

ΠΑΝΕΠΙΣΤΗΜΙΟ
ΠΑΤΡΩΝ
UNIVERSITY OF PATRAS

ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ
ΠΡΟΓΡΑΜΜΑ ΔΙΔΑΚΤΟΡΙΚΩΝ ΣΠΟΥΔΩΝ

Προηγμένες Τεχνικές Μηχανικής Μάθησης για Πρόβλεψη και Ταξινόμηση με Χρήση Συναισθηματικής Ανάλυσης και Ενεργής Μάθησης

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Αικατερίνη Χ. Καρανικόλα

Επιβλέπων: Κωτσιαντής Σωτήριος
Αναπληρωτής Καθηγητής Πανεπιστημίου Πατρών

Πάτρα, 2025

Πανεπιστήμιο Πατρών
Σχολή Θετικών Επιστημών
Τμήμα Μαθηματικών

Προηγμένες Τεχνικές Μηχανικής Μάθησης για Πρόβλεψη και Ταξινόμηση με Χρήση Συναισθηματικής Ανάλυσης και Ενεργής Μάθησης

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Αικατερίνη Χ. Καρανικόλα

Επιβλέπων: Σωτήριος Κωτσιαντής – Αναπληρωτής Καθηγητής

Εγκρίθηκε από την επταμελή εξεταστική επιτροπή στις 05/05/2025

Σ. Κωτσιαντής Αναπληρωτής Καθηγητής Πανεπιστήμιο Πατρών	Θ. Γράψα Ομότιμη Καθηγήτρια Πανεπιστήμιο Πατρών	Δ. Καββαδίας Επίκουρος Καθηγητής Πανεπιστήμιο Πατρών
Β. Βερύκιος Καθηγητής Ελληνικό Ανοικτό Πανεπιστήμιο	Χ. Μακρής Αναπληρωτής Καθηγητής Πανεπιστήμιο Πατρών	Δ. Καλές Καθηγητής Ελληνικό Ανοικτό Πανεπιστήμιο
	Κ. Σγάρμπας Αναπληρωτής Καθηγητής, Πανεπιστήμιο Πατρών	

Πανεπιστήμιο Πατρών, Τμήμα Μαθηματικών
Αικατερίνη Χ. Καρανικόλα
© 2025 - Με την επιφύλαξη παντός δικαιώματος.

"Nothing ever begins. There is no first moment; no single word or place from which this or any other story springs."

— **Weaveworld**, Clive Barker

"There is no such thing as a happy ending. I never met a single one to equal 'Once upon a time'.
Endings are heartless.
Ending is just another word for goodbye."

— **The Dark Tower**, Stephen King

"La vida no es la que uno vivió, sino la que uno recuerda y cómo la recuerda para contarla."

— **Pedro Páramo**, Juan Rulfo

Στους γονείς μου

Υπεύθυνη Δήλωση

Δηλώνω υπεύθυνα ότι το κείμενο της παρούσας Διδακτορικής Διατριβής αποτελεί πρωτότυπο έργο, το οποίο έχει συγγραφεί αποκλειστικά από εμένα, ενώ η σύνταξη, η οργάνωση και η επιστημονική τεκμηρίωση του κειμένου αποτελούν προσωπική μου ευθύνη. Όλες οι πηγές, τα δεδομένα, οι ιδέες και τα αποσπάσματα που χρησιμοποιήθηκαν από το έργο τρίτων, έχουν αναφερθεί με σαφήνεια και σύμφωνα με τους κανόνες της επιστημονικής δεοντολογίας.

Επιπλέον, δηλώνω ότι η εργασία αυτή δεν έχει κατατεθεί, ούτε έχει χρησιμοποιηθεί στο παρελθόν για την απόκτηση οποιουδήποτε ακαδημαϊκού τίτλου ή άλλης διάκρισης σε οποιοδήποτε ίδρυμα. Αναγνωρίζω ότι, σε περίπτωση που διαπιστωθεί λογοκλοπή ή χρήση ξένων πνευματικών έργων χωρίς τη δέουσα αναφορά, υπόκειμαι στις κυρώσεις που προβλέπονται από την κείμενη νομοθεσία και τους κανονισμούς του Πανεπιστημίου.

ΕΥΧΑΡΙΣΤΙΕΣ

Αυτή η σελίδα είναι αφιερωμένη σε όλους εκείνους, χωρίς τη στήριξη των οποίων η εκπόνηση αυτής της Διδακτορικής Διατριβής θα ήταν – αν όχι ακατόρθωτη – σίγουρα πολύ πιο δύσκολη. Η αναφορά ίσως να είναι σύντομη, είναι όμως απόλυτα ειλικρινής και προέρχεται από τα βάθη της καρδιάς μου.

Πρωτίστως, θα ήθελα να ευχαριστήσω τον επιβλέποντά μου, κ. Σωτήρη Κωτσιαντή, όχι μόνο για την πολύτιμη καθοδήγησή του καθ' όλη τη διάρκεια των μεταπτυχιακών και, στη συνέχεια, των διδακτορικών σπουδών μου, αλλά και επειδή, πέρα από παράδειγμα επιστήμονα και δασκάλου, αποτέλεσε και παράδειγμα ανθρώπου. Χωρίς τις ουσιαστικές συμβουλές του και την ορθή κρίση του, η ιδέα ενός τέτοιου εγχειρήματος θα είχε παραμείνει απλώς μια ιδέα.

Θα ήταν μεγάλη παράλειψη να μην αναφερθώ και στα υπόλοιπα μέλη της ΤΣΕ, την κα. Θεοδούλα Γράψα, Ομότιμη Καθηγήτρια, και τον κ. Δημήτρη Καββαδία, Επίκουρο Καθηγητή του Τμήματος Μαθηματικών, για τις πολύτιμες συμβουλές τους καθ' όλη τη διάρκεια της εκπόνησης της διατριβής.Και οι δύο υπήρξαν καθηγητές μου κατά τη διάρκεια των σπουδών μου, και τους ευχαριστώ θερμά για όσα μου προσέφεραν μέσα από τη διδασκαλία τους. Ευχαριστώ επίσης θερμά τα μέλη της Επταμελούς Εξεταστικής Επιτροπής για τα εύστοχα σχόλιά τους, που συνέβαλαν στη βελτίωση της διατριβής.

Το εγχείρημα αυτό θα ήταν αδύνατο χωρίς την υποστήριξη – συναισθηματική, ηθική και υλική - της οικογένειάς μου. Πρωτίστως ευχαριστώ τον πατέρα μου Χρήστο, που από την πρώτη μέρα στηρίζει κάθε απόφαση και βήμα μου με αγάπη και αυταπάρνηση. Η παρουσία και η δύναμή του είναι για μένα σταθερό σημείο αναφοράς. Ευχαριστώ επίσης τον αδερφό μου Δημήτρη, που πάντα υπήρξε υποστηρικτικός και υπομονετικός, παρέχοντας μια ορθολογική ματιά σε κάθε δυσκολία. Δε θα μπορούσα να μην ευχαριστήσω την μητέρα μου Ζωή, η οποία άφησε κληρονομία σε μένα και τον αδερφό μου την περιέργεια και την αγάπη για τη μάθηση. Από εκείνη ξεκίνησε όλο αυτό το ταξίδι.

Υπάρχουν δύο άτομα στα οποία νιώθω την ανάγκη να αναφερθώ ξεχωριστά. Ο κ. Σταμάτης Κάρλος υπήρξε συνεργάτης μου στο πρώτο στάδιο εκπόνησης της διατριβής και είναι αυτός που με μύησε στις διαδικασίες και τις μεθόδους της έρευνας. Ο κ. Χαράλαμπος Λιάπης στάθηκε πολύτιμος συνεργάτης κατά το μεγαλύτερο μέρος της φοίτησής μου, συμβάλλοντας ουσιαστικά με τη γνώση, τη διάθεση για συνεργασία και τη δημιουργικότητά του. Η παρακίνηση, οι ιδέες, η πρακτικότητα και ο ενθουσιασμός του υπήρξαν για μένα κινητήριος μοχλός – ιδιαίτερα τις στιγμές που οι απαιτήσεις της καθημερινότητας γίνονταν εξαντλητικές. Είναι πραγματικά σπάνιο να συναντά κανείς συνεργάτες με τους οποίους η συνεργασία προκύπτει αβίαστα και ευχάριστα – νιώθω τυχερή που είχα δύο και ακόμη πιο τυχερή που πλέον τους θεωρώ φίλους. Σταμάτη και Μπάμπη, σας ευχαριστώ θερμά.

Θα ήθελα επίσης να ευχαριστήσω τους φίλους μου, που ήταν δίπλα μου σε κάθε χαρά και απογοήτευση. Τη Γεωργία και την Αθανασία, τις δύο φίλες που το Μαθηματικό μας έφερε κοντά, σε εκείνο το γραφείο κάπου στον δεύτερο όροφο. Τη Βάσω και τη Νέλλη, τις παιδικές μου φίλες που γνωριζόμαστε πλέον πάνω από τη μισή μας ζωή και που δεν έχουν πάψει ποτέ να με στηρίζουν. Τη Μαρία και την Άντα, που, αν και μακριά, πάντα καταφέρνουν να τις νιώθω κοντά μου.

Τέλος, θα ήθελα να ευχαριστήσω τον καλύτερό μου φίλο και μεγαλύτερο υποστηρικτή, τον σύντροφό μου Νίκο, για την καλοσύνη και την αγάπη του όλα αυτά τα χρόνια.

Σας ευχαριστώ όλους, για όλα. Δεν θεωρώ τίποτα δεδομένο.

ΠΕΡΙΛΗΨΗ

Η παρούσα διδακτορική διατριβή εστιάζει στη διερεύνηση τεχνικών Μηχανικής Μάθησης οι οποίες αξιοποιούν την Ανάλυση Συναισθήματος και την Ενεργή Μάθηση, με κεντρικό στόχο τη βελτίωση υπαρχουσών, αποδοτικών μεθόδων. Τα ζητήματα που εξετάζονται μπορούν να διακριθούν σε δύο βασικές ενότητες.

