ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΠΛΗΡΟΦΟΡΙΚΗ ΚΑΙ ΥΠΟΛΟΓΙΣΤΙΚΗ ΒΙΟΙΑΤΡΙΚΗ

Συγκριτική Αποτίμηση Μεθόδων που Υιοθετούνται για Sentiment Analysis

Ματίκας Αθανάσιος

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Επιβλέπων

Σταμούλης Γεώργιος

Λαμία, 2020

UNIVERSITY OF THESSALY

SCHOOL OF SCIENCE

INFORMATICS AND COMPUTATIONAL BIOMEDICINE

Comparative Evaluation of Methods Adopted for Sentiment Analysis

Matikas Athanasios

Master thesis

Supervisor

Stamoulis Georgios

Lamia, 2020

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΔΙΑΤΜΗΜΑΤΙΚΟ ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΠΛΗΡΟΦΟΡΙΚΗ ΚΑΙ ΥΠΟΛΟΓΙΣΤΙΚΗ ΒΙΟΙΑΤΡΙΚΗ
ΚΑΤΕΥΘΥΝΣΗ

«ΠΛΗΡΟΦΟΡΙΚΗ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗΝ ΑΣΦΑΛΕΙΑ, ΔΙΑΧΕΙΡΙΣΗ ΜΕΓΑΛΟΥ ΟΓΚΟΥ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΠΡΟΣΟΜΟΙΩΣΗ»

Συγκριτική Αποτίμηση Μεθόδων που Υιοθετούνται για Sentiment Analysis

Ματίκας Αθανάσιος

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Επιβλέπων

Stamoulis Georgios

Λαμία, 2020

«Υπεύθυνη Δήλωση μη λογοκλοπής και ανάληψης προσωπικής ευθύνης»

Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, και γνωρίζοντας τις συνέπειες της λογοκλοπής, δηλώνω υπεύθυνα και ενυπογράφως ότι η παρούσα εργασία με τίτλο [«τίτλος εργασίας»] αποτελεί προϊόν αυστηρά προσωπικής εργασίας και όλες οι πηγές από τις οποίες χρησιμοποίησα δεδομένα, ιδέες, φράσεις, προτάσεις ή λέξεις, είτε επακριβώς (όπως υπάρχουν στο πρωτότυπο ή μεταφρασμένες) είτε με παράφραση, έχουν δηλωθεί κατάλληλα και ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή. Αναλαμβάνω πλήρως, ατομικά και προσωπικά, όλες τις νομικές και διοικητικές συνέπειες που δύναται να προκύψουν στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δεν μου ανήκει διότι είναι προϊόν λογοκλοπής.

Ο ΔΗΛΩΝ

Ματίκας Αθανάσιος

Ημερομηνία

05/02/2020

Υπογραφή

Συγκριτική Αποτίμηση Μεθόδων που Υιοθετούνται για Sentiment Analysis

Ματίκας Αθανάσιος

Τριμελής Επιτροπή:

Σταμούλης Γεώργιος

Κολομβάτσος Κωνσταντίνος

Λουκόπουλος Αθανάσιος

Επιστημονικός Σύμβουλος:

Κολομβάτσος Κωνσταντίνος

Περίληψη

Η μεγάλη ανάπτυξη του κοινωνικού ιστού τα τελευταία χρόνια παράγει τεράστιο όγκο περιεχομένου, όπως κριτικές, σχόλια και απόψεις. Αυτό το περιεχόμενο που δημιουργείται από τον χρήστη μπορεί να αφορά προϊόντα, άτομα, εκδηλώσεις κλπ. Οι πληροφορίες αυτές είναι πολύ χρήσιμες για επιχειρήσεις, κυβερνήσεις και ιδιώτες. Παρόλο που το περιεχόμενο αυτό θα ήταν χρήσιμο να αναλυθεί, το μεγαλύτερο μέρος του περιεχομένου που παράγει ο χρήστης είναι δύσκολο και χρονοβόρο στην επεξεργασία του. Υπάρχει λοιπόν η ανάγκη να αναπτυχθούν έξυπνα συστήματα τα οποία να εξορύγνουν αυτόματα αυτό το τεράστιο περιεχόμενο και να το κατηγοριοποιούν σε μία θετική, αρνητική ή ουδέτερη κατηγορία. Η ανάλυση συναισθημάτων είναι η αυτοματοποιημένη εξόρυξη στάσεων, απόψεων και συναισθημάτων από πηγές κειμένου, λόγου και βάσεων δεδομένων μέσω της επεξεργασίας φυσικής γλώσσας.

Η παρούσα διπλωματική εργασία παρουσιάζει μια συγκριτική μελέτη των τεχνικών της σε αυτόν τον τομέα.

Λέξεις Κλειδιά

Ανάλυση Συναισθήματος, Μηχανική Μάθηση, Επεξεργασίας Φυσικής Γλώσσας, Κατηγοριοποίηση.

Abstract

The increasing growth of social web on this days contributes vast amount of generated content such as reviews, comments and opinions. This user generated content can be about products, people, events, etc. This information is very useful for businesses, governments and individuals. While this content meant to be helpful to be analyzed, this bulk of user generated content is difficult and time consuming. So there is a need to develop intelligent systems which automatically mine such huge content and classify them into positive, negative and neutral category. Sentiment analysis is the automated mining of attitudes, opinions, and emotions from text, speech, and database sources through Natural Language Processing.

The aim of this thesis is to presents a comparative study of its techniques in this field.

Keywords

Sentiment Analysis, Machine Learning, Natural Language Processing, Classification.

Περιεχόμενα

Πίνακας Σχημάτων

Πίνακας Εξισώσεων

1. Εισαγωγή

1.1 Εισαγωγή στην ανάκτηση πληροφορίας (Information Retrieval - IR)

Ο όρος «ανάκτηση πληροφορίας» μπορεί να έχει ευρεία έννοια. Απλά η πληκτρολόγηση του αριθμού της πιστωτικής κάρτας για την πληρωμή ενός λογαριασμού, είναι μια μορφή ανάκτησης πληροφοριών. Ωστόσο, στον ερευνητικό τομέα, η ανάκτηση πληροφορίας μπορεί να οριστεί ως εξής:

> Η ανάκτηση πληροφορίας (IR) βρίσκει υλικό (συνήθως έγγραφα) αδόμητης φύσης (συνήθως κειμένου) από μεγάλες συλλογές (συνήθως αποθηκευμένες σε υπολογιστές) που ικανοποιεί μια ανάγκη πληροφόρησης [1].

Όπως ορίστηκε παραπάνω, η ανάκτηση πληροφοριών αποτελούσε δραστηριότητα που μόνο λίγοι άνθρωποι ασχολούνταν, όπως: βιβλιοθηκονόμοι, νομικοί βοηθοί και παρόμοιοι επαγγελματίες ερευνητές. Στις μέρες μας, εκατοντάδες εκατομμύρια άνθρωποι ασχολούνται με την ανάκτηση πληροφοριών κάθε μέρα όταν χρησιμοποιούν μια μηχανή αναζήτησης ιστού ή κάνοντας αναζήτηση στο ηλεκτρονικό τους ταχυδρομείο. Η «ανάκτηση πληροφορίας» έχει γίνει γρήγορα η κυρίαρχη μορφή πρόσβασης στις πληροφορίες, ξεπερνώντας την παραδοσιακή αναζήτηση σε σύστημα αναζήτησης δεδομένων (π.χ. αυτό που συμβαίνει όταν ένας υπάλληλος μπορεί να αναζητήσει την παραγγελία μόνο αν δοθεί το αναγνωριστικό παραγγελίας). Το IR μπορεί επίσης να καλύπτει και άλλα είδη δεδομένων και προβλημάτων πληροφόρησης πέραν αυτού που καθορίζεται στον ανωτέρω βασικό ορισμό, όπως για την πρόσβαση σε πολυμεσικά δεδομένα (video, ήχος, εικόνα) ή σε δεδομένα μέσω του παγκόσμιου ιστού. Ωστόσο για τη διπλωματική θα θεωρήσουμε τον όρο έγγραφο ισοδύναμο με ένα έγγραφο κειμένου. Ο όρος "αδόμητα δεδομένα" αναφέρεται σε δεδομένα που δεν έχουν σαφή, σημασιολογικά εμφανή και εύκολη δομή για υπολογιστή. Είναι το αντίθετο των δομημένων δεδομένων, όπως για παράδειγμα είναι μια σχεσιακή βάση δεδομένων μιας εταιρείας που τη χρησιμοποιεί για την απογραφή προϊόντων και αρχείου προσωπικού. Στην πραγματικότητα, σχεδόν κανένα στοιχείο δεν είναι πραγματικά "αδόμητο". Αν αναλογιστεί κανείς ότι το μεγαλύτερο μέρος ενός κειμένου έχει δομή, όπως επικεφαλίδες, παραγράφους και υποσημειώσεις.

Στο πεδίο της ανάκτησης πληροφορίας επίσης, είναι πολύ σημαντικό το φιλτράρισμα και ή ομαδοποίηση συλλογών εγγράφων για την περαιτέρω επεξεργασία μιας σειράς ανακτημένων εγγράφων. Δεδομένης μιας δέσμης εγγράφων, η κατηγοριοποίηση είναι το καθήκον να καταλήξουμε σε μια καλή ομαδοποίηση των εγγράφων βάσει του περιεχομένου τους. Είναι παρόμοιο με την οργάνωση βιβλίων σε μια βιβλιοθήκη ανάλογα με το θέμα τους. Δεδομένης μιας σειράς θεμάτων, σταθερών αναγκών πληροφόρησης ή άλλων κατηγοριών (όπως η καταλληλότητα των κειμένων για διαφορετικές ηλικιακές ομάδες), η κατηγοριοποίηση είναι η διεργασία να αποφασίζει ποιες τάξεις, αν υπάρχουν, ανήκουν σε κάθε ένα από τα έγγραφα. Συχνά προσεγγίζεται με την κατηγοριοποίηση ορισμένων εγγράφων με μη αυτόματο τρόπο και με την ελπίδα να είναι δυνατή η αυτόματη κατηγοριοποίηση νέων εγγράφων [1].

1.2 Εισαγωγή στην Ανάλυση Συναισθήματος (Sentiment Analysis - SA)

Η ανάλυση συναισθημάτων (Sentiment Analysis), που ονομάζεται επίσης και εξόρυξη γνώμης (opinion mining), είναι το πεδίο μελέτης που αναλύει τις απόψεις, τις αισθήσεις, τις αξιολογήσεις, τις εκτιμήσεις, τις στάσεις και τα συναισθήματα των ανθρώπων, σε προϊόντα, υπηρεσίες, οργανισμούς, άτομα, θέματα, γεγονότα και τις ιδιότητές τους. Αντιπροσωπεύει ένα μεγάλο χώρο προβλημάτων. Επίσης υπάρχουν πολλά ονόματα για πολλά ελαφρώς διαφορετικά καθήκοντα, όπως π.χ. sentiment analysis, opinion mining, opinion extraction, sentiment mining, subjectivity analysis, affect analysis, emotion analysis, review mining, κλπ. Ωστόσο, είναι όλα υπό την αιγίδα της ανάλυσης συναισθημάτων ή εξόρυξης γνώμης. Ενώ, ο όρος συναισθηματική ανάλυση χρησιμοποιείται πιο συχνά στη βιομηχανία, στον ακαδημαϊκό χώρο χρησιμοποιούνται τόσο η ανάλυση συναισθημάτων όσο και η εξόρυξη γνώμης. Βασικά αντιπροσωπεύουν τον ίδιο τομέα μελέτης. Ο όρος συναισθηματική ανάλυση ίσως εμφανίστηκε για πρώτη φορά στο (Nasukawa and Yi, 2003) και ο όρος εξόρυξη γνώμης εμφανίστηκε για πρώτη φορά στο (Dave, Lawrence and Pennock, 2003). Ωστόσο, η έρευνα σχετικά με τα συναισθήματα και τις απόψεις εμφανίστηκε νωρίτερα (Das and Chen, 2001; Morinaga et al., 2002; Pang, Lee and Vaithyanathan, 2002; Tong, 2001; Turney, 2002; Wiebe, 2000). Σε αυτή τη διπλωματική, χρησιμοποιούμε περισσότερο τον όρο "ανάλυση συναισθημάτων". Η έννοια της ίδιας της γνώμης εξακολουθεί να έχει πολύ μεγάλο εύρος. Η ανάλυση του συναισθήματος και η εξόρυξη γνώμης επικεντρώνεται κυρίως στις απόψεις που εκφράζουν ή υποδηλώνουν θετικά ή αρνητικά συναισθήματα.

Παρόλο που η γλωσσολογία και η επεξεργασία φυσικής γλώσσας (Natural Language Processing - NLP) έχουν μακρά ιστορία, λίγες έρευνες είχαν γίνει για τις απόψεις και τα συναισθήματα των ανθρώπων πριν από το 2000. Από τότε, ο τομέας έχει γίνει ένας πολύ ενεργός ερευνητικός χώρος. Υπάρχουν διάφοροι λόγοι για αυτό. Πρώτον, έχει ένα ευρύ φάσμα εφαρμογών, σχεδόν σε κάθε τομέα. Η βιομηχανία που περιβάλλει την ανάλυση συναισθημάτων έχει επίσης ακμάσει λόγω του πολλαπλασιασμού των εμπορικών εφαρμογών. Αυτό παρέχει ισχυρό κίνητρο για έρευνα. Δεύτερον, προσφέρει πολλά προκλητικά ερευνητικά προβλήματα, τα οποία δεν είχαν μελετηθεί ποτέ πριν. Τρίτον, για πρώτη φορά στην ανθρώπινη ιστορία, έχουμε ένα τεράστιο όγκο εκτιμημένων δεδομένων στα κοινωνικά μέσα στο διαδίκτυο. Χωρίς αυτά τα δεδομένα, δεν θα ήταν

δυνατή η διεξαγωγή πολλών ερευνών. Δεν αποτελεί έκπληξη το γεγονός ότι η έναρξη και η ταχεία ανάπτυξη της ανάλυσης συναισθημάτων συμπίπτουν με εκείνες των κοινωνικών μέσων δικτύωσης. Στην πραγματικότητα, η ανάλυση των συναισθημάτων είναι πλέον στο επίκεντρο της έρευνας των κοινωνικών μέσων. Ως εκ τούτου, η έρευνα στην ανάλυση του συναισθήματος δεν έχει μόνο σημαντικό αντίκτυπο στην NLP, αλλά μπορεί επίσης να έχει βαθιές επιπτώσεις στις επιστήμες της διαχείρισης (management), τις πολιτικές επιστήμες, την οικονομία και τις κοινωνικές επιστήμες, καθώς όλες επηρεάζονται από τις απόψεις των ανθρώπων [2].

2. Επεξεργασία φυσικής γλώσσας (Natural Language Processing - NLP)

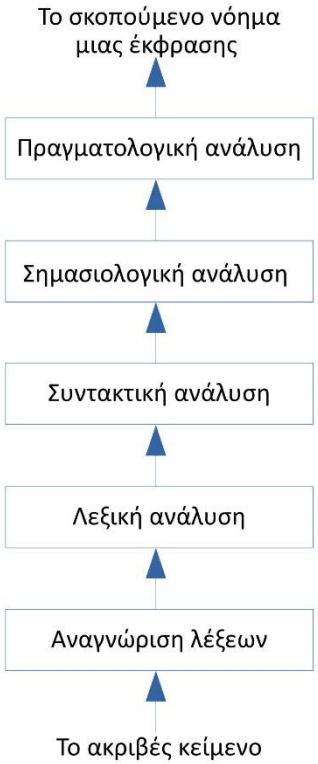
2.1 Εισαγωγή

Η ανάλυση συναισθημάτων είναι πρόβλημα NLP και επηρεάζεται από κάθε πτυχή του. Ωστόσο, είναι επίσης χρήσιμο να συνειδητοποιηθεί ότι η ανάλυση συναισθημάτων είναι ένα πολύ περιορισμένο πρόβλημα NLP επειδή το σύστημα δεν χρειάζεται να κατανοήσει πλήρως τη σημασιολογία κάθε φράσης ή εγγράφου, αλλά χρειάζεται να κατανοήσει κάποιες πτυχές του, δηλαδή θετικά ή αρνητικά συναισθήματα και τις οντότητες ή τα θέματα στόχων τους.

Παραδοσιακά, η επεξεργασία φυσικής γλώσσας έχει την τάση να βλέπει τη διαδικασία της γλωσσικής ανάλυσης ως αποσυντιθέμενη σε διάφορα στάδια, αντικατοπτρίζοντας τις θεωρητικές γλωσσικές διακρίσεις μεταξύ Σύνταξης, Σημασιολογίας και Πραγματολογίας (Syntax, Semantic και Pragmatics). Η απλή άποψη είναι ότι οι προτάσεις ενός κειμένου αναλύονται πρώτα μέσω της σύνταξής τους, έτσι παρέχει μια τάξη και μια δομή που υπάγεται περισσότερο σε μια ανάλυση όσον αφορά τη σημασιολογία, ή την κυριολεκτική έννοια και αυτό ακολουθείται από ένα στάδιο ρεαλιστικής ανάλυσης με το οποίο προσδιορίζεται η έννοια της έκφρασης ή του κειμένου. Στο τελευταίο στάδιο (Πραγματολογία) θεωρείται συχνά ότι αφορά την ομιλία (DISCOURSE), ενώ οι προηγούμενες δύο αφορούν γενικά προτασιακά θέματα. Αυτή η απόπειρα μιας συσχέτισης μεταξύ μιας διαστρωματικής διάκρισης (σύνταξη, σημασιολογία και πραγματολογία) και μιας διάκρισης ως προς τη λεπτομέρεια (φράση έναντι λόγου), προκαλεί μερικές φορές κάποια σύγχυση όσον αφορά τα ζητήματα που σχετίζονται με τη φυσική επεξεργασία των γλωσσών. Είναι ευρέως αναγνωρισμένο ότι στην πραγματικότητα δεν είναι τόσο εύκολο να χωριστεί η επεξεργασία της γλώσσας με τακτοποιημένο τρόπο σε πλαίσια που αντιστοιχούν σε κάθε ένα από τα παρακάτω στρώματα:

- Το σκοπούμενο νόημα μιας έκφρασης (Speaker’s intended meaning)
- Πραγματολογική ανάλυση(Pragmatic analysis)
- Σημασιολογική ανάλυση (Semantic analysis)
- Συντακτική ανάλυση (Syntactic analysis)

- Λεξική ανάλυση (Lexical analysis)
- Αναγνώριση λέξεων (Tokenization)
- Το ακριβές κείμενο (Surface text)



Σχήμα 2.1 Τα στάδια της ανάλυσης στην επεξεργασία φυσικής γλώσσας

Ωστόσο, ένας τέτοιος διαχωρισμός χρησιμεύει ως χρήσιμο παιδαγωγικό βοήθημα και αποτελεί τη βάση για αρχιτεκτονικά μοντέλα που καθιστούν το έργο της φυσικής ανάλυσης πιο εύχρηστο από άποψη μηχανικής λογισμικού. Παρ 'όλα αυτά, η τριμερής διάκριση σε σύνταξη, σημασιολογία και πραγματολογία χρησιμεύει, στην καλύτερη περίπτωση, ως αφετηρία όταν εξετάζεται η επεξεργασία πραγματικών κειμένων φυσικής γλώσσας. Μια λεπτομερέστερη αποσύνθεση της διαδικασίας είναι χρήσιμη όταν λαμβάνεται υπόψη η τρέχουσα κατάσταση της τεχνολογίας σε συνδυασμό με την ανάγκη αντιμετώπισης πραγματικών γλωσσικών δεδομένων. Αυτό αντικατοπτρίζεται στο Σχήμα 2.1. Αναγνωρίζεται εδώ το στάδιο του tokenization και της κατάτμησης των προτάσεων (αναγνώριση λέξεων) ως ένα κρίσιμο πρώτο βήμα. Το κείμενο της φυσικής γλώσσας δεν αποτελείται γενικά από τις σύντομες, τακτοποιημένες, καλά διαμορφωμένες και καλά οριοθετημένες προτάσεις που βρίσκουμε στα εγχειρίδια. Σε γλώσσες όπως η Κινέζικη,

τα Ιαπωνική ή η Ταϊλανδική, οι οποίες δεν χρησιμοποιούν την φαινομενικά εύκολη οριοθέτηση με κενό που μπορεί να πιστεύουμε ότι είναι γενικό χαρακτηριστικό γλωσσών όπως τα αγγλικά, η ικανότητα αντιμετώπισης των ζητημάτων του tokenization είναι απαραίτητη για να ξεκινήσει η ανάλυση. Επίσης αντιμετωπίζουμε τη λεξική ανάλυση ως ξεχωριστό βήμα στη διαδικασία. Σε κάποιο βαθμό αυτή η λεπτομερέστερη αποσύνθεση αντικατοπτρίζει την τρέχουσα κατάσταση γνώσης μας σχετικά με την επεξεργασία της γλώσσας. Γνωρίζουμε αρκετά σχετικά με τις γενικές τεχνικές για το tokenization, τη λεξική ανάλυση και τη συντακτική ανάλυση, αλλά πολύ λιγότερα για τη σημασιολογία και την επεξεργασία σε επίπεδο λόγου. Αντανακλά επίσης το γεγονός ότι το γνωστό είναι το «Surface text» και οποιαδήποτε βαθύτερη ανάλυση είναι μια αντιπροσωπευτική αφαίρεση που είναι πιο δύσκολο να αποτυπωθεί, οπότε δεν είναι τόσο περίεργο που υπάρχουν καλύτερα αναπτυγμένες τεχνικές στα τελευταία στάδια της επεξεργασίας.

Φυσικά, η επεξεργασία φυσικής γλώσσας είναι μόνο το ήμισυ της ιστορίας. Πρέπει επίσης να εξεταστεί η δημιουργία φυσικών γλωσσών, όπου ασχολείται με τη χαρτογράφηση από κάποια (συνήθως μη γλωσσική) εσωτερική αναπαράσταση σε ένα κείμενο. Στην ιστορία του πεδίου μέχρι τώρα, υπήρξε πολύ λιγότερη δουλειά στη παραγωγή φυσικής γλώσσας από ότι στην επεξεργασία της φυσικής γλώσσας. Αυτό συμβαίνει επειδή γενικά πιστεύεται ότι η παραγωγή φυσικών γλωσσών είναι ευκολότερη. Αυτό απέχει πολύ από την αλήθεια. Υπάρχουν πολλά περίπλοκα στοιχεία που πρέπει να αντιμετωπιστούν για τη δημιουργία ευέλικτων και συνεκτικών πολυεθνικών κειμένων από μια υποκείμενη πηγή πληροφοριών. Ο πιο πιθανός λόγος για τη σχετική έλλειψη εργασίας στη παραγωγή είναι ακριβώς ο συσχετισμός της παρατήρησης που έγινε στο τέλος της προηγούμενης παραγράφου: είναι σχετικά απλό να δημιουργηθούν θεωρίες γύρω από την επεξεργασία ενός γνωστού στοιχείου (όπως μια ακολουθία λέξεων) αλλά πολύ πιο δύσκολο όταν η είσοδος στη διαδικασία παραμένει περισσότερο ή λιγότερο στη φαντασία. Πολλές έρευνες στην παραγωγή φυσικής γλώσσας ασχολούνται με την αντιμετώπιση αυτών των ζητημάτων. Η εργασία στην κατανόηση της φυσικής γλώσσας μπορεί τελικά να ωφελήσει για τη δημιουργία του σημείου εκκίνησης της παραγωγής ως τελικού της στόχου [3].

2.2 Προεπεξεργασία Κειμένου (Text Preprocessing)

2.2.1 Εισαγωγή

Στο αρχικό στάδιο της ανάλυσης συναισθημάτων, δηλαδή την γλωσσολογική ανάλυση ενός ψηφιακού κειμένου φυσικής γλώσσας, είναι απαραίτητο να οριστούν σαφώς οι χαρακτήρες, οι λέξεις και οι προτάσεις. Ο καθορισμός αυτών των μονάδων παρουσιάζει διαφορετικές προκλήσεις ανάλογα με τη γλώσσα και την πηγή των εγγράφων και η εργασία δεν είναι τετριμμένη, ειδικά λαμβάνοντας υπόψη την ποικιλία των ανθρώπινων γλωσσών και των συστημάτων γραφής. Οι φυσικές γλώσσες εμπεριέχουν εγγενείς αμφισημίες, και τα συστήματα γραφής συχνά ενισχύουν τις ασάφειες και δημιουργούν πρόσθετες αμφισημίες. Μεγάλο μέρος της πρόκλησης της Επεξεργασίας Φυσικής Γλώσσας (NLP) συνεπάγεται την επίλυση αυτών των αμφιβολιών. Οι πρώιμες μελέτες στην NLP επικεντρώθηκαν σε ένα μικρό αριθμό καλά διαμορφωμένων σωμάτων κειμένων σε μικρό αριθμό γλωσσών, αλλά έχει σημειωθεί σημαντική πρόοδο τα τελευταία χρόνια χρησιμοποιώντας μεγάλα και ποικίλα σώματα κειμένων από ένα ευρύ φάσμα πηγών, συμπεριλαμβανομένου ενός τεράστιου και συνεχώς αυξανόμενου δυναμικά παραγόμενου κειμένου παρεχόμενο από το Διαδίκτυο. Αυτή η έκρηξη στο μέγεθος και την ποικιλία των κειμένων απαιτούσε τεχνικές για την αυτόματη συγκομιδή και την προετοιμασία τους για την NLP.

Σε αυτό το κεφάλαιο, παρουσιάζουμε κάποιες προκλήσεις που θέτει η προεπεξεργασία κειμένου, το καθήκον μετατροπής ενός ακατέργαστου αρχείου κειμένου, ουσιαστικά μιας ακολουθίας ψηφιακών δυαδικών ψηφίων, σε μια σαφώς καθορισμένη ακολουθία γλωσσολογικά ουσιαστικών μονάδων: στα χαμηλότερα επίπεδα οι χαρακτήρες που αντιπροσωπεύουν τα μεμονωμένα γραφήματα στο γραπτό σύστημα μιας γλώσσας, λέξεις που αποτελούνται από έναν ή περισσότερους χαρακτήρες και προτάσεις που αποτελούνται από μία ή περισσότερες λέξεις. Η προεπεξεργασία κειμένου αποτελεί ουσιαστικό μέρος της διαδικασίας της ανάλυσης συναισθημάτων, καθώς οι χαρακτήρες, οι λέξεις και οι προτάσεις που προσδιορίζονται σε αυτό το στάδιο είναι οι θεμελιώδεις μονάδες που διαβιβάζονται σε όλα τα περαιτέρω στάδια επεξεργασίας, όπως η αναγνώριση συναισθήματος, η επιλογή χαρακτηριστικών και η συναισθηματική κατηγοριοποίηση.

Η προεπεξεργασία κειμένου μπορεί να χωριστεί σε δύο στάδια: κατηγοριοποίηση εγγράφων και τμηματοποίηση κειμένου. Η κατηγοριοποίηση εγγράφων είναι η διαδικασία μετατροπής ενός συνόλου ψηφιακών αρχείων σε σαφώς καθορισμένα έγγραφα κειμένου. Για τις πρώιμες δομές κειμένων, αυτή ήταν μια αργή και χειρωνακτική διαδικασία, ενώ αυτές οι πρώιμες δομές σπάνια ήταν περισσότερες από μερικές εκατομμύρια λέξεις.

