. Το κέρδος πληροφορίας βασίζεται στην ιδέα της εντροπίας, μιας έννοιας που συναντάται στη θεωρία πληροφοριών. Για κάθε κόμβο του δέντρου, το κέρδος πληροφορίας αναπαριστά την αναμενόμενη ποσότητα πληροφορίας που θα χρειαστεί για να αποφασιστεί σε ποια από τις δύο κλάσεις θα ταξινομηθεί μία νέα παρατήρηση, δεδομένου ότι αυτή έχει φτάσει στον αντίστοιχο κόμβο.

Δ΄2.3 Τεχνητά Νευρωνικά Δίκτυα (Artificial Neural Networks – ANNs)

Τα τεχνητά νευρωνικά δίκτυα είναι μία δημοφιλής μέθοδος για την επίλυση σύνθετων προβλημάτων και χρησιμοποιούνται εκτενώς σε προβλήματα Επεξεργασίας Φυσικής Γλώσσας. Το πιο διαδεδομένο μοντέλο επιβλεπόμενης μάθησης είναι ο perceptron πολλών επιπέδων (MLP), ο οποίος είναι ένα μοντέλο πρόσθιας τροφοδότησης (feed-forward) που εκπαιδεύεται με τον αλγόριθμο backpropagation.

Ο MLP αποτελείται από πολλά επίπεδα, όπου κάθε επίπεδο περιέχει κόμβους, οι οποίοι συνθέτουν συνολικά έναν κατευθυνόμενο γράφο. Εκτός από τους κόμβους εισόδου, κάθε κόμβος στα υπόλοιπα επίπεδα είναι ένας τεχνητός νευρώνας, ο οποίος χαρακτηρίζεται από μία μη-γραμμική συνάρτηση ενεργοποίησης. Κατά τη διαδικασία της μάθησης, τα βάρη των συνδέσεων μεταξύ των κόμβων αναπροσαρμόζονται σε κάθε επανάληψη, με βάση το ποσοστό σφάλματος μεταξύ της πραγματικής και της επιθυμητής εξόδου.

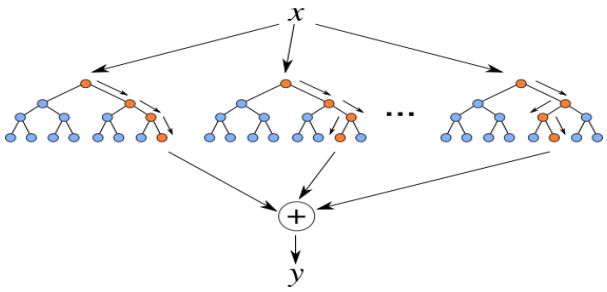


Εικόνα 11 Παράδειγμα τεχνητού νευρωνικού δικτύου
Πηγή: https://en.wikipedia.org/wiki/Artificial_neural_network

Δ΄.2.4 Τυχαία Δάση (Random Forests)

Η μέθοδος Random Forests [32] αποτελεί παράδειγμα συλλογικής μάθησης και χρησιμοποιείται σε προβλήματα ταξινόμησης, όπου κατασκευάζει ένα πλήθος δέντρων απόφασης κατά τη διαδικασία της εκπαίδευσης και εξάγει σαν αποτέλεσμα την κλάση που εμφανίζεται πιο συχνά στο σύνολο των παραγόμενων δέντρων. Η μέθοδος αυτή χρησιμοποιεί τη bootstrap συνάθροιση, έναν αλγόριθμο συλλογικής μάθησης, που χρησιμοποιείται για τη βελτίωση της σταθερότητας και της ακρίβειας των αλγορίθμων μηχανικής μάθησης.

Τα τυχαία δάση έρχονται να συμπληρώσουν τα δέντρα απόφασης, καθώς εξαλείφουν την τάση των δεύτερων να υπερεκπαιδεύονται.



Εικόνα 12 Παράδειγμα της μεθόδου Random Forest

Πηγή: https://kgpdag.wordpress.com/

Δ΄.2.5 Μέθοδος Naïve Bayes

Οι ταξινομητές Naïve Bayes ανήκουν στην οικογένεια των πιθανοτικών μοντέλων, τα οποία βασίζονται στο θεώρημα Μπέυζ, σύμφωνα με το οποίο τα χαρακτηριστικά που περιγράφουν τα δεδομένα πρέπει να είναι ανεξάρτητα μεταξύ τους. Αυτό σημαίνει ότι η παρουσία (ή απουσία) ενός χαρακτηριστικού μίας κλάσης δε σχετίζεται με την παρουσία (ή απουσία) οποιουδήποτε άλλου χαρακτηριστικού.

Από μαθηματική σκοπιά, ο Naïve Bayes ταξινομητής υπολογίζει την πιθανότητα ένα δείγμα εισόδου να ανήκει σε μία συγκεκριμένη κλάση [33]. Δεδομένου ενός δείγματος X, το οποίο αποτελείται από ένα διάνυσμα χαρακτηριστικών $\{x_1, \dots, x_n\}$, η πιθανότητα της κλάσης y_j μπορεί να δοθεί από την εξίσωση:

$$p(y_j|X) = p(X|y_j)p(y_j) = p(x_1, \dots, x_n|y_j)p(y_j)$$

(1)

όπου $p(y_j)$ η προηγούμενη πιθανότητα του y_j. Ωστόσο, η μέθοδος Naïve Bayes υποθέτει, όπως είπαμε, ότι οι δεσμευμένες πιθανότητες των ανεξάρτητων μεταβλητών είναι στατιστικά ανεξάρτητες, οπότε προκύπτει ότι:

$$p(X|y_j) = \prod_{i=1}^{n} p(x_i|y_j)$$

(2)

και η πιθανότητα της κλάσης y_j δίνεται, τελικά, από την εξίσωση:

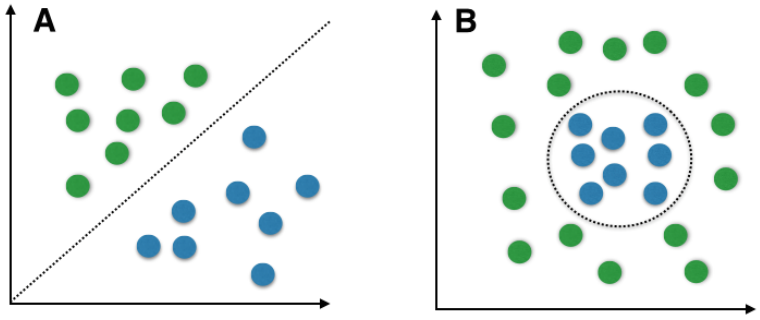
$$p(y_j|X) = p(y_j) \prod_{i=1}^{n} p(x_i|y_j)$$

(3)

Δεδομένου ενός αριθμού κλάσεων $Y = \{y_1, \dots, y_k\}$, ο Naïve Bayes ταξινομεί ένα νέο δείγμα X με βάση τη σχέση:

$$c = \underset{y_j \in Y}{\operatorname{argmax}}\, p(y_j|X)$$

(4)

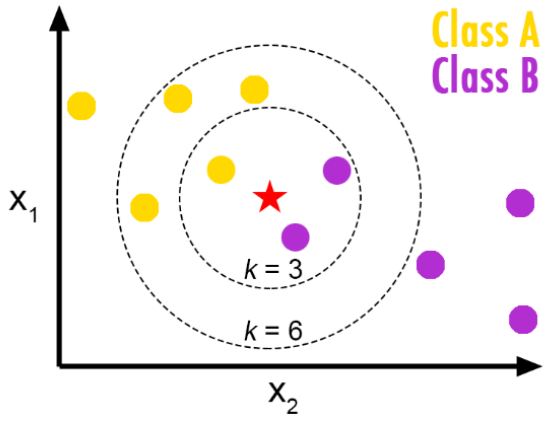


Εικόνα 13 Παράδειγμα αλγορίθμου Naïve Bayes

Πηγή: https://provalisresearch.com/blog/exploring-naive-bayes-classifier-maybe-not-naive/

Δ΄.2.6 Αλγόριθμος κ-κοντινότερων γειτόνων (k-Nearest Neighbors - k-NN)

Ο αλγόριθμος k-NN είναι μία μη-παραμετρική μέθοδος, όπου για κάθε νέα παρατήρηση, η οποία περιγράφεται σαν ένα διάνυσμα N χαρακτηριστικών στο N-διάστατο χώρο, υπολογίζονται τα k κοντινότερα σημεία, με βάση τις ευκλείδειες αποστάσεις. Αυτά τα σημεία συνιστούν τους k κοντινότερους γείτονες. Τελικά, η κλάση εξόδου που αποδίδεται στο αντικείμενο αντιστοιχεί στην κλάση της πλειοψηφίας των γειτόνων. Γι' αυτό το λόγο, το k συνήθως λαμβάνεται ως ένας μικρός περιττός αριθμός, ώστε να αποφευχθεί το ενδεχόμενο της ισοψηφίας. Βέβαια, η τιμή που θα έχει το k εξαρτάται και από τα δεδομένα. Γενικά, μεγαλύτερες τιμές του k μειώνουν την επίδραση θορύβου, ωστόσο κάνουν λιγότερο διακριτά τα σύνορα μεταξύ των κλάσεων. Ένας αποτελεσματικός τρόπος για την εύρεση του k είναι η χρήση ευρετικών αλγορίθμων (heuristics).



Εικόνα 14 Παράδειγμα αλγορίθμου k-NN

Πηγή: http://bdewilde.github.io/blog/blogger/2012/10/26/classification-of-hand-written-digits-3/

Δ΄.2.7 Αλγόριθμος ενίσχυσης (boosting)

Ο βασικός στόχος του αλγορίθμου ενίσχυσης είναι η βελτίωση της απόδοσης της ταξινόμησης, μέσω του συνδυασμού των προβλέψεων πολλών διαφορετικών μοντέλων ταξινόμησης, τα οποία ονομάζονται αδύναμοι ταξινομητές (weak classifiers).