Η πρώτη ενότητα εστιάζει στη χρήση της Ανάλυσης Συναισθήματος σε οικονομικές εφαρμογές. Αρχικά, μελετάται η πρόβλεψη οικονομικών χρονοσειρών αξιοποιώντας δεδομένα από τα μέσα κοινωνικής δικτύωσης και συγκεκριμένα από το Twitter, προκειμένου να αποτυπωθεί και να εκτιμηθεί η άποψη της κοινής γνώμης. Για το σκοπό αυτό, σχεδιάστηκαν δύο εκτεταμένες πειραματικές διαδικασίες: η πρώτη βασίζεται κυρίως σε κλασικές μεθόδους Μηχανικής Μάθησης, ενώ η δεύτερη σε τεχνικές Βαθιάς Μάθησης. Και στις δύο περιπτώσεις, διερευνάται τόσο η απόδοση πλήθους αλγορίθμων όσο και η επίδραση διαφορετικών προσεγγίσεων συναισθηματικής αποτίμησης (TextBlob, VADER, FinBERT). Τα ευρήματα αναδεικνύουν τη θετική συμβολή της Συναισθηματικής Ανάλυσης στη βελτίωση των οικονομικών προβλέψεων. Στη συνέχεια, η μελέτη εστιάζει στην Οικονομική Συναισθηματική Ανάλυση (Financial Sentiment Analysis - FSA) ως πρόβλημα ταξινόμησης, συγκρίνοντας αφενός πλήθος κλασικών μεθόδων κι ομαδοποιητές τους, αφετέρου σύγχρονα προεκπαιδευμένα μοντέλα που βασίζονται στους transformers και συγκεκριμένα διάφορες εκδοχές του μοντέλουBERT. Παράλληλα, εξετάζεται η αποτελεσματικότητα τεχνικών εξισορρόπησης κλάσεων (SMOTE) σε συνδυασμό με διαφορετικές γλωσσικές αναπαραστάσεις (BoW, TF-IDF). Τα αποτελέσματα υποδεικνύουν την υπεροχή των μοντέλων BEPT, με κυρίαρχο το RoBERTa, καθώς και ότι ο συνδυασμός SMOTE και TF-IDF βελτιστοποιεί περαιτέρω τις επιδόσεις.

Η δεύτερη ενότητα αφορά τη χρήση της Ενεργής Μάθησης με στόχο τη μείωση των απαιτήσεων σε επισημασμένα δεδομένα κατά την επίλυση προβλημάτων ταξινόμησης. Υλοποιήθηκαν δύο διακριτά σενάρια. Στο πρώτο, εξετάζεται η εφαρμογή της Ενεργής Μάθησης στο πρόβλημα της Πρόβλεψης Σφαλμάτων Λογισμικού , τόσο με χρήση μεμονωμένων μεθόδων Μηχανικής Μάθησης όσο και με ομαδοποιητές τους. Στο δεύτερο, διερευνάται το κατά πόσον μπορούν να επιτευχθούν συγκρίσιμες επιδόσεις στις διάφορες μετρικές αποτίμησης, όταν η Ανάλυση Συναισθήματος υποστηρίζεται από μοντέλα Transformer τα οποία αξιοποιούν Ενεργή Μάθηση. Και στα δύο σενάρια, η στοχευμένη επιλογή των παραδειγμάτων που θα ετικετοποιηθούν καταλήγει σε μοντέλα με απόδοση ανάλογη εκείνης που προκύπτει όταν χρησιμοποιείται ολόκληρο το διαθέσιμο σύνολο δεδομένων. Τέλος, προτείνεται ένας «μετα-δείκτης» αξιολόγησης των σεναρίων Ενεργής Μάθησης, κατάλληλος για συναφή εκπαιδευτικά σενάρια, με στόχο την πληρέστερη αποτίμηση της αποτελεσματικότητας των προτεινόμενων μεθόδων.

ABSTRACT

This doctoral dissertation focuses on exploring Machine Learning techniques that incomporate Sentiment Analysis and Active Learning, with the primary goal of enhancing existing, efficient methods. The issues addressed are grouped into two main sections.

In the first section, we concentrate on the use of Sentiment Analysis in financial domain applications. Initially, we investigate the scenario of stock market forecasting by utilizing data from social networks—specifically Twitter—to capture and assess public opinion. To this end, two extensive experimental procedures were designed: the first relies primarily on classical Machine Learning methods, whereas the second employs Deep Learning techniques. In both cases, we examine not only the performance of multiple algorithms but also the effect of different sentiment configurations (TextBlob, VADER, FinBERT). The findings highlight the positive contribution of Sentiment Analysis to improving economic forecasts. Subsequently, the study focuses on Finanancial Sentiment Analysis as a classification problem, comparing on the one hand numerous classical methods and their ensembles, and on the other hand various BERT variants. Concurrently, the effectiveness of class-balancing techniques (e.g., SMOTE) is examined in combination with different language representations (BoW, TF-IDF). The results highlight the superiority of BERT-based models—particularly RoBERTa—and show that the SMOTE and TF-IDF combination further optimizes performance.

The second section addresses the use of Active Learning to reduce the requirement for labeled data in classification tasks. Two distinct scenarios were implemented. In the first, Active Learning is applied to the problem of Software Defect Prediction, employing both individual Machine Learning methods and their ensemble versions. In the second, we investigate whether comparable performance across various metrics can be achieved when Sentiment Analysis is supported by Transformer models that incorporate Active Learning. In both scenarios, the targeted selection of examples to be labeled yields models with performance analogous to that obtained by training on the entire dataset. Finally, a "meta-index" for evaluating Active Learning scenarios is proposed, suited to related experimental contexts, with the goal of providing a more comprehensive assessment of the effectiveness of the proposed methods.

Contents

I Θεωρητικό υπόβαθρο **30**

1 Μηχανική Μάθηση **31**
1.1 Ορισμοί της Μηχανικής Μάθησης 31
1.2 Σύνδεση της Μηχανικής Μάθησης με άλλα πεδία της επιστήμης 33
1.3 Εφαρμογές της Μηχανικής Μάθησης 34
1.4 Κύριες Προκλήσεις και Περιορισμοί της Μηχανικής Μάθησης 36
1.5 Ανθρώπινη μάθηση vs μάθηση μηχανής 37

2 Κατηγορίες Μηχανικής Μάθησης **40**
2.1 Επιβλεπόμενη Μηχανική Μάθηση 40
2.1.1 Ταξινόμηση 41
2.1.1.1 Ορισμός προβλήματος ταξινόμησης και εφαρμογές . . . 41
2.1.1.2 Είδη προβλημάτων ταξινόμησης 42
2.1.1.3 Προκλήσεις των προβλημάτων ταξινόμησης 43
2.1.1.4 Μετρικές Αξιολόγησης 44
2.1.2 Παλινδρόμηση 47
2.1.2.1 Μέθοδοι παλινδρόμησης 49
2.1.2.2 Μείωση της Υπερπροσαρμογής - Κανονικοποίηση . . . 52
2.1.2.3 Μετρικές αξιολόγησης της παλινδρόμησης 54
2.2 Μη - Επιβλεπόμενη Μηχανική Μάθηση 56
2.2.1 Τύποι μη επιβλεπόμενης μάθησης 56
2.3 Ημι - Επιβλεπόμενη Μάθηση 58
2.4 Ενισχυτική Μάθηση 59

3 Ενεργή Μάθηση **61**
3.1 Διαδικασία Ενεργής Μάθησης 62
3.2 Σενάρια Ενεργής Μάθησης 63
3.3 Στρατηγικές επιλογής ερωτημάτων 65
3.4 Πρακτικές παρατηρήσεις και προβλήματα 72

4 Ανάλυση Συναισθήματος **74**
4.1 Είδη και μέθοδοι Ανάλυσης Συναισθήματος 75
4.2 Μέθοδοι βασισμένες σε λεξικά (Lexicon-based Methods) 76
4.2.1 Dictionary-based Method: 76
4.2.2 Corpus-based Method: 77
4.2.3 TextBlob και VADER 77
4.3 Μέθοδοι από το χώρο της Μηχανικής Μάθησης 78

4.4 Μέθοδοι από το χωρο της Βαθιάς Μάθησης 79
4.5 Υβριδικές Προσεγγίσεις 80

II Πειραματικά Αποτελέσματα σε Πρακτικές εφαρμογές 81

5 Πρόβλεψη χρονοσειρών με χρήση ανάλυσης συναισθήματος 82
5.1 Βασικές έννοιες χρονοσειρών και σύντομη ιστορική αναδρομή 82
5.2 Συγκριτική μελέτη μεθόδων για προβλέψεις σε χρηματοοικονομικά δεδομένα 86
5.3 Οικογένειες προβλεπτικών μοντέλων 86
5.3.1 Μοντέλα ARIMA 86
5.3.2 Μοντέλο Prophet 88
5.3.3 Εποχικά Μοντέλα Holt - Winters 89
5.3.4 Προσθετικό Μοντέλο Holt-Winters (HWAAS) 89
5.3.5 Πολλαπλασιαστικό Μοντέλο Holt-Winters (HWAMS) 90
5.3.6 Το μοντέλο NBEATS 90
5.3.7 Το μοντέλο DeepAR 91
5.3.8 Trigonometric BAT 92
5.4 Πειραματική διαδικασία 92
5.4.1 Δεδομένα 93
5.4.2 Αλγόριθμοι 93
5.4.3 Αποτελέσματα 95
5.4.3.1 Αποτελέσματα της μετρικής MAE 97
5.4.3.2 Αποτελέσματα της μετρικής RMSE 98
5.4.3.3 Αποτελέσματα της μετρικής MAPE 100
5.4.4 Σύνοψη Αποτελεσμάτων και Συμπεράσματα 101
5.5 Πρόβλεψη μονομεταβλητών χρηματοοικονομικών χρονοσειρών με χρήση Ανάλυσης Συναισθήματος 104
5.5.1 Πειραματική διαδικασία 105
5.5.1.1 Προεπεξεργασία δεδομένων και εξαγωγή συναισθηματικών πολικοτήτων 105
5.5.2 Αλγόριθμοι 108
5.5.2.1 Long-Short Term Memory - (LSTM) 111
5.5.2.2 Bidirectional Long Short-Term Memory (BiLSTM) . 112
5.5.3 Αποτελέσματα 113
5.5.3.1 Από τη σκοπιά των αλγορίθμων 115
5.5.3.2 Από τη σκοπιά των συναισθηματικών διαμορφώσεων . 117
5.5.3.3 Χρονικός ορίζοντας μιας εβδομάδας 122
5.5.3.4 Χρονικός ορίζοντας δυο εβδομάδων 126
5.5.3.5 Συμπεράσματα 128
5.5.4 Χρήση Αλγορίθμων Βαθιάς Μάθησης 130
5.5.4.1 Αλγόριθμοι 130
5.5.4.2 Temporal Convolutional Networks (TCNs) 133
5.5.5 Αποτελέσματα 134
5.5.5.1 Σύγκριση μεθόδων 135
5.5.5.2 Αποτελέσματα συναισθηματικών διαμορφώσεων 141
5.5.5.3 Συμπεράσματα 146