Αντίθετα, οι σημερινές δομές κειμένων που συλλέγονται από το Διαδίκτυο μπορούν να περιλαμβάνουν δισεκατομμύρια λέξεις την ημέρα, πράγμα που απαιτεί μια πλήρως αυτοματοποιημένη διαδικασία κατηγοριοποίησης εγγράφων. Αυτή η διαδικασία μπορεί να περιλαμβάνει διάφορα βήματα, ανάλογα με την προέλευση των προς επεξεργασία αρχείων. Πρώτον, για να είναι αναγνώσιμο από μια μηχανή το κείμενο φυσικής γλώσσας, πρέπει να χρησιμοποιείται μια κωδικοποίηση χαρακτήρων, στην οποία ένα ή περισσότερα bytes σε ένα αρχείο να αντιστοιχούν σε ένα γνωστό χαρακτήρα. Δεύτερον, προκειμένου να γνωρίζουμε ποιοι γλωσσικοί αλγόριθμοι πρέπει να εφαρμοστούν σε ένα έγγραφο, η αναγνώριση γλώσσας καθορίζει τη φυσική γλώσσα ενός εγγράφου. Αυτό το βήμα συνδέεται στενά με την κωδικοποίηση χαρακτήρων, αλλά όχι με μοναδικό τρόπο. Τέλος, ο τμηματικός χωρισμός του κειμένου προσδιορίζει το πραγματικό περιεχόμενο μέσα σε ένα αρχείο ενώ απορρίπτει ανεπιθύμητα στοιχεία, όπως εικόνες, πίνακες, κεφαλίδες, συνδέσμους και μορφοποίηση HTML. Η έξοδος του σταδίου της διαλογής των εγγράφων είναι ένα καλά καθορισμένο σώμα κειμένου, οργανωμένο ανά γλώσσα, κατάλληλο για κατακερματισμό κειμένου και περαιτέρω ανάλυση.

Η κατάτμηση κειμένου (*Text segmentation*) είναι η διαδικασία μετατροπής μιας σαφώς καθορισμένης δομής κειμένου στις συνθετικές λέξεις και προτάσεις του. Η κατάτμηση λέξεων (*Word segmentation*) διαχωρίζει την ακολουθία των χαρακτήρων σε ένα κείμενο, εντοπίζοντας τα όρια των λέξεων, τα σημεία δηλαδή όπου μια λέξη τελειώνει και αρχίζει η επόμενη. Για λόγους υπολογιστικής γλωσσολογίας, οι λέξεις που προσδιορίζονται με τον τρόπο αυτό αναφέρονται συχνά ως «tokens» και η κατάτμηση των λέξεων είναι γνωστή ως «tokenization». Η κανονικοποίηση του κειμένου (*Text normalization***)** είναι ένα σχετικό βήμα που συνεπάγεται τη συγχώνευση διαφορετικών μορφών γραφής ενός συμβόλου σε κανονική ομαλοποιημένη μορφή. Για παράδειγμα, ένα έγγραφο μπορεί να περιέχει τα ισοδύναμα αναγνωριστικά: "κ.", "Κος", "κύριος" και "Κύριος", τα οποία όλα θα ομαλοποιηθούν σε μια ενιαία μορφή. Η κατάτμηση φράσεων (*Sentence segmentation*) είναι η διαδικασία προσδιορισμού των μεγαλύτερων μονάδων επεξεργασίας που

αποτελούνται από μία ή περισσότερες λέξεις. Αυτή η εργασία περιλαμβάνει τον προσδιορισμό των ορίων των φράσεων μεταξύ λέξεων που ανήκουν σε διαφορετικές προτάσεις.

Στην πράξη, η κατάτμηση φράσεων και λέξεων δεν μπορεί να εκτελεστεί με επιτυχία ανεξάρτητα το ένα από το άλλο. Για παράδειγμα, μια σημαντική υποεργασία κατάτμησης λέξεων και φράσεων για τις περισσότερες ευρωπαϊκές γλώσσες είναι η ταυτοποίηση των συντομογραφιών, επειδή μια περίοδος μπορεί να χρησιμοποιηθεί για να σηματοδοτήσει μια συντομογραφία καθώς και για να σημάνει το τέλος μιας φράσης. Στην περίπτωση μιας περιόδου που χαρακτηρίζει μια συντομογραφία, η περίοδος θεωρείται συνήθως μέρος του συμβόλου συντομογραφίας, ενώ μια περίοδος στο τέλος μιας φράσης θεωρείται συνήθως ως token από μόνη της. Σε περίπτωση σύντμησης στο τέλος μιας φράσης, η περίοδος σηματοδοτεί τόσο τη συντομογραφία όσο και το όριο της φράσης [3].

2.2.2 Προκλήσεις της προεπεξεργασίας κειμένων

Υπάρχουν πολλά προβλήματα που προκύπτουν στην προεπεξεργασία κειμένου που πρέπει να αντιμετωπιστούν κατά το σχεδιασμό συστήματος NLP και πολλά μπορούν να αντιμετωπιστούν ως μέρος της κατηγοριοποίησης εγγράφων για την προετοιμασία μίας δομής κειμένου για την ανάλυση συναισθήματος.

Ο τύπος του συστήματος γραφής που χρησιμοποιείται για μια γλώσσα είναι ο πιο σημαντικός παράγοντας για τον προσδιορισμό της καλύτερης προσέγγισης στην προεπεξεργασία κειμένου. Τα συστήματα γραφής μπορούν να είναι λογογραφικά, όπου ένας μεγάλος αριθμός (συχνά χιλιάδες) μεμονωμένων συμβόλων αντιπροσωπεύει λέξεις. Αντίθετα, τα συστήματα γραφής μπορούν να είναι συλλαβικά, στα οποία τα μεμονωμένα σύμβολα αντιπροσωπεύουν συλλαβές, ή αλφαβητικά στα οποία μεμονωμένα σύμβολα (περισσότερο ή λιγότερο) αντιπροσωπεύουν ήχους. Αντίθετα με τα λογογραφικά συστήματα, τα συλλαβικά και αλφαβητικά συστήματα έχουν συνήθως λιγότερα από 100 σύμβολα. Σύμφωνα με τους Comrie et al. (1996), οι περισσότερες από τις γραπτές γλώσσες χρησιμοποιούν ένα αλφαβητικό ή συλλαβικό σύστημα. Ωστόσο, στην πράξη, κανένα σύγχρονο σύστημα γραφής δεν χρησιμοποιεί σύμβολα μόνο ενός είδους, επομένως κανένα σύστημα γραφής φυσικής γλώσσας δεν μπορεί να κατηγοριοποιηθεί ως καθαρά λογογραφικό, συλλαβικό ή αλφαβητικό. Ακόμη και τα αγγλικά, με το σχετικά απλό σύστημα γραφής που βασίζεται στο ρωμαϊκό αλφάβητο, χρησιμοποιούν λογοτεχνικά σύμβολα που περιλαμβάνουν αραβικούς αριθμούς (0-9), σύμβολα νομισμάτων ($, £) και άλλα σύμβολα (%, &, #). Ωστόσο, τα αγγλικά είναι κατά κύριο λόγο αλφαβητικά όπως και τα περισσότερα άλλα συστήματα γραφής και αποτελούνται από σύμβολα που είναι κυρίως ενός τύπου [3].

2.2.3 Σύνολα χαρακτήρων

Στο χαμηλότερο επίπεδο, ένα κείμενο ή έγγραφο βασισμένο σε υπολογιστή είναι απλά μια ακολουθία δυαδικών ψηφίων σε ένα αρχείο. Το πρώτο βασικό καθήκον είναι η ερμηνεία αυτών των δυαδικών ψηφίων ως χαρακτήρων ενός συστήματος γραφής μιας φυσικής γλώσσας.

Ιστορικά, η ερμηνεία των ψηφιακών αρχείων κειμένου ήταν ασήμαντη, καθώς σχεδόν όλα τα κείμενα κωδικοποιήθηκαν στο σύνολο χαρακτήρων των 7 bit ASCII, το οποίο επέτρεπε μόνο 128 χαρακτήρες και περιλάμβανε μόνο το αλφάβητο της ρωμαϊκής (ή λατινικής) γλώσσας και χαρακτήρες για τη συγγραφή αγγλικών. Αυτός ο περιορισμός απαιτούσε τη "λατινοποίηση" πολλών κειμένων, όπου καθορίστηκαν ισοδύναμα ASCII για χαρακτήρες που δεν καθοριζόταν στο σύνολο των χαρακτήρων. Ένα παράδειγμα αυτής της τεχνικής είναι η προσαρμογή πολλών ευρωπαϊκών γλωσσών που περιέχουν διαλυτικά και τόνους, όπου τα διαλυτικά αντικαθίστανται από ένα διπλό εισαγωγικό ή το γράμμα «e» και οι τόνοι σημειώνονται με ένα μόνο εισαγωγικό ή ακόμη και με έναν αριθμό. Στο σύστημα αυτό, η γερμανική λέξη über θα γράφεται ως u”ber και η γαλλική λέξη déjà θα γράφεται ως de'ja’ ή de1ja2. Οι γλώσσες που δεν χρησιμοποιούν το λατινικό αλφάβητο, όπως το ρωσικό και το αραβικό, απαιτούσαν πολύ πιο περίτεχνα συστήματα λατινοποίησης, συνήθως βασισμένα σε μια φωνητική χαρτογράφηση των χαρακτήρων πηγής στους λατινικούς χαρακτήρες. Το σύστημα λατινοποίησης Pinyin της κινεζικής γραφής είναι ένα άλλο παράδειγμα της συσσώρευσης ενός πιο περίπλοκου συστήματος γραφής. Αυτές οι προσαρμογές είναι ακόμα κοινές λόγω της ευρείας οικειότητας με τους λατινικούς χαρακτήρες. Επιπλέον, ορισμένες εφαρμογές ηλεκτρονικών υπολογιστών εξακολουθούν να περιορίζονται σε αυτήν την κωδικοποίηση των 7-bit [3].

2.3 Λεξικολογική Ανάλυση (Lexical Analysis)

Το επίπεδο της λεξικολογικής ανάλυσης αφορά τις τεχνικές και το μηχανισμό για την ανάλυση κειμένου στο επίπεδο της λέξης, τη λεξική ανάλυση. Μια λέξη μπορεί να θεωρηθεί με δύο τρόπους, είτε ως μια συμβολοσειρά στο τρέχον κείμενο, για παράδειγμα, το ρήμα «τρέχει» είτε ως ένα πιο αφηρημένο αντικείμενο που έχει ρίζα βάσης (ριζικές ή πρωτότυπες λέξεις) ένα σύνολο συμβολοσειρών. Έτσι, το ρήμα «τρέχω» είναι η ριζική λέξη για το σύνολο των λέξεων {τρέχει, έτρεχα, τρέχοντας}. Ένα βασικό καθήκον της λεξικής ανάλυσης είναι να συσχετίσει τις μορφολογικές παραλλαγές με το λήμμα τους που βρίσκεται σε ένα λεξικό, το οποίο έχει συσσωρευμένες τις αμετάβλητες σημασιολογικές και συντακτικές του πληροφορίες. Η λημματοποίηση χρησιμοποιείται με διαφορετικούς τρόπους ανάλογα με το έργο του συστήματος επεξεργασίας φυσικής γλώσσας (NLP).

Μπορούμε να σκεφτούμε τη χαρτογράφηση της συμβολοσειράς με το λήμμα ως μόνο μία πλευρά της λεξικής ανάλυσης, της πλευράς της ανάλυσης. Η άλλη πλευρά χαρτογραφεί από το λήμμα σε μια συμβολοσειρά, μορφολογική παραγωγή. Στην ανάκτηση πληροφορίας (IR), η ανάλυση και η παραγωγή εξυπηρετούν διαφορετικούς σκοπούς. Για την αυτόματη δημιουργία μιας λίστας βασικών όρων, είναι λογικό να συνδέονται μορφολογικές παραλλαγές κάτω από ένα λήμμα. Αυτό επιτυγχάνεται στην πράξη κατά τη διάρκεια του *stemming*, μιας διαδικασίας επεξεργασίας κειμένου όπου εντοπίζονται μορφολογικά πολύπλοκες συμβολοσειρές, αποσυντίθενται σε ριζική λέξη (= κανονική μορφή του λήμματος) και επίθεμα και στη συνέχεια τα επιθέματα διαγράφονται. Το αποτέλεσμα είναι τα κείμενα ως αντικείμενα αναζήτησης που αποτελούνται από ριζικές λέξεις μόνο, μπορούν να αναζητηθούν μέσω λίστας λημμάτων. Η μορφολογική παραγωγή παίζει επίσης ρόλο στο IR, όχι στο στάδιο προεπεξεργασίας, αλλά ως μέρος της αντιστοίχισης ερωτήματος. Δεδομένου ότι ένα λήμμα έχει αμετάβλητη σημασιολογία, η εύρεση μιας μορφολογικής παραλλαγής του ικανοποιεί τις σημασιολογικές απαιτήσεις μιας αναζήτησης. Τα ορθογραφικά λεξικά κάνουν επίσης χρήση της μορφολογικής παραγωγής για τον ίδιο λόγο, για να λογαριάσουν τόσο τις εισηγμένες όσο και τις «δυνητικές» λέξεις [3].

2.4 Συντακτική ανάλυση (Syntactic Parsing)

Η συντακτική ανάλυση είναι το καθήκον της αναγνώρισης μιας πρότασης και της ανάθεσης μιας συντακτικής δομής. Στις περισσότερες περιπτώσεις, αυτό δεν αποτελεί στόχο από μόνο του αλλά μάλλον ένα ενδιάμεσο βήμα για το σκοπό της περαιτέρω επεξεργασίας, όπως η εκχώρηση ενός νοήματος στην πρόταση. Η δομική ασάφεια είναι ίσως το πιο σοβαρό πρόβλημα που αντιμετωπίζουν οι συντακτικοί αναλυτές που προκύπτει από πολλούς κανόνες που χρησιμοποιούνται στις γραμματικές φράσης-δομής [3].

2.5 Σημασιολογική Ανάλυση (Semantic Analysis)

Γενικά στη γλωσσολογία, η σημασιολογική ανάλυση αναφέρεται στην ανάλυση των εννοιών των λέξεων, των σταθερών εκφράσεων, των ολόκληρων προτάσεων και των δηλώσεων στο περιεχόμενο. Στην πράξη, αυτό σημαίνει μετάφραση των πρωτότυπων εκφράσεων σε κάποιο είδος σημασιολογικής μετάφρασης. Επομένως, τα κύρια θεωρητικά ζητήματα της σημασιολογικής ανάλυσης αναδεικνύουν τη φύση της μεταγλώσσας ή ισοδύναμου αντιπροσωπευτικού συστήματος.

Ο σκοπός της σημασιολογικής ανάλυσης είναι να εξαγάγει την ακριβή έννοια από το κείμενο. Το έργο του σημασιολογικού αναλυτή είναι να ελέγξει το κείμενο για νόημα.

Η λεξικολογική ανάλυση ασχολείται επίσης με την έννοια των λέξεων, αλλά βασίζεται σε μικρότερο κομμάτι - λήμμα, από την άλλη πλευρά η σημασιολογική ανάλυση επικεντρώνεται σε μεγαλύτερα κομμάτια. Αυτός είναι ο λόγος για τον οποίο η σημασιολογική ανάλυση μπορεί να χωριστεί στα ακόλουθα δύο μέρη.

- Μελέτη της έννοιας μίας μεμονωμένης λέξης: Είναι το πρώτο μέρος της σημασιολογικής ανάλυσης στην οποία εκτελείται η μελέτη της σημασίας των επιμέρους λέξεων. Αυτό το μέρος ονομάζεται λεξικολογική σημασιολογία.
- Μελέτη συνδυασμών μεμονωμένων λέξεων: Στο δεύτερο μέρος, οι μεμονωμένες λέξεις θα συνδυαστούν για να δώσουν νόημα σε προτάσεις.

Το πιο σημαντικό καθήκον της σημασιολογικής ανάλυσης είναι να αποκτηθεί η σωστή έννοια της φράσης. Για παράδειγμα, στη φράση «τον έπιασε στα πράσα», ο ομιλητής μιλάει είτε για ένα συγκεκριμένο μέρος του πιάστηκε κάποιος είτε για επ’ αυτοφώρω σύλληψη. Αυτός είναι ο λόγος που η σημασιολογική ανάλυση είναι σημαντική [3].

2.6 Πραγματολογική Ανάλυση (Pragmatic Analysis)

Στην πραγματολογική ανάλυση μελετάται η σημασία του κειμένου όπως αυτό χρησιμοποιείται για επικοινωνία από τον γράφοντα και το πώς ερμηνεύεται από τον αναγνώστη. Αυτό σημαίνει ότι έχει να κάνει περισσότερο με αυτό που εννοούν οι άνθρωποι με τις λέξεις και τις φράσεις που χρησιμοποιούν παρά με το τι μπορεί να σημαίνουν αυτές. Σε αυτή την προσέγγιση εξετάζεται και το περιβάλλον στο οποίο γίνεται η ανάλυση, όπως ο χώρος, ο χρόνος, η κοινωνική κατάσταση, κτλ. Η πραγματολογία είναι ο μόνος τύπος ανάλυσης που επιτρέπει να μπει ο ανθρώπινος παράγοντας στην ανάλυση. Το πλεονέκτημα είναι ότι αναζητείται η επιδιωκόμενη σημασία του κειμένου από τον γράφοντα, ενώ το μεγαλύτερο μειονέκτημα είναι η δυσκολία να αναλυθούν με συνεπή και αντικειμενικό τρόπο οι ανθρώπινες έννοιες. Το παρακάτω παράδειγμα είναι μία τέτοια προβληματική περίπτωση.

Ομιλητής 1: Τελικά… πήγες;

Ομιλητής 2: Έλα τώρα… και ποιος δεν θα πήγαινε;

Η πραγματολογική ανάλυση είναι ένα από τα πιο δύσκολα στάδια στην γλωσσική ανάλυση και την ανάλυση συναισθημάτων καθώς το σύστημα πρέπει να κατανοήσει τους ανθρώπους και αυτό που έχουν στο μυαλό τους [3].

3. Μηχανική Μάθηση (Machine Learning - ML)

Η μηχανική μάθηση είναι μια εφαρμογή της Τεχνητής Νοημοσύνης (AI) που παρέχει στα συστήματα τη δυνατότητα να μαθαίνουν και να βελτιώνονται αυτόματα από την εμπειρία χωρίς να έχουν προγραμματιστεί ρητά. Η μηχανική μάθηση επικεντρώνεται στην ανάπτυξη προγραμμάτων που μπορούν να έχουν πρόσβαση σε δεδομένα και να τα χρησιμοποιούν για να μάθουν.

Σύμφωνα με αυτή την τεχνική, υπάρχουν δύο σύνολα δεδομένων, δηλαδή ένα σύνολο εκπαίδευσης και ένα σύνολο δοκιμών. Γενικά, τα σύνολα δεδομένων που συλλέγονται από διαφορετικές πηγές και των οποίων η συμπεριφορά και οι τιμές εξόδου είναι γνωστές σε εμάς εμπίπτει στην κατηγορία των συνόλων δεδομένων εκπαίδευσης. Σε αντίθεση με αυτό, τα σύνολα δεδομένων των οποίων οι τιμές ή η συμπεριφορά είναι άγνωστες σε εμάς καλούνται ως σύνολα δεδομένων δοκιμής. Για την ανάλυση συναισθήματος τα διάφορα μοντέλα εκπαιδεύονται με δεδομένα εκπαίδευσης και στη συνέχεια άγνωστα δεδομένα ώσπου για να επιτευχθούν τα επιθυμητά αποτελέσματα [4].

3.1 Κατηγορίες μηχανικής μάθησης

3.1.1 Εποπτευόμενη μηχανική μάθηση

Η πλειονότητα των εφαρμογών της μηχανικής μάθησης χρησιμοποιεί εποπτευόμενη μάθηση. Η εποπτευόμενη μάθηση χρησιμοποιεί μεταβλητές εισόδου (X) και μεταβλητές εξόδου (Y) και έναν αλγόριθμο για να προσδιοριστεί η συσχέτιση μεταξύ της εισόδου και της εξόδου.

$$Y = f(X) \qquad (3.1)$$

Ο στόχος είναι να προσεγγιστεί η λειτουργία της συσχέτισης τόσο καλά, ώστε όταν εισαχθούν νέα δεδομένα εισόδου (X) να μπορούν να προβλεφθούν οι μεταβλητές εξόδου (Y) για αυτά τα δεδομένα. Ονομάζεται εποπτευόμενη μάθηση επειδή η διαδικασία ενός αλγορίθμου που μαθαίνει από το σύνολο δεδομένων εκπαίδευσης μπορεί να θεωρηθεί ως δάσκαλος που επιβλέπει τη διαδικασία μάθησης. Γνωρίζοντας τις σωστές απαντήσεις, ο αλγόριθμος κάνει επανειλημμένα προβλέψεις για τα δεδομένα εκπαίδευσης και διορθώνεται από τον «δάσκαλο». Η μάθηση σταματά όταν ο αλγόριθμος επιτυγχάνει ένα αποδεκτό επίπεδο απόδοσης. Τα προβλήματα εποπτευόμενης μάθησης μπορούν να ομαδοποιηθούν περαιτέρω σε προβλήματα παλινδρόμησης και κατηγοριοποίησης.

- Κατηγοριοποίηση (Classification): Ένα πρόβλημα κατηγοριοποίησης είναι όταν η μεταβλητή εξόδου είναι μια κατηγορία, όπως κόκκινο ή μπλε, υγιείς ή μη υγιείς.
- Παλινδρόμηση (Regression): Ένα πρόβλημα παλινδρόμησης είναι όταν η μεταβλητή εξόδου είναι μια πραγματική τιμή, όπως δολάρια ή βάρος.

Μερικά δημοφιλή παραδείγματα εποπτευόμενων αλγορίθμων μηχανικής μάθησης είναι:

- Γραμμική παλινδρόμηση για προβλήματα παλινδρόμησης.
- Τυχαία δάση (Random forest) για προβλήματα κατηγοριοποίησης και παλινδρόμησης.
- Μηχανές Διανυσμάτων Υποστήριξης (SVM) για προβλήματα κατηγοριοποίησης [5][6].

3.1.2 Μη εποπτευόμενη μηχανική μάθηση

Η μη εποπτευόμενη μάθηση είναι όπου υπάρχουν μόνο δεδομένα εισόδου (X) και καμία αντίστοιχη μεταβλητή εξόδου. Ο στόχος για την μη εποπτευόμενη μάθηση είναι να μοντελοποιήσει την βασική δομή ή τη διανομή στα δεδομένα, προκειμένου να μάθει περισσότερα για τα δεδομένα. Ονομάζεται μη εποπτευόμενη μάθηση, διότι σε αντίθεση με την εποπτευόμενη παραπάνω δεν υπάρχουν σωστές απαντήσεις και δεν υπάρχει «δάσκαλος». Οι αλγόριθμοι αφήνονται στις δικές τους επινοήσεις για να ανακαλύψουν και να παρουσιάσουν την δομή των δεδομένων. Τα προβλήματα μη εποπτευόμενης μάθησης μπορούν να ομαδοποιηθούν περαιτέρω σε προβλήματα ομαδοποίησης και συσχέτισης.

- Ομαδοποίηση (Clustering): Ένα πρόβλημα ομαδοποίησης είναι όπου υπάρχει ανάγκη να ανακαλυφθούν οι εγγενείς ομαδοποιήσεις των δεδομένων, όπως ομαδοποίηση πελατών σύμφωνα με την αγοραστική συμπεριφορά τους.
- Συσχέτιση (Association): Ένα πρόβλημα μάθησης κατά κανόνα συσχέτισης είναι όπου υπάρχουν κανόνες που περιγράφουν μεγάλα τμήματα των δεδομένων, όπως τα άτομα που αγοράζουν ένα προϊόν Α έχουν επίσης την τάση να αγοράζουν και το Β.

Ορισμένα δημοφιλή παραδείγματα αλγόριθμων μάθησης χωρίς επίβλεψη είναι:

- k-μέσα (k-means) *MacQueen, J. B. (1967)* για προβλήματα ομαδοποίησης.
- Αλγόριθμος Apriori για προβλήματα μάθησης κατά κανόνα συσχέτισης [5][6].

3.1.3 Ημι-εποπτευόμενη μηχανική μάθηση

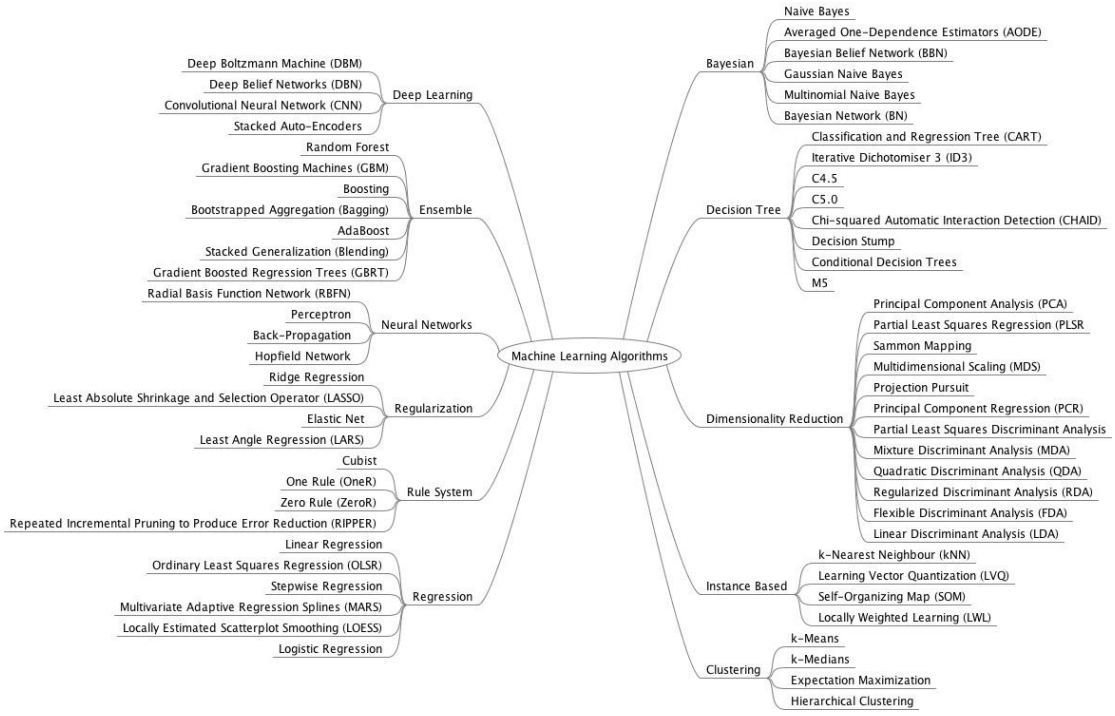
Μεταξύ της εποπτευόμενης και της μη εποπτευόμενης μάθησης είναι η ημι-εποπτευόμενη μάθηση, όπου οι αλγόριθμοι λειτουργούν με σύνολο εκπαίδευσης μέσα στο οποίο υπάρχουν παραδείγματα με μη γνωστές εξόδους. Ένα καλό παράδειγμα είναι ένα αρχείο φωτογραφιών όπου μόνο μερικές από αυτές επισημαίνονται (π.χ. σκύλος, γάτα, πρόσωπο) και η πλειοψηφία δεν φέρει ετικέτα. Πολλά προβλήματα μηχανικής μάθησης στον πραγματικό κόσμο εμπίπτουν σε αυτήν την περιοχή. Αυτό οφείλεται στο γεγονός ότι μπορεί να είναι δαπανηρή ή χρονοβόρα η επισήμανση των δεδομένων, καθώς μπορεί να απαιτεί πρόσβαση σε ειδικούς τομείς. Ενώ τα μη επισημασμένα δεδομένα είναι φθηνά και εύκολα συλλέγονται και αποθηκεύονται.

Μοντέλα ημι-εποπτευόμενης μάθησης που είναι ουσιαστικά μια παραλλαγή των μοντέλων εποπτευόμενης μάθησης είναι κατάλληλα για προβλήματα πρόβλεψης. Εδώ μπορούν να χρησιμοποιηθούν τεχνικές εποπτευόμενης μάθησης για να γίνουν οι καλύτερες δυνατές προβλέψεις για τα μη επισημασμένα δεδομένα, που τροφοδοτούνται πάλι πίσω στον αλγόριθμο εποπτευόμενης μάθησης ως δεδομένα εκπαίδευσης, ώστε να χρησιμοποιηθεί το μοντέλο για να κάνει προβλέψεις με νέα άγνωστα δεδομένα [5][6].