Οι αδύναμοι ταξινομητές χρησιμοποιούνται σαν υπορουτίνες, οι οποίες συνδυάζονται ώστε να συνθέσουν έναν απίστευτα ακριβή ταξινομητή. Για κάθε αδύναμο ταξινομητή, ο αλγόριθμος ενίσχυσης διατηρεί μία κατανομή των βαρών στα πρότυπα του συνόλου εκπαίδευσης, ώστε κάθε πρότυπο να έχει τη δυνατότητα να συνεισφέρει με διαφορετικό τρόπο στο τελικό σφάλμα εκπαίδευσης. Αρχικά, όλα τα βάρη λαμβάνονται ίσα, ωστόσο, σε κάθε επανάληψη τα βάρη των δειγμάτων που ταξινομούνται λανθασμένα αυξάνονται, έτσι ώστε να αναγκάσουν τον αδύναμο ταξινομητή να επικεντρωθεί σε αυτές τις δύσκολες περιπτώσεις. Μόλις ολοκληρωθεί η διαδικασία, οι απλοί ταξινομητές συνδυάζονται για να προκύψει ένα τελικό μοντέλο, το οποίο, συνήθως, επιτυγχάνει μεγάλο ποσοστό ακρίβειας (accuracy) στο σύνολο των δεδομένων ελέγχου.

Δ΄.3 Βιβλιογραφική επισκόπηση προσεγγίσεων στο πρόβλημα της αυτόματης αναγνώρισης ορθογραφικών λαθών

Σε αυτήν την ενότητα θα γίνει μία εκτενής αναφορά στις μεθοδολογίες μηχανικής μάθησης που έχουν προταθεί για το πρόβλημα της αυτόματης αναγνώρισης ορθογραφικών λαθών από διαφορετικούς ερευνητές.

Αφού αναλυθεί κάθε προσέγγιση ξεχωριστά, θα γίνει μία σύνοψη αυτών, ώστε να παρουσιαστούν τα βασικά χαρακτηριστικά του προβλήματος και να σχολιαστεί η αποτελεσματικότητα κάθε προσέγγισης.

Οι συγγραφείς στο [21] περιγράφουν μία μέθοδο η οποία εκτελεί αυτόματη αναγνώριση ορθογραφικών λαθών σε λέξεις της ελληνικής γλώσσας, και πιο συγκεκριμένα σε ομόφωνα, χρησιμοποιώντας μόνο τις καταλήξεις που περιβάλλουν τη λέξη, η ορθογραφία της οποίας πρέπει να ελεγχθεί. Η προσέγγισή τους απαιτεί ελάχιστη πληροφορία, καθώς λαμβάνονται υπόψη μόνο οι καταλήξεις των λέξεων που προηγούνται και έπονται της λέξης-στόχου, γεγονός που την καθιστά κατάλληλη ώστε να ενσωματωθεί σε επεξεργαστές κειμένου για ορθογραφικό έλεγχο πραγματικού χρόνου.

Όπως σε κάθε πρόβλημα μηχανικής μάθησης, το πρώτο βήμα είναι η επιλογή των χαρακτηριστικών που θα τροφοδοτούνται στο εκάστοτε μοντέλο. Οι συγγραφείς στο [21] επέλεξαν να πειραματιστούν με διαφορετικούς αριθμούς χαρακτηριστικών, όπου επιλέγονται οι n λέξεις πριν και μετά της αμφίσημης λέξεις, με το n να είναι μικρότερο ή ίσο του πέντε. Για κάθε τέτοια λέξη λαμβάνονται το τελευταίο γράμμα και τα δύο και τρία τελευταία γράμματα αντίστοιχα. Το διάνυσμα, επομένως, των χαρακτηριστικών έχει μέγεθος ίσο με 2*3*n. Στην περίπτωση που οι λέξεις πριν ή μετά της αμφίσημης λέξης είναι λιγότερες από πέντε, τοποθετείται μία παύλα (‘-’) στη θέση του αντίστοιχου χαρακτηριστικού.

Μετά την επιλογή των χαρακτηριστικών, οι συγγραφείς κατασκεύασαν τα σύνολα δεδομένων τα οποία χρησιμοποιήθηκαν στα πειράματα. Δημιουργήθηκαν δύο σύνολα δεδομένων, ένα για κάθε ζευγάρι αμφίσημων λέξεων.

Πιο συγκεκριμένα, χρησιμοποιήθηκαν τα ομόφωνα με καταλήξεις ‘-αι’ και ‘-ε’ και ‘-οι’ και ‘-η’ αντίστοιχα. Το ελληνικό κείμενο που χρησιμοποιήθηκε από τους συγγραφείς για την κατασκευή των συνόλων δεδομένων αντλήθηκε από την ηλεκτρονική εφημερίδα ‘Ελευθεροτυπία’ (http://www.elda.fr/catalogue/en/text/W0022.html). Προφανώς, θα πρέπει τα δεδομένα που θα χρησιμοποιηθούν να ελεγχθούν πρωτίστως ως προς την ορθογραφία τους, καθώς αυτά θα καθορίσουν την ορθότητα του μοντέλου.

Ένα παράδειγμα του διανύσματος χαρακτηριστικών φαίνεται στο παρακάτω σχήμα, όπου έχουν επιλεγεί οι 3 προηγούμενες και επόμενες λέξεις μεταξύ της λέξης-στόχου:

1w-2	2w-2	3w-2	1w-1	2w-1	3w-1	1w+1	2w+1	3w+1	1w+2	2w+2	3w+3	c
ς	ές	τές	ι	αι	ναι	-	-	-	-	-	-	οι

Το παραπάνω διάνυσμα χαρακτηριστικών προέκυψε από την πρόταση ‘αυτοί οι μαθητές είναι καλοί’, όπου η αμφίσημη λέξη σε αυτήν την περίπτωση είναι η λέξη ‘καλοί’. Παρατηρούμε ότι κάθε χαρακτηριστικό έχει ένα όνομα aw±b, όπου το a δηλώνει τον αριθμό των γραμμάτων από το τέλος προς την αρχή της λέξης. Όπως αναφέρθηκε παραπάνω, λαμβάνονται το πολύ τρία γράμματα από το τέλος της λέξης. Το σύμβολο + χρησιμοποιείται για τις λέξεις που βρίσκονται δεξιά της αμφίσημης, ενώ το – για όσες βρίσκονται αριστερά. Το b δηλώνει τη θέση της γειτονικής λέξης αναφορικά με τη λέξη στόχο. Επομένως, το χαρακτηριστικό ‘2w-1’ έχει τιμή ίση με ‘αι’, καθώς πρόκειται για τα δύο τελευταία γράμματα της πρώτης λέξης που βρίσκεται πριν της αμφίσημης, δηλαδή της λέξης ‘είναι’. Επειδή η αμφίσημη λέξη είναι η τελευταία στην υπό μελέτη πρόταση, τα χαρακτηριστικά που αντιστοιχούν στις λέξεις που έπονται αυτής δεν έχουν τιμή, οπότε τοποθετείται παύλα.

Παρατηρούμε στο παραπάνω διάνυσμα το χαρακτηριστικό με όνομα ‘c’. Αυτό αντιστοιχεί στην κλάση εξόδου του διανύσματος χαρακτηριστικών και δηλώνει τη σωστή ορθογραφία της υπό μελέτη λέξης. Οι συγγραφείς παρατήρησαν ότι στην περίπτωση του ζεύγους των αμφίσημων ρημάτων (-ε/-αι) υπήρχε μεροληψία ως προς τη μία κλάση στην περίπτωση της ύπαρξης ερωτηματικού. Γι’ αυτό χρησιμοποίησαν ένα επιπλέον χαρακτηριστικό που δήλωνε αν η πρόταση ήταν κατάφαση ή ερώτηση.

Τα πειράματα εκτελέστηκαν στο πρόγραμμα WEKA, όπου χρησιμοποιήθηκαν οι παρακάτω αλγόριθμοι ταξινόμησης:

- Id3
- C4.5
- Nearest Neighbor (k-NN)
- NaiveBayes
- RandomForest

Επειδή τα σύνολα δεδομένων ήταν ανομοιόμορφα ως προς το μέγεθος, χρησιμοποιήθηκε το φίλτρο SMOTE του WEKA, όπου η κλάση με τις λιγότερες εγγραφές υπερ-δειγματοληπτείται λαμβάνοντας κάθε δείγμα και συνθέτοντας παραδείγματα μέσω της ένωσης οποιωνδήποτε από τους k κοντινότερους γείτονες της κλάσης. Για τη διαδικασία ελέγχου κάθε αλγορίθμου εφαρμόστηκε 10-fold cross validation. Για κάθε πείραμα χρησιμοποιήθηκε ένας από τους προαναφερθέντες αλγορίθμους και διαφορετικοί συνδυασμοί των χαρακτηριστικών. Πιο συγκεκριμένα, ξεκινώντας από όλα τα διαθέσιμα χαρακτηριστικά (5 λέξεις εκατέρωθεν της αμφίσημης λέξης), κάθε φορά αφαιρούσαν μία λέξη από το διάνυσμα των χαρακτηριστικών, μέχρι να καταλήξουν, τελικά, σε μία μόνο λέξη. Ξεκινώντας, λοιπόν, από 5 λέξεις πριν και μετά της εξεταζόμενης λέξης (το αναφέρουν σαν [-5,5] παράθυρο) και αφαιρώντας τα χαρακτηριστικά ένα προς ένα, συρρικνώνουν το παράθυρο, ώστε να πλησιάζει τη λέξη-στόχο ως εξής: [-5,4], [-5,3], …, [-5,0] αριστερά της λέξης και [-4,5], [-4,4], …, [-4,0] δεξιά της λέξης, καταλήγοντας τελικά στο [0,1].

Ο συνολικός αριθμός των διαφορετικών συνδυασμών χαρακτηριστικών που χρησιμοποιήθηκαν στα πειράματα ήταν 35 (6*6 εξαιρώντας το παράθυρο [0,0]) και για κάθε συνδυασμό χαρακτηριστικών εξετάστηκαν και οι πέντε αλγόριθμοι, οπότε εκτελέστηκαν συνολικά 13*35 πειράματα (ο αλγόριθμος των κοντινότερων γειτόνων kNN χρησιμοποιήθηκε για k=1,3,5.