6 Η ανάλυση συναισθήματος σε χρηματοοικονομικές προβλέψεις από τη σκοπιά της ταξινόμησης – Κλασικές vs. μέθοδοι Βαθιάς Μάθησης 151
6.1 Περιγραφή της πειραματικής διαδικασίας . . . 153
6.1.1 Σύνολο Δεδομένων . . . 153
6.1.2 Γλωσσικές αναπαραστάσεις . . . 155
6.1.3 Μοντέλα Ταξινόμησης . . . 156
6.1.3.1 Κλασικές μέθοδοι ταξινόμησης . . . 157
6.1.3.2 Προεκπαιδευμένα γλωσσικά μοντέλα . . . 159
6.1.3.3 Αρχιτεκτονική των transformers . . . 159
6.2 Αποτελέσματα . . . 163
6.2.1 Κλασσικές μέθοδοι ταξινόμησης + BoW + SMOTE . . . 163
6.2.2 Κλασσικές μέθοδοι ταξινόμησης + TF-IDF + SMOTE . . . 165
6.2.3 Κλασσικές μέθοδοι ταξινόμησης + BoW . . . 165
6.2.4 Κλασσικές μέθοδοι ταξινόμησης + TF-IDF . . . 168
6.2.5 Deep learning pre-trained models . . . 170
6.3 Συμπεράσματα . . . 171

7 Ελαχιστοποιώντας την Ανάγκη Ετικετοποιημένων Δεδομένων στην Κατηγοριοποίηση Κειμένων, με χρήση Ενεργής Μάθησης σε BERT-based transformers 173
7.1 Εισαγωγή . . . 173
7.2 Πειραματική Διαδικασία . . . 174
7.2.1 Υλοποίηση της Πειραματικής Διάταξης . . . 174
7.2.1.1 Σύνολα δεδομένων . . . 175
7.2.1.2 Μοντέλα BERT . . . 177
7.2.1.3 Αξιολόγηση . . . 178
7.3 Αποτελέσματα . . . 180
7.3.1 Σύγκριση μοντέλων χωρίς τη χρήση Ενεργής Μάθησης . . . 180
7.3.2 Συγκρίσεις όπου εμπλέκεται Ενεργή Μάθηση . . . 183
7.4 Συμπεράσματα . . . 189

8 Πρόβλεψη Σφαλμάτων Λογισμικού: Βελτιστοποίηση Προβλέψεων με Ενεργή Μάθηση και Ομαδοποιητές Ταξινομητών 190
8.1 Εισαγωγή . . . 190
8.2 Πειραματική διαδικασία . . . 191
8.2.1 Δεδομένα . . . 191
8.2.2 Ταξινομητές . . . 193
8.2.3 Πρώτο στάδιο πειραματικής διαδικασίας: Σύγκριση απλών μεθόδων και δημιουργία ομαδοποιητών . . . 194
8.2.4 Δεύτερο στάδιο πειραματικής διαδικασίας: Σχήματα Ενεργητικής Μάθησης . . . 195
8.2.5 Αξιολόγηση . . . 195
8.3 Αποτελέσματα . . . 197
8.3.1 Σύγκριση μεμονωμένων ταξινομητών . . . 198
8.3.2 Σύγκριση μεμονωμένων ταξινομητών και ομαδοποιητών . . . 198
8.3.3 Αποτελέσματα σχετικά με τα σχήματα Ενεργής Μάθησης . . . 200
8.4 Συμπεράσματα . . . 208

9 Επίλογος **211**
9.1 Κύρια συμπεράσματα της διδακτορικής διατριβής 211
9.2 Μελλοντικές επεκτάσεις της διδακτορικής διατριβής 212

List of Tables

3.1 Σύγκριση πλεονεκτημάτων και μειονεκτημάτων της μεθόδου δειγματοληψίας με βάση την αβεβαιότητα. 67
3.2 Σύγκριση πλεονεκτημάτων και μειονεκτημάτων της μεθόδου επιλογής ερωτήσεων με βάση την ψηφοφορία επιτροπής 69
3.3 Σύγκριση πλεονεκτημάτων και μειονεκτημάτων της μεθόδου επιλογής ερωτημάτων με βάση την Αναμενόμενη Αλλαγή Μοντέλου 70
3.4 Σύγκριση πλεονεκτημάτων και μειονεκτημάτων της μεθόδου επιλογής ερωτημάτων με βάση την Αναμενόμενη Μείωση Σφάλματος 71
3.5 Σύγκριση πλεονεκτημάτων και μειονεκτημάτων της μεθόδου επιλογής ερωτημάτων με βάση τη Μείωση Διακύμανσης Σφάλματος 72

5.1 Ειδικές περιπτώσεις του μοντέλου ARIMA 88
5.2 Οι μετοχές και οι αντίστοιχες συντομογραφίες τους 94
5.3 N/I/H για τη μετρική αξιολόγησης MAE 97
5.4 Κατάταξη Friedman για τις μετρικές MAE, RMSE, MAPE 97
5.5 N/I/H για τη μετρική αξιολόγησης RMSE 98
5.6 N/I/H για τη μετρική αξιολόγησης MAPE 100
5.7 Διαμορφώσεις συναισθηματικών αποτιμήσεων 106
5.8 Λίστα μετοχών 107
5.9 Λίστα αλγορίθμων 110
5.10 Κατάταξη Friedman (Μέθοδοι) - Ορίζοντας πρόβλεψης μιας ημέρας . . 116
5.11 Κατάταξη Friedman (Μέθοδοι) - Ορίζοντας πρόβλεψης μιας εβδομάδας 118
5.12 Κατάταξη Friedman (Μέθοδοι) - Ορίζοντας πρόβλεψης δύο εβδομάδων 120
5.13 Κατάταξη Friedman (Συναισθηματικές διαμορφώσεις) - Ορίζοντας πρόβλεψης μιας ημέρας 123
5.14 Κατάταξη Friedman (Συναισθηματικές διαμορφώσεις) - Ορίζοντας πρόβλεψης μιας εβδομάδας 125
5.15 Κατάταξη Friedman (Συναισθηματικές διαμορφώσεις) - Ορίζοντας πρόβλεψης δύο εβδομάδων 129
5.16 Αλγόριθμοι Βαθιάς Μάθησης 132
5.17 Κατάταξη Friedman (Μέθοδοι) — Χρονικός ορίζοντας μιας ημέρας . . 136
5.18 Κατάταξη Friedman (Μέθοδοι) — Χρονικός ορίζοντας μιας εβδομάδας 138
5.19 Κατάταξη Friedman (Μέθοδοι) — Χρονικός ορίζοντας δύο εβδομάδων 141
5.20 Κατάταξη Friedman (Συναισθηματικές Διαμορφώσεις) — Χρονικός ορίζοντας μιας ημέρας 142
5.21 Κατάταξη Friedman (Συναισθηματικές Διαμορφώσεις) — Χρονικός ορίζοντας μιας εβδομάδας 144
5.22 Κατάταξη Friedman (Συναισθηματικές Διαμορφώσεις) — Χρονικός ορίζοντας δύο εβδομάδων 146

6.1 Διαφορές των Bow και TF-IDF 156
6.2 Αλγόριθμοι Μηχανικής Μάθησης 157
6.3 Κλασσικές μέθοδοι + BoW + SMOTE 164
6.4 Κλασσικές μέθοδοι + TF-IDF + SMOTE 166
6.5 Κλασσικές μέθοδοι + BoW 167
6.6 Κλασσικές μέθοδοι + TF-IDF 169
6.7 Απόδοση προεκπαιδευμένων μοντέλων 171

7.1 Σύνολα Δεδομένων 177
7.2 Κατάταξη Friedman με fine-tuning στο πλήρες σύνολο δεδομένων . . . 181

8.1 Χαρακτηριστικά των συνόλων δεδομένων 192
8.2 Λίστα των μεθόδων ταξινόμησης που χρησιμοποιήθηκαν 194
8.3 Κατάταξη Friedman - Μεμονωμένες μέθοδοι 198
8.4 Κατάταξη Friedman - Μεμονωμένες μέθοδοι και ομαδοποιητές 199
8.5 Η απόδοση των μεθόδων σύμφωνα με τη μετρική EAPI 209

List of Figures

1.1 Διάγραμμα Venn για τη σύνδεση της Μηχανικής Μάθησης με άλλα επιστημονικά πεδία 34

2.1 Ο πίνακας σύγχυσης σε ένα δυαδικό πρόβλημα ταξινόμησης 45

3.1 Query synthesis: Ο εκμαθητής μπορεί να ζητά την ετικέτα οποιουδήποτε δείγματος - αυτό σημαίνει ότι είναι δυνατή η δημιουργία δυγμάτων de novo 64
3.2 Stream-Based Selective Sampling: Ο εκμαθητής αποφασίζει αν θα αιτηθεί την ετικέτα ενός δείγματος ή αν θα το απορρίψει για να περάσει στο επόμενο. 65
3.3 Pool-based selective sampling: Ο εκμαθητής αξιολογεί όλα τα διαθέσιμα δείγματα (ή ένα υποσύνολό τους) χρησιμοποιώντας ένα κριτήριο βασισμένο στην πληροφορία και επιλέγει εκείνο με την περισσότερη πληροφορία για επισήμανση. 66

5.1 Γράφημα που απεικονίζει την τιμή κλεισίματος της μετοχής της Microsoft Corporation από τον Ιανουάριο 2018 έως τον Δεκέμβριο 2018 . . 84
5.2 Θηκογράμματα της απόδοσης των μεθόδων στην μετρική MAE 98
5.3 Διάγραμμα κρίσιμων διαφορών της μετρικής MAE 99
5.4 Θηκογράμματα της απόδοσης των μεθόδων στην μετρική RMSE 99
5.5 Διάγραμμα κρίσιμων διαφορών της μετρικής RMSE 100
5.6 Θηκογράμματα της απόδοσης των μεθόδων στην μετρική MAPE 101
5.7 Διάγραμμα κρίσιμων διαφορών της μετρικής MAPE 102
5.8 In-sample προβλέψεις της μετοχής F 103
5.9 In-sample προβλέψεις της μετοχής SPOT 103
5.10 In-sample προβλέψεις της μετοχής ORCL 103
5.11 Διαγραμμάτική αναπαράσταση της προεπεξεργασίας 107
5.12 Διαγραμματική αναπαράσταση της διαδικασίας εξαγωγής των 22 συναισθηματικών διαμορφώσεων 109
5.13 Θηκογράμματα για την απόδοση των αλγορίθμων - Χρονικός ορίζοντας μιας ημέρας 115
5.14 CD διαγράμματα για την απόδοση των αλγορίθμων - Χρονικός ορίζοντας μιας ημέρας 117
5.15 Θηκογράμματα για την απόδοση των αλγορίθμων - Χρονικός ορίζοντας μιας εβδομαδας 119
5.16 CD διαγράμματα για την απόδοση των αλγορίθμων - Χρονικός ορίζοντας μιας εβδομαδας 119
5.17 Θηκογράμματα για την απόδοση των αλγορίθμων - Χρονικός ορίζοντας δύο εβδομάδων 121