3.2 Μέθοδοι Μηχανικής Μάθησης

Η μηχανική μάθηση στηρίζεται στους αλγορίθμους της μηχανικής μάθησης, ώστε να αντιμετωπίσει και να επιλύσει τα προβλήματα της ανάλυσης συναισθήματος. Στην προσέγγιση αυτή ένα πρόβλημα συναισθηματικής ανάλυσης αντιμετωπίζεται ως ένα απλό πρόβλημα κατηγοριοποίησης κειμένου, κάνοντας χρήση των συντακτικών και γλωσσικών του χαρακτηριστικών.

Το σχήμα 3.1 παρουσιάζει μια λεπτομερή κατηγοριοποίηση των υπαρχουσών μεθόδων μηχανικής μάθησης, τις βασικότερες από τις οποίες χρησιμοποιούνται στην ανάλυση συναισθημάτων όπου θα αναλύσουμε σε επόμενο κεφάλαιο. Αυτή η κατηγοριοποίηση δεν είναι αποκλειστική. Μια λύση μπορεί να χωρέσει σε περισσότερες από μία κατηγορίες.



Σχήμα 3.1: Μέθοδοι μηχανικής μάθησης. datasciencecentral (2015).

Μια περαιτέρω κατηγοριοποίηση των αλγόριθμων μηχανικής μάθησης, είναι οι παραμετρικοί και οι μη παραμετρικοί αλγόριθμοι.

3.2.1 Παραμετρικοί Αλγόριθμοι Μηχανικής Μάθησης

Οι παράμετροι μπορούν να απλοποιήσουν σε μεγάλο βαθμό τη διαδικασία μάθησης, αλλά μπορούν επίσης και να περιορίσουν το τι μπορεί να μάθει. Αλγόριθμοι που απλοποιούν τη λειτουργία σε μια γνωστή μορφή ονομάζονται αλγόριθμοι παραμετρικής μηχανικής μάθησης. Οι αλγόριθμοι αυτοί κάνουν χρήση αλγεβρικών εξισώσεων για να προσδιοριστεί η συσχέτιση που υπάρχει μεταξύ της εισόδου και της εξόδου. Οι εξισώσεις αυτές χρησιμοποιούν παραμέτρους οι οποίες είναι απροσδιόριστες και εκτιμώνται μέσω εκπαιδευτικών παραδειγμάτων που δίνονται στην είσοδο.

Οι αλγόριθμοι περιλαμβάνουν δύο βήματα:

1. Επιλογή της μορφής της εξίσωσης.

2. Εκμάθηση των συντελεστών από τα δεδομένα εκπαίδευσης.

Μια εύκολη στην κατανόηση μορφή εξίσωσης για τη λειτουργία χαρτογράφησης είναι μια γραμμή, όπως χρησιμοποιείται στη γραμμική οπισθοδρόμηση:

$$B0 + B1 \times X1 + B2 \times X2 = 0 \qquad (3.2)$$

Όπου οι B0, B1 και B2 είναι οι συντελεστές της γραμμής που ελέγχει τη τομή και την κλίση, και οι X1 και X2 είναι δύο μεταβλητές εισόδου. Υποθέτοντας τη μορφή της εξίσωσης ως μιας γραμμής, απλουστεύεται πολύ η διαδικασία μάθησης. Επομένως, το μόνο που απομένει είναι να υπολογιστούν οι συντελεστές της εξίσωσης γραμμής και δημιουργείται ένα πρότυπο για το πρόβλημα.

Συχνά η θεωρητική μορφή μιας εξίσωσης είναι ένας γραμμικός συνδυασμός των μεταβλητών εισόδου και ως εκ τούτου οι αλγόριθμοι παραμετρικής μηχανικής μάθησης συχνά ονομάζονται αλγόριθμοι γραμμικής μηχανικής μάθησης. Το πρόβλημα είναι ότι η πραγματική άγνωστη εξίσωση μπορεί να μην είναι μια γραμμική συνάρτηση όπως μια γραμμή. Εδώ υπάρχουν δυο περιπτώσεις, όπου στην πρώτη, θα μπορούσε να είναι σχεδόν μια γραμμή και απαιτείται κάποια μικρή μετατροπή των δεδομένων εισόδου για να λειτουργήσει σωστά, ή θα μπορούσε να μην έχει καμία σχέση με μια γραμμή, στην οποία περίπτωση η παραδοχή είναι λάθος και η προσέγγιση θα οδηγήσει σε λάθος αποτελέσματα.

Ορισμένα παραδείγματα παραμετρικών αλγορίθμων μηχανικής μάθησης είναι:

- Λογιστική παλινδρόμηση
- Γραμμική ανάλυση διακρίσεων
- Ο νευρώνας Perceptron ή Αντίληπτρο

Οφέλη των παραμετρικών αλγορίθμων μηχανικής μάθησης:

- Απλούστεροι: Είναι πιο εύκολο να κατανοηθούν και να ερμηνεύουν τα αποτελέσματα.
- Ταχύτητα: Τα παραμετρικά μοντέλα είναι πολύ γρήγορα για να μάθουν από τα δεδομένα.
- Λιγότερα δεδομένα: Δεν απαιτούν πολλά δεδομένα εκπαίδευσης και μπορούν να λειτουργήσουν καλά ακόμα και αν η προσαρμογή στα δεδομένα δεν είναι τέλεια.

Περιορισμοί των παραμετρικών αλγορίθμων μηχανικής μάθησης:

- Περιορισμένοι: Επιλέγοντας μια λειτουργική μορφή, αυτές οι μέθοδοι είναι εξαιρετικά περιορισμένες στην καθορισμένη μορφή.
- Περιορισμένη πολυπλοκότητα: Οι μέθοδοι είναι πιο κατάλληλες για απλούστερα προβλήματα.
- Κακή προσαρμογή: Στην πράξη, οι μέθοδοι είναι απίθανο να ταιριάζουν με τη βασική λειτουργία χαρτογράφησης.

3.2.2 Μη Παραμετρικοί Αλγόριθμοι Μηχανικής Μάθησης

Αλγόριθμοι που δεν χρησιμοποιούν παραμέτρους για να προσδιορίσουν τη μορφή της λειτουργίας χαρτογράφησης ονομάζονται μη παραμετρικοί αλγόριθμοι μηχανικής μάθησης. Το μοντέλο δεν χρησιμοποιεί εξισώσεις αλλά προσαρμόζεται από τα δεδομένα εκπαίδευσης. Οι μη παραμετρικές τεχνικές είναι καλύτερες όταν υπάρχουν πολλά δεδομένα και δεν υπάρχει προηγούμενη γνώση και είναι περισσότερο κατάλληλες για εφαρμογές της μηχανικής μάθησης.

Οι μη παραμετρικές μέθοδοι αποσκοπούν στην καλύτερη προσαρμογή των δεδομένων εκπαίδευσης κατά την κατασκευή του μοντέλου, διατηρώντας παράλληλα κάποια ικανότητα γενίκευσης σε δεδομένα που δεν βλέπουν. Ως εκ τούτου, έχουν τη δυνατότητα δυναμικής εκμάθησης με την είσοδο νέων δεδομένων. Ένα εύκολο στην κατανόηση μη παραμετρικό μοντέλο είναι ο αλγόριθμος κ-πλησιέστερων γειτόνων (k-nearest neighbors) που κάνει προβλέψεις με βάση γνωστές παρατηρήσεις του συνόλου εκπαίδευσης για μια νέα είσοδο δεδομένων. Ορισμένα παραδείγματα δημοφιλών μη παραμετρικών αλγορίθμων μηχανικής μάθησης είναι:

- Δέντρα Αποφάσεων όπως το CART
- Naive Bayes
- Support Vector Machines
- Νευρωνικά δίκτυα

Οφέλη των μη παραμετρικών αλγορίθμων μηχανικής μάθησης:

- Ευελιξία: Δυνατότητα διεύρυνσης του μοντέλου με την είσοδο νέων δεδομένων.
- Απόδοση: Μπορεί να οδηγήσει σε υψηλότερης απόδοσης μοντέλα πρόβλεψης.

Περιορισμοί των μη παραμετρικών αλγορίθμων μηχανικής μάθησης:

- Περισσότερα δεδομένα: Απαιτούνται πολύ περισσότερα δεδομένα εκπαίδευσης για την δημιουργία του μοντέλου.
- Ταχύτητα: Μεγαλύτερος χρόνος εκπαίδευσης, δεδομένου ότι συχνά χρειάζονται πολλά περισσότερα δεδομένα για την εκπαίδευση.

- Overfitting: Υπάρχει περισσότερος κίνδυνος να γίνει υπερπροσαρμογή του μοντέλου στα δεδομένα εκπαίδευσης και είναι πιο δύσκολο να εξηγηθούν συγκεκριμένες προβλέψεις [5][6].

4. Κατηγοριοποίηση στην Ανάκτηση Πληροφορίας (Classification in Information Retrieval)

4.1 Εισαγωγή

Η ανάκτηση πληροφοριών (IR) είναι η πειθαρχία της απόκτησης πληροφοριών σχετικών με τις απαιτούμενες πληροφορίες από μια συλλογή πληροφοριακών πόρων. Το IR ασχολείται με την οργάνωση, την αποθήκευση, την αντιπροσώπευση και την πρόσβαση σε μη δομημένα και ημιδομημένα αρχεία πληροφοριών, όπως έγγραφα, ηλεκτρονικούς καταλόγους, ιστοσελίδες και πολυμέσα. Τελευταία, υπάρχουν αξιοσημείωτες ανάγκες για αποτελεσματικές τεχνικές που αυτοματοποιούν τη διαδικασία ανάκτησης πληροφοριών, καθώς οι περισσότερες πηγές δεδομένων παρέχουν μη δομημένες ομάδες δεδομένων. Δεδομένου ότι ο όγκος των πληροφοριών αυξάνεται δυναμικά, είναι απαραίτητο να δημιουργηθεί μια εξειδικευμένη δομή δεδομένων για γρήγορη αναζήτηση, η οποία καλείται ευρετήριο. Τα ευρετήρια είναι πυρήνας για κάθε σύγχρονο σύστημα ανάκτησης πληροφοριών. Τα ευρετήρια επιτρέπουν γρήγορη πρόσβαση στα δεδομένα και επιτρέπουν την επιτάχυνση της επεξεργασίας ερωτήματος. Στη συνέχεια, η επεξεργασία ανάκτησης έρχεται να παράγει τον κατάλογο επιτυχίας, ο οποίος αποτελείται από την ανάκτηση εγγράφων που ικανοποιούν ένα ερώτημα ενός χρήστη.

4.2 Ανάκτηση πληροφορίας (IR)

Τα συστήματα IR συνίστανται κυρίως στη δημιουργία αποτελεσματικών ευρετηρίων που οργανώνουν δεδομένα στοιχείων με οργανωμένο τρόπο, επεξεργάζονται ερωτήματα χρηστών και αναπτύσσουν αλγόριθμους κατάταξης για την ενίσχυση των αποτελεσμάτων με την εμφάνιση των πιο σχετικών. Η πρώτη φάση της δημιουργίας ενός συστήματος IR είναι η συγκέντρωση της συλλογής εγγράφων και η αποθήκευση σε αποθετήριο, το οποίο διαμορφώνει το σώμα του συστήματος. Στη δεύτερη φάση, τα έγγραφα πρέπει να οργανωθούν και να αναπροσαρμοστούν για γρήγορη ανάκτηση και κατάταξη. Η πιο δημοφιλής δομή ευρετηριοποίησης είναι το ανεστραμμένο ευρετήριο (inverted index), το οποίο αποτελείται από όλες τις ξεχωριστές λέξεις της συλλογής και για κάθε λέξη υπάρχει μια λίστα συνδεδεμένων που υποδεικνύει τα έγγραφα που το περιέχουν.

Ένα ανεστραμμένο αρχείο (ανεστραμμένο ευρετήριο) είναι μια τεχνική προσανατολισμένη στις λέξεις για την ευρετηριοποίηση μιας συλλογής κειμένου για την επιτάχυνση της αναζήτησης. Η δομή των ανεστραμμένων αρχείων αποτελείται από δύο στοιχεία: το λεξιλόγιο και τις εμφανίσεις. Το λεξιλόγιο είναι το σύνολο όλων των διαφορετικών λέξεων (όρων) στο κείμενο. Για κάθε όρο υπάρχει μία λίστα με την εμφάνισή του σε κάποιο κείμενο. Το σχήμα 4.1 απεικονίζει το ανεστραμμένο ευρετήριο.

Μπορεί να χρησιμοποιηθεί και άλλη ευρετηριοποίηση, όπως τα δέντρα επιθεμάτων (suffix trees) και οι συστοιχίες επιθεμάτων (suffix arrays).

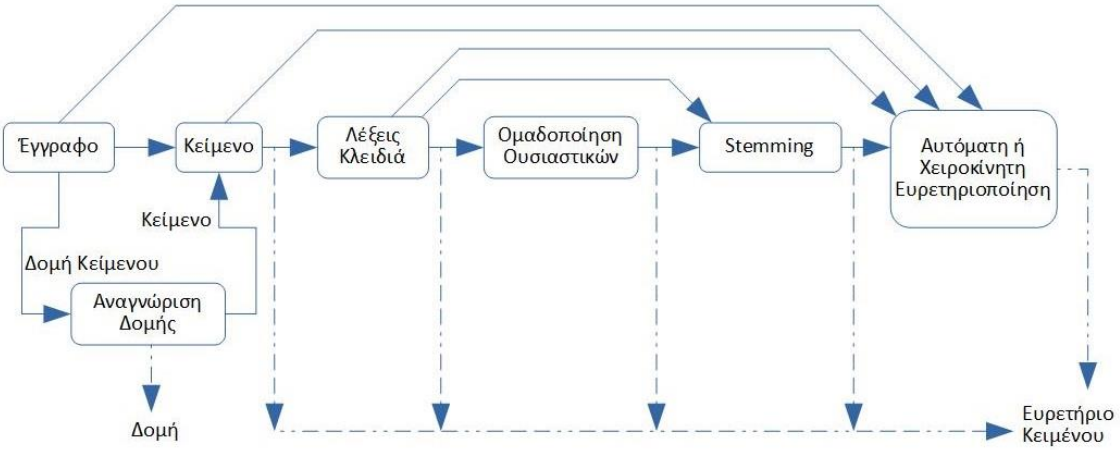
Λεξιλόγιο	Εμφανίσεις
Μια	4
Λέξη	4
Αποτελείται	1
Από	1
Πολλά	3
Γράμματα	2

Σχήμα 4.1. Ανεστραμμένο ευρετήριο

Το IR συνήθως επιδιώκει να βρει έγγραφα σε κάθε συλλογή που αφορούν ένα συγκεκριμένο θέμα ή που ικανοποιεί μια συγκεκριμένη ανάγκη πληροφόρησης. Ο χρήστης εκφράζει το απαραίτητο θέμα ή πληροφορίες δημιουργώντας ένα ερώτημα. Έγγραφα που ικανοποιούν το εκφρασμένο ερώτημα - κατά την κρίση του χρήστη, περιγράφονται ως "σχετικά". Τα έγγραφα που δεν σχετίζονται με το δεδομένο ερώτημα περιγράφονται ως "μη σχετικά". Μία σημαντική λειτουργία μιας μηχανής IR είναι η κατηγοριοποίηση των εγγράφων στις καθορισμένες ομάδες τους για τη διευκόλυνση της εμφάνισης των αποτελεσμάτων στους χρήστες. Φυσικά, η κατηγοριοποίηση των εγγράφων παίζει σημαντικό ρόλο στην ενίσχυση του αποτελέσματος. Με άλλα λόγια, όσο καλύτερα είναι τα χαρακτηριστικά κατηγοριοποίησης, τόσο υψηλότερο είναι το ποσοστό των εγγράφων που επιστρέφονται στον χρήστη που θα κριθούν σχετικά.

Οι ερευνητές ανέφεραν ότι η ευρετηριοποίηση είναι πολύ σημαντικό βήμα για τη βελτίωση της αποτελεσματικότητας και της αποτελεσματικότητας της έρευνας. Το ευρετήριο είναι μια δομή δεδομένων που βασίζεται σε ένα κείμενο για να επιταχύνει την αναζήτηση. Όπως φαίνεται στο σχήμα 4.2, η ευρετηριοποίηση θα μπορούσε να περάσει από πολλά στάδια:

1. Αφαίρεση λέξεων-κλειδιών: Αφαίρεση των λέξεων που είναι περιττές σε κάθε έγγραφο.
2. Αποκοπή Καταλήξεων (Stemming): Επιστροφή της κάθε λέξης στη γραμματική ρίζα της για να μειωθούν οι ξεχωριστές λέξεις.
3. Αφαίρεση κενών και σύμβολων (π.χ. -, _, *,?)
4. Ομαδοποίηση ουσιαστικών: Ένταξη των ουσιαστικών σε ομάδες για να κατηγοριοποιηθούν.



Σχήμα 4.2. Στάδια ευρετηριοποίησης

Ακόμη, θα μπορούσε να είναι ευρετηριοποίηση πλήρους κειμένου. Εδώ το σύστημα είναι έτοιμο για αναζήτηση σε αυτό. Στο στάδιο επιλογής, για να κατηγοριοποιηθούν τα αποτελέσματα υπάρχουν αρκετές μέθοδοι που θα μπορούσαν να χρησιμοποιηθούν:

1. Συχνότητα των εγγράφων: Μια απλή και αποτελεσματική διαδικασία επιλογής είναι να χρησιμοποιηθούν μόνο οι όροι των οποίων η συχνότητα εγγράφων υπερβαίνει ένα προκαθορισμένο όριο συχνότητας που αντιπροσωπεύει τα έγγραφα.
2. Βάρη TF-IDF: Παρόμοιες αλλά ελαφρώς πιο εξελιγμένες διαδικασίες για την υιοθέτηση διαδικασίας επιλογής όρων που διατηρεί τους όρους υψηλότερων TF-IDF σταθμίσεων σε κάθε έγγραφο.

Ως τελικό βήμα, οι μηχανές IR ανακατατάσσουν το πλήθος των εγγράφων σύμφωνα με ένα δεδομένο χαρακτηριστικό κατάταξης. Για παράδειγμα, εάν ένα έγγραφο D1 έλαβε υψηλότερη κατάταξη από το D2, αυτό σημαίνει ότι το D1 είναι πιθανότερο να είναι πιο σχετικό με ένα δεδομένο ερώτημα του χρήστη από το D2.

Το IR ασχολείται με την προσέγγιση του εγγράφου σε ένα δεδομένο ερώτημα, το οποίο είναι επίσης ένα σημαντικό ζήτημα. Η ανάκτηση πληροφοριών που σχετίζονται με την ανάγκη ή το ερώτημα ενός χρήστη απαιτεί την αντιστοίχιση κάθε εγγράφου με ένα δεδομένο ερώτημα. Ορισμένα από τα συστήματα IR που ασχολούνται με την

υποκειμενικότητα της συνάφειας, πρέπει να δημιουργούν προφίλ χρηστών, δηλ. προτιμήσεις χρηστών. Στόχος αυτού του προφίλ είναι η παροχή πρόσθετων πληροφοριών σύμφωνα με προκαθορισμένα κριτήρια. Άλλη προσέγγιση για την ενίσχυση του αποτελέσματος των ανακτηθέντων εγγράφων είναι η κατηγοριοποίηση της συλλογής. Κάνοντας κάθε ομάδα της συλλογής που μιλάει για την ίδια γενική ετικέτα σε ένα ενιαίο όνομα για όλη την ομάδα [7].

4.3 Κατηγοριοποίηση κειμένου και συναισθηματική ανάλυση

Η κατηγοριοποίηση κειμένου είναι η διαδικασία κατηγοριοποίησης εγγράφων κάτω από ένα συγκεκριμένο σύμπλεγμα ή κατηγορία, χρησιμοποιώντας πλήρως εποπτευόμενη διαδικασία μάθησης. Η κατηγοριοποίηση μπορεί να γίνει χειροκίνητα από αναλυτές ή χρησιμοποιώντας αυτόματα γνωστούς και ευρέως χρησιμοποιούμενους αλγορίθμους κατηγοριοποίησης.

Πρώτη τακτική για την κατηγοριοποίηση εγγράφων είναι η ανάθεση μιας ετικέτας σε κάθε έγγραφο, αλλά αυτό επιλύει το πρόβλημα μόνο όταν οι χρήστες γνωρίζουν τις ετικέτες των εγγράφων που αναζητούν. Αυτή η τακτική δεν επιλύει το γενικότερο πρόβλημα εύρεσης εγγράφων σε συγκεκριμένο θέμα ή αντικείμενο. Για την περίπτωση αυτή, καλύτερη λύση είναι η ομαδοποίηση εγγράφων με κοινά γενικά θέματα και η ετικέτα κάθε ομάδας με ένα νόημα. Κάθε ομάδα με ετικέτα ονομάζεται κατηγορία ή κλάση.

Η συναισθηματική ανάλυση είναι μία υποπεριοχή της κατηγοριοποίησης κειμένου. Είναι η διαδικασία εξόρυξης πληροφοριών, ώστε να καταγραφεί η συναισθηματική κατάσταση ενός χρήστη μέσα από το γραπτό του λόγο. Η εργασία ανάλυσης συναισθημάτων συνήθως διαμορφώνεται ως πρόβλημα κατηγοριοποίησης όπου ένας κατηγοριοποιητής τροφοδοτείται με ένα κείμενο και επιστρέφει την αντίστοιχη κατηγορία (π.χ. θετικό, αρνητικό ή ουδέτερο σε περίπτωση που πραγματοποιείται ανάλυση πολικότητας). Για την κατηγοριοποίηση του κειμένου σε κλάσεις που εκφράζουν συναίσθημα χρησιμοποιούνται οι τεχνικές που έχουμε αναφέρει σε προηγούμενα κεφάλαια, η επεξεργασία φυσικής γλώσσας, στατιστικές μέθοδοι και μηχανική μάθηση.

Μια ανάλυση που μπορεί να γίνει είναι η γενικότερη συναισθηματική κατάσταση που προσπαθεί να μεταδώσει ο συγγραφέας μέσω του κειμένου του είτε στην άποψη που μεταδίδει πάνω σε ένα θέμα. Στην πρώτη περίπτωση η κατηγοριοποίηση μπορεί να γίνει σε τάξεις συναισθήματος και στη δεύτερη περίπτωση σε γενικότερες τάξεις, συνήθως δύο ή τρεις , θετική, αρνητική και ουδέτερη.

Μια άλλη ανάλυση που γίνεται είναι με βάση το μέγεθος του κειμένου που εξετάζεται το συναίσθημα. Μπορεί να αναζητηθεί η πολικότητα ολόκληρου κειμένου ή μιας πρότασης ή μίας έκφρασης ή χαρακτηριστικού μιας οντότητας [7].

4.4 Κατηγοριοποιητής

Υπάρχουν διάφορες μέθοδοι που μπορούν να χρησιμοποιηθούν στην κατηγοριοποίηση κειμένου. Στις απλούστερες περιπτώσεις μια φράση εκφράζει μόνο ένα συναίσθημα και το συνολικό συναίσθημα της φράσης αντικατοπτρίζει το συναίσθημα προς οποιεσδήποτε από τις οντότητες. Δυστυχώς, οι περισσότερες προτάσεις είναι λίγο πιο περίπλοκες από αυτό.

Οι προτάσεις περιλαμβάνουν συχνά αναφορές σε πολλαπλές οντότητες για λόγους σύγκρισης. Στις περιπτώσεις αυτές, η διάσπαση της φράσης και η ανάλυση του συναισθήματος μέσα σε κάθε κομμάτι είναι συχνά αρκετή για να εκχωρήσει σωστά το συναίσθημα σε κάθε οντότητα. Αυτή είναι η μέθοδος της διαίρεσης. Για παράδειγμα, στην πρόταση «Μου αρέσουν τα hotdog, αλλά σιχαίνομαι τη μουστάρδα», χρησιμοποιώντας τη μέθοδο της διαίρεσης θα αποδοθεί συναίσθημα στα «hotdog» με βάση την έκφραση «Μου αρέσουν τα hotdog» και θα αποδοθεί συναίσθημα στη «μουστάρδα» με βάση την έκφραση «σιχαίνομαι τη μουστάρδα».

Η μέθοδος των πλησιέστερων γειτόνων βασίζεται στη μέτρηση της απόστασης μεταξύ ενός κέντρου βάρους της ομάδας και κάθε επόμενου αντικειμένου. Όσο μικρότερη είναι η απόσταση συνεπάγεται μεγαλύτερη συσχέτιση. Στην πρόταση, «Αγαπώ το ποδόσφαιρο, ακόμα και αν οι διαιτητές είναι απαράδεκτοι», η κοντινότερή λέξη στο «ποδόσφαιρο» που εκφράζει συναίσθημα είναι η λέξη «αγαπώ» και η κοντινότερη στους «διαιτητές» είναι η «απαράδεκτοι» [7].

Ο αλγόριθμος κατηγοριοποίησης Naïve Bayes έχει σχεδιαστεί πάνω από ένα πιθανολογικό πλαίσιο, στο οποίο υπολογίζεται η πιθανότητα να ανήκει ένα συγκεκριμένο αντικείμενο σε μια συγκεκριμένη κατηγορία. Όσο μεγαλύτερη είναι η πιθανότητα τόσο μεγαλύτερη είναι η πιθανότητα να εκχωρηθεί το αντικείμενο σε ένα δεδομένο σύμπλεγμα [7].

Η δομή δέντρου απόφασης προσομοιώνει μια προσέγγιση διαίρει και βασίλευε, όπου τα χαρακτηριστικά χρησιμοποιούνται για την κατασκευή ενός δέντρου αποφάσεων και τα φύλλα αντιπροσωπεύουν ετικέτες ή κατηγορίες τάξεων.

Οι μέθοδοι που περιγράφονται παραπάνω, μπορούν να επιστρέψουν τα σωστά αποτελέσματα υπό ορισμένες συνθήκες. Κάθε μέθοδος έχει τα δικά της πλεονεκτήματα και αδυναμίες. Συνδυάζοντας τις μεθόδους, είναι δυνατό να δημιουργηθεί ένας πολύ πιο

ισχυρός αλγόριθμος. Αυτός ονομάζεται συλλογικός αλγόριθμος και επιτυγχάνεται λαμβάνοντας το μέσο αποτέλεσμα πολλών μεθόδων.

Μετρήσεις και αξιολόγηση στην ανάλυση συναισθημάτων

Υπάρχουν πολλοί τρόποι με τους οποίους μπορούν να γίνουν μετρήσεις απόδοσης για την αξιολόγηση ενός κατηγοριοποιητή ώστε να φανεί η ακρίβεια ενός μοντέλου ανάλυσης συναισθημάτων. Ένα από τους πιο συχνά χρησιμοποιούμενους τρόπους είναι το cross-validation.

Αυτό που κάνει το cross-validation είναι η διάσπαση των δεδομένων εκπαίδευσης σε ένα σετ εκπαίδευσης (π.χ. με το 75% των δεδομένων εκπαίδευσης) και τον ίδιο αριθμό δοκιμαστικό σετ (με το 25% των δεδομένων εκπαίδευσης) και χρησιμοποιεί το σετ εκπαίδευσης για να εκπαιδεύσει τον κατηγοριοποιητή και το δοκιμάζει έναντι σετ δοκιμών για να λάβει μετρήσεις απόδοσης. Η διαδικασία επαναλαμβάνεται πολλές φορές και υπολογίζεται ένας μέσος όρος για κάθε μία από τις μετρήσεις.

Εάν το σετ δοκιμών είναι πάντα το ίδιο, ενδέχεται ο κατηγοριοποιητής να υπερπροσαρμοστεί σε αυτό το σετ δοκιμών, πράγμα που σημαίνει ότι μπορεί να προσαρμόσει την ανάλυση σε ένα συγκεκριμένο σύνολο δεδομένων, ώστε να μην μπορεί να αναλυθεί διαφορετικό σετ. Το cross-validation βοηθάει στην πρόληψη αυτού.