Τα αποτελέσματα των πειραμάτων τους οδήγησαν τους συγγραφείς στο συμπέρασμα ότι οι μέθοδοι Id3 και Random Forest είχαν την καλύτερη απόδοση, ωστόσο τα αποτελέσματα του Id3 ήταν παραπλανητικά, καθώς ένα σημαντικό ποσοστό των παραδειγμάτων δεν ταξινομήθηκε. Επομένως, κατέληξαν ότι ο καλύτερος αλγόριθμος ήταν ο Random Forest. Να σημειωθεί ότι η μέθοδός τους χρησιμοποιεί χαμηλού επιπέδου πληροφορία, καθώς λαμβάνονται υπόψη οι καταλήξεις των λέξεων, σε αντίθεση με άλλες μεθόδους που ενσωματώνουν πιο υψηλού επιπέδου πληροφορία (PoS tags). Επίσης, η μέθοδός τους έχει εφαρμοστεί μόνο σε δύο ζεύγη αμφίσημων λέξεων, οπότε, μελλοντικά, θα μπορούσε να εξεταστεί και σε άλλα παραδείγματα αμφισημίας, καθώς και να χρησιμοποιηθούν και άλλοι συνδυασμοί χαρακτηριστικών ή διαφορετικοί αλγόριθμοι επιβλεπόμενης μάθησης.

Οι συγγραφείς στο [25] προτείνουν ένα εργαλείο βασισμένο σε χαρακτηριστικά πολλαπλών επιπέδων το οποίο εκτελεί διόρθωση ορθογραφικών λαθών χρησιμοποιώντας τεχνικές μηχανικής μάθησης. Ο στόχος τους είναι να ενσωματώσουν όσο το δυνατόν περισσότερη πληροφορία στα μοντέλο τους και, γι' αυτό το λόγο, χρησιμοποιούν χαρακτηριστικά από το επίπεδο των χαρακτήρων, των φωνημάτων και των λέξεων, καθώς και από το συντακτικό και σημασιολογικό επίπεδο.

Τα χαρακτηριστικά αυτά τροφοδοτούνται σε ένα μοντέλο Support Vector Machine το οποίο προβλέπει τη σωστή ορθογραφία των λέξεων. Οι συγγραφείς υποστηρίζουν ότι η μέθοδός τους μπορεί να ανιχνεύει και να διορθώνει τα ορθογραφικά λάθη τόσο σε μη-λέξεις (λέξεις που δεν περιέχονται στο λεξικό) όσο και σε πραγματικές λέξεις, χρησιμοποιώντας τις ίδιες μεθόδους εξαγωγής χαρακτηριστικών και εξαλείφοντας το χάσμα μεταξύ των παραδοσιακών μεθόδων αναγνώρισης λαθών και αυτών που βασίζονται στα συμφραζόμενα.

Το εργαλείο τους διακρίνεται σε δύο μέρη: το αυτόνομο μέρος που εκτελεί τη διόρθωση των ορθογραφικών λαθών και το μέρος που εκτελεί διόρθωση των λαθών βασισμένη στα συμφραζόμενα. Στο αυτόνομο μέρος πραγματοποιείται η γενίκευση των υποψήφιων λέξεων από το λεξικό, μέσω της εφαρμογής δύο διαφορετικών αλγορίθμων. Μία εκ των υποψήφιων λέξεων αποτελεί και τη σωστή λέξη που πρέπει να αντικαταστήσει τη λέξη-στόχο με τη λανθασμένη ορθογραφία. Ο πρώτος αλγόριθμος του αυτόνομου μέρους εντάσσεται στο επίπεδο των φωνημάτων και επιλέγει εκείνες τις λέξεις από το λεξικό που προφέρονται με παρόμοιο τρόπο. Ο δεύτερος αλγόριθμος προσθέτει στο σύνολο των υποψήφιων λέξεων όλες τις λέξεις που αν γίνει η αλλαγή ενός χαρακτήρα (προσθήκη, διαγραφή, εναλλαγή, αντικατάσταση), η υποψήφια λέξη θα ταυτιστεί με τη λέξη στόχο. Στη συνέχεια, το σύνολο των υποψήφιων λέξεων βαθμολογείται μέσω ενός μοντέλου λαθών και κρατούνται μόνο οι πέντε πρώτες λέξεις με την υψηλότερη βαθμολογία. Οι λέξεις αυτές, προωθούνται στο μέρος της διόρθωσης λαθών βασισμένης στα συμφραζόμενα. Εκεί, συλλέγεται επιπλέον πληροφορία από το λεξιλογικό, συντακτικό και σημασιολογικό επίπεδο. Όλη αυτή η πληροφορία τροφοδοτείται σε ένα μοντέλο SVM το οποίο ξαναβαθμολογεί τις πέντε υποψήφιες λέξεις, οπότε η λέξη με την υψηλότερη βαθμολογία αποτελεί τη διόρθωση της λέξης-στόχου.

Όσον αφορά το επίπεδο των φωνημάτων, ο στόχος είναι να βρεθούν οι λέξεις που προφέρονται παρόμοια με τη λέξη-στόχο. Αυτό επιτυγχάνεται μέσω ενός αλγορίθμου, γνωστού με το όνομα Soundex, ο οποίος αντιστοιχεί κάθε αλληλουχία χαρακτήρων σε ένα φωνητικό κωδικό. Υπολογίζοντας, επομένως, το φωνητικό κωδικό της λέξης-στόχου, οι συγγραφείς αναζητούν στο λεξικό όλες τις λέξεις με τον ίδιο κωδικό. Στο επίπεδο των λέξεων, επιλέγονται bigram τα οποία προστίθενται στο διάνυσμα χαρακτηριστικών. Σαν bigram ορίζεται το ζεύγος των λέξεων αριστερά και δεξιά της υπό μελέτη λέξης. Τα χαρακτηριστικά που προέρχονται από το συντακτικό επίπεδο αντιστοιχούν σε PoS tags των λέξεων που περιβάλλουν τη λέξη στόχο. Τέλος, στο διάνυσμα χαρακτηριστικών ενσωματώνεται πληροφορία από το σημασιολογικό επίπεδο.

Η πληροφορία αυτή είναι χρήσιμη στις περιπτώσεις λέξεων, η διαφοροποίηση των οποίων δεν επιτυγχάνεται από τα biagram ή τα PoS tags, όπως για παράδειγμα οι λέξεις της αγγλικής γλώσσας ‘dessert’ και ‘desert’. Ωστόσο, στο σημασιολογικό επίπεδο, οι αναγνώστες μπορούν να επιλέξουν τη σωστή υποψηφία λέξη, καθώς ακολουθούν μία λεξιλογική αλυσίδα και μπορούν να αντιληφθούν από τα συμφραζόμενα ποια είναι η κατάλληλη λέξη για την εκάστοτε περίπτωση. Για να το επιτύχουν αυτό οι συγγραφείς εκπαιδεύουν ένα vector space μοντέλο, γνωστό και ως bag-of-words μοντέλο, για κάθε λήμμα-στόχο. Το bag-of-words μοντέλο είναι μία αναπαράσταση που χρησιμοποιείται στην Επεξεργασία Φυσικής Γλώσσας, όπου ένα κείμενο (που μπορεί να είναι μία λέξη ή ολόκληρο έγγραφο) αναπαριστάται σαν ένα σύνολο από λέξεις, όπου παραβλέπεται η γραμματική ή και η σειρά των λέξεων, αλλά διατηρείται η πολλαπλότητα αυτών.

Μετά, λοιπόν, την κατασκευή του διανύσματος των χαρακτηριστικών που προέρχονται από πολλαπλά επίπεδα, οι συγγραφείς εκπαιδεύουν το SVM μοντέλο πάνω στα δεδομένα εκπαίδευσης, όπου η κλάση εξόδου είναι +1 για τη διόρθωση και -1 (δυαδική ταξινόμηση) για τους λανθασμένους υποψηφίους. Στη φάση της πρόβλεψης των πέντε καλύτερων υποψηφίων, ο SVM θέτει στην έξοδο τιμές από -1 έως +1 έτσι ώστε να είναι δυνατή η βαθμολόγηση των υποψήφιων λέξεων. Παρατηρούμε ότι σε αυτήν την περίπτωση έχουμε ένα πρόβλημα παλινδρόμησης, καθώς η έξοδος παίρνει συνεχόμενες τιμές από το διάστημα [-1,1].

Η συγκεκριμένη μέθοδος, σε αντίθεση με αυτήν που περιγράφηκε στο [21], δίνει βαρύτητα στην ενσωμάτωση πολλαπλών επιπέδων χαρακτηριστικών, έτσι ώστε να καλύπτονται όσο το δυνατόν περισσότερες περιπτώσεις ορθογραφικά λανθασμένων λέξεων. Η πολυπλοκότητα του μοντέλου τους οφείλεται στο γεγονός ότι θέλουν να καλύψουν ένα ευρύτερο εύρος ορθογραφικών λαθών, ενώ στο [21] περιορίζονται στην περίπτωση της αμφισημίας λέξεων. Επίσης, χρησιμοποιούν ένα SVM μοντέλο, δηλαδή μία kernel μηχανή, στο πρόβλημα της αναγνώρισης ορθογραφικών λαθών.

Στην παρούσα διπλωματική εργασία επιλέχθηκε η μελέτη για την περίπτωση των αμφίσημων λέξεων και, πιο συγκεκριμένα, το ζεύγος των ρημάτων με κατάληξη –τε και –ται.

Η αναλυτική παρουσίαση των πειραμάτων και των αποτελεσμάτων αυτών ακολουθεί στο επόμενο κεφάλαιο.

Ε΄ **Πειραματισμοί με Αλγόριθμους και Σετ χαρακτηριστικών**

Ε΄.1 Περιβάλλον WEKA

Πριν την παρουσίαση της διαδικασίας των πειραμάτων, θα γίνει μία σύντομη αναφορά στο περιβάλλον WEKA και στους αλγορίθμους μηχανικής μάθησης που χρησιμοποιήθηκαν.