5.18 CD διαγράμματα για την απόδοση των αλγορίθμων - Χρονικός ορίζοντας δύο εβδομάδων . . . 121
5.19 Θηκογράμματα για την απόδοση των συναισθηματικών διαμορφώσεων - Χρονικός ορίζοντας μιας ημέρας . . . 122
5.20 CD διαγράμματα για την απόδοση των συναισθηματικών διαμορφώσεων - Χρονικός ορίζοντας μιας ημέρας . . . 124
5.21 Θηκογράμματα για την απόδοση των συναισθηματικών διαμορφώσεων - Χρονικός ορίζοντας μιας εβδομάδας . . . 126
5.22 CD διαγράμματα για την απόδοση των συναισθηματικών διαμορφώσεων - Χρονικός ορίζοντας μιας εβδομάδας . . . 127
5.23 Θηκογράμματα για την απόδοση των συναισθηματικών διαμορφώσεων - Χρονικός ορίζοντας δύο εβδομάδων . . . 127
5.24 CD διαγράμματα για την απόδοση των συναισθηματικών διαμορφώσεων - Χρονικός ορίζοντας δύο εβδομάδων . . . 128
5.25 Διαγραμματική αναπαράσταση της διαδικασίας εξαγωγής των 67 συναισθηματικών διαμορφώσεων . . . 131
5.26 Θηκογράμματα για την απόδοση των αλγορίθμων - Χρονικός ορίζοντας μιας ημέρας . . . 137
5.27 CD διαγράμματα για την απόδοση των αλγορίθμων - Χρονικός ορίζοντας μιας ημέρας . . . 137
5.28 Θηκογράμματα για την απόδοση των αλγορίθμων - Χρονικός ορίζοντας μιας εβδομάδας . . . 139
5.29 CD διαγράμματα για την απόδοση των αλγορίθμων - Χρονικός ορίζοντας μιας εβδομάδας . . . 139
5.30 Θηκογράμματα για την απόδοση των αλγορίθμων - Χρονικός ορίζοντας 2 εβδομάδων . . . 140
5.31 CD διαγράμματα για την απόδοση των αλγορίθμων - Χρονικός ορίζοντας δύο εβδομάδων . . . 140
5.32 Θηκογράμματα για την απόδοση των συναισθηματικών διαμορφώσεων - Χρονικός ορίζοντας μιας ημέρας . . . 143
5.33 CD διαγράμματα για την απόδοση των αλγορίθμων - Χρονικός ορίζοντας μιας ημέρας . . . 143
5.34 Θηκογράμματα για την απόδοση των συναισθηματικών διαμορφώσεων - Χρονικός ορίζοντας μιας εβδομάδας . . . 145
5.35 CD διαγράμματα για την απόδοση των αλγορίθμων - Χρονικός ορίζοντας μιας εβδομάδας . . . 145
5.36 Θηκογράμματα για την απόδοση των συναισθηματικών διαμορφώσεων - Χρονικός ορίζοντας δύο εβδομάδων . . . 147
5.37 CD διαγράμματα για την απόδοση των αλγορίθμων - Χρονικός ορίζοντας δύο εβδομάδων . . . 147
5.38 Σχετική κατάταξη των TCN, TSTPlus και XCMPlus . . . 148
5.39 Σχετική κατάταξη των κλάσεων συναισθηματικών διαμορφώσεων TextBlob, VADER και FinBERT . . . 150

6.1 Κατανομή των δειγμάτων σε κλάσεις . . . 154
6.2 Διαγραμματική απεικόνιση της προεπεξεργασίας του συνόλου δεδομένων 155
6.3 Αρχιτεκτονική ενός transformer . . . 161
6.4 Απόδοση προεκπαιδευμένων μοντέλων . . . 168

6.5 Απόδοση προεκπαιδευμένων μοντέλων 170
6.6 Σύγκριση της μεθόδου RoBERTa με την επικρατούσα κλασική μέθοδο σε κάθε σενάριο 172

7.1 Απόσοδη αλγορίθμων στο σενάριο όπου δε χρησιμοποιείται μηχανική μάθηση 182
7.2 Διάγραμμα κρίσιμων διαφορών της σύγκρισης των μεθόδων στο σενάριο όπου δε χρησιμοποιείται Ενεργή Μάθηση 183
7.3 Ποσοστό χρήσης του αρχικού συνόλου δεδομένων στο σενάριο της Ενεργής Μάθησης στο οποίο επιτυγχάνεται παρόμοια ακρίβεια με απόκλιση 1%, 2% και 3% αντίστοιχα 184
7.4 Ραβδογράμματα κατανομής των περιπτώσεων σε κλάσεις ποσοστού χρήσης του συνολικού συνόλου δεδομένου στις οποίες το σενάριο Μηχανικής Μάθησης επιτυγχάνει με απόκλιση μικρότερη του 1% (πράσινο) ή μεγαλύτερη του 1% (ροζ) παραπλήσιο σκορ στην κάθε μετρική σε σχέση με το σενάριο όπου αυτή δε χρησιμοποιείται 186
7.5 Οπτική απεικόνηση του ποσοστού του πλήρους συνόλου δεδομένων για κάθε ζεύγος μεθόδου - συνόλου δεδομένων, στο οποίο η χρήση Ενεργής Μάθησης έχει παρόμοιο αποτέλεσμα με όταν αυτή δε χρησιμοποιείται με απόκλιση μικρότερη του 1% (πράσινο) και μεγαλύτερη του 1% (ροζ) . 187
7.6 Βαθμολογία EAPI για κάθε ζεύγος μεθόδου - συνόλου δεδομένων ανά μετρική 188

8.1 Κατανομή των δειγμάτων σε κλάσεις 193
8.2 Διαγράμματα κρίσιμων διαφορών της καλύτερης δεκάδας μεθόδων . . . 199
8.3 Σύγκλιση των μεθόδων 201
8.4 Απόδοση των μεθόδων 202
8.5 Θηκογράμματα που απεικονίζουν το ποσοστό του αρχικού συνόλου που χρησιμοποιήθηκε στο σενάριο Ενεργής Μάθησης 204
8.6 Ποσοστό χρήσης δεδομένων χωρισμένο σε κλάσεις ίσου πλάτους 205
8.7 Οπτική απεικόνηση πειραμάτων που πετυχαίνουν την μείωση των επικεττοποιημένων δεδομένων εκπαίδευσης. 206
8.8 Απεικόνηση των περιπτώσεων στατιστικής και μη ανεξαρτησίας ανάμεσα στο σενάριο μη χρήσης ενεργητικής μάθησης και στις παραλλαγές όπου αυτή χρησημοποιείται 207
8.9 Στατιστική ανεξαρτησία των παραλλαγών στην περίπτωση του rf+gbc+lda207

Πρόλογος

Εν είδει προλόγου θα παρατεθούν τα σημεία συνεισφοράς την παρούσας διδακτορικής διατριβής, καθώς και θα παρουσιαστεί εν τάχει το περίγραμμά της.

Συνεισφορά της διατριβής

Η παρούσα διατριβή προτείνει καινοτόμες προσεγγίσεις στην περιοχή της Μηχανικής Μάθησης, αξιοποιώντας την Ανάλυση Συναισθήματος και την Ενεργή Μάθηση για τη βελτίωση της ακρίβειας και της αποτελεσματικότητας προβλεπτικών μοντέλων. Τα κυριότερα σημεία συνεισφοράς της μπορούν να συνοψιστούν ως εξής:

Βελτιστοποίηση Προβλέψεων μέσω Συναισθηματικής Ανάλυσης

- Διερευνάται ο ρόλος της Ανάλυσης Συναισθήμτος ως βοηθητικού μηχανισμού για την περαιτέρω βελτίωση των προβλέψεων, σε σενάρια που εμπλέκουν κλασικές μεθόδους Μηχανική Μάθησης, αλλά και μεθόδους Βαθιάς Μάθησης.

- Προτείνονται νέες μέθοδοι ενσωμάτωσης συναισθηματικών χαρακτηριστικών σε μοντέλα πρόβλεψης, επιτυγχάνοντας υψηλότερη ακρίβεια σε σχέση με παραδοσιακά συστήματα, όπου η πρόβλεψη βασίζεται μόνο στις ιστορικές τιμές της εκάστοτε χρονοσειράς.

Ανάπτυξη Υβριδικών Μεθόδων Ενεργής Μάθησης

- Παρουσιάζεται ένας συνδυασμός Ενεργής Μάθησης και σύγχρονων τεχνικών μάθησης, επιτρέποντας αποδοτικότερη επιλογή δεδομένων προς επισήμανση και συνεπώς επιτυγχάνεται μείωση του συνολικού αριθμού των επισημασμένων δεδομένων

- Διερευνώνται στρατηγικές επιλογής δειγμάτων που μειώνουν σημαντικά την ανάγκη για μεγάλο όγκο επισημασμένων δεδομένων, μειώνοντας έτσι τον απαιτούμενο χρόνο εκπαίδευσης και τους υπολογιστικούς πόρους.

Καινοτόμες Προσεγγίσεις στη Συναισθηματική Ανάλυση

- Εφαρμόζονται προηγμένα μοντέλα Επεξεργασίας Φυσικής Γλώσσας, όπως οι BERT-based transformers, για την επίτευξη ακριβέστερης και ευέλικτης κατηγοριοποίησης συναισθημάτων.

- Μελετάται η επίδραση της συναισθηματικής πληροφορίας σε διάφορες κατηγορίες προβλημάτων, επισημαίνοντας περιπτώσεις όπου η ενσωμάτωση συναισθημάτων ενισχύει την απόδοση των μοντέλων.

Εφαρμογή σε Ρεαλιστικά Σενάρια και Επικύρωση των Μεθόδων

- Οι προτεινόμενες τεχνικές αξιολογούνται ενδελεχώς σε πραγματικά σύνολα δεδομένων, αναδεικνύοντας την πρακτική τους χρησιμότητα.

- Πραγματοποιείται σύγκριση με παραδοσιακές και σύγχρονες μεθόδους, υπογραμμίζοντας την ανωτερότητα και την καινοτομία των προτεινόμενων προσεγγίσεων σε πραγματικές συνθήκες. Επιπλέον, η υλοποίηση εκτεταμένων πειραματικών διαδικασιών σε δημοσίως διαθέσιμα δεδομένα θα κάνει δυνατή τη σύγκριση των αποτελεσμάτων με άλλες έρευνες που θα δημοσιευθούν μελλοντικά

Με τον τρόπο αυτό, η διατριβή συμβάλλει ουσιαστικά στη διερεύνηση και στην ανάπτυξη προηγμένων εργαλείων Μηχανικής Μάθησης, τόσο σε θεωρητικό όσο και σε πρακτικό επίπεδο, προσφέροντας σημαντικές προοπτικές για μελλοντική έρευνα και εφαρμογές σε ευρύ φάσμα τομέων.