Precision, Recall και Accuracy

Το Precision, το recall και το accuracy είναι τυπικές μετρήσεις που χρησιμοποιούνται για την αξιολόγηση της απόδοσης ενός κατηγοριοποιητή.

Το precision μετράει πόσα κείμενα είχαν προβλεφθεί σωστά ως ανήκοντα σε μια δεδομένη κατηγορία από όλα τα κείμενα που είχαν προβλεφθεί (σωστά και εσφαλμένα) ως μέλη της κατηγορίας.

$$precision = \frac{true\ positives}{true\ positives + false\ positives} \quad (4.1)$$

Το recall μετράει πόσα κείμενα είχαν προβλεφθεί σωστά ως ανήκοντα σε μια δεδομένη κατηγορία από όλα τα κείμενα που θα έπρεπε να είχαν προβλεφθεί ως ανήκοντα στην

κατηγορία. Από αυτό προκύπτει ότι όσο περισσότερα δεδομένα θα παρέχονται στον κατηγοριοποιητή, τόσο καλύτερο θα είναι το recall.

$$recall = \frac{true\ positives}{true\ positives + false\ negatives} \quad (4.2)$$

Το accuracy μετράει πόσα κείμενα είχαν προβλεφθεί σωστά (ανήκουν σε μια κατηγορία και δεν έχουν κατηγοριοποιηθεί στην σωστή κατηγορία) από όλα τα κείμενα.

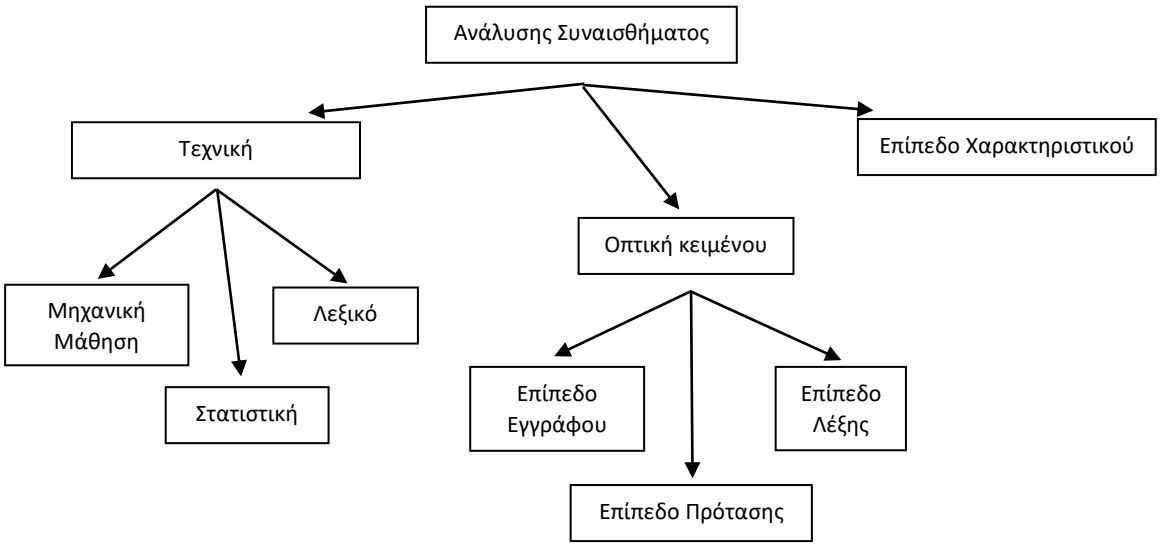
Πιο συχνά, χρησιμοποιούνται το precision και το recall για τη μέτρηση της απόδοσης, καθώς μόνο με το accuracy δεν μπορεί να εκτιμηθεί σωστά για το πόσο καλός ή κακός είναι ο κατηγοριοποιητής.

Για ένα δύσκολο έργο όπως η ανάλυση του συναισθήματος, τα επίπεδα precision και recall πιθανότατα να είναι χαμηλά στην αρχή. Καθώς τροφοδοτείται ο κατηγοριοποιητής με περισσότερα δεδομένα, η απόδοση θα βελτιωθεί [8].

5. Ανάλυση συναισθήματος (Sentiment Analysis)

5.1 Επίπεδα Ανάλυσης Συναισθήματος

Οι υπάρχουσες διαδικασίες για την ανάλυση συναισθημάτων μπορούν να ταξινομηθούν από διάφορες οπτικές γωνίες όπως: τεχνικής που χρησιμοποιείται, οπτικής του κειμένου, επίπεδο λεπτομέρειας της ανάλυσης κειμένου, βαθμολόγηση κλπ. Από την οπτική γωνία της τεχνικής, εντοπίζουμε την μηχανική μάθηση, τις τεχνικές βασισμένες σε λεξικό και τις στατιστικές τεχνικές (σχήμα 5.1).



Σχήμα 5.1: Ταξινόμηση Ανάλυσης Συναισθήματος

Η προσέγγιση της μηχανικής μάθησης χρησιμοποιεί διάφορους αλγορίθμους μάθησης για να προσδιορίσει το συναίσθημα εκπαιδευόμενο από ένα γνωστό σύνολο δεδομένων.

Η προσέγγιση που βασίζεται στο λεξικό περιλαμβάνει τον υπολογισμό της πολικότητας των συναισθημάτων ενός κειμένου χρησιμοποιώντας τον σημασιολογικό προσανατολισμό των λέξεων ή των προτάσεων του κειμένου. Ο «σημασιολογικός προσανατολισμός» είναι ένα μέτρο υποκειμενικότητας και γνώμης στο κείμενο.

Η προσέγγιση που βασίζεται στη στατιστική ανάλυση έχει ως σκοπό να αντιμετωπίσει τις δυσκολίες που ανακύπτουν από την αδυναμία διαχείρισης κάποιων λέξεων ή κάποιων κειμένων από τις μεθόδους που βασίζονται στη μηχανική μάθηση ή στα λεξικά. Η προσέγγιση αυτή περιλαμβάνει τη δημιουργία ενός «συναισθηματικού» λεξικού από ένα μεγάλο σύνολο εγγράφων ώστε να χρησιμοποιηθεί για να αποδώσει ένα σημασιολογικό προσανατολισμό σε κάθε λέξη ανάλογα με τη συχνότητα που εμφανίζεται στα έγγραφα, τα οποία έχουν ήδη χαρακτηριστεί θετικού ή αρνητικού προσανατολισμού.

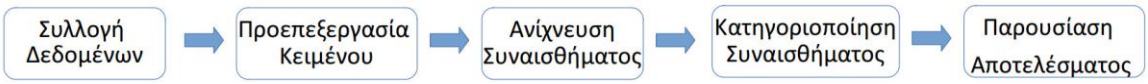
Μια άλλη ταξινόμηση προσανατολίζεται περισσότερο στη δομή του κειμένου: επίπεδο εγγράφου, επίπεδο προτάσεων ή ταξινόμηση επιπέδου λέξεων ή χαρακτηριστικών. Η ταξινόμηση σε επίπεδο εγγράφου στοχεύει να βρει μια πολικότητα συναισθημάτων για ολόκληρο το κείμενο, ενώ η ταξινόμηση σε επίπεδο προτάσεων ή σε επίπεδο λέξεων μπορεί να εκφράσει μια πολικότητα συναισθημάτων για κάθε πρόταση ενός κειμένου ή ακόμη και για κάθε λέξη. Το επίπεδο χαρακτηριστικού έχει ως στόχο να κατατάξει το συναίσθημα ενός κειμένου σύμφωνα με το συναίσθημα κάθε χαρακτηριστικού που εντοπίζεται μέσα στο κείμενο.

Σύγκριση των προσεγγίσεων

Οι μέθοδοι που βασίζονται στα λεξικά χαρακτηρίζονται από την μεγάλη ευχρηστία, ειδικά σε εφαρμογές με μεγάλα σύνολα δεδομένων. Επίσης η αυτόματη προσθήκη ετικέτας σε κάθε λέξη που περιέχουν, ενισχύει την ανάλυση του σημασιολογικού τους προσανατολισμού. Ωστόσο, η αποκλειστική εξάρτηση από συναισθηματικά λεξικά μπορεί να επηρεάσει την ικανότητα ενός συστήματος συναισθηματικής ανάλυσης ανάλογα με την εξειδίκευση του λεξικού σε συγκεκριμένους τομείς και τον τομέα έρευνας για τον οποίο χρησιμοποιείται. Αντίθετα με τη χρήση λεξικών, οι προσεγγίσεις που βασίζονται στη μηχανική μάθηση μπορούν να λειτουργήσουν με μεγάλη επιτυχία χρησιμοποιώντας μικρά σύνολα δεδομένων εκπαίδευσης, πετυχαίνοντας προβλέψεις σε πρωτοεμφανιζόμενα δεδομένα ακόμα και αν κάποιες λέξεις δεν υπάρχουν στη βάση τους. Οι μέθοδοι που βασίζονται στις παραδοσιακές στατιστικές μεθόδους είναι ικανές να κατηγοριοποιήσουν συναισθηματικά ένα κείμενο με αποδεκτή ακρίβεια μόνο αν τους δοθεί μια μεγάλη συλλογή κειμένων ως είσοδος και αυτό θα αποδώσει καλά σε επίπεδο παραγράφου και όχι σε επίπεδο πρότασης [9].

5.2 Στάδια Ανάλυσης Συναισθήματος

Υπάρχουν πέντε στάδια για την ανάλυση συναισθημάτων όπως φαίνεται και στο σχήμα 5.2.



Σχήμα 5.2 Στάδια ανάλυσης συναισθήματος.

Συλλογή δεδομένων

Στο πρώτο στάδιο γίνεται η συλλογή των δεδομένων. Για παράδειγμα, οι καταναλωτές συνήθως εκφράζουν τα συναισθήματά τους σε δημόσια φόρουμ όπως τα blogs, στα σχόλια των προϊόντων καθώς και στα ιδιωτικά τους αρχεία καταγραφής logs - ιστότοποι κοινωνικού δικτύου όπως το Facebook και το Twitter. Οι απόψεις και τα συναισθήματα εκφράζονται με διαφορετικό τρόπο, με διαφορετικό λεξιλόγιο, πλαίσιο γραφής, χρήση συντομογραφιών και αργκό, καθιστώντας τα δεδομένα τεράστια και αποδιοργανωμένα. Η χειρωνακτική ανάλυση των δεδομένων συναισθημάτων είναι σχεδόν αδύνατη. Ως εκ τούτου, ειδικές γλώσσες προγραμματισμού όπως η 'R' χρησιμοποιούνται για την επεξεργασία και την ανάλυση των δεδομένων.

Προεπεξεργασία κειμένου

Η προεπεξεργασία κειμένου δεν είναι παρά το φιλτράρισμα των εξαγόμενων δεδομένων πριν την ανάλυση. Περιλαμβάνει τον εντοπισμό και την εξάλειψη του περιεχομένου που δεν είναι κείμενο και του περιεχομένου που δεν σχετίζεται με τον τομέα της μελέτης των δεδομένων. Στην προεπεξεργασία, για την ανάλυση συναισθημάτων χρησιμοποιούνται διάφορες τεχνικές της Επεξεργασίας Φυσικής Γλώσσας και της Ανάκτησης Πληροφορίας. Τα βασικότερα στάδια της προεπεξεργασίας είναι το tokenization, η αφαίρεση stop words, το stemming, το POS tagging και η εξαγωγή χαρακτηριστικών.

Ανίχνευση συναισθήματος

Σε αυτό το στάδιο, κάθε πρόταση του κειμένου εξετάζεται για την υποκειμενικότητα. Οι προτάσεις με υποκειμενικές εκφράσεις διατηρούνται και αυτές που φέρουν αντικειμενικές εκφράσεις απορρίπτονται. Η ανάλυση του συναισθήματος γίνεται σε διαφορετικά επίπεδα χρησιμοποιώντας κοινές υπολογιστικές τεχνικές.

Κατηγοριοποίηση συναισθήματος

Τα συναισθήματα μπορούν να κατηγοριοποιηθούν ευρέως σε δύο ομάδες, θετικές και αρνητικές. Σε αυτό το στάδιο της μεθοδολογίας ανάλυσης συναισθημάτων, κάθε υποκειμενική φράση που ανιχνεύεται κατηγοριοποιείται σε θετικές ομάδες, αρνητικές, καλές ή κακές. Οι τεχνικές κατηγοριοποίησης συναισθημάτων είναι η προσέγγιση μηχανικής μάθησης, η προσέγγιση που βασίζεται στη χρήση λεξικού και η υβριδική που είναι συνδυασμός των δυο προηγούμενων.

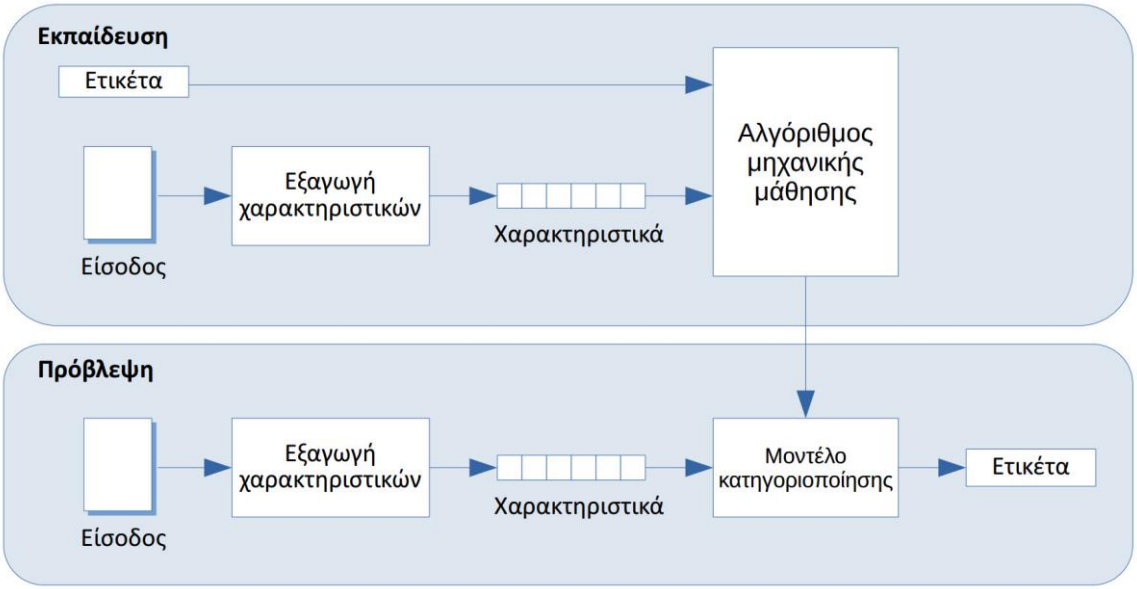
Παρουσίαση του αποτελέσματος

Η βασική ιδέα της ανάλυσης συναισθημάτων είναι να μετατρέψει το αδόμητο κείμενο σε χρήσιμες πληροφορίες. Μετά την ολοκλήρωση της ανάλυσης, τα αποτελέσματα κειμένου εμφανίζονται σε γραφήματα όπως το διάγραμμα πίτας, το διάγραμμα ράβδων και τα γραμμικά γραφήματα [10].

5.3 Τεχνικές Ανάλυσης Συναισθημάτων

5.3.1 Ανίχνευση συναισθήματος βασισμένη σε μηχανική μάθηση

Οι τεχνικές μηχανικής μάθησης χρησιμοποιούνται γενικά για τη δυαδική κατηγοριοποίηση και τις προβλέψεις των συναισθημάτων ως θετικές ή αρνητικές. Ένας κατηγοριοποιητής μηχανικής μάθησης μπορεί συνήθως να υλοποιηθεί με τα ακόλουθα βήματα, σχήμα 5.3:



Σχήμα 5.3. Βήματα υλοποίησης ενός κατηγοριοποιητή.

5.3.1.1 Οι διαδικασίες εκπαίδευσης και πρόβλεψης

Στη διαδικασία εκπαίδευσης, το μοντέλο μαθαίνει να συσχετίζει μια συγκεκριμένη είσοδο (δηλαδή ένα κείμενο) με την αντίστοιχη έξοδο (ετικέτα) βάσει των δειγμάτων δοκιμής που χρησιμοποιούνται για την εκπαίδευση. Ο εξαγωγέας χαρακτηριστικών μετατρέπει την είσοδο κειμένου σε ένα διάνυσμα χαρακτηριστικών. Ζεύγη χαρακτηριστικών και ετικετών (π.χ. θετικών, αρνητικών ή ουδέτερων) τροφοδοτούνται στον αλγόριθμο μηχανικής μάθησης για τη δημιουργία ενός μοντέλου.

Στη διαδικασία πρόβλεψης, ο εξαγωγέας χαρακτηριστικών χρησιμοποιείται για να μετασχηματίσει το εισερχόμενο κείμενο σε διανύσματα χαρακτηριστικών. Αυτά τα διανύσματα χαρακτηριστικών τροφοδοτούνται στη συνέχεια στο μοντέλο, το οποίο τα κατηγοριοποιεί στις προβλεπόμενες ετικέτες (πάλι θετικές, αρνητικές ή ουδέτερες).

5.3.1.2 Εξαγωγή χαρακτηριστικών από κείμενο

Το πρώτο βήμα σε έναν κατηγοριοποιητή κειμένου για την μηχανική μάθηση είναι η μετατροπή του κειμένου σε μια αριθμητική αναπαράσταση, συνήθως ένα διάνυσμα. Κάθε συνιστώσα του διανύσματος αντιπροσωπεύει τη συχνότητα μιας λέξης ή μιας έκφρασης σε ένα προκαθορισμένο λεξικό (π.χ. ένα λεξικό πολωμένων λέξεων). Αυτή η διαδικασία είναι γνωστή ως εξαγωγή χαρακτηριστικών ή διάνυσμα κειμένου.

5.3.1.3 Αλγόριθμοι κατηγοριοποίησης

Η κατηγοριοποίηση του συναισθήματος γίνεται με τη χρήση αλγόριθμων μηχανικής μάθησης, οι οποίοι κατηγοριοποιούνται περαιτέρω στις ακόλουθες κατηγορίες, εποπτευόμενη, μη εποπτευόμενη και ημι-εποπτευόμενη μηχανική μάθηση και σε παραμετρικούς και μη παραμετρικούς αλγόριθμους μηχανικής μάθησης [8].

5.3.2 Ανίχνευση συναισθήματος βασισμένη σε λεξικά

Οι προσεγγίσεις βασισμένες σε κανόνες ορίζουν ένα σύνολο κανόνων σε κάποιο είδος scripting γλώσσας που προσδιορίζει την υποκειμενικότητα, την πολικότητα ή το θέμα μιας γνώμης. Οι πιο σημαντικοί δείκτες συναισθημάτων είναι οι συναισθηματικές λέξεις, που ονομάζονται και λέξεις γνώμης. Αυτές είναι λέξεις που χρησιμοποιούνται συνήθως για να εκφράσουν θετικά ή αρνητικά συναισθήματα. Για παράδειγμα, *καλό*, *υπέροχο* και *εκπληκτικό* είναι θετικές συναισθηματικές λέξεις, και *κακό*, *άσχημο* και *τρομερό* είναι αρνητικές συναισθηματικές λέξεις. Εκτός από μεμονωμένες λέξεις, υπάρχουν επίσης φράσεις και ιδιώματα, π.χ., «*κόστισε ο κούκος αηδόνι*». Οι συναισθηματικές λέξεις και φράσεις είναι καθοριστικής σημασίας για την ανάλυση συναισθήματος για προφανείς λόγους. Μια λίστα με αυτές τις λέξεις και φράσεις ονομάζεται συναισθηματικό λεξικό (ή λεξικό γνώμης). Παρόλο που το συναισθηματικό λεξικό είναι απαραίτητο δεν αρκεί για την ανάλυση συναισθημάτων. Παρακάτω, επισημαίνουμε διάφορα θέματα:

- Μια θετική ή αρνητική συναισθηματική λέξη μπορεί να έχει αντίθετο προσανατολισμό σε διαφορετικούς τομείς εφαρμογής.
- Μια πρόταση που περιέχει συναισθηματική λέξη μπορεί να μην εκφράζει κανένα συναίσθημα. Αυτό το φαινόμενο συμβαίνει συχνά σε διάφορους τύπους προτάσεων. Οι ερωτήσεις (ερωτηματικές προτάσεις) και οι υποθετικές προτάσεις είναι δύο τέτοιοι τύποι.
- Οι σαρκαστικές προτάσεις με ή χωρίς συναισθηματικές λέξεις είναι δύσκολο να αντιμετωπιστούν.
- Πολλές προτάσεις χωρίς συναισθηματικές λέξεις μπορούν επίσης να υποδηλώνουν γνώμη[2].

6. Αλγόριθμοι Μηχανικής Μάθησης

Η μηχανική μάθηση χρησιμοποιείται σε πολλά στάδια της ανάλυσης συναισθημάτων. Στο στάδιο της προεπεξεργασίας με NLP μπορεί να χρησιμοποιηθεί στο tokenization, στο POS Tagging, στη συντακτική ανάλυση κ.α. Επίσης, μέθοδοι μηχανικής μάθησης εφαρμόζονται και στα στάδια της ανίχνευσης συναισθήματος όπως και στην κατηγοριοποίηση συναισθήματος. Παρακάτω περιγράφονται μερικοί από τους βασικότερους αλγορίθμους μηχανικής μάθησης που χρησιμοποιούνται στα στάδια της ανάλυσης συναισθημάτων.

6.1 Overfitting και Underfitting

Γενίκευση στη Μηχανική Μάθηση

Στη μηχανική μάθηση περιγράφουμε την εκμάθηση της λειτουργίας στόχου από τα δεδομένα εκπαίδευσης ως επαγωγική μάθηση. Η επαγωγή αναφέρεται στην εκμάθηση γενικών εννοιών από συγκεκριμένα παραδείγματα που είναι ακριβώς το πρόβλημα που επιδιώκουν τα προβλήματα εποπτευόμενης μηχανικής μάθησης να λύσουν. Αυτό είναι διαφορετικό από την αφαίρεση, που είναι το αντίστροφο και επιδιώκει να μάθει συγκεκριμένες έννοιες από τους γενικούς κανόνες. Η γενίκευση αναφέρεται στο πόσο καλά εφαρμόστηκαν οι έννοιες που αποκτήθηκαν από ένα μοντέλο μηχανικής μάθησης σε συγκεκριμένα παραδείγματα που δεν είδε το μοντέλο όταν αυτό εκπαιδευόταν. Ο στόχος ενός καλού μοντέλου μηχανικής μάθησης είναι να γενικεύσει καλά από τα δεδομένα εκπαίδευσης σε οποιαδήποτε δεδομένα από τον τομέα προβλημάτων. Αυτό μας επιτρέπει να κάνουμε προβλέψεις στο μέλλον για δεδομένα που το μοντέλο δεν έχει δει ποτέ. Η ορολογία που χρησιμοποιείται στη μηχανική μάθηση, όταν μιλάμε για το πόσο καλά μαθαίνει ένα μοντέλο μηχανικής μάθησης και γενικεύει σε νέα δεδομένα, είναι το Overfitting και το Underfitting. Αυτές είναι οι δύο μεγαλύτερες αιτίες της κακής απόδοσης των αλγορίθμων μηχανικής μάθησης.

Στατιστική προσαρμογή (Statistical Fit)

Στα στατιστικά στοιχεία, η προσαρμογή αναφέρεται στο πόσο καλά ένα μοντέλο ερμηνεύει τα δεδομένα. Αυτή είναι μια καλή ορολογία που χρησιμοποιείται στη μηχανική μάθηση, επειδή οι εποπτευόμενοι αλγόριθμοι μηχανικής μάθησης επιδιώκουν την προσέγγιση της άγνωστης συνάρτησης συσχέτισης (mapping function) για τις μεταβλητές εξόδου λαμβάνοντας υπόψη τις μεταβλητές εισόδου.

Τα στατιστικά καλής προσαρμογής, τα οποία αναφέρονται στα μέτρα που χρησιμοποιούνται για την εκτίμηση του βαθμού στον οποίο η προσέγγιση της λειτουργίας ταιριάζει με τη συνάρτηση στόχου. Ορισμένες από αυτές τις μεθόδους είναι χρήσιμες στη μηχανική μάθηση (π.χ. υπολογισμός των υπολειπόμενων σφαλμάτων), αλλά μερικές από αυτές τις τεχνικές υποθέτουν ότι γνωρίζουμε τη μορφή της συνάρτησης στόχου που προσεγγίζουμε, κάτι που δεν συμβαίνει στη μηχανική μάθηση.

Overfitting

Το Overfitting αναφέρεται σε ένα μοντέλο που διαμορφώνει τα δεδομένα εκπαίδευσης πολύ καλά. Το Overfitting συμβαίνει όταν ένα μοντέλο μαθαίνει την λεπτομέρεια και το θόρυβο στα δεδομένα εκπαίδευσης, στο βαθμό που επηρεάζει αρνητικά την απόδοση του μοντέλου σε νέα δεδομένα. Αυτό σημαίνει ότι ο θόρυβος ή οι τυχαίες διακυμάνσεις στα δεδομένα εκπαίδευσης προωθούνται και μαθαίνονται ως έννοιες από το μοντέλο. Το πρόβλημα είναι ότι αυτές οι έννοιες δεν ισχύουν για νέα δεδομένα και επηρεάζουν αρνητικά την ικανότητα γενικευμένων μοντέλων.

Το Overfitting είναι πιο πιθανό με μη παραμετρικά και μη γραμμικά μοντέλα που έχουν μεγαλύτερη ευελιξία όταν μαθαίνουν μια λειτουργία. Ως εκ τούτου, πολλοί μη παραμετρικοί αλγόριθμοι μηχανικής μάθησης περιλαμβάνουν επίσης παραμέτρους ή τεχνικές που περιορίζουν πόση λεπτομέρεια μαθαίνει το μοντέλο. Για παράδειγμα, τα δέντρα αποφάσεων είναι ένας μη παραμετρικός αλγόριθμος εκμάθησης μηχανών που είναι πολύ ευέλικτος και υπόκειται σε Overfitting δεδομένων εκπαίδευσης.

Underfitting

Το underfitting αναφέρεται σε ένα μοντέλο που δεν μπορεί ούτε να μοντελοποιήσει τα δεδομένα εκπαίδευσης, ούτε να γενικεύσει σε νέα δεδομένα. Ένα μοντέλο μηχανικής μάθησης που υπόκειται underfitting δεν είναι ένα κατάλληλο μοντέλο και θα είναι προφανές καθώς θα έχει κακή απόδοση στα δεδομένα εκπαίδευσης. Το underfitting είναι όρος ο οποίος δεν χρησιμοποιείται συχνά, καθώς είναι εύκολο να εντοπιστεί δεδομένης μιας καλής μέτρησης επιδόσεων. Η λύση είναι να δοκιμαστούν εναλλακτικοί αλγόριθμοι μηχανικής μάθησης [5].

6.2 Γραμμικοί Αλγόριθμοι

6.2.1 Επικλινής κάθοδος (Gradient Descent)

Η επικλινής κάθοδος είναι ένας αλγόριθμος βελτιστοποίησης που χρησιμοποιείται για την εύρεση των τιμών των παραμέτρων (coefficients) μιας συνάρτησης (f) που ελαχιστοποιεί μια συνάρτηση κόστους (cost). Ο αλγόριθμος επικλινούς καθόδου χρησιμοποιείται καλύτερα όταν οι παράμετροι δεν μπορούν να υπολογιστούν αναλυτικά (π.χ. χρησιμοποιώντας γραμμική άλγεβρα) και πρέπει να αναζητηθούν από έναν αλγόριθμο βελτιστοποίησης. Η μέθοδος αυτή, είναι μία απλή διαδικασία βελτιστοποίησης που μπορεί να χρησιμοποιηθεί σε πολλούς αλγορίθμους μηχανικής μάθησης.