Το WEKA (Waikato Environment for Knowledge Analysis) αποτελεί λογισμικό μηχανικής μάθησης που είναι γραμμένο σε Java και αναπτύχθηκε στο πανεπιστήμιο Waikato της Νέας Ζηλανδίας. Πρόκειται για ελεύθερο λογισμικό υπό την άδεια GNU General Public License.

Εικόνα 15 Λογότυπο WEKA
Πηγή: www.google.gr

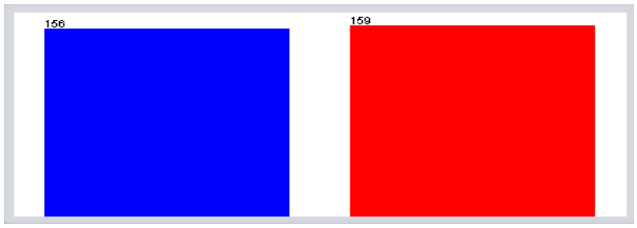
Το WEKA υποστηρίζει διάφορες εργασίες της εξόρυξης δεδομένων και της μηχανικής μάθησης, όπως είναι η προεπεξεργασία των δεδομένων, η συσταδοποίηση, η ταξινόμηση, η παλινδρόμηση, η οπτικοποίηση των δεδομένων και η επιλογή χαρακτηριστικών. Όλες οι τεχνικές του WEKA λειτουργούν με την προϋπόθεση ότι τα δεδομένα είναι διαθέσιμα σε κάποιο προσπελάσιμο αρχείο, όπου κάθε εγγραφή περιγράφεται από ένα συγκεκριμένο αριθμό χαρακτηριστικών, αριθμητικών ή μη.

Ε΄.2 Σύνολο δεδομένων που χρησιμοποιήθηκε

Για την εκτέλεση των πειραμάτων χρησιμοποιήθηκαν, όπως έχει αναφερθεί, ομόηχα ρήματα με κατάληξη –τε και –ται. Δηλαδή, πρόκειται για λέξεις που έχουν το ίδιο θέμα, αλλά διαφοροποιούνται ως προς την κατάληξη.

Οι λέξεις αυτές προήλθαν από κείμενα ηλεκτρονικών εφημερίδων και, πιο συγκεκριμένα, των εφημερίδων ‘Ελευθεροτυπία’, ‘Τα ΝΕΑ’ και ‘Το ΒΗΜΑ’. Ανατρέχοντας σε διαφορετικά άρθρα, έγινε η αναζήτηση προτάσεων που περιείχαν λέξεις που μας ενδιέφεραν. Οι προτάσεις αυτές συνέθεσαν το τελικό κείμενο το οποίο αποτελείται από 3754 λέξεις. Η αναζήτηση διήρκησε αρκετά, καθώς έπρεπε να βρεθούν ομόηχα ρήματα τα οποία συναντώνται και με τις δύο καταλήξεις.

Ωστόσο, επειδή η κλάση ‘ται’ υπερτερούσε έναντι της ‘τε’ (156 εγγραφές με κλάση εξόδου ‘ται’ και 94 εγγραφές με κλάση εξόδου ‘τε’), εφαρμόστηκε το φίλτρο SMOTE του WEKA, έτσι ώστε να υπάρξει μία ισορροπία μεταξύ των δύο κλάσεων και να αυξηθεί κατά ένα μικρό ποσοστό ο αριθμός των συνολικών δεδομένων που θα τροφοδοτηθούν στα μοντέλα. Το φίλτρο SMOTE του WEKA εφαρμόζει υπερδειγματοληψία (oversampling) στα δεδομένα μέσω της χρήσης ενός μέτρου μεροληψίας (bias), έτσι ώστε να επιλεγούν περισσότερα δείγματα από την κλάση που είναι σε μειονότητα. Πριν την εφαρμογή του φίλτρου SMOTE, δίνεται στο χρήστη η δυνατότητα να αλλάξει τις προεπιλεγμένες παραμέτρους της μεθόδου. Για να επιτευχθεί σχεδόν τέλεια ισορροπία μεταξύ των δύο κλάσεων, ορίστηκε το ποσοστό κατά το οποίο θα αυξηθεί η κλάση σε μειονότητα ίσο με 70%. Τελικά, μετά την εφαρμογή του φίλτρου SMOTE, το σύνολο δεδομένων αποτελείται από 315 εγγραφές και οι αναλογίες των δύο κλάσεων φαίνονται στη γραφική απεικόνιση που ακολουθεί:



	Ετικέτα	*Αριθμός εγγραφών*
1	ται	156
2	τε	159
Σύνολο		315

Η επιλογή των χαρακτηριστικών βασίστηκε στα συμφραζόμενα της πρότασης, καθώς πρόκειται για λέξεις με σωστή ορθογραφία, οι οποίες, όμως, χρησιμοποιούνται λανθασμένα με βάση τους γραμματικούς κανόνες της ελληνικής γλώσσας.

Επιλέχθηκαν οι τρεις λέξεις πριν και οι τις τρεις λέξεις μετά της λέξης-στόχου, καθώς και ο αριθμός της πρώτης και δεύτερης λέξης πριν τη λέξη στην οποία εστιάζουμε.

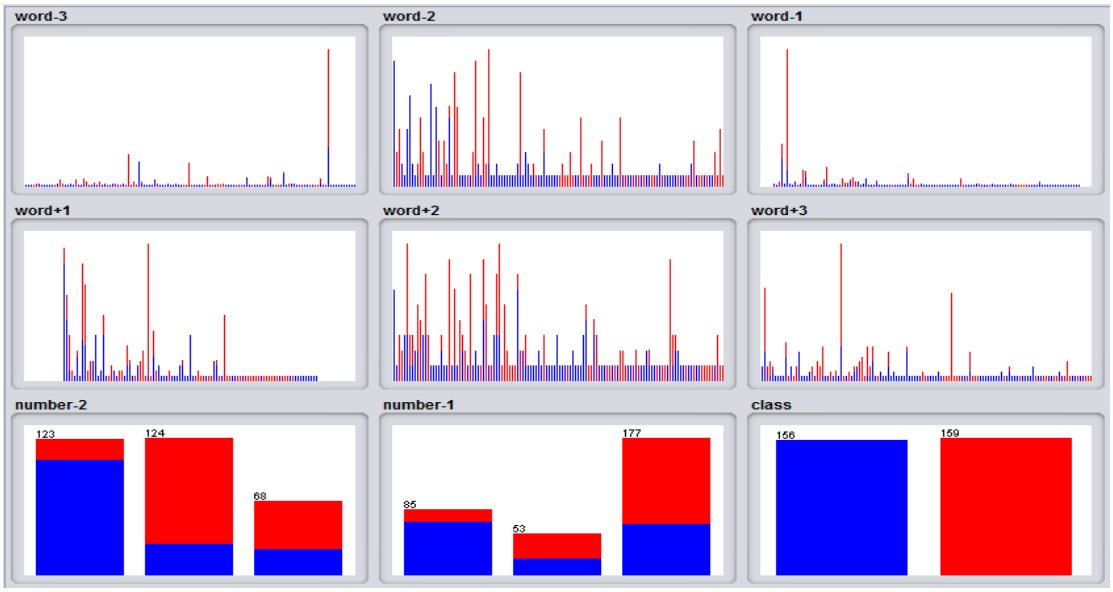
Ο αριθμός των δύο προηγούμενων λέξεων έχει επιλεγεί να προστεθεί στο διάνυσμα των χαρακτηριστικών, γιατί, στις περισσότερες περιπτώσεις, η σύνταξη μίας πρότασης στην οριστική έγκλιση ακολουθεί την εξής διαδοχή λέξεων: άρθρο, ουσιαστικό, ρήμα. Οπότε, συνήθως, ο αριθμός των δύο προηγούμενων λέξεων καθορίζει και τον αριθμό του ρήματος. Ωστόσο, στα παραδείγματα που χρησιμοποιήθηκαν, επειδή οι προτάσεις προέρχονται από ειδησεογραφικά κείμενα, όπου χρησιμοποιούνται συχνά η προστακτική, η υποτακτική αλλά και ερωτήσεις, για να δοθεί αμεσότητα στο λόγο, οι προαναφερθέντες κανόνες ανατρέπονται.

Οπότε, στην περίπτωση αυτή, θα πρέπει το μοντέλο να τροφοδοτηθεί με κατάλληλα δεδομένα, ώστε να αποκτήσει την ικανότητα να αναγνωρίζει τις διαφορετικές περιπτώσεις, δηλαδή να γίνεται αποτελεσματικά η γενίκευσή του. Παρακάτω δίνεται ένα παράδειγμα του διανύσματος χαρακτηριστικών. Στην πρώτη γραμμή βλέπουμε τα ονόματα των χαρακτηριστικών και στη δεύτερη γραμμή τις αντίστοιχες τιμές τους.

λέξη-3	λέξη-2	λέξη-1	λέξη+1	λέξη+2	λέξη+3	αριθμός-2	αριθμός-1	ταιτε
το	ίδιο	φαινόμενο	και	Για	τις	ΕΝ	ΕΝ	ται

Όπως φαίνεται παραπάνω, το διάνυσμα των χαρακτηριστικών αποτελείται από εννιά χαρακτηριστικά, εκ των οποίων το τελευταίο αποτελεί την κλάση εξόδου. Η τιμή της κλάσης εξόδου μπορεί να πάρει δύο τιμές, τις ‘ται’ και ‘τε’ οι οποίες αποτελούν και τις καταλήξεις των ρημάτων. Τα χαρακτηριστικά που αφορούν τον αριθμό κάθε λέξης μπορούν να πάρουν τις τιμές ‘ΕΝ’ για τον ενικό αριθμό και ‘ΠΛ’ για τον πληθυντικό. Οι τιμές των υπόλοιπων χαρακτηριστικών είναι ίσες με τις λέξεις που προηγούνται ή έπονται της λέξης-στόχου. Να σημειωθεί ότι σε ορισμένες περιπτώσεις υπάρχει παύλα στη θέση του χαρακτηριστικού, η οποία υποδηλώνει ότι δεν υπάρχει η αντίστοιχη λέξη πριν ή μετά της λέξης-στόχου.