Διάρθρωση της διδακτορικής διατριβής

Η διάρθρωση της διατριβής έχει ως εξής:

Στο πρώτο μέρος παρέχονται εν συντομία τα βασικά στοιχεία του θεωρητικού υποβάθρου στο οποίο έχει τις ρίζες της η παρούσα εργασία. Καθώς στο πλαίσιο της διατριβής αξιοποιείται πληθώρα μεθόδων και τεχνικών από διάφορα πεδία της Μηχανική Μάθησης, μια πλήρης και ενδελεχής αναφορά είναι πρακτικά αδύνατη. Συνεπώς το εγχείρημα αυτό επικεντρώνεται στα κύρια σημεία της θεωρίας, τα οποία ο μη εξοικειωμένος αναγνώστης πρέπει να έχει υπόψιν κατά την ανάγνωσή της.

Στο Κεφάλαιο 1, επιχειρείται ο ορισμός του αντικειμένου της Μηχανικής Μάθησης, αναφέρεται η σύνδεσή της με άλλα επιστημονικά πεδία καθώς και εφαρμογές της, εκθέτονται κάποιοι από τους βασικούς περιορισμούς της και τονίζεται η διαφορά της Μηχανικής Μάθησης με την ανθρώπινη.

Στο Κεφάλαιο 2, γίνεται μια αναφορά στα είδη της Μηχανικής Μάθησης με ιδιαίτερη έμφαση στην Επιβλεπόμενη Μάθηση, καθώς οι μέθοδοι που απαντώνται σε αυτή την Διατριβή ανήκουν σε αυτή την κατηγορία.

Το Κεφάλαιο 3 επικεντρώνεται στην Ενεργή Μάθηση. Γίνεται αναφορά στο αντικείμενό της, στα διάφορα σενάρια Ενεργής Μάθησης και στις στρατηγικές επιλογής ερωτημάτων, ενώ επίσης αναφέρονται οι περιορισμοί της.

Το Κεφάλαιο 4 επικεντρώνεται στην Ανάλυση Συναισθήματος όπου και δίνονται συνοπτικά πληροφορίες για τις διάφορες κατηγορίες και για μεθόδους που χρησιμοποιούνται στο πλαίσιο της παρούσας εργασίας.

Στο δεύτερο μέρος παρουσιάζονται τα πειραματικά αποτελέσματα της διατριβής, τα οποία σχετίζονται με μια σειρά διαφορετικών πραγματικών προβλημάτων. Το Κεφάλαιο 5 αφορά την πρόβλεψη χρονοσειρών χρηματοοικονομικών δεδομένων με χρήση Ανάλυσης Συναισθήματος. Στόχος είναι η βελτίωση των εξαγόμενων προβλέψεων. Το κεφάλαιο χωρίζεται πρακτικά σε τρία μέρη. Στο πρώτο, υλοποιείται μια εκτεταμένη σύγκριση κλασσικών Στατιστικών μεθόδων καθώς και σύγχρονων μεθόδων Μηχανικής Μάθησης, η οποία έχει εισαγωγικό χαρακτήρα. Τα αποτελέσματα που παρουσιάζονται στο πρώτο αυτό μέρος έχουν δημοσιευτεί στην εργασία μας με τίτλο *A comparison of contemporary methods on univariate time series fore-casting* στο Advances in Machine Learning/Deep Learning-based Technologies: Selected Papers in Honour of Professor Nikolaos G. Bourbakis– Vol. 2. Στη συνέχεια, παρουσιάζονται δύο συγκριτικές μελέτες που εμπλέκουν τη χρήση Ανάλυσης Συναισθήματος στην προβλεπτική διαδικασία και έχουν διττό σκοπό. Αφενός, τη σύγκριση των χρησιμοποιούμενων αλγορίθμων και αφετέρου την σύγκριση ενός μεγάλου πλήθους διαφορετικών συναισθηματικών διαμορφώσεων. Η πρώτη αφορά τόσο κλασικές μεθόδους Μηχανική

Μάθησης, όσο και μεθόδους Βαθιάς Μάθησης. Τα σχετικά αποτελέσματα έχουν δημοσιευτεί στην εργασία μας με τίτλο *A multi-method survey on the use of sentiment analysis in multivariate financial time series forecasting* στο περιοδικό Entropy. Η δεύτερη μελέτη αφορά μεθόδους Βαθιάς Μάθησης και δημοσιεύτηκε επίσης στο περιοδικό Entropy υπό τον τίτλο *Investigating deep stock market forecasting with sentiment analysis*.

Το Κεφάλαιο 6 πραγματεύεται την σύγκριση ενός μεγάλου αριθμού κλασσικών μεθόδων Μηχανικής Μάθησης και ομαδοποιητών αυτών καθώς και ενός συνόλου BERT-based transformers στο πρόβλημα ταξινόμησης κειμενικών δεδομένων οικονομικής φύσης. Η αντίστοιχη εργασία με τίτλο *Financial sentiment analysis: Classic methods vs. deep learning models* δημοσιεύτηκε στο περιοδικό Intelligent Decision Technologies.

Το Κεφάλαιο 7 επικεντρώνεται στη χρήση Ενεργής Μάθησης με στόχο τη μείωση της ανάγκης ετικετοποιημένων δεδομένων σε προβλήματα Ανάλυσης Συναισθήματος με χρήση BERT-based transformers. Τα εξαγόμενα αποτελέσματα είναι ιδιαίτερα ενθαρρυντικά, και βρίσκονται στο στάδιο της κρίσης ως προς δημοσίευση.

Το Κεφάλαιο 8 έχει παρόμοιο στόχο με το κεφάλαιο 7, όμως αφορά την μείωση τον ετικετοποιημένων δεδομένων στο πρόβλημα Πρόβλεψης Σφαλμάτων Λογισμικού. Τα αποτελέσματα που παρουσιάζονται δημοσιεύτηκαν στο περιοδικό Information sciences του Elsevier υπό τον τίτλο *Data-Efficient software defect prediction: a comparative analysis of active learningenhanced models and voting ensembles*. Στην ίδια εργασία προτείνεται ένας νέος μέτα-δείκτης για την αξιολόγηση σεναρίων Ενεργής Μάθησης ο οποίος λαμβάνει υπόψιν τόσο το ποσοστό μείωσης των δεδομένων, όσο και την ικανότητα του μοντέλου να επιτυγχάνει εφάμιλλή απόδοση με το σενάριο όπου αυτό εκπαιδεύεται στο σύνολο των δεδομένων.

Δημοσιεύσεις

Κατάτη διάρκεια εκπόνησης της Διδακτορικής Διατριβής δημοσιεύτηκαν οι παρακάτω εργασίες:

1. Charalampos M Liapis, Aikaterini Karanikola, and Sotiris Kotsiantis. "Enhancing sentiment analysis with distributional emotion embeddings". In: Neurocomputing (2025), p. 129822.

2. Charalampos M Liapis, Aikaterini Karanikola, and Sotiris Kotsiantis. "Emotion-Driven Energy Load Forecasting: An Ensemble Leveraging Insights from News". In: International Journal on Artificial Intelligence Tools 33.05 (2024), p. 2450013.

3. Charalampos M Liapis, Aikaterini Karanikola, and Sotiris Kotsiantis. "Data-Efficient software defect prediction: a comparative analysis of active learning-enhanced models and voting ensembles". In: Information sciences 676 (2024), p. 120786.

4. Aikaterini Karanikola et al. "Financial sentiment analysis: Classic meth- ods vs. deep learning models". In: Intelligent Decision Technologies 17.4 (2023), pp. 893–915.

5. Charalampos M Liapis, Aikaterini Karanikola, and Sotiris Kotsiantis. "A multivariate ensemble learning method for medium-term energy forecasting". In: Neural Computing and Applications 35.29 (2023), pp. 21479– 21497.

6. Charalampos M Liapis, Aikaterini Karanikola, and Sotiris Kotsiantis. "Investigating deep stock market forecasting with sentiment analysis". In: Entropy 25.2 (2023), p. 219.

7. Charalampos M Liapis, Aikaterini Karanikola, and Sotiris Kotsiantis. "Energy load forecasting: Investigating mid-term predictions with en- semble learners". In: IFIP International Conference on Artificial Intel- ligence Applications and Innovations. Springer. 2022, pp. 343–355

8. Aikaterini Karanikola, Charalampos M Liapis, and Sotiris Kotsiantis. "A comparison of contemporary methods on univariate time series fore- casting". In: Advances in Machine Learning/Deep Learning-based Tech- nologies: Selected Papers in Honour of Professor Nikolaos G. Bourbakis– Vol. 2 (2022), pp. 143–168.

9. Charalampos M Liapis, Aikaterini Karanikola, and Sotiris Kotsiantis. "A multi-method survey on the use of sentiment analysis in multivariate financial time series forecasting". In: Entropy 23.12 (2021), p. 1603.

10. Aikaterini Karanikola, Charalampos M Liapis, and Sotiris Kotsiantis. "Investigating cluster validation metrics for optimal number of clus- ters determination". In: Intelligent Decision Technologies 15.4 (2021), pp. 809–824.

11. liana Paliari, Aikaterini Karanikola, and Sotiris Kotsiantis. “A compar- ison of the optimized LSTM, XGBOOST and ARIMA in Time Series forecasting”. In: 2021 12th International Conference on Information, Intelligence, Systems & Applications (IISA). IEEE. 2021, pp. 1–7.

12. Aikaterini Karanikola, Charalampos M Liapis, and Sotiris Kotsiantis. “A comparative study of validity indices on estimating the optimal num- ber of clusters”. In: 2021 12th International Conference on Information, Intelligence, Systems & Applications (IISA). IEEE. 2021, pp. 1–8.

13. Charalampos M Liapis, Aikaterini Karanikola, and Sotiris Kotsiantis. “An ensemble forecasting method using univariate time series COVID- 19 data”. In: Proceedings of the 24th Pan-Hellenic Conference on In- formatics. 2020, pp. 50–52.

14. Aikaterini Karanikola and Sotiris Kotsiantis. “A hybrid method for missing value imputation”. In: Proceedings of the 23rd Pan-Hellenic Conference on Informatics. 2019, pp. 74–79.