Διαδικασία επικλινούς καθόδου

Η διαδικασία ξεκινάει με αρχικές τιμές για την παράμετρο ή τις παραμέτρους για τη συνάρτηση. Αυτές θα μπορούσαν να είναι 0,0 ή μια τυχαία μικρή τιμή.

$$coefficient = 0{,}0 \quad (6.1)$$

Το κόστος των παραμέτρων αξιολογείται βάζοντάς τες στη συνάρτηση και υπολογίζοντας το κόστος.

$$cost = f(coefficient) \quad (6.2)$$

$$cost = evaluate(f(coefficient)) \quad (6.3)$$

Η παράγωγος είναι μια έννοια που αναφέρεται στην κλίση της συνάρτησης σε ένα δεδομένο σημείο. Πρέπει να είναι γνωστή η κλίση έτσι ώστε να είναι γνωστή και η κατεύθυνση (σημείο) για να μετακινηθούν οι τιμές των παραμέτρων προκειμένου να βρεθεί το χαμηλότερο κόστος στην επόμενη επανάληψη.

$$delta = derivative(cost) \quad (6.4)$$

Τώρα που είναι γνωστή η κατεύθυνση προς τα κάτω, μπορούν να ενημερωθούν οι τιμές των παραμέτρων. Πρέπει να προσδιοριστεί μια παράμετρος ρυθμού εκμάθησης (*alpha*) που να ελέγχει πόσο μπορούν να αλλάξουν οι παράμετροι σε κάθε ενημέρωση.

$$coefficient = coefficient - (alpha \times delta) \quad (6.5)$$

Αυτή η διαδικασία επαναλαμβάνεται μέχρις ότου το κόστος των παραμέτρων (*cost*) είναι 0,0 ή δεν μπορούν να επιτευχθούν περαιτέρω βελτιώσεις στο κόστος. Απαιτείται η γνώση της κλίση της συνάρτησης κόστους ή της συνάρτησης που βελτιστοποιείται. Στη συνέχεια περιγράφεται πώς μπορεί να χρησιμοποιηθεί σε αλγόριθμους μηχανικής μάθησης.

Ο στόχος όλων των εποπτευόμενων αλγορίθμων μηχανικής μάθησης είναι η καλύτερη εκτίμηση μιας συνάρτησης στόχου (f) που χαρτογραφεί τα δεδομένα εισόδου (X) στις μεταβλητές εξόδου (Y). Αυτό περιγράφει όλα τα προβλήματα κατηγοριοποίησης και παλινδρόμησης. Ορισμένοι αλγόριθμοι μηχανικής μάθησης έχουν παραμέτρους που χαρακτηρίζουν την εκτίμηση αλγορίθμων για τη συνάρτηση στόχου (f). Διαφορετικοί αλγόριθμοι έχουν διαφορετικές παραστάσεις και διαφορετικές παραμέτρους, αλλά πολλοί από αυτούς απαιτούν μια διαδικασία βελτιστοποίησης για να βρουν το σύνολο των παραμέτρων που οδηγούν στην καλύτερη εκτίμηση της συνάρτησης στόχου. Τα συνηθισμένα παραδείγματα αλγορίθμων με παραμέτρους που μπορούν να βελτιστοποιηθούν χρησιμοποιώντας τη μέθοδο επικλινούς καθόδου, είναι η γραμμική παλινδρόμηση και η λογιστική παλινδρόμηση.

Η αξιολόγηση του πόσο κοντά ένα μοντέλο μηχανικής μάθησης μπορεί να εκτιμήσει τη συνάρτηση στόχου, μπορεί να υπολογιστεί με διάφορες μεθόδους, οι οποίες καθορίζονται στον αλγόριθμο μηχανικής μάθησης. Η συνάρτηση κόστους περιλαμβάνει την αξιολόγηση των παραμέτρων στο μοντέλο μηχανικής μάθησης με τον υπολογισμό μιας πρόβλεψης για κάθε στιγμιότυπο εκπαίδευσης στο σύνολο δεδομένων και τη σύγκριση των προβλέψεων με τις πραγματικές τιμές εξόδου, υπολογίζοντας ένα άθροισμα ή ένα μέσο σφάλμα (όπως το *Sum of Squared Residuals* ή SSR στην περίπτωση της γραμμικής παλινδρόμησης) [5].

Στοχαστική επικλινής κάθοδος (Stochastic Gradient Descent - SGD)

Η επικλινής κάθοδος μπορεί να είναι αργή για να τρέξει σε πολύ μεγάλα σύνολα δεδομένων. Επειδή μία επανάληψη του αλγόριθμου απαιτεί μια πρόβλεψη για κάθε περίπτωση στο σύνολο δεδομένων κατάρτισης, μπορεί να χρειαστεί πολύς χρόνος όταν υπάρχουν εκατομμύρια περιπτώσεις. Σε καταστάσεις όπου υπάρχουν μεγάλες ποσότητες δεδομένων, μπορεί να χρησιμοποιηθεί μια παραλλαγή της επικλινούς καθόδου που ονομάζεται στοχαστική επικλινής κάθοδος. Σε αυτήν την παραλλαγή, ακολουθείται η διαδικασία επικλινούς καθόδου που περιγράφεται παραπάνω, αλλά η ενημέρωση των παραμέτρων γίνεται για κάθε περίπτωση εκπαίδευσης και όχι στο τέλος του συνόλου περιπτώσεων.

6.2.2 Γραμμική Παλινδρόμηση (Linear Regression)

Η γραμμική παλινδρόμηση είναι ένα γραμμικό μοντέλο, π.χ. ένα μοντέλο που θεωρεί μια γραμμική σχέση μεταξύ των μεταβλητών εισόδου (x) και μιας μοναδικής μεταβλητής εξόδου (y)

Αναπαρίσταται ως μια γραμμική εξίσωση που συνδυάζει ένα συγκεκριμένο σύνολο τιμών εισόδου (x) με τη λύση στην οποία είναι η προβλεπόμενη έξοδος (y) για το σύνολο των τιμών εισόδου. Ως εκ τούτου, τόσο οι τιμές εισόδου (x) όσο και η τιμή εξόδου είναι αριθμητικές.

Η γραμμική εξίσωση εκχωρεί έναν συντελεστή κλήσης σε κάθε τιμή εισόδου, που ονομάζεται συντελεστής και αντιπροσωπεύεται συνήθως από το ελληνικό γράμμα Βήτα (B). Προστίθεται επίσης ένας επιπλέον συντελεστής, ο οποίος δίνει στη γραμμή ένα πρόσθετο βαθμό ελευθερίας (π.χ. μετακίνηση προς τα πάνω και προς τα κάτω σε μια δισδιάστατη γραφική παράσταση) και ονομάζεται συχνά συντελεστής μεταβλητότητας. Για παράδειγμα, σε ένα πρόβλημα απλής παλινδρόμησης (ένα μόνο x και ένα μοναδικό y), η μορφή του μοντέλου θα είναι:

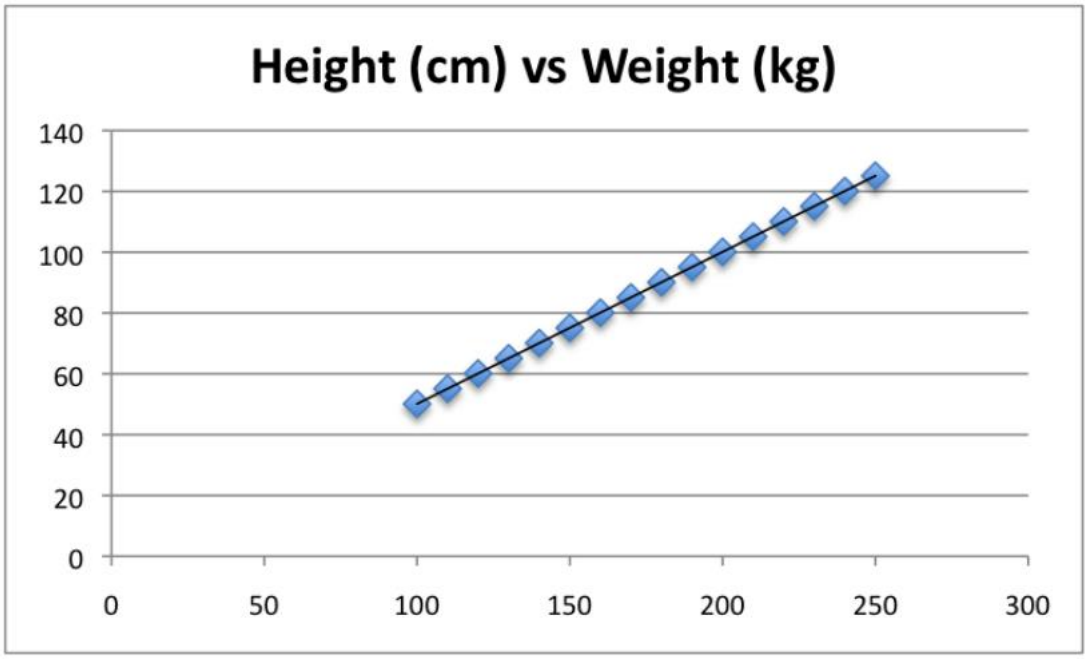
$$y = B0 + B1 \times x \qquad (6.6)$$

Σε περισσότερες διαστάσεις όταν έχουμε περισσότερες από μία είσοδους (x), η γραμμή ονομάζεται επίπεδο ή υπερ-επίπεδο. Επομένως, η αναπαράσταση είναι η μορφή της εξίσωσης με τις συγκεκριμένες τιμές που χρησιμοποιούνται για τους συντελεστές (π.χ. B0 και B1 στο παραπάνω παράδειγμα). Συχνά αναφέρεται η πολυπλοκότητα ενός μοντέλου παλινδρόμησης ως γραμμική παλινδρόμηση. Αυτό αναφέρεται στον αριθμό των συντελεστών που χρησιμοποιούνται στο μοντέλο.

Όταν ένας συντελεστής γίνει μηδέν, αφαιρεί την επίδραση της μεταβλητής εισόδου στο μοντέλο και συνεπώς από την πρόβλεψη που έγινε από το μοντέλο ($0 \times x = 0$). Αυτό ισχύει όταν εξετάζονται μέθοδοι κανονικοποίησης που αλλάζουν τον αλγόριθμο εκμάθησης για να μειωθεί η πολυπλοκότητα των μοντέλων παλινδρόμησης, πιέζοντας το απόλυτο μέγεθος των συντελεστών, οδηγώντας κάποιους στο μηδέν.

Προετοιμασία δεδομένων για γραμμική παλινδρόμηση

Η γραμμική παλινδρόμηση έχει μελετηθεί σε μεγάλο βαθμό και υπάρχει αρκετή βιβλιογραφία σχετικά με τον τρόπο με τον οποίο θα πρέπει να είναι δομημένα τα δεδομένα για να κάνουν καλύτερη χρήση του μοντέλου. Ως εκ τούτου, υπάρχει μεγάλη πολυπλοκότητα όταν μιλάμε για αυτές τις απαιτήσεις και τις προσδοκίες. Στην πράξη, αυτοί οι κανόνες μπορούν να χρησιμοποιηθούν περισσότερο ως κανόνες κατεύθυνσης όταν χρησιμοποιείται η τακτική ελαχίστων τετραγώνων, η πιο κοινή εφαρμογή γραμμικής παλινδρόμησης. Η προετοιμασία των δεδομένων βοηθάει ώστε να λειτουργεί καλύτερα σε ένα συγκεκριμένο πρόβλημα.



Σχήμα 6.1. Δείγμα γραμμικής παλινδρόμησης ύψους vs βάρους [5].

Γραμμική υπόθεση. Η γραμμική παλινδρόμηση υποθέτει ότι η σχέση μεταξύ εισόδου και εξόδου είναι γραμμική. Μπορεί να χρειαστεί να μετατραπούν τα δεδομένα για να γίνει η σχέση γραμμική.

Αφαίρεση θορύβου. Η γραμμική παλινδρόμηση υποθέτει ότι οι μεταβλητές εισόδου και εξόδου δεν έχουν θόρυβο. Η λειτουργία καθαρισμού των δεδομένων επιτρέπει την αποσαφήνιση των δεδομένων το οποίο είναι σημαντικότερο για τη μεταβλητή εξόδου και την αφαίρεση των ακραίων αποτελεσμάτων.

Αφαίρεση των συσχετισμένων εισροών. Η γραμμική παλινδρόμηση θα υπερκαλύψει τα δεδομένα όταν υπάρχουν πολύ συσχετισμένες μεταβλητές εισόδου.

Γκαουσιανή κατανομή (Gaussian Distribution). Η γραμμική παλινδρόμηση θα κάνει πιο αξιόπιστες προβλέψεις εάν οι μεταβλητές εισόδου και εξόδου έχουν Γκαουσιανή κατανομή.

Επαναδιάταξη εισόδων. Η γραμμική παλινδρόμηση θα κάνει πιο αξιόπιστες προβλέψεις εάν μετατοπιστούν οι μεταβλητές εισόδου χρησιμοποιώντας τυποποίηση ή κανονικοποίηση [5].

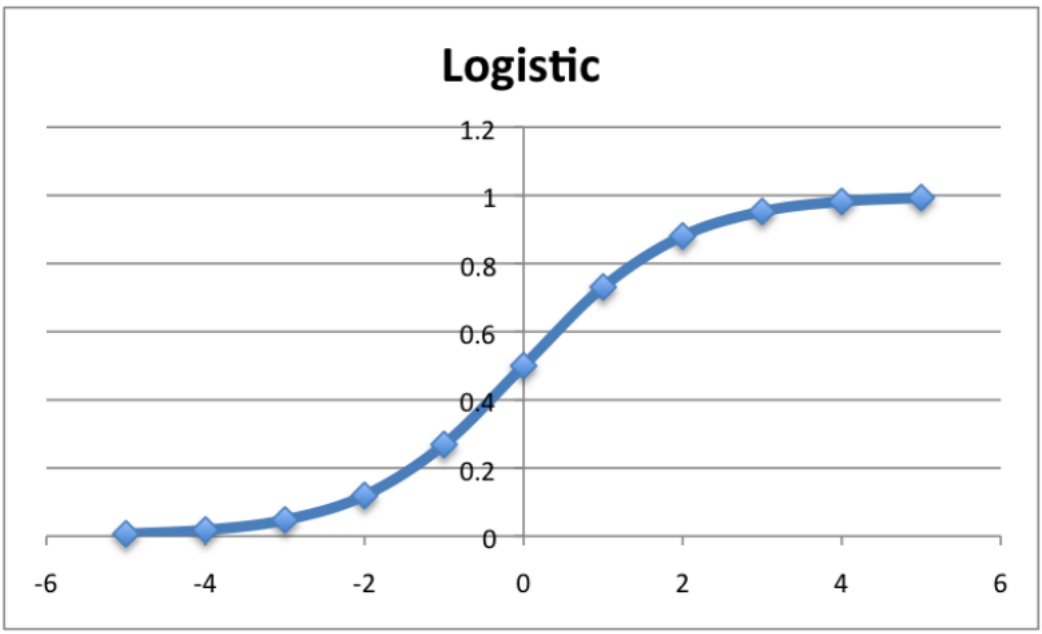
6.2.3 Λογιστική παλινδρόμηση (Logistic Regression)

Η λογιστική παλινδρόμηση που ονομάζεται και τεχνική μέγιστης εντροπίας (Maximum Entropy) είναι μια άλλη τεχνική που βασίζεται η μηχανική μάθηση από τον τομέα της στατιστικής. Είναι η καταλληλότερη μέθοδος για δυαδικά προβλήματα κατηγοριοποίησης.

Ονομάστηκε λογιστική παλινδρόμηση λόγω της συνάρτησης που χρησιμοποιείται στον πυρήνα της μεθόδου, τη λογιστική συνάρτηση. Η λογιστική συνάρτηση, που ονομάζεται επίσης και η σιγμοειδής συνάρτηση, αναπτύχθηκε από τους στατιστικολόγους για να περιγράψει τις ιδιότητες της πληθυσμιακής ανάπτυξης στην οικολογία. Είναι μια καμπύλη σχήματος S που μπορεί να πάρει οποιοδήποτε πραγματικό αριθμό και να το χαρτογραφήσει σε τιμή μεταξύ 0 και 1, αλλά ποτέ δεν είναι ακριβώς στα όρια αυτά.

$$\frac{1}{1+e-value} \qquad (6.7)$$

Στη συνάρτηση (6.7), όπου e είναι η βάση των φυσικών λογαρίθμων (ο αριθμός του Euler) και η τιμή (value) είναι η πραγματική αριθμητική τιμή που μετατρέπεται. Παρακάτω υπάρχει μια γραφική παράσταση των αριθμών μεταξύ -5 και 5 που μετασχηματίζεται στην περιοχή 0 και 1 χρησιμοποιώντας τη λογιστική συνάρτηση.



Σχήμα 6.2. Λογιστική συνάρτηση [5].

Η λογιστική παλινδρόμηση χρησιμοποιεί μια εξίσωση ως αναπαράσταση, σαν τη γραμμική παλινδρόμηση. Οι τιμές εισόδου (x) συνδυάζονται γραμμικά χρησιμοποιώντας βάρη ή συντελεστές για την πρόβλεψη μιας τιμής εξόδου (y). Μια βασική διαφορά από τη γραμμική παλινδρόμηση είναι ότι η τιμή εξόδου που διαμορφώνεται είναι δυαδική τιμή (0 ή 1) και όχι αριθμητική τιμή.

Παρακάτω παρουσιάζεται μια εξίσωση λογιστικής παλινδρόμησης:

$$y = \frac{e^{B0+B1\times x}}{1+e^{B0+B1\times x}} \qquad (6.8)$$

Όπου y είναι η προβλεπόμενη έξοδος, το B0 είναι σταθερή παράμετρος και B1 είναι το βάρος για την τιμή μιας εισόδου (x). Κάθε στήλη στα δεδομένα εισόδου έχει έναν συσχετισμένο συντελεστή B (μια σταθερή πραγματική τιμή) που πρέπει να αντληθεί από τα δεδομένα εκπαίδευσης. Η πραγματική αναπαράσταση του μοντέλου που θα αποθηκευτεί στη μνήμη ή σε ένα αρχείο είναι οι συντελεστές της εξίσωσης (η τιμή βήτα ή B).

Προετοιμασία δεδομένων για την λογιστική παλινδρόμηση

Οι παραδοχές που έγιναν από την λογιστική παλινδρόμηση σχετικά με τη διανομή και τις σχέσεις στα δεδομένα είναι παρόμοιες με τις παραδοχές που έγιναν στη γραμμική παλινδρόμηση. Για τον ορισμό αυτών των παραδοχών χρησιμοποιείται ακριβής πιθανολογική και στατιστική γλώσσα. Σε ένα project μοντέλου πρόβλεψης μηχανικής μάθησης, όπως σε ένα πρόβλημα συναισθηματικής ανάλυσης, γίνεται εστίαση στο να γίνουν ακριβείς προβλέψεις παρά για να ερμηνευτούν τα αποτελέσματα. Ως εκ τούτου, μπορούν να σπάσουν ορισμένες παραδοχές, αρκεί το μοντέλο να είναι εύρωστο και να αποδίδει καλά.

Μεταβλητή δυαδικών εξόδων. Η λογιστική παλινδρόμηση προορίζεται για δυαδικά (δύο τάξεων) κατηγορικά προβλήματα. Προβλέπει την πιθανότητα ενός στιγμιότυπου που ανήκει σε μια προεπιλεγμένη κλάση, το οποίο μπορεί να χωριστεί σε κατηγοριοποίηση 0 ή 1.

Αφαίρεση θορύβου. Η λογιστική παλινδρόμηση δεν υποθέτει λάθη στη μεταβλητή εξόδου (y), ενδεχομένως να χρειαστεί να αφαιρεθούν οι ακραίες τιμές και ενδεχομένως εσφαλμένες περιπτώσεις από τα δεδομένα εκπαίδευσης.

Γκαουσιανή κατανομή (Gaussian Distribution). Η λογιστική παλινδρόμηση είναι ένας γραμμικός αλγόριθμος (με έναν μη γραμμικό μετασχηματισμό στην έξοδο). Υποθέτει μια γραμμική σχέση μεταξύ των μεταβλητών εισόδου με την έξοδο. Τα μετασχηματισμένα δεδομένα των μεταβλητών εισόδου που εκθέτουν καλύτερα αυτή τη γραμμική σχέση μπορούν να οδηγήσουν σε ένα πιο ακριβές μοντέλο.

Αφαίρεση των συσχετισμένων εισροών. Όπως και η γραμμική παλινδρόμηση, το μοντέλο μπορεί να υπερκαλυφθεί αν υπάρχουν πολλές πολύ συσχετισμένες εισροές.

Αδυναμία σύγκλισης. Αυτό μπορεί να συμβεί αν υπάρχουν πολλά δεδομένα που σχετίζονται με τα δεδομένα εκπαίδευσης ή τα δεδομένα είναι πολύ αραιά (π.χ. πολλά μηδενικά στα δεδομένα εισόδου) [5].

6.2.4 Γραμμική Διακριτική Ανάλυση (Linear Discriminant Analysis LDA)

Περιορισμοί στη λογιστική παλινδρόμηση

Η λογιστική παλινδρόμηση είναι ένας απλός και χρήσιμος αλγόριθμος γραμμικής κατηγοριοποίησης. Έχει επίσης περιορισμούς που υποδηλώνουν την ανάγκη εναλλακτικών γραμμικών αλγορίθμων κατηγοριοποίησης.

- Προβλήματα δύο κατηγοριών. Η λογιστική παλινδρόμηση προορίζεται για προβλήματα δυο τάξεων ή δυαδικής κατηγοριοποίησης. Μπορεί να επεκταθεί για κατηγοριοποίηση σε πολλές κλάσεις, αλλά σπάνια χρησιμοποιείται για το σκοπό αυτό.
- Ασταθής με καλά διαχωρισμένες κατηγορίες. Η λογιστική παλινδρόμηση μπορεί να γίνει ασταθής όταν οι τάξεις είναι καλά διαχωρισμένες.
- Ασταθής με λίγα παραδείγματα. Η λογιστική παλινδρόμηση μπορεί να γίνει ασταθής όταν υπάρχουν λίγα παραδείγματα για την εκτίμηση των παραμέτρων.

Η Γραμμική Διακριτική Ανάλυση (LDA) αντιμετωπίζει κάθε ένα από αυτά τα σημεία και είναι η καταλληλότερη γραμμική μέθοδος για πολυταξικά προβλήματα κατηγοριοποίησης.

Αναπαράσταση του μοντέλου της Γραμμικής Διακριτικής Ανάλυσης

Η αναπαράσταση του μοντέλου της Γραμμικής Διακριτικής Ανάλυσης είναι αρκετά απλή. Αποτελείται από στατιστικές ιδιότητες των δεδομένων, που υπολογίζονται για κάθε κατηγορία. Για μια μεταβλητή εισόδου (x), αυτές είναι ο μέσος όρος και η διακύμανση της μεταβλητής για κάθε κατηγορία.

Για πολλαπλές μεταβλητές, αυτές είναι οι ίδιες ιδιότητες που υπολογίζονται με βάση τη πολυωνυμική Γκαουσιανή κατανομή, δηλαδή τον μέσο όρο και τον πίνακα συνδιασποράς (αυτός είναι μια πολυδιάστατη γενίκευση της διακύμανσης). Αυτές οι στατιστικές ιδιότητες υπολογίζονται από τα δεδομένα και συνδέονται με την εξίσωση LDA για να κάνουν προβλέψεις.

Το LDA σχετικά με τα δεδομένα κάνει κάποιες απλουστευτικές υποθέσεις:

- Ότι για τα δεδομένα ισχύει η Γκαουσιανή κατανομή.
- Ότι κάθε χαρακτηριστικό έχει την ίδια διακύμανση.

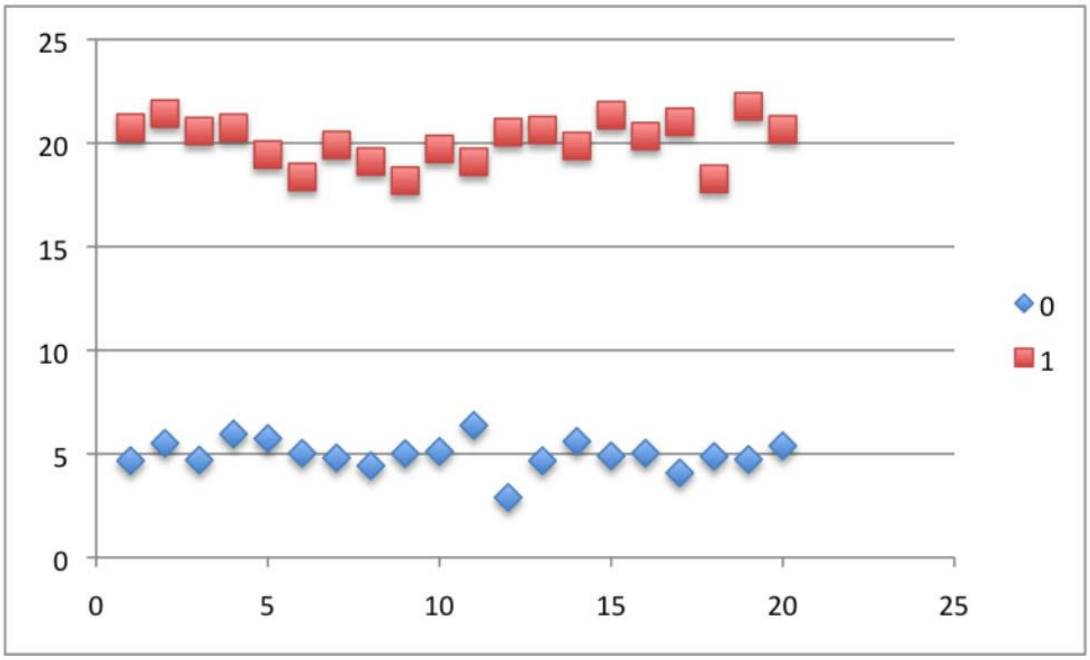
Με αυτές τις υποθέσεις, το μοντέλο LDA υπολογίζει το μέσο όρο και τη διακύμανση από τα δεδομένα για κάθε κατηγορία. Στην περίπτωση της μονοδιάστατης (μεταβλητής εισόδου) με δύο κατηγορίες, η μέση τιμή κάθε εισόδου (x) για κάθε κλάση (k) μπορεί να εκτιμηθεί με τον απλό τρόπο διαιρώντας το άθροισμα των τιμών με τον συνολικό αριθμό των τιμών, όπως φαίνεται στη παρακάτω συνάρτηση:

$$mean_k = \frac{1}{n_k} \times \sum_{i=1}^{n} x_i \qquad (6.9)$$

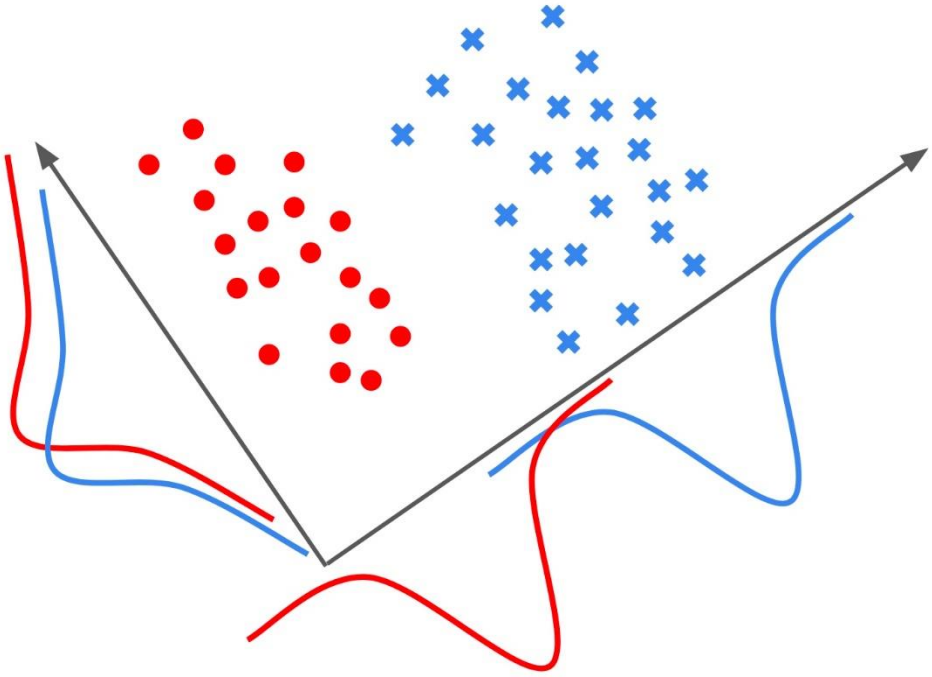
Όπου $mean_k$ είναι η μέση τιμή του x για την κλάση k, και n_k είναι ο αριθμός των στιγμιότυπων με την κλάση k. Η διακύμανση υπολογίζεται σε όλες τις κλάσεις ως η μέση τετραγωνική διαφορά κάθε τιμής από τη μέση τιμή.

$$sigma^2 = \frac{1}{n-K} \times \sum_{i=1}^{n} (x_i - mean_k)^2 \qquad (6.10)$$

Όπου το $sigma^2$ είναι η διακύμανση σε όλες τις εισόδους (x), n είναι ο αριθμός των περιπτώσεων, K είναι ο αριθμός των κλάσεων και το $mean_k$ είναι ο μέσος όρος του x για την κλάση στην οποία ανήκει το x_i. Με άλλα λόγια, υπολογίζεται η τετραγωνική διαφορά κάθε τιμής από τον μέσο όρο μέσα στις ομάδες των κατηγοριών, αλλά σαν αποτέλεσμα υπολογίζεται η μέση τιμή σε όλες τις ομάδες κατηγοριών.