Για παράδειγμα, η πρόταση «Παράλληλα αναμένεται να ληφθεί απόφαση για τη συνέχιση των κινητοποιήσεων» έχει μόνο μία λέξη πριν της ομόηχης λέξης, οπότε η τιμή των χαρακτηριστικών λέξη-3, λέξη-2 θα ισούται με '-'. Επίσης, υπάρχει περίπτωση οι δύο λέξεις που προηγούνται της λέξης που πρέπει να ελεχτούν να μην έχουν αριθμό, όπως για παράδειγμα στην περίπτωση προθέσεων ή επιρρημάτων. Για παράδειγμα, η λέξη παράλληλα στο προηγούμενο παράδειγμα δεν έχει αριθμό, οπότε το χαρακτηριστικό αριθμός-1 θα ισούται με '-'. Στην παρακάτω εικόνα φαίνονται τα ιστογράμματα των εννιά χαρακτηριστικών, όπως τα εμφάνισε το πρόγραμμα WEKA. Για κάθε χαρακτηριστικό παρατηρείται η συχνότητα εμφάνισης των διαφορετικών τιμών του αναλογικά με την εκάστοτε τιμή της κλάσης εξόδου. Το μπλε χρώμα αντιστοιχεί στην τιμή 'ται' της κλάσης εξόδου, ενώ το κόκκινο χρώμα στην τιμή 'τε' αντίστοιχα.



Εικόνα 16 Ιστογράμματα των χαρακτηριστικών του συνόλου δεδομένων

Ε΄.3 Πειραματική ανάλυση

Αφού έχει κατασκευαστεί το σύνολο δεδομένων, το επόμενο βήμα είναι να ξεκινήσουν οι πειραματισμοί στο WEKA. Να επισημανθεί ότι για να είναι δυνατόν να διαβαστούν οι ελληνικοί χαρακτήρες από το WEKA, πρέπει η κωδικοποίηση να οριστεί ίση με 'UTF-8'. Στη συνέχεια, τα δεδομένα τροφοδοτούνται με τον εξής τρόπο: στην αρχή εισάγεται ολόκληρο το διάνυσμα χαρακτηριστικών και, έπειτα, γίνεται μείωση δεξιά και αριστερά τα χαρακτηριστικά εναλλάξ, κρατώντας σταθερά τα χαρακτηριστικά που αφορούν τον αριθμό των λέξεων. Οπότε, ξεκινώντας από το αρχικό παράθυρο [-3,3,αριθμός], οι μειώσεις ακολουθούν την εξής πορεία: [-3,2,αριθμός], [-3,1,αριθμός], [-3,0,αριθμός], [-2,3,αριθμός]…[0,1,αριθμός]. Όλοι οι δυνατοί συνδυασμοί που προκύπτουν είναι ίσοι με 15 (4*4 εξαιρώντας το παράθυρο [0,0,αριθμός]). Τέλος τρέχουν τα μοντέλα στο αρχικό διάνυσμα χαρακτηριστικών, αφού αφαιρέσω τα χαρακτηριστικά αριθμός-2, αριθμός-1.

Η επιλογή των αλγορίθμων που χρησιμοποιήθηκαν για την εκτέλεση των πειραμάτων βασίστηκε στη βιβλιογραφία που μελετήθηκε για την εκπόνηση της παρούσας πτυχιακής εργασίας, όπου διαπιστώθηκε ότι οι συγκεκριμένοι αλγόριθμοι χρησιμοποιούνται ευρέως σε προβλήματα ταξινόμησης και, ειδικότερα, έχουν χρησιμοποιηθεί σε ερευνητικές μελέτες προβλημάτων αναγνώρισης ορθογραφικών λαθών, όπως στο [21]. Γι'αυτό το λόγο, καθώς το προς μελέτη πρόβλημα αποτελεί πρόβλημα δυαδικής ταξινόμησης, επιλέχθηκαν τεχνικές που έχουν χρησιμοποιηθεί διεξοδικά από ερευνητές σε μελέτες προβλημάτων μηχανικής μάθησης και έχει αποδειχθεί ότι δίνουν ικανοποιητικά αποτελέσματα. Παρακάτω, δίνεται η αναλυτική περιγραφή των μοντέλων του WEKA που χρησιμοποιήθηκαν για την εκτέλεση των πειραμάτων. Σε προηγούμενο κεφάλαιο έχει δοθεί μία πιο μαθηματική περιγραφή των αντίστοιχων αλγορίθμων.

Οι αλγόριθμοι του WEKA που θα τρέξουν στο σύνολο δεδομένων είναι οι:

- **NaiveBayes**: Αντιστοιχεί στον ταξινομητή Naïve Bayes και χρησιμοποιεί κλάσεις εκτιμητών. Πρόκειται για πιθανοτικό, μη-παραμετρικό μοντέλο.

- **RandomForest**: Η μέθοδος Random Forest, όπως αναλύθηκε στο προηγούμενο κεφάλαιο, αποτελεί παράδειγμα συλλογικής μάθησης και χρησιμοποιείται σε προβλήματα ταξινόμησης, όπου κατασκευάζει ένα πλήθος δέντρων απόφασης κατά τη διαδικασία της εκπαίδευσης και εξάγει σαν αποτέλεσμα την κλάση που εμφανίζεται πιο συχνά στο σύνολο των παραγόμενων δέντρων.

- **KStar**: Πρόκειται για έναν ταξινομητή που βασίζεται στα στιγμιότυπα (instance-based), όπου η κλάση εξόδου ενός στιγμιοτύπου βασίζεται στην κλάση των στιγμιοτύπων που είναι παρόμοια με αυτό, όπου η ομοιότητα διαπιστώνεται με τη χρήση μιας συνάρτησης ομοιότητας. Διαφέρει από άλλους αλγορίθμους βασισμένους σε στιγμιότυπα ως προς το γεγονός ότι χρησιμοποιεί μία συνάρτηση απόστασης που βασίζεται στην εντροπία.

- **DecisionStump**: Η κλάση του WEKA για τη δημιουργία ενός decision stump. Το decision stump είναι ένα μοντέλο μηχανικής μάθησης που αποτελείται από ένα δέντρο απόφασης ενός επιπέδου. Πιο συγκεκριμένα, πρόκειται για ένα δέντρο απόφασης που περιλαμβάνει έναν εσωτερικό κόμβο (τη ρίζα του δέντρου) ο οποίος συνδέεται απευθείας με τους τελικούς κόμβους (τα φύλλα του δέντρου). Ένα μοντέλο decision stump κάνει προβλέψεις βασισμένες στην τιμή ενός μόνο χαρακτηριστικού εισόδου. Γι' αυτό, μπορεί να συναντηθεί και με το όνομα 1-rules. Τα μοντέλα decision stump χρησιμοποιούνται συχνά σε συνδυασμό με έναν boosting αλγόριθμο. Χρησιμοποιούνται τόσο σε προβλήματα παλινδρόμησης (βασισμένα στο mean-squared error) όσο και σε προβλήματα κατηγοριοποίησης (βασισμένα στην εντροπία).

- **ID3**: Πρόκειται για κλάση του WEKA που χρησιμοποιείται για την κατασκευή ενός μη-κλαδεμένου (unpruned) δέντρου απόφασης βάσει του αλγορίθμου ID3. Ο αλγόριθμος ID3 ξεκινά έχοντας το αρχικό σύνολο δεδομένων στη ρίζα του δέντρου. Σε κάθε επανάληψη, στοχεύει σε κάθε χαρακτηριστικό του αρχικού συνόλου που δε χρησιμοποιείται και υπολογίζει την εντροπία του (ή το κέρδος πληροφορίας).

Έπειτα, επιλέγει το χαρακτηριστικό που έχει τη μικρότερη εντροπία (ή το μεγαλύτερο κέρδος πληροφορίας). Το αρχικό σύνολο διαχωρίζεται με βάση το επιλεγμένο χαρακτηριστικό έτσι ώστε να παραχθούν υποσύνολα των δεδομένων. Ο αλγόριθμος εξακολουθεί να επαναλαμβάνεται σε κάθε υποσύνολο, λαμβάνοντας υπόψη μόνο χαρακτηριστικά που δεν έχουν επιλεγεί στο παρελθόν. Ο αλγόριθμος ID3 του WEKA μπορεί να εφαρμοστεί μόνο σε κατηγορικές μεταβλητές και τα φύλλα του δέντρου που δεν έχουν τιμή μπορεί να οδηγήσουν σε εγγραφές που δεν έχουν ταξινομηθεί.

- **J48**: Κλάση για τη δημιουργία ενός C4.5 δέντρου απόφασης, το οποίο μπορεί να έχει ή να μην έχει υποστεί κλάδεμα (pruning). Ο C4.5, ο οποίος αποτελεί επέκταση του αλγορίθμου ID3, χρησιμοποιείται για την παραγωγή δέντρων που εφαρμόζονται σε προβλήματα κατηγοριοποίησης, γι' αυτό συναντάται και με το όνομα στατιστικός ταξινομητής. Ο συγκεκριμένος αλγόριθμος κατασκευάζει δέντρα από ένα σύνολο δεδομένων εκπαίδευσης με τον ίδιο τρόπο που χρησιμοποιεί και ο ID3, αξιοποιώντας τη βασική ιδέα της εντροπίας πληροφοριών. Σε κάθε κόμβο του δέντρου, ο C4.5 επιλέγει εκείνο το χαρακτηριστικό του συνόλου δεδομένων το οποίο διαχωρίζει με τον πιο αποτελεσματικό τρόπο το σύνολο των δειγμάτων σε υποσύνολα που υπάγονται στη μία ή την άλλη κλάση. Το κριτήριο διαχωρισμού είναι το κανονικοποιημένο κέρδος πληροφορίας (διαφορά στην εντροπία). Το χαρακτηριστικό με το υψηλότερο κέρδος πληροφορίας επιλέγεται για τη λήψη της απόφασης. Έπειτα, ο C4.5 αλγόριθμος επαναλαμβάνεται στις μικρότερες υπολίστες.