15. Aikaterini Karanikola et al. “An incrementally updateable ensemble learner”. In: Proceedings of the 22nd Pan-Hellenic Conference on In- formatics. 2018, pp. 243–248

16. Stamatis Karlos et al. “Local weighted averaged 2-dependence estima- tor”. In: Proceedings of the 10th Hellenic Conference on Artificial In- telligence. 2018, pp. 1–4.

17. Aikaterini Karanikola et al. “Active fuzzy rule induction”. In: 2018 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS). IEEE. 2018, pp. 1–8.

Μέρος Ι

Θεωρητικό υπόβαθρο

1. Μηχανική Μάθηση

1.1 Ορισμοί της Μηχανικής Μάθησης

Στην πρώτη αυτή ενότητα επιχειρείται μια σύντομη και περιεκτική παρουσίαση του αντικειμένου της Μηχανικής Μάθησης (Machine Learning) καθώς και η περιγραφή της σύνδεσής της με άλλα πεδία της επιστήμης. Επιπλέον, αναφέρονται προβλήματα στα οποία η Μηχανική Μάθηση βρίσκει εφαρμογή, αλλά και συζητώνται οι περιορισμοί της και οι κύριες προκλήσεις που καλείται να αντιμετωπίσει. Ωστόσο, δεδομένου ότι το πεδίο της Μηχανικής Μάθησης είναι ιδιαίτερα εκτενές και προσελκύει τεράστιο ερευνητικό ενδιαφέρον - γεγονός που αποδεικνύεται από τον μεγάλο αριθμό δημοσιευμένων συγγραμμάτων και εργασιών - δεν είναι εφικτό να αποδοθεί πλήρως στο πλαίσιο της παρούσας διατριβής. Σκοπός της παρούσας ενότητας είναι, πρωτίστως, να αναδειχθούν οι βασικές αρχές και η ευρύτερη φιλοσοφία της Μηχανικής Μάθησης, προσφέροντας μια πρώτη, αλλά ουσιαστική, επαφή με τον κλάδο.

Οι απαρχές του αντικειμένου που σήμερα καλείται Μηχανική Μάθηση, βρίσκονται στη δεκαετία του 1950, όταν ο Alan Turing - ο πατέρας της πληροφορικής όπως μαθαίνουμε σήμερα στους μαθητές στο μάθημα της πληροφορικής - πρότεινε το λεγόμενο Turing Test στην εργασία του με τίτλο: Computing Machinery and Intelligence [276]. Στόχος του Turing Test είναι να αξιολογήσει αν ένας υπολογιστής μπορεί να μιμηθεί την ανθρώπινη νοημοσύνη. Αν ένας υπολογιστής καταφέρει να πείσει έναν άνθρωπο ότι είναι επίσης άνθρωπος μέσω γραπτής επικοινωνίας, τότε θεωρείται ότι έχει περάσει τη δοκιμασία, κάτι που στην εποχή μοντέλων όπως το ChatGPT-4 [197] μοιάζει ήδη μακρινό.

Παρ' όλα αυτά, η δεκαετία του '50 αποτέλεσε ορόσημο για τον κλάδο της Μηχανικής Μάθησης, διότι το 1952 - ο A. Samuel αναπτύσσει το πρώτο πρόγραμμα Μηχανικής Μάθησης, ένα πρόγραμμα Ντάμας (Checkers), το οποίο βελτιώνεται μόνο του από την εμπειρία του, παίζοντας με τον εαυτό του και η αντίστοιχη εργασία δημοσιεύεται λίγα χρόνια αργότερα [240]. Στη συνέχεια, μόλις πέντε χρόνια μετά, προτείνεται ο νευρώνας perceptron [230], πρωτόλεια μορφή των Νευρωνικών Δικτύων.

Οι επόμενες δεκαετίες υπήρξαν πλούσιες στην εισαγωγή μεθόδων που σήμερα θεωρούνται κλασικές, ενώ η αυγή της χιλιετίας μας βρίσκει μπροστά σε μια έκρηξη όγκου των διαθέσιμων δεδομένων, που οδήγησε στην ανάγκη ανάπτυξης νέων, ισχυρότερων τεχνικών, όπου η έννοια της Βαθιάς Μάθησης αρχίζει να διαφαίνεται και εν τέλει να ανθίζει μέσα στα επόμενα χρόνια. Σήμερα, στεκόμαστε μπροστά σε μια πληθώρα νέων εφαρμογών της Μηχανικής Μάθησης με πιο πρόσφατα παραδείγματα τα Μεγάλα Γλωσσικά Μοντέλα (Large Language Models - LLMs), τα οποία ελκύουν το ενδιαφέρον της επιστημονικής - και όχι μόνο - κοινότητας.

Τι είναι όμως η Μηχανική Μάθηση; Θα ήταν τουλάχιστον αφελές να προσπαθήσουμε να δώσουμε έναν ορισμό που να καλύπτει κάθε πτυχή της Μηχανικής Μάθησης, καθώς περιλαμβάνει ένα μεγάλο εύρος μεθόδων και τεχνικών. Κατά συνέπεια, πολλοί

ορισμοί έχουν προταθεί στη βιβλιογραφία, καθένας από τους οποίους κάτι προσθέτει και κάτι αφαιρεί από την περιγραφή του πεδίου αυτού. Για το λόγο αυτό, θα παρουσιαστούν μερικοί από τους πιο δημοφιλείς και περιεκτικούς ορισμούς που έχουν δοθεί ως σήμερα.

Ο πρώτος ορισμός που δόθηκε, αποδίδεται στον A. Samuel και χρονολογείται το 1959. Σύμφωνα με τον A. Samuel [240]:

> «*Η Μηχανική Μάθηση είναι ο τομέας της μελέτης που δίνει στους υπολογιστές τη δυνατότητα να μαθαίνουν χωρίς να έχουν προγραμματιστεί ρητά.*»

Ο ορισμός αυτός, αν και απλός, εστιάζει στην ιδέα ότι ένας υπολογιστής μπορεί να μάθει χωρίς να του έχουν δοθεί ρητές οδηγίες. Δίνεται έμφαση στην αυτοματοποίηση της μάθησης μέσω της εμπειρίας, κάτι που αποτελεί τον ακρογωνιαίο λίθο των τεχνικών της Μηχανικής Μάθησης.

Ένας σαφώς πιο φορμαλιστικός ορισμός δόθηκε από τον T. Mitchell στο βιβλίο του Machine Learning [173]:

> «*Ένα πρόγραμμα υπολογιστή λέγεται ότι μαθαίνει από την εμπειρία E, σε σχέση με μια κατηγορία εργασιών T και ένα μέτρο απόδοσης Π, εάν η απόδοσή του στις εργασίες T, όπως μετράται από το Π, βελτιώνεται με την εμπειρία E.*»

Ο ορισμός αυτός παραμένει από τους πιο ευρέως διαδεδομένους ως και τις μέρες μας, όχι μόνο εξαιτίας της απλότητάς του, αλλά κυρίως γιατί τονίζει ότι η Μηχανική Μάθηση βασίζεται στην εμπειρία (δηλαδή στα δεδομένα) και στη βελτίωση της απόδοσης, γεγονός που την ξεχωρίζει από παραδοσιακούς αλγορίθμους που είναι αυστηρά προγραμματισμένοι. Επιπλέον, παρέχει έναν μετρήσιμο τρόπο αξιολόγησης της μάθησης, καθώς η πρόοδος μετράται βάσει της απόδοσης στο συγκεκριμένο έργο.

Δε θα μπορούσαν να λείπουν ορισμοί που συνδέουν τη Μηχανική Μάθηση με τη στατιστική, καθώς τα δύο αυτά πεδία είναι άρρηκτα συνδεδεμένα. Ένας από αυτούς είναι ο ορισμός του K. Murphy [180]:

> «*Η Μηχανική Μάθηση είναι ένα σύνολο μεθόδων που ανιχνεύουν αυτόματα πρότυπα στα δεδομένα και στη συνέχεια χρησιμοποιούν τα αποκαλυφθέντα πρότυπα για να προβλέψουν μελλοντικά δεδομένα ή να λάβουν άλλου είδους αποφάσεις υπό συνθήκες αβεβαιότητας.*»

Στον ορισμό του K. Murphy, η Μηχανική Μάθηση περιγράφεται ως μια συλλογή μεθόδων για την αναγνώριση προτύπων στα δεδομένα και τη χρήση αυτών για πρόβλεψη ή λήψη αποφάσεων. Με αυτόν τον τρόπο όχι μόνο δίνεται έμφαση στην ανίχνευση προτύπων αλλά και αναδεικνύεται η έννοια της αβεβαιότητας, που είναι κεντρική στις μεθόδους Μηχανικής Μάθησης που βασίζονται στην στατιστική.

Η σύνδεση της Μηχανικής Μάθησης με τα μαθηματικά γίνεται επίσης εμφανής στον ορισμό του E. Alpaydin [10], σύμφωνα με τον οποίο:

> «*Η Μηχανική Μάθηση είναι ο προγραμματισμός υπολογιστών ώστε να βελτιστοποιούν ένα κριτήριο απόδοσης χρησιμοποιώντας παραδείγματα ή παρελθοντική εμπειρία.*»

Ο ορισμός αυτός εστιάζει στη βελτιστοποίηση της απόδοσης ενός αλγορίθμου, χρησιμοποιώντας δεδομένα ή εμπειρία. Εισάγει επίσης μια έννοια από τα μαθηματικά και την επιστήμη των υπολογιστών – την έννοια της βελτιστοποίησης – εισάγωντας στη Μηχανική Μάθηση μετρήσιμα κριτήρια, κάνοντάς την έτσι ακριβέστερη και περισσότερο συνδεδεμένη με τις εφαρμοσμένες μαθηματικές και στατιστικές μεθόδους. Έτσι, επιτρέπει στους μηχανικούς και στους επιστήμονες δεδομένων να προσεγγίζουν τη Μηχανική Μάθηση ως ένα πρόβλημα βελτιστοποίησης, με σαφώς καθορισμένους στόχους και μετρικές απόδοσης.

Η ύπαρξη πολλών ορισμών για την περιγραφή του περιεχομένου της Μηχανικής Μάθησης είναι σημαντική γιατί αντανακλά τις διαφορετικές προσεγγίσεις, εφαρμογές και θεωρητικές βάσεις του πεδίου. Κάποιοι ορισμοί, όπως αυτός του A. Samuel, εστιάζουν στη γενική έννοια της αυτοματοποίησης της μάθησης, ενώ άλλοι, όπως του T. Mitchell, προσφέρουν μια αυστηρή μαθηματική διατύπωση που επιτρέπει ακριβή ανάλυση και αξιολόγηση της απόδοσης. Ορισμοί όπως του K. Murphy και του E. Alpaydin δίνουν έμφαση σε πτυχές όπως η ανίχνευση προτύπων και η βελτιστοποίηση, που είναι κρίσιμες για πρακτικές εφαρμογές. Αυτή η ποικιλομορφία βοηθά στην κατανόηση της Μηχανικής Μάθησης από διαφορετικές σκοπιές— θεωρητική, εφαρμοσμένη και μαθηματική —και επιτρέποντας σε ερευνητές και επαγγελματίες να επιλέγουν τον πιο κατάλληλο ορισμό ανάλογα με το πλαίσιο στο οποίο εργάζονται.