Σχήμα 6.3. LDA [5].



Σχήμα 6.4. Διακύμανση των δεδομένων για κάθε κατηγορία [11].

Προετοιμασία δεδομένων για LDA

Η προετοιμασία των δεδομένων συμβάλει στην καλύτερη λειτουργία σε ένα συγκεκριμένο πρόβλημα.

Προβλήματα κατηγοριοποίησης. Το LDA προορίζεται για προβλήματα κατηγοριοποίησης, όπου η μεταβλητή εξόδου είναι κατηγορηματική και υποστηρίζει την κατηγοριοποίηση δυαδικών αλλά και πολλαπλών ομάδων.

Κατανομή Gauss. Η τυπική υλοποίηση του μοντέλου προϋποθέτει οι μεταβλητές εισόδου να ανήκουν στην κατανομή Γκάους.

Αποφυγή ακραίων τιμών. Καλό είναι να αφαιρούνται οι ακραίες τιμές από τα δεδομένα. Αυτές μπορούν να στρεβλώνουν τα βασικά στατιστικά που χρησιμοποιούνται για την κατηγοριοποίηση, όπως η μέση τιμή και η τυπική απόκλιση.

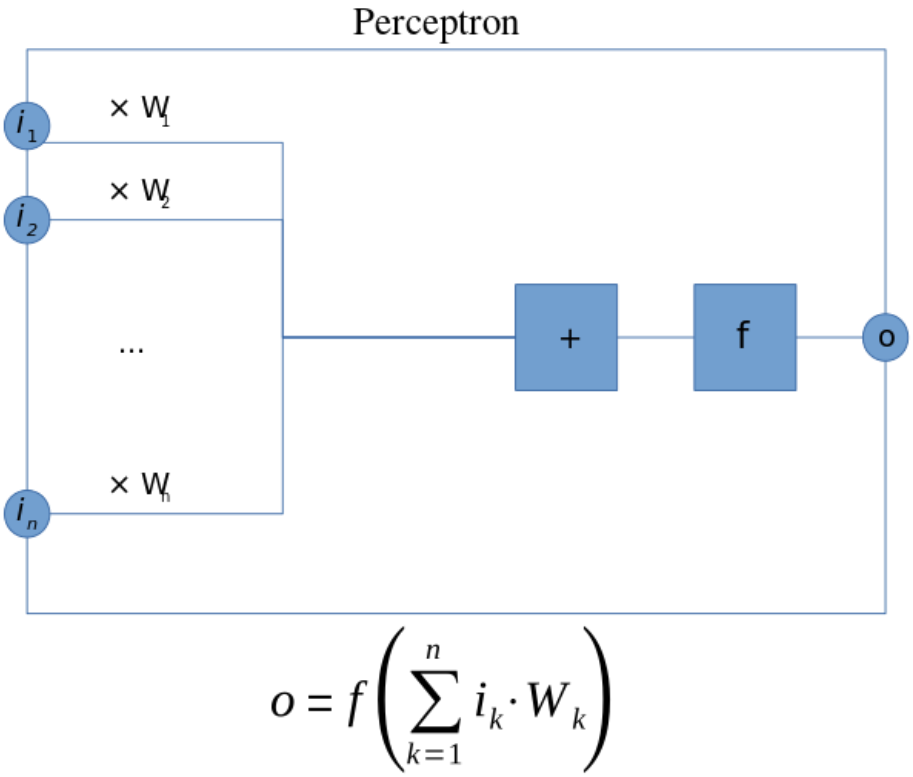
Ίδια Απόκλιση. Το LDA υποθέτει ότι κάθε μεταβλητή εισόδου έχει την ίδια διακύμανση [5].

6.2.5 Νευρώνας Perceptron (Αντίληπτρο)

Ο νευρώνας perceptron είναι ένα μονό επίπεδο νευρωνικού δικτύου και ένας perceptron πολλαπλών επιπέδων καλείται νευρωνικό δίκτυο. Ο perceptron είναι ένας γραμμικός κατηγοριοποιητής (δυαδικός), ο οποίος χρησιμοποιείται στην εποπτευόμενη μάθηση. Η συνάρτηση του, απεικονίζει την είσοδο x (ένα διάνυσμα με πραγματικές τιμές) σε μία τιμή εξόδου f (x) (μία και μοναδική δυαδική τιμή).

$$f(x) = \begin{cases} 1 \; if \; w \times x + b > 0 \\ 0 \; else \end{cases} \qquad (6.11)$$

Στη συνάρτηση (6.11), όπου w είναι ένα διάνυσμα από βάρη με πραγματικές τιμές, το x είναι διάνυσμα από εισόδους x και w × x είναι το εσωτερικό γινόμενο μεταξύ των διανυσμάτων w και x. Το b είναι το 'bias', ένας σταθερός όρος ο οποίος δεν εξαρτάται από καμία τιμή εισόδου [12][13].



Σχήμα 6.5. Η εξίσωση του perceptron. Wikipedia (2019).

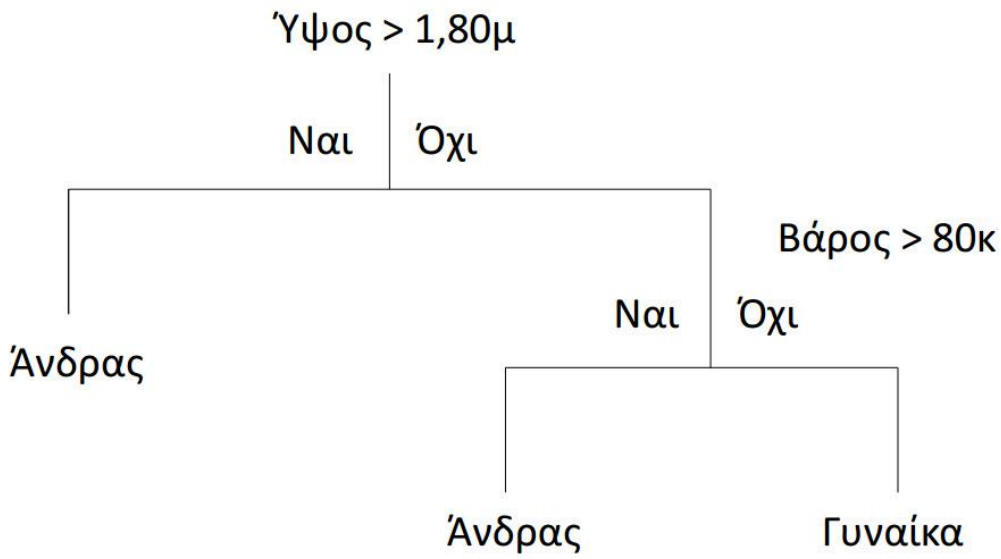
6.3 Μη Γραμμικοί Αλγόριθμοι

6.3.1 Classification and Regression Trees (CART)

Τα δέντρα κατηγοριοποίησης και παλινδρόμησης ή CART για συντομία είναι ένας όρος που εισήγαγε ο Leo Breiman για να αναφερθεί στους αλγόριθμους με δένδρο απόφασης που μπορούν να χρησιμοποιηθούν σε μοντέλα πρόβλεψης για προβλήματα κατηγοριοποίησης ή παλινδρόμησης. Αυτός ο αλγόριθμος αναφέρεται ως δέντρα απόφασης, αλλά σε ορισμένες πλατφόρμες όπως η R αναφέρονται στο πιο σύγχρονο όρο CART. Ο αλγόριθμος CART παρέχει τα θεμέλια για σημαντικούς αλγόριθμους, όπως δέντρα αποφάσεων, τυχαία δάση και ενισχυμένα δέντρα αποφάσεων.

Το μοντέλο CART

Το μοντέλο CART αναπαρίσταται ως ένα δυαδικό δέντρο. Κάθε κόμβος αντιπροσωπεύει μια μεταβλητή εισόδου (x) και ένα διαχωριστικό σημείο σε αυτήν τη μεταβλητή (υποθέτοντας ότι η μεταβλητή είναι αριθμητική). Τα φύλλα του δέντρου περιέχουν μια μεταβλητή εξόδου (y) που χρησιμοποιείται για την πρόβλεψη. Δεδομένου ενός συνόλου δεδομένων με δύο εισροές ύψους σε εκατοστά και το βάρος σε κιλά, η πρόγνωση φύλου ως αρσενικό ή θηλυκό, παρακάτω είναι ένα παράδειγμα ενός δένδρου δυαδικής απόφασης (εντελώς πλασματικό για λόγους επίδειξης μόνο) [5].



Σχήμα 6.5 CART [5].

6.3.2 Naive Bayes

Στη μηχανική μάθηση είναι πολύ σημαντικό να γίνει επιλογή της καλύτερης υπόθεσης (h) δεδομένων (d). Σε ένα πρόβλημα κατηγοριοποίησης, όπως στη συναισθηματική ανάλυση, η υπόθεση (h) μπορεί να είναι η τάξη που θα εκχωρηθούν τα νέα δεδομένα (d). Το Θεώρημα του Bayes παρέχει έναν τρόπο με τον οποίο μπορεί να υπολογιστεί η πιθανότητα μιας υπόθεσης δεδομένης της προηγούμενης γνώσης. Το Θεώρημα του Bayes αναφέρεται ως:

$$P(d|h) = \frac{P(d|h) \times P(h)}{P(d)} \quad (6.12)$$

Όπου:

- P(h|d) είναι η πιθανότητα της υπόθεσης h που δίνεται στα δεδομένα d. Αυτό ονομάζεται οπίσθια πιθανότητα.
- P(d|h) είναι η πιθανότητα των δεδομένων d δεδομένου ότι η υπόθεση h ήταν αληθής.
- P(h) είναι η πιθανότητα της υπόθεσης h να είναι αληθής (ανεξάρτητα από τα δεδομένα). Αυτό ονομάζεται η προηγούμενη πιθανότητα του h.
- P(d) είναι η πιθανότητα των δεδομένων (ανεξάρτητα από την υπόθεση)

Μπορεί να υπολογιστεί η οπίσθια πιθανότητα του P(h|d) από την προηγούμενη πιθανότητα P(h) με P(d) και P(d|h). Μετά τον υπολογισμό της οπίσθιας πιθανότητας για μια σειρά διαφορετικών υποθέσεων, μπορεί να επιλεγεί η υπόθεση με την υψηλότερη πιθανότητα. Αυτή είναι η μέγιστη πιθανή υπόθεση και μπορεί τυπικά να ονομάζεται η μέγιστη υπόθεση εκ των υστέρων (MAP). Αυτό μπορεί να γραφτεί ως εξής:

$$MAP(h) = max\big(P(h|d)\big)$$
$$MAP(h) = max\left(\frac{P(d|h) \times P(h)}{P(d)}\right) \quad (6.1)$$
$$MAP(h) = \max(P(h|d) \times P(h))$$

Κατηγοριοποιητής Naive Bayes

Ο Naive Bayes είναι ένας αλγόριθμος κατηγοριοποίησης για δυαδικά (δύο κατηγοριών) και πολυταξικά προβλήματα κατηγοριοποίησης. Η τεχνική είναι ευκολότερη στην κατανόηση όταν περιγράφεται χρησιμοποιώντας δυαδικές ή κατηγορικές τιμές εισόδου. Ονομάζεται αφελής Bayes επειδή ο υπολογισμός των πιθανοτήτων για κάθε υπόθεση είναι απλοποιημένος για να κάνει τον υπολογισμό τους ελκυστικό. Αντί να υπολογιστούν οι τιμές κάθε χαρακτηριστικού P (d1, d2, d3 | h), θεωρείται ότι είναι υπό όρους ανεξάρτητες δεδομένης της τιμής στόχου και υπολογίζονται ως P (d1 | h) × P (d2 | h) και ούτω καθεξής. Αυτή είναι μια πολύ ισχυρή υπόθεση, που είναι πολύ απίθανη σε πραγματικά δεδομένα, δηλαδή ότι τα χαρακτηριστικά δεν αλληλοεπιδρούν. Παρ 'όλα αυτά, η προσέγγιση λειτουργεί εκπληκτικά καλά σε δεδομένα όπου αυτή η υπόθεση δεν ισχύει.

Gaussian Naive Bayes

Ο Naive Bayes μπορεί να επεκταθεί σε πραγματικά αξιόπιστα χαρακτηριστικά, συνηθέστερα υποθέτοντας Γκαουσιανή (Gaussian) κατανομή. Αυτή η επέκταση των Naive Bayes ονομάζεται Gaussian Naive Bayes. Μπορούν να χρησιμοποιηθούν και άλλες λειτουργίες για την εκτίμηση της κατανομής των δεδομένων, αλλά η Gaussian (ή κανονική κατανομή) είναι η ευκολότερη, επειδή χρειάζεται μόνο να υπολογιστεί η μέση και η τυπική απόκλιση από τα δεδομένα εκπαίδευσης.

Με πραγματικές τιμές εισόδου, μπορεί να υπολογιστεί η μέση και τυπική απόκλιση των τιμών εισόδου (x) για κάθε κλάση. Αυτό σημαίνει ότι εκτός από τις πιθανότητες για κάθε κλάση, πρέπει επίσης να αποθηκευτούν οι μέσες και τυπικές αποκλίσεις της κάθε μεταβλητής εισόδου για κάθε κλάση.

Ο υπολογισμός της μέσης τιμής κάθε μεταβλητής εισόδου (x) για κάθε τιμή κλάσης γίνεται:

$$mean(x) = \frac{1}{n} \times \sum_{i=1}^{n} x_i \qquad (6.\ 2)$$

Όπου n είναι ο αριθμός των στιγμιότυπων και το x είναι οι τιμές μιας μεταβλητής εισόδου στα δεδομένα εκπαίδευσης. Η τυπική απόκλιση υπολογίζεται χρησιμοποιώντας την ακόλουθη εξίσωση:

$$StandardDeviation(x)\sqrt{\frac{1}{n} \times \sum_{i=1}^{n}(x_i - mean(x))^2} \quad (6.\ 3)$$

Προετοιμασία των δεδομένων για Naive Bayes

Γκαουσιανές (Gaussian) εισροές. Αν οι μεταβλητές εισόδου είναι πραγματικές τιμές, θεωρείται μια Γκαουσιανή κατανομή. Σε αυτήν την περίπτωση ο αλγόριθμος θα έχει καλύτερες επιδόσεις εάν οι μονοδιάστατες κατανομές των δεδομένων είναι Γκαουσιανές ή κοντά στις Γκαουσιανές. Αυτό μπορεί να απαιτεί την αφαίρεση των ακραίων τιμών (π.χ. τιμές που είναι περισσότερες από 3 ή 4 τυπικές αποκλίσεις από τη μέση τιμή).

Προβλήματα κατηγοριοποίησης. Ο Naive Bayes είναι ένας αλγόριθμος κατηγοριοποίησης κατάλληλος για δυαδική και πολυταξική κατηγοριοποίηση.

Λογαριθμικές πιθανότητες. Ο υπολογισμός της πιθανότητας διαφορετικών κλάσεων συνεπάγεται πολλαπλασιασμό πολλών μικρών αριθμών μαζί. Αυτό μπορεί να οδηγήσει σε υποβιβασμό αριθμητικής ακρίβειας. Ως εκ τούτου, είναι καλή πρακτική η χρήση ενός λογαριθμικού μετασχηματισμού των πιθανοτήτων για να αποφευχθεί αυτός ο υποβιβασμός.

Μέθοδος πυρήνα. Αντί να υποθέσουμε μια Γκαουσιανή κατανομή για αριθμητικές τιμές εισόδου, μπορούν να χρησιμοποιηθούν πιο σύνθετες κατανομές.

Πιθανότητες ενημέρωσης. Όταν υπάρχουν νέα δεδομένα, οι πιθανότητες του μοντέλου μπορούν να ενημερώνονται. Αυτό μπορεί να βοηθήσει εάν τα δεδομένα αλλάζουν συχνά [5].

6.3.3 K-Nearest Neighbors (KNN)

Το μοντέλο KNN

Η αναπαράσταση του μοντέλου για το KNN είναι το σύνολο των δεδομένων εκπαίδευσης. Το KNN δεν έχει άλλο μοντέλο εκτός από την αποθήκευση ολόκληρου του συνόλου δεδομένων, οπότε δεν απαιτείται μάθηση. Οι αποδοτικές εφαρμογές μπορούν να αποθηκεύσουν τα δεδομένα χρησιμοποιώντας σύνθετες δομές δεδομένων, όπως k-d δέντρα για να κάνουν την αναζήτηση και την αντιστοίχιση νέων μοτίβων κατά την πρόβλεψη, αποτελεσματική. Επειδή αποθηκεύεται όλο το σύνολο δεδομένων εκπαίδευσης, είναι απαραίτητη η συνέπεια των δεδομένων εκπαίδευσης καθώς και να ενημερώνονται συχνά τα νέα δεδομένα όταν είναι διαθέσιμα και να αφαιρούνται τα εσφαλμένα και ακραία δεδομένα.

Πραγματοποίηση προβλέψεων με το KNN

Το KNN κάνει προβλέψεις χρησιμοποιώντας άμεσα το σύνολο δεδομένων κατάρτισης. Οι προβλέψεις γίνονται για ένα νέο σημείο δεδομένων, αναζητώντας ολόκληρο το σετ εκπαίδευσης για τις παρόμοιες περιπτώσεις K (τους γείτονες) και συνοψίζοντας τη μεταβλητή εξόδου για αυτές τις περιπτώσεις K.

Για να προσδιοριστεί ποια από τις περιπτώσεις K στο σύνολο δεδομένων κατάρτισης είναι πιο παρόμοια με μια νέα είσοδο, χρησιμοποιείται ένα μέτρο απόστασης. Για πραγματικές μεταβλητές εισόδου, το πιο δημοφιλές μέτρο απόστασης είναι η ευκλείδεια απόσταση. Η ευκλείδεια απόσταση υπολογίζεται ως η τετραγωνική ρίζα του αθροίσματος των τετραγωνικών διαφορών μεταξύ ενός σημείου α και του σημείου β σε όλα τα χαρακτηριστικά εισόδου i.

$$EuclideanDistance(a,b) = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2} \quad (6.4)$$

Άλλα δημοφιλή μέτρα για την απόσταση περιλαμβάνουν:

- Απόσταση Hamming: Υπολογίζει την απόσταση μεταξύ δυαδικών διανυσμάτων.

- Απόσταση Manhattan: Υπολογίζει την απόσταση μεταξύ πραγματικών διανυσμάτων χρησιμοποιώντας το άθροισμα της απόλυτης διαφοράς τους. Ονομάζεται επίσης απόσταση City Block.
- Απόσταση Minkowski: Γενίκευση της Ευκλείδειας και της Μανχάταν απόστασης.

Υπάρχουν πολλά άλλα μέτρα απόστασης που μπορούν να χρησιμοποιηθούν, όπως η απόσταση Tanimoto, Jaccard, Mahalanobis και το συνημίτονο. Μπορεί να επιλεγεί η καλύτερη μετρική απόσταση βάσει των ιδιοτήτων των δεδομένων. Η Ευκλείδεια είναι ένα καλό μέτρο απόστασης για χρήση εάν οι μεταβλητές εισόδου είναι παρόμοιου τύπου (π.χ. όλα τα μετρημένα πλάτη και ύψη). Η απόσταση Μανχάταν είναι ένα καλό μέτρο για να χρησιμοποιηθεί αν οι μεταβλητές εισόδου δεν είναι παρόμοιου τύπου (όπως ηλικία, φύλο, ύψος κλπ.).

Η υπολογιστική πολυπλοκότητα του KNN αυξάνεται με το μέγεθος του συνόλου δεδομένων εκπαίδευσης. Για πολύ μεγάλα σετ εκπαίδευσης, το KNN μπορεί να γίνει στοχαστικό, λαμβάνοντας ένα δείγμα από το σύνολο δεδομένων κατάρτισης από το οποίο υπολογίζει τις K-παρόμοιες περιπτώσεις. Το KNN υπάρχει εδώ και πολύ καιρό και έχει μελετηθεί πολύ καλά. Ως εκ τούτου, το KNN σε διαφορετικούς κλάδους χρησιμοποιείται με διαφορετικά ονόματα, για παράδειγμα:

- Εκμάθηση βασισμένη σε περιστατικά (Instance-Based Learning): Οι πρώτες περιπτώσεις κατάρτισης χρησιμοποιούνται για πρόβλεψη. Ως τέτοιος ο KNN αναφέρεται συχνά ως εκμάθηση βασισμένη σε περιστατικά ή ως case-based learning (όπου κάθε εκπαιδευτικό στοιχείο είναι μια περίπτωση από τον τομέα του προβλήματος).
- Οκνηρή εκμάθηση (Lazy Learning): Δεν απαιτείται εκμάθηση του μοντέλου και όλη η εργασία συμβαίνει τη στιγμή που ζητείται η πρόβλεψη. Ως εκ τούτου, ο KNN αναφέρεται συχνά ως τεμπέλης αλγόριθμος εκμάθησης.
- Μη παραμετρικό: Ο KNN δεν κάνει υποθέσεις σχετικά με τη λειτουργική μορφή του προβλήματος που επιλύεται. Ως τέτοιος ο KNN αναφέρεται ως μη παραμετρικός αλγόριθμος μηχανικής εκμάθησης. Επίσης, μπορεί να χρησιμοποιηθεί για προβλήματα παλινδρόμησης και κατηγοριοποίησης.

Προετοιμασία των δεδομένων για KNN

Αναπροσαρμογή δεδομένων. Ο KNN αποδίδει πολύ καλύτερα εάν όλα τα δεδομένα έχουν την ίδια κλίμακα. Η κανονικοποίηση των δεδομένων στο εύρος μεταξύ 0 και 1 είναι μια καλή ιδέα. Μπορεί επίσης τα δεδομένα να κανονικοποιηθούν αν έχουν Γκαουσιανή κατανομή.

Αντιμετώπιση δεδομένων που λείπουν. Όταν λείπουν δεδομένα σημαίνει ότι η απόσταση μεταξύ των δειγμάτων δεν μπορεί να υπολογιστεί. Αυτά τα δείγματα θα μπορούσαν να αποκλειστούν ή θα μπορούσαν να εισαχθούν οι τιμές που λείπουν.

Χαμηλότερη διαστατικότητα. Ο KNN είναι κατάλληλος για δεδομένα χαμηλότερων διαστάσεων. Σε μεγάλης διαστατικότητας δεδομένα (εκατοντάδες ή χιλιάδες μεταβλητές εισόδου) μπορεί να μην λειτουργεί όπως άλλες τεχνικές. Ο KNN μπορεί να επωφεληθεί από την επιλογή λειτουργιών που μειώνουν τη διαστασιοποίηση του χώρου των χαρακτηριστικών εισόδου [5].

6.3.4 Εκπαιδευόμενος Διανυσματικός Κβαντιστής (Learning Vector Quantization - LVQ)

Η αναπαράσταση για το LVQ είναι μια συλλογή από διανύσματα λεξικών (codebook vectors). Το LVQ αναπτύχθηκε και χρησιμοποιείται καλύτερα ως αλγόριθμος κατηγοριοποίησης. Υποστηρίζει τόσο τα δυαδικά (δύο τάξεων) όσο και τα πολυταξικά προβλήματα κατηγοριοποίησης. Ένα διάνυσμα λεξικών είναι ένας κατάλογος αριθμών που έχουν τα ίδια χαρακτηριστικά εισόδου και εξόδου όπως τα δεδομένα εκπαίδευσης. Για παράδειγμα, εάν το πρόβλημα είναι μια δυαδική κατηγοριοποίηση με τάξεις 0 και 1 για τις εισόδους πλάτος μήκος και ύψος, τότε ένα διάνυσμα λεξικών θα αποτελείται από τα τέσσερα χαρακτηριστικά: πλάτος, μήκος, ύψος και κλάση. Η αναπαράσταση του μοντέλου είναι μια σταθερή ομάδα διανυσμάτων λεξικών, που αντλήθηκαν από τα δεδομένα εκπαίδευσης.

Μοιάζουν με περιπτώσεις κατάρτισης, αλλά οι τιμές κάθε χαρακτηριστικού έχουν προσαρμοστεί με βάση τη διαδικασία εκμάθησης. Στη γλώσσα των νευρωνικών δικτύων, κάθε διάνυσμα λεξικών μπορεί να ονομάζεται νευρώνας, κάθε χαρακτηριστικό ονομάζεται βάρος και η συλλογή των διανυσμάτων λεξικών ονομάζεται δίκτυο.

Οι προβλέψεις που γίνονται χρησιμοποιώντας τα διανύσματα λεξικών LVQ γίνονται με τον ίδιο τρόπο όπως οι K-Nearest Neighbours. Οι προβλέψεις για μια νέα περίπτωση γίνονται κάνοντας αναζήτηση σε όλα τα διανύσματα λεξικών για τις παρόμοιες περιπτώσεις K και συνοψίζοντας τη μεταβλητή εξόδου για αυτές τις περιπτώσεις K. Για την κατηγοριοποίηση, αυτή είναι η τιμή της κλάσης. Συνήθως οι προβλέψεις γίνονται με το K = 1 και το διάνυσμα λεξικών που ταιριάζει ονομάζεται Best Matching Unit (BMU).

Για να προσδιοριστεί ποια από τις περιπτώσεις K στο σύνολο δεδομένων κατάρτισης είναι παρόμοια με μια νέα είσοδο, χρησιμοποιείται ένα μέτρο απόστασης. Για πραγματικές μεταβλητές εισόδου, το πιο δημοφιλές μέτρο απόστασης είναι η ευκλείδεια απόσταση. Η ευκλείδεια απόσταση υπολογίζεται ως η τετραγωνική ρίζα του αθροίσματος των τετραγωνικών διαφορών μεταξύ ενός σημείου α και του σημείου β σε όλα τα χαρακτηριστικά εισόδου i.

$$EuclideanDistance(a,b) = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2} \quad (6.5)$$

Προετοιμασία δεδομένων για LVQ

Γενικά, η προετοιμασία των δεδομένων για τον LVQ γίνεται όπως και για τον K-Nearest Neighbours.