Ε΄3.1 Αποτελέσματα πειραμάτων

Στην υποενότητα αυτή παρουσιάζεται η απόδοση των διαφορετικών μοντέλων ταξινομητών στο σύνολο δεδομένων που κατασκευάστηκε, με βάση τις μετρικές Precision, Recall και F-measure. Να σημειωθεί ότι τα μοντέλα που χρησιμοποιήθηκαν έχουν συγκεκριμένες τιμές παραμέτρων, οι οποίες είναι οι προεπιλεγμένες τιμές από το πρόγραμμα WEKA. Προκειμένου να γίνει ο έλεγχος της απόδοσης των μοντέλων χρησιμοποιείται 10-fold cross-validation. Για την αξιολόγηση των μοντέλων θα δοθεί μεγαλύτερη βαρύτητα στο μέτρο F-measure, αφού, όπως ειπώθηκε και σε προηγούμενη ενότητα, αποτελεί το πιο ολοκληρωμένο και σημαντικό μέτρο αξιολόγησης. Πέρα από το F-measure, χρησιμοποιούνται τα μέτρα Precison και Recall, τα οποία χρησιμοποιούνται ευρέως για την αξιολόγηση μοντέλων μηχανικής μάθησης.

Για τα πειράματα έχουν χρησιμοποιηθεί διαφορετικοί συνδυασμοί των χαρακτηριστικών, όπως αναλύθηκε παραπάνω. Στους πίνακες που ακολουθούν, η στήλη παράθυρο θα αναφέρεται στον αντίστοιχο συνδυασμό χαρακτηριστικών που εφαρμόστηκε στους διαφορετικούς αλγορίθμους. Για παράδειγμα, το παράθυρο [-3,3,αριθμός] αναφέρεται και στα οκτώ χαρακτηριστικά του συνόλου δεδομένων, ενώ το παράθυρο [-3,3] στα έξι χαρακτηριστικά που αφορούν μόνο τις λέξεις πριν και μετά της ομόηχης λέξης (έχουν αφαιρεθεί τα δύο χαρακτηριστικά που αφορούν τον αριθμό των δύο προηγούμενων λέξεων).

Στον πίνακα που ακολουθεί παρουσιάζεται η απόδοση των διαφορετικών αλγορίθμων στο παράθυρο [-3,3,αριθμός], δηλαδή όταν λαμβάνονται υπόψη όλα τα χαρακτηριστικά του συνόλου δεδομένων.

Αλγόριθμος	Παράθυρο	Κλάσεις	Precision	Recall	F-Measure
NaiveBayes	[-3,3,αριθμός]	-ται	0,902	0,769	0,830
		-τε	0,802	0,918	0,856
RandomForest	[-3,3,αριθμός]	-ται	0,885	0,744	0,808
		-τε	0,783	0,906	0,840
KStar	[-3,3,αριθμός]	-ται	0,919	0,795	0,852
		-τε	0,822	0,931	0,873
DecisionStump	[-3,3,αριθμός]	-ται	0,846	0,667	0,746
		-τε	0,729	0,881	0,798
ID3	[-3,3,αριθμός]	-ται	0,901	0,889	0,895
		-τε	0,938	0,945	0,941
J48	[-3,3,αριθμός]	-ται	0,846	0,667	0,746
		-τε	0,729	0,881	0,798

Πίνακας 3 Αξιολόγηση μοντέλων όταν λαμβάνονται υπόψη όλα τα χαρακτηριστικά του συνόλου δεδομένων

Στον επόμενο πίνακα παρουσιάζονται οι τιμές απόδοσης των μοντέλων όταν αυτά εφαρμόζονται στα χαρακτηριστικά που αντιστοιχούν στις τρεις λέξεις πριν και μετά της λέξης-στόχου. Έχουν αφαιρεθεί, δηλαδή, τα χαρακτηριστικά που αφορούν τον αριθμό των δύο προηγούμενων λέξεων.

Αλγόριθμος	Παράθυρο	Κλάσεις	Precision	Recall	F-Measure
NaiveBayes	[-3,3]	-ται	0,939	0,788	0,857
		-τε	0,821	0,950	0,880
RandomForest	[-3,3]	-ται	0,913	0,737	0,816
		-τε	0,783	0,931	0,851
KStar	[-3,3]	-ται	0,929	0,756	0,834
		-τε	0,798	0,943	0,865
DecisionStump	[-3,3]	-ται	0,629	0,936	0,753

		-τε	0,880	0,459	0,603
ID3	[-3,3]	-ται	0,901	0,889	0,895
		-τε	0,938	0,945	0,941
J48	[-3,3]	-ται	0,926	0,481	0,633
		-τε	0,654	0,962	0,779

Πίνακας 4 Αξιολόγηση μοντέλων όταν δε λαμβάνονται υπόψη τα χαρακτηριστικά που αφορούν τον αριθμό των δύο προηγούμενων λέξεων

Στη συνέχεια, ακολουθούν τα αποτελέσματα του καλύτερου συνδυασμού συνόλου χαρακτηριστικών και αλγορίθμου. Δηλαδή, έγινε τροφοδότηση σε κάθε αλγόριθμο όλοι οι δυνατοί συνδυασμοί χαρακτηριστικών (συνολικά 15 συν 1 στην περίπτωση που αφαιρούμε τα δύο χαρακτηριστικά που αφορούν τον αριθμό). Αυτό σημαίνει ότι εκτελέστηκαν συνολικά 16*6=96 πειράματα, όπου κάθε πείραμα περιλαμβάνει την εφαρμογή ενός αλγορίθμου στο σετ χαρακτηριστικών με σκοπό να αξιολογηθεί η απόδοσή του. Αφού έτρεξα κάθε αλγόριθμο σε όλα τα δυνατά σετ χαρακτηριστικών, επέλεξα την καλύτερη απόδοση και συμπερίλαβα τις τιμές στον πίνακα που ακολουθεί. Το παράθυρο υποδηλώνει το σετ χαρακτηριστικών στο οποίο παρατηρήθηκε η βέλτιστη απόδοση.

Αλγόριθμος	Παράθυρο	Κλάσεις	Precision	Recall	F-Measure
NaiveBayes	[-3,3]	-ται	0,939	0,788	0,857
		-τε	0,821	0,950	0,880
RandomForest	[-3,3]	-ται	0,913	0,737	0,816
		-τε	0,783	0,931	0,851
KStar	[-3,3,αριθμός]	-ται	0,919	0,795	0,852
		-τε	0,822	0,931	0,873

DecisionStump	[-3,3,αριθμός]	-ται	0,846	0,667	0,746
		-τε	0,729	0,881	0,798
ID3	[-2,3,αριθμός]	-ται	0,930	0,904	0,917
		-τε	0,945	0,960	0,953
J48	[-3,3,αριθμός]	-ται	0,846	0,667	0,746
		-τε	0,729	0,881	0,798

Πίνακας 5 Αξιολόγηση μοντέλων όταν λαμβάνεται ο καλύτερος συνδυασμός χαρακτηριστικών και αλγορίθμου

Στον παραπάνω πίνακα, παρατηρώντας τις μετρικές Precision, Recall και F-Measure, διαπιστώνουμε ότι η καλύτερη απόδοση επιτυγχάνεται από το μοντέλο ID3.

Γι’αυτό, στον πίνακα που ακολουθεί θα δοθούν αναλυτικά οι τιμές των μετρικών για κάθε σετ χαρακτηριστικών, για να δούμε πώς αυτά επηρεάζουν την απόδοση του συγκεκριμένου αλγορίθμου.

Παράθυρο	Κλάσεις	Precision	Recall	F-Measure
[-3,3,αριθμός]	-ται	0,901	0,889	0,895
	-τε	0,938	0,945	0,941
[-3,2,αριθμός]	-ται	0,901	0,889	0,895
	-τε	0,938	0,945	0,941
[-3,1,αριθμός]	-ται	0,901	0,889	0,895
	-τε	0,938	0,945	0,941
[-3,0,αριθμός]	-ται	0,900	0,887	0,894
	-τε	0,937	0,944	0,940
[-2,3,αριθμός]	-ται	0,930	0,904	0,917
	-τε	0,945	0,960	0,953
[-2,2,αριθμός]	-ται	0,930	0,904	0,917
	-τε	0,945	0,960	0,953

[-2,1,αριθμός]	-ται	0,930	0,904	0,917
	-τε	0,945	0,960	0,953
[-2,0,αριθμός]	-ται	0,926	0,900	0,913
	-τε	0,946	0,961	0,953
[-1,3,αριθμός]	-ται	0,808	0,778	0,792
	-τε	0,881	0,899	0,890
[-1,2,αριθμός]	-ται	0,778	0,778	0,778
	-τε	0,881	0,881	0,881
[-1,1,αριθμός]	-ται	0,789	0,818	0,804
	-τε	0,914	0,898	0,906
[-1,0,αριθμός]	-ται	0,753	0,810	0,780
	-τε	0,891	0,854	0,872
[0,3,αριθμός]	-ται	0,841	0,828	0,835
	-τε	0,885	0,895	0,890
[0,2,αριθμός]	-ται	0,803	0,828	0,815
	-τε	0,885	0,867	0,876
[0,1,αριθμός]	-ται	0,861	0,818	0,839
	-τε	0,835	0,874	0,854
[-3,3]	-ται	0,901	0,889	0,895
	-τε	0,938	0,945	0,941

Πίνακας 6 Αξιολόγηση του μοντέλου ID3 για όλους τους συνδυασμούς των χαρακτηριστικών

Ε΄3.2 Σχολιασμός αποτελεσμάτων

Μελετώντας τα αποτελέσματα που παρουσιάστηκαν στην προηγούμενη ενότητα βγήκε το συμπέρασμα ότι ο αλγόριθμος ID3 παρουσίασε την καλύτερη απόδοση στο παράθυρο [-2,3,αριθμός] (την ίδια απόδοση παρουσιάζει και στα παράθυρα [-2,2,αριθμός], [-2,1,αριθμός]. Παρατηρώντας, μάλιστα, τον πίνακα 6 διαπιστώνεται ότι η απόδοση του μοντέλου σημειώνει ενίοτε σημαντικές διακυμάνσεις όταν αυτό τροφοδοτείται με διαφορετικά σετ χαρακτηριστικών.