1.2 Σύνδεση της Μηχανικής Μάθησης με άλλα πεδία της επιστήμης

Είναι εμφανές μέσα από τους παραπάνω ορισμούς, ότι η Μηχανική Μάθηση είναι ένα πολυδιάστατο επιστημονικό πεδίο που αντλεί στοιχεία από διάφορες επιστήμες, συμπεριλαμβανομένων της Στατιστικής, της Πληροφορικής, της Τεχνητής Νοημοσύνης και της Επιστήμης των Δεδομένων. Το διάγραμμα Venn στο Σχήμα 1.1 απεικονίζει τη σχέση της Μηχανικής Μάθησης με αυτούς τους τέσσερις κλάδους, καταδεικνύοντας πώς συνδυάζει γνώσεις και τεχνικές από κάθε έναν εξ αυτών. Στη συνέχεια παρουσιάζεται η σύνδεση της Μηχανικής Μάθησης με κάθε κάθε ένα από τα παραπάνω επιστημονικά παιδία.

- **Στατιστική (Statistics)**: Η Στατιστική παρέχει τη θεωρητική βάση για τη Μηχανική Μάθηση, καθώς πολλές από τις μεθόδους που χρησιμοποιούνται στη διαδικασία της μάθησης βασίζονται σε στατιστικά μοντέλα και στη θεωρία πιθανοτήτων. Τεχνικές όπως η γραμμική και λογιστική παλινδρόμηση, η ανάλυση διακύμανσης (ANOVA) και οι πιθανοτικές μέθοδοι της οικογένειας Bayes αποτελούν βασικά εργαλεία τόσο για τη Στατιστική όσο και για τη Μηχανική Μάθηση.

- **Επιστήμη των Δεδομένων (Data Science)**: Η Επιστήμη των Δεδομένων επικεντρώνεται στη διαχείριση, ανάλυση και εξαγωγή γνώσης από μεγάλες ποσότητες δεδομένων. Η Μηχανική Μάθηση αποτελεί βασικό εργαλείο της, καθώς επιτρέπει την αυτοματοποίηση της ανάλυσης δεδομένων και την ανάπτυξη προβλεπτικών μοντέλων. Οι πρακτικές της Επιστήμης των Δεδομένων, όπως η προεπεξεργασία και η οπτικοποίηση δεδομένων, αλλά και η ανάλυση χαρακτηριστικών, είναι αναπόσπαστο μέρος των μοντέλων Μηχανικής Μάθησης.

- **Τεχνητή Νοημοσύνη (Artificial Intelligence - AI)**: Η Τεχνητή Νοημοσύνη αποτελεί ένα ευρύτερο πεδίο που στοχεύει στη δημιουργία υπολογιστικών συστημάτων που μπορούν να εκτελούν καθήκοντα τα οποία απαιτούν νοημοσύνη. Η Μηχανική Μάθηση αποτελεί μία από τις κύριες τεχνικές που χρησιμοποιούνται στην ανάπτυξη έξυπνων συστημάτων, καθώς της επιτρέπει να αναγνωρίζει πρότυπα και να λαμβάνει αποφάσεις βασισμένες σε δεδομένα. Η Βαθιά Μάθηση, η οποία συγκεντρώνει τεράστιο όγκο έρευνας στις μέρες μας, αποτελεί υποσύνολο της τομής της Μηχανικής Μάθησης και της Τεχνητής Νοημοσύνης, το οποίο έχει συμβάλει σημαντικά στην ανάπτυξη εφαρμογών όπως η αναγνώριση εικόνας και η επεξεργασία φυσικής γλώσσας.
- **Πληροφορική (Computer Science)**: Η Μηχανική Μάθηση έχει άμεση σχέση με την Πληροφορική, καθώς απαιτεί την ανάπτυξη αποδοτικών αλγορίθμων και την κατανόηση της θεωρίας της πολυπλοκότητας. Οι κλάδοι των δομών δεδομένων, της βελτιστοποίησης αλγορίθμων και της κατανεμημένης επεξεργασίας διαδραματίζουν κρίσιμο ρόλο στην δημιουργία και την βελτιστοποίηση των μοντέλων Μηχανικής Μάθησης.

1.3 Εφαρμογές της Μηχανικής Μάθησης

Η Μηχανική Μάθηση έχει αναδειχθεί ως ένας από τους πιο καθοριστικούς τεχνολογικούς τομείς του 21ου αιώνα, επιτρέποντας την αυτοματοποίηση πολύπλοκων διαδικασιών και τη βελτίωση της απόδοσης σε πλήθος εφαρμογών. Με την εκθετική αύξηση των δεδομένων και των υπολογιστικών δυνατοτήτων, η Μηχανική Μάθηση έχει συμβάλει καθοριστικά σε πληθώρα εφαρμογών που αφορούν τη βιομηχανία, την επιστήμη και την καθημερινή ζωή.



Σχήμα 1.1: Διάγραμμα Venn για τη σύνδεση της Μηχανικής Μάθησης με άλλα επιστημονικά πεδία

Συνοπτικά, στη συνέχεια παρουσιάζονται κάποιοι μόνο από τους τομείς στους οποίους η Μηχανική Μάθηση βρίσκει εφαρμογή:

- **Υγεία και Ιατρική Διάγνωση**: Στον τομέα της υγείας, η Μηχανική Μάθηση έχει επιφέρει σημαντικές βελτιώσεις στην ακρίβεια και την ταχύτητα των διαγνωστικών διαδικασιών. Παραδείγματα αποτελούν η πρόβλεψη της καρδιοπάθειας [175] και της καρδιακής ανεπάρκειας [275], η διάγνωση και η κατάλληλη επιλογή θεραπείας κατά του καρκίνου [308, 265] και η πρόγνωση της εξέλιξης του διαβήτη [203]. Επιπλέον, κατά την πρόσφατη πανδημία του COVID, η Μηχανική Μάθηση χρησιμοποιήθηκε για την πρόβλεψη της θνησιμότητας [179], τον εντοπισμό κρουσμάτων [3] και περιπτώσεων long COVID [211], αλλά και εν γένει για τη διαχείριση της πανδημίας [117].

- **Οικονομία**: Στον τομέα της οικονομίας, η συμβολή της Μηχανικής Μάθησης είναι επίσης σημαντική. Εφαρμογές περιλαμβάνουν την πρόβλεψη των τιμών κλεισίματος του χρηματιστηρίου [209, 79], τον εντοπισμό οικονομικής απάτης [171, 185] και τον έλεγχο του οικονομικού ρίσκου [91].

- **Εκπαίδευση**: Στην εκπαίδευση, η ενσωμάτωση της Μηχανικής Μάθησης έχει επιφέρει σημαντικές αλλαγές, βελτιώνοντας τη μαθησιακή εμπειρία και ενισχύοντας τις διδακτικές πρακτικές. Συγκεκριμένα, οι αλγόριθμοι Μηχανικής Μάθησης χρησιμοποιούνται για την ανάπτυξη εξατομικευμένων προγραμμάτων μάθησης, τα οποία προσαρμόζονται στις ανάγκες και τον ρυθμό κάθε μαθητή, επιτρέποντας μια πιο στοχευμένη και αποτελεσματική εκπαιδευτική διαδικασία [47]. Παράλληλα, η ενσωμάτωση εργαλείων Μηχανικής Μάθησης στην τάξη, όπως οι εικονικοί δάσκαλοι [150] και τα διαδραστικά συστήματα διδασκαλίας [72], ενισχύει την εμπλοκή των μαθητών και προωθεί την ενεργή μάθηση. Επιπλέον, πέρα από τη βελτίωση της διδασκαλίας και της εμπειρίας των μαθητών, η Μηχανική Μάθηση έχει συμβάλει και σε άλλους τομείς της εκπαίδευσης, όπως η πρόβλεψη και η πρόληψη της σχολικής διαρροής [183].

- **Βιομηχανία**: Στη βιομηχανία, η χρήση μεθόδων Μηχανικής Μάθησης εντοπίζεται επίσης σε πολλά στάδια. Μία από τις κύριες εφαρμογές της είναι ο εντοπισμός βλαβών [201] και η πρόβλεψη για ανάγκη συντήρησης μηχανημάτων [132], όπου οι αλγόριθμοι Μηχανικής Μάθησης αναλύουν δεδομένα από αισθητήρες εξοπλισμού για να προβλέψουν πιθανές βλάβες, επιτρέποντας την έγκαιρη παρέμβαση και μειώνοντας τον χρόνο διακοπής της παραγωγής. Επιπλέον, στον τομέα της βιομηχανικής όρασης [310], οι τεχνικές Μηχανικής Μάθησης χρησιμοποιούνται για την αναγνώριση και κατηγοριοποίηση αντικειμένων με υψηλή ακρίβεια, βελτιώνοντας τον ποιοτικό έλεγχο και την αυτοματοποίηση της παραγωγής. Παράλληλα, η Μηχανική Μάθηση συμβάλλει στη βελτιστοποίηση της αλυσίδας εφοδιασμού [82], προβλέποντας τη ζήτηση και διαχειρίζοντας τα αποθέματα με μεγαλύτερη αποτελεσματικότητα.

- **Κυβερνοασφάλεια**: Η Μηχανική Μάθηση διαδραματίζει καθοριστικό ρόλο και στην ενίσχυση της κυβερνοασφάλειας, προσφέροντας προηγμένα εργαλεία για την ανίχνευση και αντιμετώπιση απειλών στον κυβερνοχώρο. Συγκεκριμένα, οι τεχνικές Μηχανικής Μάθησης εφαρμόζονται σε συστήματα ανίχνευσης εισβολών [133] για την αναγνώριση μη εξουσιοδοτημένων προσβάσεων και την προστασία

δικτύων και συστημάτων. Επιπλέον, χρησιμοποιούνται για την ανίχνευση κακόβουλου λογισμικού [4] μέσω της ανάλυσης συμπεριφοράς εφαρμογών, καθώς και για την πρόληψη επιθέσεων phishing [16] με την αναγνώριση ύποπτων προτύπων σε ηλεκτρονικές επικοινωνίες. Η ενσωμάτωση της Μηχανικής Μάθησης στην κυβερνοασφάλεια επιτρέπει την προσαρμογή σε νέες και εξελισσόμενες απειλές, βελτιώνοντας συνεχώς την αποτελεσματικότητα των διαθέσιμων αμυντικών μηχανισμών.