Κατηγοριοποίηση. Ο LVQ είναι ένας αλγόριθμος κατηγοριοποίησης που λειτουργεί τόσο για δυαδικά (δύο τάξεις) όσο και για πολυταξικά προβλήματα.

Πολλαπλές διελεύσεις (Multiple-Passes). Μια καλή τεχνική με τον LVQ είναι η εκτέλεση πολλαπλών περασμάτων του συνόλου δεδομένων εκπαίδευσης μέσω των διανυσμάτων λεξικών (π.χ. πολλαπλές διαδικασίες εκμάθησης). Το πρώτο πέρασμα με υψηλότερο ρυθμό εκμάθησης για την εγκατάσταση της ομάδας των διανυσμάτων λεξικών και το δεύτερο με ένα μικρότερο ρυθμό εκμάθησης για την καλύτερη ρύθμιση των διανυσμάτων.

Πολλαπλές βέλτιστες αντιστοιχίες. Οι επεκτάσεις του LVQ επιλέγουν πολλαπλές βέλτιστες μονάδες που ταιριάζουν, ώστε να τροποποιηθούν κατά τη διάρκεια της μάθησης. Άλλες επεκτάσεις χρησιμοποιούν ένα προσαρμοσμένο ρυθμό εκμάθησης για κάθε διάνυσμα. Αυτές οι επεκτάσεις μπορούν να βελτιώσουν τη διαδικασία εκμάθησης.

Κανονικοποίηση εισόδων. Παραδοσιακά, οι είσοδοι είναι κανονικοποιημένες (ανακατανεμημένες) σε τιμές μεταξύ 0 και 1. Αυτό είναι για να αποφευχθεί η κυριαρχία ενός χαρακτηριστικού στο μέτρο απόστασης. Εάν τα δεδομένα εισόδου είναι κανονικοποιημένα, τότε οι αρχικές τιμές για τα διανύσματα λεξικών μπορούν να επιλεγούν ως τυχαίες τιμές μεταξύ 0 και 1.

Επιλογή χαρακτηριστικών. Η επιλογή χαρακτηριστικών που μπορεί να μειώσει τη διαστασιολογία των μεταβλητών εισόδου μπορεί να βελτιώσει την ακρίβεια της μεθόδου. Ο LVQ πάσχει από το ίδιο πρόβλημα της διαστατικότητας, όπως ο K-Nearest Neighbours [5].

6.3.5 Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines - SVM)

Οι Μηχανές Διανυσμάτων Υποστήριξης, χάριν συντομίας SVM, είναι ίσως ένας από τους πιο δημοφιλείς αλγόριθμους μηχανικής μάθησης. Ήταν εξαιρετικά δημοφιλής την εποχή που αναπτύχθηκε τη δεκαετία του 1990 και εξακολουθεί να είναι η μέθοδος που προτιμάται από πολλούς καθώς παράγει σημαντική ακρίβεια με λιγότερη υπολογιστική ισχύ. Το SVM, μπορεί να χρησιμοποιηθεί τόσο για προβλήματα παλινδρόμησης όσο και κατηγοριοποίησης, αλλά χρησιμοποιείται ευρέως σε θέματα κατηγοριοποίησης.

Ο στόχος του αλγόριθμου SVM είναι να βρεθεί ένα υπερεπίπεδο σε ένα χώρο N-διαστάσεων (N - ο αριθμός των χαρακτηριστικών), που κατηγοριοποιεί ευδιάκριτα τα σημεία δεδομένων.

Κατηγοριοποιητής μέγιστου περιθωρίου

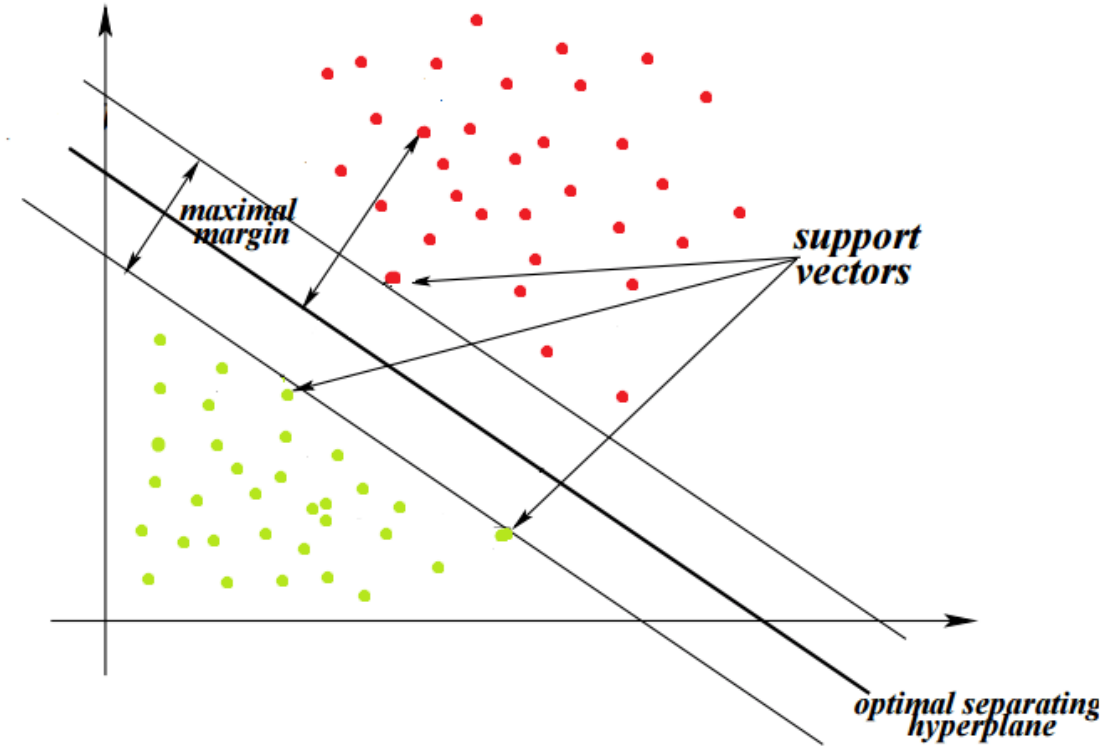
Ο κατηγοριοποιητής μέγιστου περιθωρίου είναι ένας υποθετικός κατηγοριοποιητής που εξηγεί καλύτερα πώς λειτουργεί το SVM στην πράξη. Οι αριθμητικές μεταβλητές εισόδου (x) στα δεδομένα (οι στήλες) σχηματίζουν έναν n-διαστάσεων χώρο. Για παράδειγμα, εάν είχαμε δύο μεταβλητές εισόδου, αυτό θα αποτελούσε έναν δισδιάστατο χώρο. Ένα υπερεπίπεδο είναι μια γραμμή που χωρίζει τον χώρο μεταβλητών εισόδου. Στο SVM, επιλέγεται ένα υπερεπίπεδο για να διαχωρίσει καλύτερα τα σημεία στο χώρο εισαγωγής μεταβλητών από την τάξη τους είτε στην κλάση 0 είτε στην κλάση 1. Σε δύο διαστάσεις μπορεί να απεικονιστεί ως γραμμή υποθέτοντας ότι όλα τα σημεία εισόδου μπορούν να διαχωρίζονται πλήρως από αυτή τη γραμμή. Για παράδειγμα:

$$B0 + (B1 \times X1) + (B2 \times X2) = 0 \qquad (6.\ 6)$$

Όπου οι συντελεστές B1 και B2 που καθορίζουν την κλίση της γραμμής και το σημείο τομής B0 εντοπίζονται από τον αλγόριθμο εκμάθησης και τα X1 και X2 είναι οι δύο μεταβλητές εισόδου. Μπορούν να γίνουν κατηγοριοποιήσεις χρησιμοποιώντας αυτή τη γραμμή. Εισάγοντας τις τιμές εισόδου στην εξίσωση γραμμής, μπορεί να υπολογιστεί εάν ένα νέο σημείο είναι πάνω ή κάτω από τη γραμμή.

- Πάνω από τη γραμμή, η εξίσωση επιστρέφει μια τιμή μεγαλύτερη από 0 και το σημείο ανήκει στην πρώτη τάξη (κλάση 0).
- Κάτω από τη γραμμή, η εξίσωση επιστρέφει μια τιμή μικρότερη από 0 και το σημείο ανήκει στη δεύτερη τάξη (κλάση 1).
- Μια τιμή κοντά στη γραμμή επιστρέφει μια τιμή κοντά στο μηδέν και το σημείο μπορεί να είναι δύσκολο να κατηγοριοποιηθεί.
- Εάν το μέγεθος της τιμής είναι μεγάλο, το μοντέλο έχει μεγαλύτερη αξιοπιστία στην πρόβλεψη.

Η απόσταση μεταξύ της γραμμής και των πλησιέστερων σημείων δεδομένων αναφέρεται ως περιθώριο. Η καλύτερη ή βέλτιστη γραμμή που μπορεί να χωρίσει τις δύο κατηγορίες είναι η γραμμή που έχει το μεγαλύτερο περιθώριο. Αυτό ονομάζεται μέγιστο περιθώριο υπερεπιπέδου. Το περιθώριο υπολογίζεται ως η κάθετη απόσταση από τη γραμμή μόνο στα πλησιέστερα σημεία. Μόνο αυτά τα σημεία είναι σημαντικά για τον ορισμό της γραμμής και την κατασκευή του κατηγοριοποιητή. Αυτά τα σημεία καλούνται διανύσματα υποστήριξης. Υποστηρίζουν ή καθορίζουν το υπερεπίπεδο που υπολογίζεται από τα δεδομένα εκπαίδευσης χρησιμοποιώντας μια διαδικασία βελτιστοποίησης που μεγιστοποιεί το περιθώριο.



Σχήμα 6.6. Κατηγοριοποιητής μέγιστου περιθωρίου. hackerearth (2017).

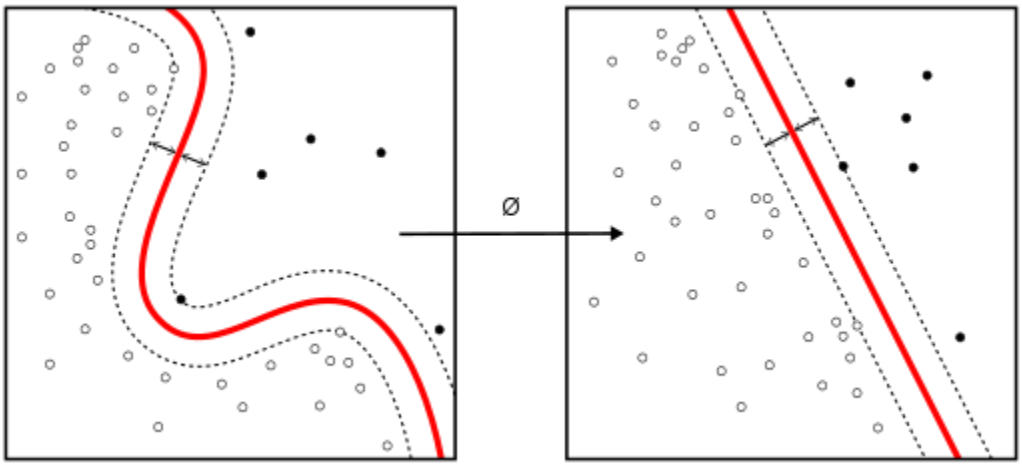
Κατηγοριοποιητής ήπιου περιθωρίου

Στην πράξη, τα πραγματικά δεδομένα είναι ακατάστατα και δεν μπορούν να χωριστούν τέλεια με ένα υπερεπίπεδο. Ο περιορισμός μεγιστοποίησης του περιθωρίου της γραμμής που χωρίζει τις κλάσεις πρέπει να είναι χαλαρός. Αυτό ονομάζεται συχνά κατηγοριοποιητής ήπιου περιθωρίου. Αυτή η αλλαγή επιτρέπει ορισμένα σημεία στα δεδομένα εκπαίδευσης να παραβιάζουν τη διαχωριστική γραμμή. Δημιουργείται ένα επιπλέον σύνολο συντελεστών που δίνουν το περιθώριο περιθωριοποίησης σε κάθε διάσταση. Αυτοί οι συντελεστές μερικές φορές ονομάζονται ήπιες μεταβλητές. Αυτό αυξάνει την πολυπλοκότητα του μοντέλου, καθώς υπάρχουν περισσότεροι παράμετροι για το μοντέλο ώστε να ταιριάζει στα δεδομένα για να παρέχει αυτή την πολυπλοκότητα.

Εισάγεται μια παράμετρος ρύθμισης που ονομάζεται C που ορίζει το μέγεθος της επιτρεπόμενης διαδρομής σε όλες τις διαστάσεις. Οι παράμετροι C ορίζουν το μέγεθος της παραβίασης του επιτρεπόμενου περιθωρίου. Ένα C = 0 δεν αποτελεί παραβίαση και επιστρέφουμε στον κατηγοριοποιητή μέγιστου περιθωρίου που περιγράφεται παραπάνω.

Όσο μεγαλύτερη είναι η τιμή του C, τόσο περισσότερες παραβιάσεις του υπερεπιπέδου επιτρέπονται. Κατά την εκμάθηση του υπερεπιπέδου από τα δεδομένα, όλες οι περιπτώσεις εκπαίδευσης που βρίσκονται εντός του περιθωρίου θα επηρεάσουν την τοποθέτηση του υπερεπιπέδου και αναφέρονται ως διανύσματα υποστήριξης. Καθώς το C επηρεάζει τον αριθμό των περιπτώσεων που επιτρέπεται να πέσουν εντός του περιθωρίου, το C επηρεάζει επίσης και τον αριθμό των διανυσμάτων υποστήριξης που χρησιμοποιούνται από το μοντέλο.

- Όσο μικρότερη είναι η τιμή του C, τόσο πιο ευαίσθητος είναι ο αλγόριθμος στα δεδομένα εκπαίδευσης (υψηλότερη διακύμανση και ακριβέστερη πρόβλεψη).
- Όσο μεγαλύτερη είναι η τιμή του C, τόσο λιγότερο ευαίσθητος είναι ο αλγόριθμος στα δεδομένα εκπαίδευσης (χαμηλότερη διακύμανση και ανακριβής πρόβλεψη).



Εικόνα 6.7 Μη γραμμικό SVM. wikipedia (2019).

Προετοιμασία δεδομένων για SVM

Αριθμητικές εισόδους. Το SVM υποθέτει ότι οι είσοδοι είναι αριθμητικές τιμές. Εάν υπάρχουν κατηγορικές είσοδοι, ίσως χρειαστεί να μετατραπούν σε δυαδικές εικονικές μεταβλητές (μία μεταβλητή για κάθε κατηγορία).

Δυαδική κατηγοριοποίηση. Το βασικό SVM για δυαδικά (δύο τάξεων) ταξινομικά προβλήματα, αν και έχουν αναπτυχθεί επεκτάσεις για παλινδρόμηση και πολυταξική κατηγοριοποίηση[5].

6.4 Συλλογικοί αλγόριθμοι (Ensemble Algorithms)

6.4.1 Bagging και Random Forest

Το Random Forest είναι ένας από τους πιο δημοφιλείς και ισχυρότερους αλγόριθμους μάθησης. Ανήκει στην κατηγορία συλλογικών αλγορίθμων μηχανικής μάθησης και ονομάζεται Bootstrap Aggregation ή bagging.

Μέθοδος Bootstrap

Το bootstrap είναι μια ισχυρή στατιστική μέθοδος για την εκτίμηση μιας ποσότητας από ένα δείγμα δεδομένων. Αυτό είναι ευκολότερο να κατανοηθεί εάν η ποσότητα είναι ένα περιγραφικό στατιστικό στοιχείο, όπως μια μέση τιμή ή μια τυπική απόκλιση. Για μία εκτίμηση του μέσου όρου από ένα υποθετικό δείγμα 100 τιμών (x), ο υπολογισμός γίνεται ως εξής:

$$mean(x) = \frac{1}{100} \times \sum_{i=1}^{100} x_i \qquad (6.\ 19)$$

Γνωρίζοντας ότι το δείγμα είναι μικρό και ότι ο μέσος όρος είναι λάθος, μπορεί να βελτιωθεί η εκτίμηση της μέσης τιμής χρησιμοποιώντας τη διαδικασία bootstrap:

- Δημιουργώντας πολλά (π.χ. 1000) τυχαία υποδείγματα του συνόλου δεδομένων με αντικατάσταση (δηλαδή μπορεί να επιλεγεί την ίδια τιμή πολλές φορές).
- Υπολογίζοντας τη μέση τιμή για κάθε υπο-δείγμα.
- Υπολογίζοντας το μέσο όρο όλων των συλλεγόμενων μέσων τιμών και χρησιμοποιώντας τον ως εκτιμώμενο μέσο όρο για τα δεδομένα.

Αυτή η διαδικασία μπορεί να χρησιμοποιηθεί για την εκτίμηση άλλων ποσοτήτων όπως η τυπική απόκλιση, ακόμη και των ποσοτήτων που χρησιμοποιούνται στους αλγόριθμους μηχανικής μάθησης.

Bootstrap Aggregation (Bagging)

Το Bootstrap Aggregation (ή το Bagging για συντομία) είναι μια απλή και πολύ ισχυρή συλλογική (ensemble) μέθοδος. Μια συλλογική μέθοδος είναι μια τεχνική που συνδυάζει τις προβλέψεις από πολλούς αλγορίθμους μηχανικής μάθησης μαζί για να κάνει ακριβέστερες προβλέψεις από οποιοδήποτε μεμονωμένο μοντέλο. Το Bagging είναι μια γενική διαδικασία που μπορεί να χρησιμοποιηθεί για να μειώσει τη διακύμανση για κάποιους αλγόριθμους. Ένας αλγόριθμος που έχει μεγάλη διακύμανση είναι τα δέντρα απόφασης, όπως το CART.

Τα δέντρα απόφασης είναι ευαίσθητα στα συγκεκριμένα δεδομένα στα οποία έχουν εκπαιδευτεί. Εάν αλλάξουν τα δεδομένα εκπαίδευσης (π.χ. ένα δέντρο εκπαιδεύεται σε ένα υποσύνολο των δεδομένων εκπαίδευσης), το δέντρο απόφασης που θα προκύψει μπορεί να είναι αρκετά διαφορετικό και με τη σειρά τους οι προβλέψεις μπορεί να είναι τελείως διαφορετικές. Το bagging είναι η εφαρμογή της διαδικασίας Bootstrap σε έναν αλγόριθμο μηχανικής μάθησης υψηλής διακύμανσης, συνήθως δέντρα αποφάσεων. Υποθέτοντας ένα σύνολο δεδομένων 1000 περιπτώσεων και χρησιμοποιώντας τον αλγόριθμο CART, το bagging του αλγορίθμου CART θα λειτουργήσει ως εξής:

- Δημιουργία πολλών (π.χ. 100) τυχαίων υποδειγμάτων του συνόλου δεδομένων με αντικατάσταση.
- Εκπαίδευση ενός μοντέλου CART σε κάθε δείγμα.
- Με δεδομένο ένα νέο σύνολο δεδομένων, υπολογίζεται η μέση πρόβλεψη από κάθε μοντέλο.

Τυχαία Δάση (Random Forest)

Ο Random Forest είναι μια βελτίωση των «bagged» δέντρων αποφάσεων. Ένα πρόβλημα με τα δέντρα αποφάσεων όπως το CART είναι ότι επιλέγουν ποια μεταβλητή θα χωριστεί (split) χρησιμοποιώντας έναν άπληστο αλγόριθμο που ελαχιστοποιεί το σφάλμα. Ως εκ τούτου, ακόμη και με το Bagging, τα δέντρα αποφάσεων μπορούν να έχουν πολλές δομικές ομοιότητες και με τη σειρά τους να οδηγήσουν σε υψηλή συσχέτιση στις προβλέψεις τους. Ο συνδυασμός προβλέψεων από πολλαπλά συλλογικά μοντέλα λειτουργεί καλύτερα εάν οι προβλέψεις από τα υπο-μοντέλα είναι ασυσχέτιστες ή στην καλύτερη περίπτωση συσχετίζονται ελάχιστα.

Ο Random Forest αλλάζει τον αλγόριθμο για τον τρόπο με τον οποίο μαθαίνουν τα υπο-δέντρα, έτσι ώστε οι προβλέψεις που προκύπτουν από όλα τα υπο-δέντρα να έχουν μικρότερη συσχέτιση. Στο CART, κατά την επιλογή ενός σημείου διαχωρισμού, ο αλγόριθμος μάθησης επιτρέπεται να εξετάσει όλες τις μεταβλητές για να επιλέξει το βέλτιστο διαχωριστικό σημείο. Ο αλγόριθμος του Random Forest αλλάζει αυτή τη διαδικασία έτσι ώστε ο αλγόριθμος εκμάθησης να περιορίζεται σε ένα τυχαίο δείγμα χαρακτηριστικών που μπορεί να αναζητήσει. Ο αριθμός των χαρακτηριστικών που μπορούν να αναζητηθούν σε κάθε σημείο διαίρεσης (m) πρέπει να οριστεί ως παράμετρος του αλγορίθμου.

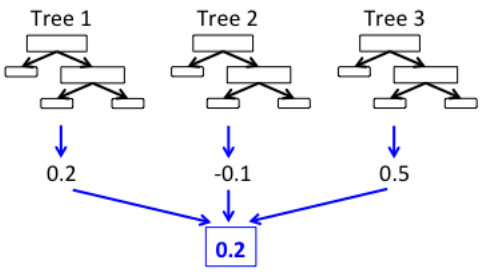
Για την κατηγοριοποίηση μια καλή προεπιλογή είναι: $m = \sqrt{p}$ *(6. 20)*

Για την παλινδρόμηση μια καλή προεπιλογή είναι: $m = \frac{p}{3}$ *(6. 21)*

Όπου m είναι ο αριθμός των τυχαία επιλεγμένων χαρακτηριστικών που μπορούν να αναζητηθούν σε ένα διαχωριστικό σημείο και το p είναι ο αριθμός των μεταβλητών εισόδου. Για παράδειγμα, εάν ένα σύνολο δεδομένων είχε 25 μεταβλητές εισόδου για ένα πρόβλημα κατηγοριοποίησης, τότε:

$m = \sqrt{25}$

$m = 5$



Σχήμα 6.8. Random Forests [14].

Εκτιμώμενη απόδοση

Για κάθε δείγμα bootstrap που λαμβάνεται από τα δεδομένα εκπαίδευσης, θα παραμείνουν δείγματα που δεν είχαν συμπεριληφθεί. Αυτά τα δείγματα καλούνται δείγματα Out-Of-Bag ή OOB. Η απόδοση κάθε μοντέλου στα δείγματα που έμειναν εκτός όταν υπολογίζεται ο μέσος όρος, μπορεί να παρέχει μια εκτιμώμενη ακρίβεια των μοντέλων με bagging. Αυτή η εκτιμώμενη απόδοση συχνά ονομάζεται εκτίμηση OOB.

Προετοιμασία δεδομένων για το Bagged CART

Το Bagged CART δεν απαιτεί ιδιαίτερη προετοιμασία δεδομένων εκτός από την καλή αναπαράσταση του προβλήματος [5].

6.4.2 Boosting and AdaBoost

Συλλογική Μέθοδος Ενίσχυσης

Η ενίσχυση είναι μια γενική μέθοδος του συνόλου που δημιουργεί έναν ισχυρό κατηγοριοποιητή από έναν αριθμό αδύναμων κατηγοριοποιητών. Αυτό γίνεται με την οικοδόμηση ενός μοντέλου από τα δεδομένα εκπαίδευσης και στη συνέχεια δημιουργώντας ένα δεύτερο μοντέλο επιχειρεί να διορθώσει τα σφάλματα από το πρώτο μοντέλο. Τα μοντέλα προστίθενται έως ότου το σετ εκπαίδευσης προβλέπεται τέλεια ή προστίθεται ένας μέγιστος αριθμός μοντέλων. Το AdaBoost ήταν ο πρώτος πολύ επιτυχημένος αλγόριθμος ενίσχυσης που αναπτύχθηκε για τη δυαδική κατηγοριοποίηση. Οι σύγχρονες μέθοδοι ενίσχυσης βασίζονται στο AdaBoost.

Ο AdaBoost χρησιμοποιείται καλύτερα για να ενισχύσει την απόδοση των δέντρων αποφάσεων σε δυαδικά προβλήματα κατηγοριοποίησης. Αρχικά ονομάστηκε AdaBoost.M1 από τους προγραμματιστές της τεχνικής. Σήμερα μπορεί να αναφέρεται και ως διακεκριμένος AdaBoost επειδή χρησιμοποιείται για κατηγοριοποίηση και όχι για παλινδρόμηση. Το AdaBoost μπορεί να χρησιμοποιηθεί για να ενισχύσει την απόδοση οποιουδήποτε αλγορίθμου μηχανικής εκμάθησης. Χρησιμοποιείται καλύτερα με τους αδύναμους αλγορίθμους μάθησης.

Αυτά είναι μοντέλα που επιτυγχάνουν ακρίβεια πάνω σε τυχαία πιθανότητα σε ένα πρόβλημα κατηγοριοποίησης. Ο πιο κατάλληλος και επομένως πιο κοινός αλγόριθμος που χρησιμοποιείται με το AdaBoost είναι τα δέντρα απόφασης με ένα επίπεδο. Επειδή αυτά τα δέντρα είναι τόσο σύντομα και περιέχουν μόνο μία απόφαση για κατηγοριοποίηση, συχνά ονομάζονται κούτσουρα απόφασης. Κάθε περίπτωση στο σύνολο δεδομένων εκπαίδευσης είναι σταθμισμένο. Το αρχικό βάρος ορίζεται σε:

$$weight(x_i) = \frac{1}{n} \qquad (6.\ 22)$$

Όπου x_i είναι η i-οστή εκπαίδευση και n είναι ο αριθμός των εκπαιδευτικών περιπτώσεων.