Θα περίμενε κανείς ότι ο αλγόριθμος J48 θα παρουσίαζε καλύτερη απόδοση από τον ID3, καθώς ο πρώτος αποτελεί βελτιωμένη έκδοση του δεύτερου.

Πιο συγκεκριμένα, ο J48 διαχειρίζεται τις τιμές που απουσιάζουν από το σύνολο δεδομένων όπως, επίσης, εκτελεί πιο ολοκληρωμένο διαχωρισμό κατά την κατασκευή του δέντρου, αφού χρησιμοποιεί το ποσοστό κέρδους σε συνδυασμό με διαφορετικούς ευριστικούς αλγορίθμους. Ωστόσο, στην περίπτωσή αυτή δεν απουσιάζουν τιμές από το σύνολο δεδομένων και το διάνυσμα χαρακτηριστικών περιέχει αρκετή πληροφορία, ώστε να γίνεται αποτελεσματικά ο διαχωρισμός των δεδομένων κατά τη δημουργία του δέντρου. Γι'αυτό το λόγο, ο ID3 παρουσιάζει αρκετά καλύτερη απόδοση από τον J48. Ωστόσο, επειδή ο J48 χρησιμοποιείται για να ξεπεραστούν φαινόμενα υπερεκπαίδευσης (overfitting) που παρατηρούνται ορισμένες φορές με τον ID3, μελλοντικά θα ήταν ενδιαφέρον να μελετηθεί αν ο ID3 υπερεκπαιδεύεται, δηλαδή μαθαίνει πολύ καλά τα δεδομένα εκπαίδευσης αλλά αδυνατεί να αναγνωρίζει αποτελεσματικά καινούρια παραδείγματα. Αυτό μπορεί να γίνει τροφοδοτώντας με καινούρια δεδομένα τον ταξινομητή και συγκρίνοντας την απόδοσή του με αυτήν που παρατηρήθηκε πάνω στα δεδομένα ελέγχου.

Τα μοντέλα NaïveBayes, RandomForest και KStar απέδωσαν σε βαθμό συγκρίσιμο μεταξύ τους, αλλά πολύ χειρότερα συγκριτικά με το μοντέλο ID3. Ακολουθούν τα μοντέλα J48 και DecisionStump τα οποία είχαν παρόμοια απόδοση. Κατά την εκτέλεση των πειραμάτων με διαφορετικό συνδυασμό χαρακτηριστικών, τις περισσότερες φορές παρατηρήθηκε ότι, καθώς μειωνόταν το μέγεθος του διανύσματος χαρακτηριστικών, γινόταν χειρότερη και η απόδοση των μοντέλων. Κάποια μοντέλα, ωστόσο, εμφάνισαν καλύτερη συμπεριφορά σε υποσύνολα των χαρακτηριστικών. Επίσης, τα χαρακτηριστικά που αφορούν τον αριθμό των δύο λέξεων που προηγούνται της λέξης-στόχου μείωναν ορισμένες φορές την απόδοση των μοντέλων. Αυτό μπορεί να αποδοθεί στο γεγονός ότι το κείμενο προέρχεται από εφημερίδες, οπότε για να δοθεί περισσότερη αμεσότητα στο λόγο, χρησιμοποιούνται ερωτήσεις, προτάσεις στην προστακτική έγκλιση και σύνταξη που δεν ακολουθεί πάντα την παραδοσιακή σύνταξη των προτάσεων της ελληνικής γλώσσας.

Παρακάτω δίνονται ορισμένα παραδείγματα τα οποία δυσκόλεψαν τον ταξινομητή, με αποτέλεσμα να ταξινομηθούν λανθασμένα (misclassified instances) οι αντίστοιχες εγγραφές. Μελετάται η περίπτωση του αλγορίθμου ID3 στο παράθυρο [-3,3,αριθμός].

λέξη-3	**λέξη-2**	**λέξη-1**	**λέξη+1**	**λέξη+2**	**λέξη+3**	**αριθμός-2**	**αριθμός-1**	**Πραγματική κλάση**	**Πρόβλεψη**
ενώ	οι	βάσεις	να	εκδοθούν	το	ΠΛ	ΠΛ	ται	τε
-	να	μη	πολιτικό	άλλοθι	σε	-	-	ται	τε
-	-	-	τις	μνημονιακές	υποχρεώσεις	-	-	τε	ται
-	γνωρίζετε	ότι	λάθος	προς	τα	ΠΛ	-	τε	ται

Πίνακας 7 Παραδείγματα που ταξινομήθηκαν λανθασμένα από τον ταξινομητή ID3

Παρατηρώντας το πρώτο παράδειγμα του παραπάνω πίνακα, φαίνεται ότι η κλάση εξόδου ‘ται’ δεν μπορεί να προβλεφθεί σωστά. Η αδυναμία του ταξινομητή να προβλέψει σωστά την κλάση εξόδου οφείλεται στο γεγονός ότι τα χαρακτηριστικά αριθμός-2 και αριθμός-1 παίρνουν την τιμή ‘ΠΛ’. Παρατηρήθηκε ότι στην πλειονότητα των περιπτώσεων όπου η κλάση εξόδου ισούται με ‘ται’, ο αριθμός των δύο προηγούμενων λέξεων είναι ο ενικός. Επομένως, ο ταξινομητής μαθαίνει να προβλέπει σωστά αντίστοιχες περιπτώσεις, ωστόσο αποτυγχάνει στην περίπτωση που ο αριθμός των προηγούμενων λέξεων είναι ο πληθυντικός. Στα επόμενα δύο παραδείγματα του πίνακα 7, η αστοχία του ταξινομητή οφείλεται στην απουσία των τιμών αρκετών χαρακτηριστικών. Τέλος, στο τελευταίο παράδειγμα του πίνακα, η κλάση ‘τε’ και πάλι δεν μπορεί να προβλεφθεί λόγω της έλλειψης επαρκούς πληροφορίας στο διάνυσμα χαρακτηριστικών.

Προκειμένου να αυξηθεί η ακρίβεια και η αποτελεσματικότητα του ταξινομητή, θα πρέπει να ληφθούν υπόψη όλες οι περιπτώσεις κατά τις οποίες αστοχεί ο ταξινομητής να προβλέψει σωστά την κλάση εξόδου, οι πιο χαρακτηριστικές από τις οποίες αναφέρθηκαν παραπάνω. Αρχικά, μελλοντικά, είναι θεμιτό να τροφοδοτούνται τα μοντέλα με μεγαλύτερο αριθμό δεδομένων, έτσι ώστε να είναι, αναλογικά, πολύ μικρός ο αριθμός των εγγραφών από τις οποίες απουσιάζουν οι τιμές ορισμένων χαρακτηριστικών.

Αρκετό ενδιαφέρον παρουσιάζει η περίπτωση των λέξεων με κατάληξη ‘τε’, όπου ο ταξινομητής αστοχούσε συχνά λόγω της ιδιαιτερότητας των ρημάτων με αυτήν την κατάληξη. Αυτό οφείλεται στο γεγονός ότι τα ρήματα που βρίσκονται στο β’ πρόσωπο πληθυντικού αριθμού, όταν χρησιμοποιούνται σε ειδησεογραφικά κείμενα, εντοπίζονται, κυρίως, σε ερωτήσεις ή στην προστακτική και υποτακτική έγκλιση. Γι’αυτό το λόγο, θα γινόταν σε μία μελλοντική επέκταση της πτυχιακής εργασίας, να προστεθούν επιπλέον χαρακτηριστικά τα οποία να αφορούν το είδος της πρότασης, δηλαδή αν πρόκειται για κατάφαση ή ερώτηση, καθώς και την έγκλιση, δηλαδή αν έχουμε οριστική, προστακτική ή υποτακτική. Τέλος, παρατηρήθηκε ότι τα σημεία στίξης, και πιο συγκεκριμένα το κόμμα, έπαιζε σημαντικό ρόλο αν υπήρχε πριν τη λέξη-στόχο, κυρίως σε ρήματα με κατάληξη ‘τε’, επομένως θα μπορούσε να συμπεριληφθεί σαν ένα επιπλέον χαρακτηριστικό.

Στο παρακάτω πίνακα μπορούμε να δούμε την χρησιμότητα του SMOTE.

Αλγόριθμος	SMOTE	Κλάσεις	Precision	Recall	F-Measure
NaiveBayes	No	-ται -τε	0,899 0,669	0,744 0,862	0,814 0,753
NaiveBayes	Yes	-ται -τε	0,902 0,802	0,769 0,918	0,830 0,856
ID3	No	-ται -τε	0,914 0,922	0,941 0,887	0,928 0,904
ID3	Yes	-ται -τε	0,901 0,938	0,889 0,945	0,895 0,941

Πίνακας8 Αποτελέσματα πριν και μετά της χρήσης SMOTE (Παράθυρο (-3,3,Αριθμός))

ΣΤ΄ **Συμπεράσματα και Μελλοντική Έρευνα**

Στην παρούσα διπλωματική εργασία μελετήθηκε το πρόβλημα της αυτόματης αναγνώρισης ορθογραφικών λαθών με τη χρήση τεχνικών μηχανικής μάθησης. Η περίπτωση των λέξεων που εξετάστηκε είναι αρκετά πολύπλοκη, καθώς πρόκειται για λέξεις με σωστή ορθογραφία, όπου η επιλογή της κατάλληλης λέξης βασίζεται στα συμφραζόμενα της πρότασης. Αυτό κρίνει πολύ σημαντική την επιλογή του κατάλληλου διανύσματος χαρακτηριστικών, ώστε να κατασκευαστούν μοντέλα με καλή ικανότητα γενίκευσης. Μέχρι στιγμής, τα παραδοσιακά εργαλεία αυτόματης διόρθωσης σε κείμενα της ελληνικής γλώσσας βασίζονται στη διόρθωση λέξεων που δεν είναι συμβατές με λέξεις που υπάρχουν σε κάποιο λεξικό. Επομένως, η μέθοδος που προτείνεται έρχεται να δώσει λύση στην περίπτωση της διόρθωσης λέξεων οι οποίες έχουν λανθασμένη ορθογραφία με βάση τα συμφραζόμενα, ωστόσο αποτελούν έγκυρες λέξεις που μπορούν να βρεθούν σε οποιοδήποτε λεξικό. Στην παρούσα πτυχιακή εργασία μελετήθηκε μόνο ένα ζεύγος τέτοιων λέξεων, ωστόσο θα μπορούσαν να εξεταστούν μελλοντικά και άλλες περιπτώσεις ομόφωνων, δηλαδή λέξεων που προφέρονται ακριβώς το ίδιο αλλά έχουν διαφορετική κατάληξη.