Εκπαίδευση ενός μοντέλου

Ένας αδύναμος κατηγοριοποιητής προετοιμάζεται για τα δεδομένα εκπαίδευσης χρησιμοποιώντας τα σταθμισμένα δείγματα. Υποστηρίζονται μόνο δυαδικά προβλήματα κατηγοριοποίησης, επομένως κάθε στήλη απόφασης παίρνει μία απόφαση σε μία μεταβλητή εισόδου και εξάγει τιμή +1.0 ή -1.0 για την τιμή της πρώτης ή της δεύτερης κλάσης. Το ποσοστό εσφαλμένης κατηγοριοποίησης υπολογίζεται για το εκπαιδευμένο μοντέλο ως εξής:

$$error = \frac{correct - N}{N} \qquad (6.23)$$

Όπου το error είναι ο ρυθμός εσφαλμένης κατηγοριοποίησης, correct είναι ο αριθμός των περιπτώσεων εκπαίδευσης που προβλέφθηκαν σωστά από το μοντέλο και ο N είναι ο συνολικός αριθμός των περιπτώσεων εκπαίδευσης. Για παράδειγμα, αν το μοντέλο προέβλεπε 78 από 100 περιπτώσεις εκπαίδευσης, το ποσοστό σφάλματος ή εσφαλμένης κατηγοριοποίησης θα ήταν (78-100) / 100 ή 0,22. Αυτό τροποποιείται για να χρησιμοποιηθεί η στάθμιση των περιπτώσεων εκπαίδευσης:

$$error = \frac{\sum_{i=n}^{n}(w_i \times perror_i)}{\sum_{i=n}^{n} w} \qquad (6.\ 24)$$

Το οποίο είναι το σταθμισμένο άθροισμα του ποσοστού λανθασμένης κατηγοριοποίησης, όπου w είναι το βάρος για την εκπαίδευση της περίπτωσης i και perror είναι το σφάλμα πρόβλεψης για την περίπτωση εκπαίδευσης i το οποίο είναι 1 αν έχει κατηγοριοποιηθεί εσφαλμένα και 0 αν είναι σωστά κατηγοριοποιημένο. Για παράδειγμα, αν είχαμε 3 εκπαιδευτικές περιπτώσεις με τα βάρη 0,01, 0,5 και 0,2, οι προβλεπόμενες τιμές ήταν -1, -1 και -1 και οι πραγματικές μεταβλητές εξόδου στις περιπτώσεις ήταν -1, 1 και -1, τότε οι τιμές perror θα ήταν 0, 1 και 0. Σε αυτή την περίπτωση ο ρυθμός εσφαλμένης κατηγοριοποίησης υπολογίζεται ως :

$$error = \frac{0{,}01 \times 0 + 0{,}5 \times 1 + 0{,}2 \times 0}{0{,}01 + 0{,}5 + 0{,}2}$$

$$error = 0{,}704$$

Μια τιμή σταδίου υπολογίζεται για το εκπαιδευμένο μοντέλο που παρέχει μια στάθμιση για τυχόν προβλέψεις που κάνει το μοντέλο. Η τιμή σταδίου για ένα εκπαιδευμένο μοντέλο υπολογίζεται ως εξής:

$$stage = ln(\frac{1-error}{error}) \quad (6.\ 25)$$

Όπου stage είναι η τιμή σταδίου που χρησιμοποιείται για την πρόβλεψη βάρους από το μοντέλο, ο ln είναι ο φυσικός λογάριθμος και το error είναι το λάθος κατηγοριοποίησης του μοντέλου. Το αποτέλεσμα του βάρους του σταδίου είναι ότι τα ακριβέστερα μοντέλα έχουν περισσότερο βάρος ή συμβολή στην τελική πρόβλεψη. Τα βάρη της εκπαίδευσης ενημερώνονται δίνοντας μεγαλύτερο βάρος σε λανθασμένα προβλεπόμενες περιπτώσεις και λιγότερο βάρος σε σωστά προβλεπόμενες περιπτώσεις. Για παράδειγμα, το βάρος ενός στιγμιότυπου εκπαίδευσης (w) ενημερώνεται χρησιμοποιώντας:

$$w = w \times e^{stage \times perror} \quad (6.\ 26)$$

Όπου w είναι το βάρος για ένα συγκεκριμένο παράδειγμα εκπαίδευσης, το e είναι η σταθερά του Euler υψωμένη σε μια δύναμη, το stage είναι το ποσοστό λανθασμένης κατηγοριοποίησης για τον αδύναμο κατηγοριοποιητή και το perror είναι το λάθος που ο αδύναμος κατηγοριοποιητής έκανε προβλέποντας την μεταβλητή εξόδου για την εκπαίδευση, παίρνοντας τις τιμές:

- perror = 0 για y == p
- perror = 1 για y != p

Όπου y είναι η μεταβλητή εξόδου για την περίπτωση εκπαίδευσης και p είναι η πρόβλεψη από τον αδύναμο κατηγοριοποιητή. Αυτό έχει ως αποτέλεσμα να μην αλλάζει το βάρος, εάν η περίπτωση κατάρτισης κατηγοριοποιήθηκε σωστά και κάνοντας το βάρος ελαφρώς μεγαλύτερο, εάν ο αδύναμος κατηγοριοποιητής κατηγοριοποιεί εσφαλμένα την περίπτωση.

Προετοιμασία δεδομένων για AdaBoost

Ποιοτικά δεδομένα. Επειδή η συλλογική μέθοδος συνεχίζει να προσπαθεί να διορθώσει τα σφάλματα κατηγοριοποίησης στα δεδομένα εκπαίδευσης, πρέπει τα δεδομένα εκπαίδευσης να είναι υψηλής ποιότητας.

Ακραίες τιμές. Οι ακραίες τιμές θα αναγκάσουν τον αλγόριθμο να εργαστεί σκληρά για να διορθώσει περιπτώσεις που δεν είναι ρεαλιστικές. Αυτές θα μπορούσαν να αφαιρεθούν από το σύνολο δεδομένων κατάρτισης.

Θορυβώδη Δεδομένα. Τα θορυβώδη δεδομένα, ειδικά ο θόρυβος στη μεταβλητή εξόδου, μπορεί να είναι προβληματικά. Εάν είναι δυνατόν, η απομόνωση και ο καθαρισμός από το σύνολο δεδομένων κατάρτισης είναι μια καλή πρακτική για την αποφυγή του προβλήματος [5].

7. Συγκριτική Αποτίμηση Μεθόδων

Υπάρχουν πολλές μεθοδολογίες και πολλοί αλγόριθμοι με τους οποίους μπορούν να υλοποιηθούν συστήματα συναισθηματικής ανάλυσης. Όπως έχει αναφερθεί και στα προηγούμενα κεφάλαια, οι τρείς κατηγορίες είναι: τα συστήματα που βασίζονται στους κανόνες (NLP, λεξικά), τα αυτόματα συστήματα (μηχανική μάθηση) και τα υβριδικά συστήματα (συνδυασμός των δύο προηγούμενων).

7.1 Προσέγγιση βασισμένη σε κανόνες

Ένα από τα μεγαλύτερα πλεονεκτήματα αυτής της προσέγγισης είναι ότι μπορεί να έχει εξαιρετική απόδοση σε ένα περιορισμένο τομέα. Στη μηχανική μάθηση αυτό είναι γνωστό ως overfitting και είναι ένα μεγάλο πρόβλημα που η μηχανική μάθηση προσπαθεί να αποφύγει. Επίσης είναι ανεξάρτητη από τα σύνολα δεδομένων. Τα μειονεκτήματα των συστημάτων που είναι βασισμένα σε αυτή την προσέγγιση είναι ότι τείνουν να έχουν πολύ κακή γενίκευση και δεν μπορούν να διαχειριστούν νέους όρους. Επίσης απαιτούν μεγάλο κόπο για να δημιουργηθούν. Στην δημοσίευση των Αραμπατζή κ.α. [15] για συναισθηματική ανάλυση Ελληνικών tweet και hashtag με χρήση συναισθηματικού λεξικού, συμπερασματικά αναφέρεται η ανάγκη για χρήση ενός ειδικά κατασκευασμένου λεξικού ώστε να βελτιωθούν τα αποτελέσματα της ανάλυσης.

7.2 Προσέγγιση βασισμένη στη μηχανική μάθηση

Οι αυτόματες μέθοδοι, αντίθετα με τα συστήματα που βασίζονται σε κανόνες, δεν βασίζονται σε χειρονακτικά επεξεργασμένους κανόνες, αλλά στις τεχνικές μηχανικής μάθησης οι οποίες έχουν πολύ καλή γενίκευση και μπορούν να διαχειριστούν νέους όρους. Το μοντέλο όμως έχει μεγάλη εξάρτηση από τα σύνολα δεδομένων και μπορεί να υποστεί overfitting ή underfitting ανάλογα με την ποιότητα των δεδομένων εκπαίδευσης.

Αλγόριθμοι μηχανικής μάθησης

Ένα σημαντικό θέμα που τίθεται όταν αντιμετωπίζεται ένα πρόβλημα συναισθηματικής ανάλυσης είναι η επιλογή από μια μεγάλη ποικιλία αλγορίθμων μηχανικής μάθησης. Η επιλογή ποικίλλει ανάλογα με πολλούς παράγοντες, όπως [16]:

Το μέγεθος του σετ εκπαίδευσης. Αυτός ο παράγοντας είναι ένας σημαντικός παράγοντας στην επιλογή του αλγορίθμου. Για ένα μικρό σετ εκπαίδευσης, οι κατηγοριοποιητές υψηλής μεροληψίας / χαμηλής διακύμανσης (π.χ. Naive Bayes) έχουν πλεονέκτημα έναντι των κατηγοριοποιητών χαμηλής μεροληψίας / υψηλής διακύμανσης (π.χ., KNN), αφού οι τελευταίοι θα κάνουν overfit. Ωστόσο, οι κατηγοριοποιητές χαμηλής μεροληψίας / υψηλής διακύμανσης αρχίζουν να κερδίζουν, καθώς αυξάνεται το σετ κατάρτισης (έχουν χαμηλότερο ασυμπτωτικό σφάλμα), αφού οι κατηγοριοποιητές υψηλής μεροληψίας δεν είναι αρκετά ισχυροί για να παρέχουν ακριβή μοντέλα.

Ο χρόνος εκπαίδευσης. Διάφοροι αλγόριθμοι έχουν διαφορετικό χρόνο λειτουργίας. Ο χρόνος εκπαίδευσης συνήθως εξαρτάται από το μέγεθος του συνόλου δεδομένων και την ακρίβεια του αποτελέσματος.

Η γραμμικότητα. Πολλοί αλγόριθμοι μηχανικής μάθησης, όπως η γραμμική παλινδρόμηση, η λογιστική παλινδρόμηση και οι μηχανές διανυσμάτων υποστήριξης κάνουν χρήση γραμμικότητας. Αυτές οι προσεγγίσεις μειώνουν την ακρίβεια κατά πολύ. Παρά τους κινδύνους, οι γραμμικοί αλγόριθμοι είναι πολύ δημοφιλείς ως πρώτη επιλογή. Έχουν την τάση να είναι αλγοριθμικά απλοί και γρήγοροι για την εκπαίδευση.

Ο αριθμός των παραμέτρων. Οι παράμετροι επηρεάζουν τη συμπεριφορά του αλγορίθμου, όπως η ανοχή στα σφάλματα ή ο αριθμός των επαναλήψεων. Παρόλο που πολλές παράμετροι παρέχουν συνήθως μεγαλύτερη ευελιξία, ο χρόνος εκπαίδευσης και η ακρίβεια του αλγορίθμου μπορεί μερικές φορές να είναι αρκετά ευαίσθητοι στη σωστή ρύθμιση των παραμέτρων.

Ο αριθμός των χαρακτηριστικών. Ο αριθμός των χαρακτηριστικών σε ορισμένα σύνολα δεδομένων μπορεί να είναι πολύ μεγάλος. Αυτό συμβαίνει συχνά με τα δεδομένα κειμένου. Ο μεγάλος αριθμός χαρακτηριστικών μπορεί να τελματώσει μερικούς αλγορίθμους μάθησης, καθιστώντας το χρόνο εκπαίδευσης ανέφικτα μεγάλο. Μερικοί αλγόριθμοι, όπως οι Μηχανές Διανύσματος Υποστήριξης, είναι ιδιαίτερα κατάλληλοι για αυτήν την περίπτωση.

Αλγόριθμος	**Γραμμική παλινδρόμηση**	**Λογιστική παλινδρόμηση**	**Naïve Bayes**	**KNN**	**Δέντρα αποφάσεων**	**Τυχαία δάση**	**AdaBoost**	**Νευρωνικά Δίκτυα**
Τύπος προβλήματος	Παλινδρόμηση	Κατηγοριοποίηση	Κατηγοριοποίηση	Παλινδρόμηση / Κατηγοριοποίηση	Παλινδρόμηση / Κατηγοριοποίηση	Παλινδρόμηση / Κατηγοριοποίηση	Παλινδρόμηση / Κατηγοριοποίηση	Παλινδρόμηση / Κατηγοριοποίηση
Εύκολος στην κατανόηση	Ναι	Κάπως	Κάπως	Ναι	Κάπως	Λίγο	Λίγο	Όχι
Μέση ακρίβεια πρόγνωσης	Χαμηλή	Χαμηλή	Χαμηλή	Χαμηλή	Χαμηλή	Υψηλή	Υψηλή	Υψηλή
Ταχύτητα εκπαίδευσης	Υψηλή	Υψηλή	Υψηλή	Υψηλή	Υψηλή	Χαμηλή	Χαμηλή	Χαμηλή
Ταχύτητα πρόβλεψης	Υψηλή	Υψηλή	Υψηλή	Εξαρτάται από το k	Υψηλή	Χαμηλή	Χαμηλή	Χαμηλή
Παραμετρικός	Ναι	Ναι	Ναι	Όχι	Όχι	Όχι	Όχι	Όχι
Λειτουργεί καλά με μικρό αριθμό δεδομένων εκπαίδευσης	Ναι	Ναι	Ναι	Όχι	Όχι	Όχι	Όχι	Όχι
Χρειάζεται προσαρμογή στα δεδομένα του προβλήματος	Όχι	Όχι	Λίγο	Ελάχιστη	Λίγο	Λίγο	Λίγο	Πολύ
Διαχωρισμός εισόδου – θορύβου	Όχι	Όχι	Ναι	Όχι	Όχι	Ναι (εκτός πολύ υψηλού θορύβου)	Ναι	Ναι

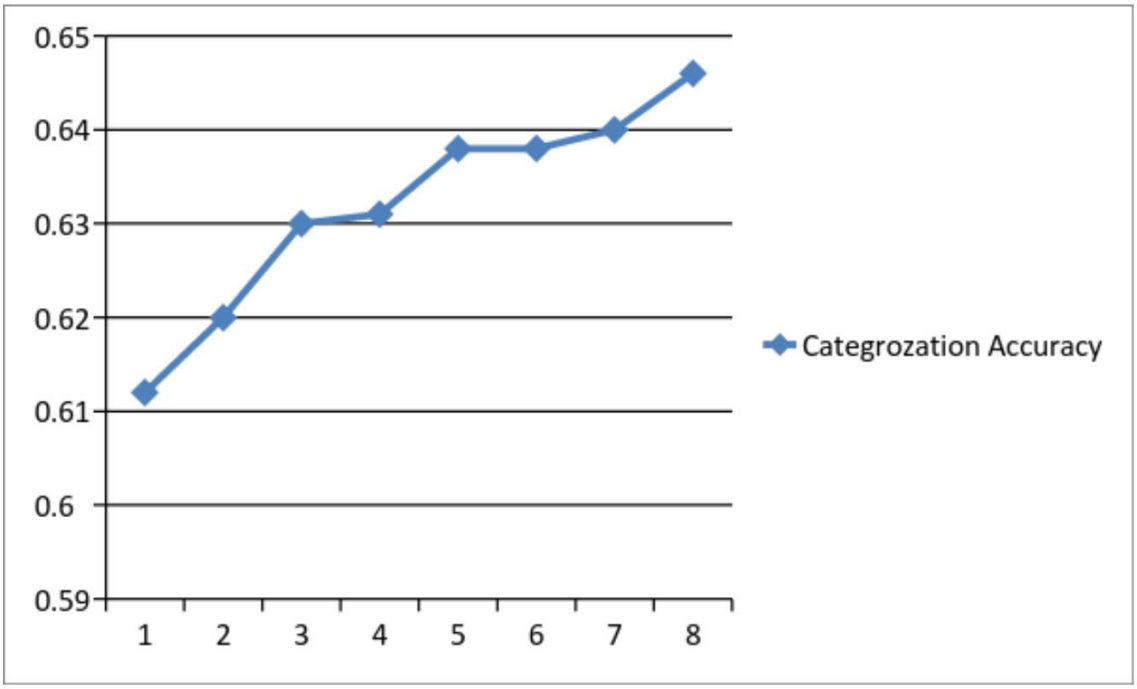
Σχήμα 7.1. Σύγκριση εποπτευόμενων αλγορίθμων μηχανικής μάθησης

7.3 Προσέγγιση βασισμένη σε υβριδικά συστήματα

Τα υβριδικά συστήματα συνδυάζουν μεθόδους από τις δύο προηγούμενες προσεγγίσεις. Συνήθως ο συνδυασμός μεθόδων μπορεί να βελτιώσει το accuracy και το precision.

7.4 Μετρήσεις απόδοσης τεχνικών και αλγορίθμων μηχανικής μάθησης

Στη δημοσίευση του Wijesinghe [17], αναφέρονται τα αποτελέσματα της σύγκρισης οκτώ αλγορίθμων μηχανικής μάθησης (SVM, SGD with SVM, Logistic Regression, SGD with Logistic Regression, KNN Classifier, Random Forest, Naive Bayesian, Adaboost) για ανάλυση συναισθήματος σε δεδομένα κριτικής ταινιών. Για την υποστήριξη της ανάλυσης έγινε χρήση δυο διαφορετικών βιβλιοθηκών (Vowpal Wabbit, Stanford Core NLP) και άλλων διαθέσιμων εργαλείων. Τα αποτελέσματα της σύγκρισης εμφανίζονται στο σχήμα 7.2.

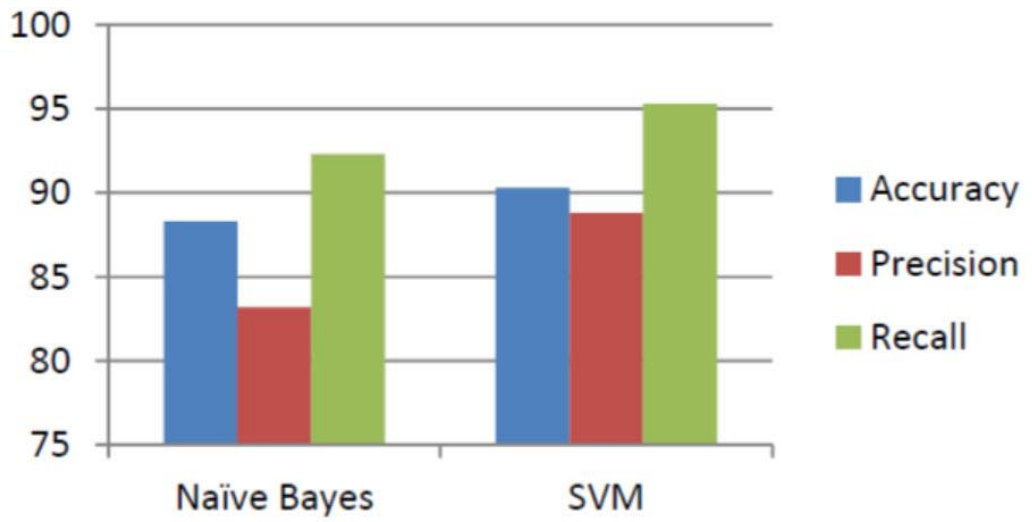


1. SVM
2. SGD-SVM
3. Logistic regression
4. Random forest
5. K nearest neighbour
6. SGD-logistic regression
7. Adaboost
8. Naive Bayes

Σχήμα 7.2. Ακρίβεια κατηγοριοποίησης με τον αντίστοιχο αλγόριθμο [17].

Στα συμπεράσματά του αναφέρεται στη σωστή προεπεξεργασία των δεδομένων με τεχνικές που βασίζονται στην NLP ώστε να αυξηθεί η απόδοση των μοντέλων.

Στη δημοσίευση τους οι Amolik et al. [18], συμπεραίνουν ότι οι τεχνικές μηχανικής μάθησης υπερτερούν σε σχέση με τις τεχνικές που βασίζονται σε κανόνες σε ευκολία και αποδοτικότητα. Αυτές οι τεχνικές εφαρμόστηκαν για συναισθηματική ανάλυση σε ανασκοπήσεις ταινιών με δεδομένα από το tweeter. Αναφέρονται και στη παρόμοια απόδοση των αλγορίθμων που χρησιμοποιήθηκαν (σχήμα 7.3 - Naivy Bayes, SVM) και αποδίδουν τις επιδόσεις του συστήματος στην χρήση της τεχνικής εξαγωγής χαρακτηριστικών, ως την πιο κατάλληλη για το συγκεκριμένο σύνολο δεδομένων και την ποιότητα αυτών. Επίσης παρατηρούν ότι η ακρίβεια της κατηγοριοποίησης αυξάνεται όσο αυξάνονται τα δεδομένα εκπαίδευσης.



Σχήμα 7.3. Απόδοση των κατηγοριοποιητών σε συναισθηματική ανάλυση στο Tweeter [18].

Οι Günther και Furrer [19] στο πείραμα συναισθηματικής ανάλυσης σύντομων μηνυμάτων με χρήση γλωσσολογικών χαρακτηριστικών και σχολαστικής επικλινούς καθόδου (Stochastic Gradient Descent - SGD), δείξανε ότι η απόδοση για αυτό το πείραμα μπορεί να βελτιωθεί χρησιμοποιώντας γλωσσολογικά χαρακτηριστικά καθώς και την προσεκτική επιλογή του αλγορίθμου μάθησης και των ρυθμίσεων των υπερπαραμέτρων του.

Στο άρθρο των Kharde & Sonawane [20], παρέχεται μια έρευνα και μια συγκριτική μελέτη των υφιστάμενων τεχνικών εξόρυξης γνώμης, συμπεριλαμβανομένης της μηχανικής μάθησης και των προσεγγίσεων που βασίζονται σε λεξικό, μαζί με μεθόδους cross-domain και cross-lingual και μερικές μετρήσεις αξιολόγησης (σχήμα 7.4). Τα αποτελέσματα της έρευνας δείχνουν ότι οι μέθοδοι μηχανικής μάθησης, όπως οι SVM και Naive Bayes, έχουν την υψηλότερη ακρίβεια και μπορούν να θεωρηθούν ως βασικές μέθοδοι μάθησης, ενώ οι μέθοδοι που βασίζονται σε λεξικό είναι πολύ αποτελεσματικές σε ορισμένες περιπτώσεις, οι οποίες απαιτούν λίγη ανθρώπινη προσπάθεια στην επισήμανση των εγγράφων. Συμπεραίνεται επίσης ότι όσο καθαρότερα είναι τα δεδομένα τόσο ακριβέστερα είναι τα αποτελέσματα. Η χρήση του μοντέλου bigram παρέχει καλύτερη συναισθηματική ακρίβεια σε σύγκριση με άλλα μοντέλα. Σαν τελικό συμπέρασμα αναφέρεται η ανάγκη για βαθύτερη μελέτη του συνδυασμού της μεθόδου μηχανικής μάθησης με τη μέθοδο λεξικών γνώμης, προκειμένου να βελτιωθεί η ακρίβεια της κατηγοριοποίησης των συναισθημάτων και της προσαρμοστικής ικανότητας σε ποικίλους τομείς και διαφορετικές γλώσσες.

	Method	Data Set	Acc.	Author
Machine Learning	SVM	Movie reviews	86.40%	Pang, Lee
	CoTraining SVM	Twitter	82.52%	Liu
	Deep learning	Stanford Sentiment Treebank	80.70%	Richard
Lexical based	Corpus	Product reviews	74.00%	Turkey
	Dictionary	Amazon's Mechanical Turk	---	Taboada
Cross-lingual	Ensemble	Amazon	81.00%	Wan,X
	Co-Train	Amazon, ITI68	81.30%	Wan,X.
	EWGA	IMDb movie review	>90%	Abbasi,A.
	CLMM	MPQA,NTCIR,ISI	83.02%	Mengi
Cross-domain	Active Learning	Book, DVD, Electronics, Kitchen	80% (avg)	Li, S
	Thesaurus			Bollegala
	SFA			Pan S J

Σχήμα 7.4. Σύγκριση απόδοσης των μεθόδων συναισθηματικής ανάλυσης [20].

7.5 Συμπεράσματα και μελλοντικές επεκτάσεις

Στην παρούσα διπλωματική εργασία, περιγράφονται οι βασικότερες τεχνικές και αλγόριθμοι που μπορούν να χρησιμοποιηθούν στην ανάλυση συναισθήματος. Έπειτα αναφέρθηκαν κάποιες έρευνες και κάποιες συγκριτικές μελέτες που αφορούν τεχνικές εξόρυξης γνώμης σε διάφορα σετ δεδομένων. Από τα παραπάνω, συμπεραίνεται η ανάγκη να επικεντρωθούμε στη μελέτη συνδυασμών μεθόδων και τεχνικών ανάλυσης που θα αυξήσουν τις ικανότητες των συστημάτων τεχνητής νοημοσύνης.

Σε μελλοντική επέκταση μπορούν να υλοποιηθούν προγράμματα συναισθηματικής ανάλυσης σε γνωστά dataset, τα οποία θα κάνουν χρήση των τεχνικών που αναφέραμε. Έπειτα με μετρήσεις της απόδοσης των συστημάτων μπορεί να γίνει επιλογή της καταλληλότερης μεθόδου για έναν τομέα και σε τελική φάση να γίνει βελτίωση των τεχνικών που χρησιμοποιήθηκαν.

8. Βιβλιογραφικές Αναφορές

[1] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze. (2009). An Introduction to Information Retrieval. Online edition Cambridge UP.

[2] Bing Liu. (2012). Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers.

[3] Indurkhya N, Damerau F. J. (2010). Handbook of Natural Language Processing. CRC Press.

[4] medium (2018). https://medium.com/datadriveninvestor/twitter-sentiment-analysis-of-movie-reviews-using-machine-learning-techniques-23d4724e7b05 (πρόσβαση 28 Δεκεμβρίου 2019).

[5] Brownlee J. (2016). Master Machine Learning Algorithms. Goodreads.

[6] Βλαχάβας Ι, Βασιλειάδης Ν, Κόκκορας Φ, Σακελλαρίου Η, Κεφάλας Π. (2011). Τεχνητή νοημοσύνη. Εκδόσεις Πανεπιστημίου Μακεδονίας.

[7] Abdullah, Maher & al Zamil, Mohammed. (2018). The Effectiveness of Classification on Information Retrieval System (Case Study).

[8] Towardsdatascience (2019). https://towardsdatascience.com/deep-learning-for-sentiment-analysis-7da8006bf6c1 (πρόσβαση 29 Δεκεμβρίου 2019).

[9] Vohra S. M. & Teraiya J. (2013). "A Comparative Study of Sentiment Analysis Techniques". Journal of Information, Knowledge and Research in Computer Engineering, 02:02, pp. 313-317.

[10] Al-Asmari, Salma & Dahab, Mohamed. (2017). Sentiment Detection, Recognition and Aspect Identification. International Journal of Computer Applications. 177. 975-8887. 10.5120/ijca2017915675.

[11] Eigenfoo (2017). https://eigenfoo.xyz/lda/ (πρόσβαση 25 Νοεμβρίου 2019).

[12] Towardsdatascience (2017). https://towardsdatascience.com/what-the-hell-is-perceptron-626217814f53 (πρόσβαση 27 Δεκεμβρίου 2019).

[13] Wikipedia (2019). https://el.wikipedia.org/wiki/Perceptron (πρόσβαση 19 Δεκεμβρίου 2019).

[14] Databricks (2015). https://databricks.com/blog/2015/01/21/random-forests-and-boosting-in-mllib.html (πρόσβαση 27 Νοεμβρίου 2019).

[15] Αραμπατζής Α, Καλαματιανός Γ, Μάλλης Δ, Νικολαράς Δ. (2015).

Συναισθηματική ανάλυση ελληνικών tweets και hashtags με χρήση λεξικού συναισθημάτων. 8ο Συνέδριο Φοιτητών Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, ΣΦΗΜΜΥ 8, σελ. 56-61, 3-13 Απριλίου, Πανεπιστήμιο Πατρών.

[16] Medium (2018). https://medium.com/@aravanshad/how-to-choose-machine-learning-algorithms-9a92a448e0df (πρόσβαση 30 Δεκεμβρίου 2019).

[17] Wijesinghe, Isuru. (2015). Sentiment Analysis. Researchgate.

[18] Amolik, Akshay & Jivane, Niketan & Bhandari, Mahavir & Venkatesan, M.. (2016). Twitter Sentiment Analysis of Movie Reviews using Machine Learning Techniques.. International Journal of Engineering and Technology. 7. 2038-2044.

[19] Günther, Tobias & Furrer, Lenz. (2013). GU-MLT-LT: Sentiment Analysis of Short Messages using Linguistic Features and Stochastic Gradient Descent. 328-332.

[20] Kharde, Vishal & Sonawane, Sheetal. (2016). Sentiment Analysis of Twitter Data: A Survey of Techniques. International Journal of Computer Applications. 139. 5-15. 10.5120/ijca2016908625.

[21] Jurafsky D & Martin J. (2008). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall PTR, NJ.

[22] Ricardo A. Baeza-Yates & Berthier Ribeiro-Neto. (1999). Modern Information Retrieval, Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA.

[23] Dataschool (2015). https://www.dataschool.io/comparing-supervised-learning-algorithms/ (πρόσβαση 29 Νοεμβρίου 2019).