Μελλοντικά, επίσης θα μπορούσαν να γίνουν πειραματισμοί και με άλλους τύπους χαρακτηριστικών που αφορούν τα συμφραζόμενα. Για παράδειγμα, επειδή η σύνταξη της πρότασης αλλάζει στην περίπτωση της ερώτησης, θα μπορούσε να προστεθεί ένα χαρακτηριστικό που να προσδιορίζει αν η πρόταση είναι καταφατική ή ερωτηματική. Επιπλέον, σαν μελλοντική επέκταση της παρούσας εργασίας προτείνεται η μελέτη και εφαρμογή και άλλων αλγορίθμων μηχανικής μάθησης, όπως είναι τα μοντέλα SVM και τα νευρωνικά δίκτυα. Η περίπτωση ομόηχων λέξεων που εξετάστηκε είναι αυτή των ρημάτων –ται και –τε. Αυτό σημαίνει ότι τα μοντέλα μας μπορούν να αναγνωρίσουν μόνο αυτά τα δύο ρήματα. Μελλοντικά, θα ήταν αρκετά ενδιαφέρον αν ενσωματώνονταν και άλλα παραδείγματα αντίστοιχων λέξεων.

Τέλος, κατά την εκπαίδευση των μοντέλων μηχανικής μάθησης, είναι πολύ σημαντικό να τροφοδοτούνται αυτά με διαφορετικά σετ δεδομένων έτσι ώστε να ‘μάθουν’ να αναγνωρίζουν όσο το δυνατόν πιο αποτελεσματικά νέα δεδομένα, όπου πλέον είναι άγνωστη η κλάση εξόδου. Θα μπορούσε, λοιπόν, το παρόν πρόβλημα να μελετηθεί πάνω και σε άλλα σύνολα δεδομένων, ώστε να εξασφαλιστεί ότι τα μοντέλα έχουν καλή ικανότητα γενίκευσης και να αποκλειστεί το ενδεχόμενο υπερεκπαίδευσης. Έχει ξαναγίνει κάτι παρόμοιο στην Ελληνική γλώσσα και συγκεκριμένα σε πτυχιακή εργασία η οποία έχει τίτλο: “Αναγνώριση μερών του λόγου σε ελληνικά κείμενα με τεχνικές μηχανικής μάθησης”, που ωστόσο έχει σκοπό την κατασκευή ενός συστήματος το οποίο αναγνωρίζει αυτόματα τα μέρη του λόγου της κάθε λέξης στο κείμενο [32]. Επιπροσθέτως, η διδακτορική διατριβή: “Μηχανική μάθηση στην επεξεργασία φυσικής γλώσσας” έχει ως σκοπό την χρήση τεχνικών μηχανικής μάθησης σε ποικίλα στάδια της επεξεργασίας φυσικής γλώσσας προκειμένου να εξαχθεί πληροφορία από το κείμενο [33].

Βιβλιογραφία

[1] T. Mitchell, «Machine Learning,» 1997.

[2] T. Mitchell, «The Discipline of Machine Learning,» αρ. School of Computer Science, Carnegie Mellon University, 2006.

[3] «Data mining,» [Ηλεκτρονικό]. Available: https://en.wikipedia.org/wiki/Data_mining.

[4] P.N. Tan, M. Steinbach, V. Kumar, «Introduction to Data Mining,» 2006.

[5] «Forbes,» [Ηλεκτρονικό]. Available: https://www.forbes.com/sites/bernardmarr/2016/09/30/what-are-the-top-10-use-cases-for-machine-learning-and-ai/#298c74b794c9.

[6] «Machine learning,» [Ηλεκτρονικό]. Available: https://en.wikipedia.org/wiki/Machine_learning.

[7] «Supervised learning,» [Ηλεκτρονικό]. Available: https://en.wikipedia.org/wiki/Supervised_learning.

[8] «Reinforcement learning,» [Ηλεκτρονικό]. Available: https://en.wikipedia.org/wiki/Reinforcement_learning.

[9] S. Kotsiantis, «Supervised Machine Learning: A Review of Classification Techniques,» *Informatica,* τόμ. 31, pp. 249-268, 2007.

[10] A.K. Jain, M.N. Murty, P.J. Flynn, «Data clustering: a review,» *ACM computing surveys (CSUR),* 1999.

[11] «Confusion matrix,» [Ηλεκτρονικό]. Available: https://en.wikipedia.org/wiki/Confusion_matrix.

[12] A.A. Lunsford, K.J. Lunsford, «Mistakes Are a Fact of Life: A National Comparative Study,» *College Composition and Communication,* τόμ. vol. 59, αρ. no. 4, pp. 781-806, 2008.

[13] Ch. Desmet, R. Balthazor, «Finding Patterns in Textual Corpora: Data Mining, Research, and Assessment in First-year Composition,» *Conference Proceedings, Computers and Writing 2005: New Writing and Computer Technologies,* pp. 1-8, 2005.

[14] M. Flor, Y. Futagi, «On Using Context for Automatic Correction of Non-word Misspellings in Student Essays,» *The 7th Workshop on the Innovative Use of NLP for Building Educational Applications, BEA-7 (at NAACL HLT 2012), Montreal, Canada,* pp. 105-115, 2012.

[15] M. Flor, «Four types of context for automatic spelling correction,» *Educational Testing Service, Rosedale Road, Princeton, NJ, 08541, USA,* 2012.

[16] R. Nagata, E. Whittaker, V. Sheinman, «Creating a Manually Error-tagged and Shallow-parsed Learner Corpus,» *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, Portland, Oregon, USA,* pp. 1210-1219, 2011.

[17] R. De Felice, S.G. Pulman, «A Classifier-Based Approach to Preposition and Determiner Error Correction in L2 English,» *In Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008),* pp. 69-176.

[18] A. Rozovskaya, M. Sammons, D. Roth, «The UI System in the HOO 2012 Shared Task on Error Correction,» *The 7th Workshop on the Innovative Use of NLP for Building Educational Applications, BEA-7, Montreal, Canada,* pp. 272-280, 2012.

[19] J.Z. Sukkarieh, J. Blackmore, «C-rater: Automatic Content Scoring for Short Constructed Responses,» *In Proceedings of the 22nd International Florida Artificial Intelligence Research Society Conference,* pp. 290-295, 2009.

[20] T. Brants, A. Franz, «Web 1T 5-gram Version 1,» *LDC2006T13, 2006, Philadelphia, PA, USA: Linguistic Data Consortium.*

[21] S. Sagiadinos, P. Gasteratos, V. Dragonas, A. Kalamara, A. Spyridonidou, K. Kermanidis, «Knowledge-poor Context-sensitive Spelling Correction for Modern Greek,» *Hellenic Conference on Artificial Intelligence, SETN 2014: Artificial Intelligence: Methods and Applications ,* pp. 360-369, 2014.

[22] RC. Angell, GE. Freund, P. Willett, «Automatic Spelling Correction Using A Trigram Similarity Measure,» *Information Processing and Management, Vol. 19, No. 4,* pp. 255-261, 1983.

[23] R.L Kashyap, B.J. Oommen, «Spelling correction using probabilistic methods,» *Pattern Recognition Letters 2,* pp. 147-154, 1984.

[24] E. Mays, FJ. Damerau, RL. Mercer, «Context Based Spelling Correction,» *Information Processing and Management Vol. 27, No. 5,* pp. 517-522, 1991.

[25] J. Schaback, F. Li, «Multi-Level Feature Extraction for Spelling Correction,» *IJCAI-07,* 2007.

[26] A. Carlson, I. Fette, «Memory-Based Context-Sensitive Spelling Correction at Web Scale,» *In Proceedings of the IEEE International Conference on Machine Learning and Applications (ICMLA) ,* 2007.

[27] AR. Golding, D. Roth, «A Winnow-Based Approach to Context-Sensitive Spelling Correction,» *Machine Learning, 34,* pp. 107-130, 1999.

[28] G. Chowdhury, «Natural language processing,» *Annual Review of Information Science and Technology,* τόμ. 37, αρ. 1, p. 3–540, 2003.

[29] E. Cambria, B. White, «Jumping NLP Curves: A Review of Natural Language Processing Research,» *IEEE Computational Intelligence Magazine,* τόμ. 9, αρ. 2, pp. 48 - 57, 2014.

[30] «Natural Language Processing,» [Ηλεκτρονικό]. Available: https://en.wikipedia.org/wiki/Natural_language_processing.

[31] W. Khan, A. Daud, J. Nasir, T. Amjad, «A survey on the state-of-the-art machine learning models in the context of NLP,» *Kuwait J. Sci. 43 (4),* pp. 95-113, 2016.

[32] http://www2.aueb.gr/users/ion/docs/tsichlas_final_report.pdf

[33] Γεώργιος Πέτρου Πετάσης, «Μηχανική Μάθηση στην Επεξεργασία Φυσικής Γλώσσας», Σχολή θετικών επιστημών, τμήμα πληροφορικής και τηλεπικοινωνιών, Αθήνα, 2011.