1.4 Κύριες Προκλήσεις και Περιορισμοί της Μηχανικής Μάθησης

Η Μηχανική Μάθηση, παρά τη ραγδαία ανάπτυξή της και τις πολυάριθμες εφαρμογές της, εξακολουθεί να αντιμετωπίζει σημαντικές προκλήσεις και περιορισμούς που σχετίζονται με την ποιότητα των δεδομένων, την ερμηνευσιμότητα των μοντέλων, τις υπολογιστικές απαιτήσεις, καθώς και με ηθικά και δεοντολογικά ζητήματα. Ένα από τα κύρια ζητήματα που ανακύπτουν αφορά την εξάρτηση των μοντέλων από τα δεδομένα. Η απόδοση των αλγορίθμων επηρεάζεται καθοριστικά από την ποιότητα και την ποσότητα των δεδομένων που χρησιμοποιούνται κατά την εκπαίδευσή τους. Η ανεπαρκής ή μεροληπτική συλλογή δεδομένων μπορεί να οδηγήσει σε σφάλματα και αναξιόπιστες προβλέψεις [24, 38], ενώ σε ορισμένους τομείς, όπως η ιατρική διάγνωση, οι περιορισμοί στην πρόσβαση σε δεδομένα λόγω ζητημάτων ιδιωτικότητας καθιστούν την ανάπτυξη ακριβών μοντέλων ιδιαίτερα δύσκολη. Επιπλέον, η μεροληψία που ενυπάρχει σε πολλά σύνολα δεδομένων μπορεί να μεταφέρεται στα αποτελέσματα των αλγορίθμων, δημιουργώντας άδικες ή μη αντικειμενικές αποφάσεις, γεγονός που έχει οδηγήσει σε ανησυχίες σχετικά με την αξιοπιστία των μοντέλων αυτών, εγείροντας ερωτήματα για το κατά πόσον είναι «δίκαια» [172].

Ένα δεύτερο ζήτημα που αναδεικνύεται είναι η έλλειψη ερμηνευσιμότητας και διαφάνειας, ιδιαίτερα στις περιπτώσεις όπου χρησιμοποιούνται πολύπλοκα μοντέλα, όπως τα νευρωνικά δίκτυα Βαθιάς Μάθησης [234]. Τα μοντέλα αυτά λειτουργούν συχνά ως «μαύρα κουτιά», γεγονός που καθιστά δύσκολη την εξήγηση των αποφάσεών τους, ειδικά σε κρίσιμους τομείς όπως η ιατρική και τα χρηματοοικονομικά. Παρά τις προσπάθειες ανάπτυξης τεχνικών που επιχειρούν να προσδώσουν διαφάνεια, όπως οι μέθοδοι SHAP [154] και LIME [226], η πλήρης κατανόηση του τρόπου με τον οποίο λαμβάνονται οι αποφάσεις των αλγορίθμων παραμένει πρόκληση. Η απουσία επεξηγηματικότητας δημιουργεί προβλήματα εμπιστοσύνης και αποδοχής των μοντέλων, ιδιαίτερα σε περιβάλλοντα όπου απαιτείται ανθρώπινη επίβλεψη.

Ένας ακόμη περιορισμός αφορά το υψηλό υπολογιστικό κόστος και την ενεργειακή κατανάλωση που απαιτείται για την εκπαίδευση και τη λειτουργία προηγμένων μοντέλων. Οι σύγχρονοι αλγόριθμοι, ιδιαίτερα οι αλγόριθμοι Βαθιάς Μάθησης, απαιτούν εξαιρετικά ισχυρό υλικό, όπως τη χρήση GPU και TPU, γεγονός που αυξάνει το κόστος ανάπτυξης και περιορίζει τη δυνατότητα πρόσβασης σε τέτοιες τεχνολογίες για μικρότερες ερευνητικές ομάδες και επιχειρήσεις. Επιπλέον, το περιβαλλοντικό αποτύπωμα της εκπαίδευσης μεγάλων μοντέλων είναι αξιοσημείωτο, καθώς η εκπαίδευση ενός μόνο προηγμένου γλωσσικού μοντέλου μπορεί να καταναλώσει μεγάλη ποσότητα ενέργειας [260], εγείροντας ανησυχίες για τη βιωσιμότητα αυτών των τεχνολογιών.

Εκτός από τις πρακτικές προκλήσεις, η Μηχανική Μάθηση παρουσιάζει και σημαντικές αδυναμίες στον τομέα της ασφάλειας, καθώς τα μοντέλα είναι επιρρεπή σε

επιθέσεις αντιπαραδείγματος (adversarial attacks), όπου εισάγονται τεχνητές μεταβολές στα δεδομένα εισόδου με σκοπό να παραπλανηθεί το σύστημα [101]. Τέτοιες επιθέσεις μπορεί να έχουν σοβαρές επιπτώσεις σε εφαρμογές όπως η κυβερνοασφάλεια, η αυτόνομη οδήγηση και η βιομετρική ταυτοποίηση. Επιπλέον, εγείρονται ανησυχίες σχετικά με την προστασία των δεδομένων, καθώς επιτιθέμενοι μπορούν να εκμεταλλευτούν αλγόριθμους Μηχανικής Μάθησης για να εξάγουν προσωπικές πληροφορίες από μοντέλα που έχουν εκπαιδευτεί σε ευαίσθητα δεδομένα [253].

Τέλος, η ίδια η χρήση της Μηχανικής Μάθησης εγείρει σοβαρά ηθικά και δεοντολογικά ζητήματα. Τα μοντέλα που χρησιμοποιούνται για τη λήψη αποφάσεων ενδέχεται να αναπαράγουν ή και να ενισχύσουν κοινωνικές προκαταλήψεις, εφόσον εκπαιδεύονται σε δεδομένα που δεν είναι αντικειμενικά ή περιλαμβάνουν διακρίσεις. Ορισμένες εφαρμογές, όπως η ανάλυση προσώπου και η επιτήρηση μέσω αλγορίθμων, έχουν προκαλέσει ανησυχίες για τη διατήρηση της ιδιωτικότητας και των ανθρωπίνων δικαιωμάτων, ενώ η χρήση αυτόνομων συστημάτων σε στρατιωτικά περιβάλλοντα θέτει ερωτήματα σχετικά με τον έλεγχο και την ευθύνη για τις αποφάσεις που λαμβάνουν αυτά τα μοντέλα. Η απουσία σαφούς νομικού πλαισίου για τη ρύθμιση αυτών των τεχνολογιών καθιστά ακόμη πιο δύσκολη τη διαχείριση των ηθικών προκλήσεων που προκύπτουν από τη χρήση τους [131].

Λαμβάνοντας υπόψιν όλα τα παραπάνω, η Μηχανική Μάθηση παρουσιάζει τεράστιες δυνατότητες, αλλά ταυτόχρονα συνοδεύεται από σοβαρές προκλήσεις που πρέπει να αντιμετωπιστούν προκειμένου να διασφαλιστεί η ασφαλής, δίκαιη και υπεύθυνη ανάπτυξή της. Η εξέλιξη του πεδίου απαιτεί τη συνεχή βελτίωση των μεθόδων διαχείρισης δεδομένων, τη μεγαλύτερη διαφάνεια στη λήψη αποφάσεων, την ανάπτυξη ενεργειακά αποδοτικών υπολογιστικών μεθόδων και τη διαμόρφωση ενός αυστηρότερου κανονιστικού πλαισίου που θα προστατεύει τόσο την ιδιωτικότητα όσο και τα δικαιώματα των χρηστών.

1.5 Ανθρώπινη μάθηση vs μάθηση μηχανής

Η διαδικασία μάθησης στους υπολογιστές διαφέρει ριζικά από την ανθρώπινη μάθηση, καθώς δεν βασίζεται στη βιωματική εμπειρία, τη διαίσθηση ή τη δημιουργικότητα, αλλά στην επεξεργασία δεδομένων και τη βελτιστοποίηση αλγορίθμων [173, 235], όπως ήδη έγινε εμφανές από τους ορισμούς της Μηχανικής Μάθησης. Στην ουσία, η Μηχανική Μάθηση επιτρέπει στους υπολογιστές να αναγνωρίζουν μοτίβα και να εξάγουν χρήσιμα συμπεράσματα από δεδομένα, βελτιώνοντας την απόδοσή τους σε συγκεκριμένα καθήκοντα μέσω της εμπειρίας [29, 180]. Η μαθησιακή διαδικασία μπορεί να συγκριθεί με τον τρόπο με τον οποίο ένα παιδί μαθαίνει να αναγνωρίζει αντικείμενα ή να επιλύει προβλήματα, αλλά αντί να βασίζεται στην ανθρώπινη γνώση, χρησιμοποιεί στατιστικά μοντέλα και αλγοριθμικές τεχνικές [99].

Η μάθηση ενός υπολογιστή βασίζεται σε δύο βασικές αρχές: την αναγνώριση μοτίβων μέσα από μεγάλα σύνολα δεδομένων και τη βελτιστοποίηση της απόδοσης με την πάροδο του χρόνου [11]. Όταν ένας αλγόριθμος λαμβάνει ως είσοδο ένα σύνολο δεδομένων, ξεκινά αναζητώντας επαναλαμβανόμενες σχέσεις και συσχετίσεις μεταξύ των στοιχείων αυτών. Μέσα από αυτή τη διαδικασία, το μοντέλο «μαθαίνει» να προβλέπει αποτελέσματα ή να κατηγοριοποιεί νέες εισόδους βασιζόμενο σε παλαιότερες παρατηρήσεις. Για παράδειγμα, ένας αλγόριθμος που έχει εκπαιδευτεί να αναγνωρίζει χειρόγραφα ψηφία μπορεί, μέσω της σταδιακής προσαρμογής των παραμέτρων του, να αυξήσει την ακρίβειά του στην αναγνώριση νέων χαρακτήρων.

Η διαδικασία μάθησης των υπολογιστών μπορεί να διαχωριστεί σε διάφορες κατηγορίες, ανάλογα με τον τρόπο με τον οποίο ο αλγόριθμος λαμβάνει και επεξεργάζεται την πληροφορία. Στην επιβλεπόμενη μάθηση (η οποία συζητάται εκτενέστερα στη συνέχεια, καθώς στην κατηγορία αυτή ανήκουν οι προβληματικές που εξετάζονται στην παρούσα διατριβή), το μοντέλο εκπαιδεύεται με δεδομένα που συνοδεύονται από ετικέτες, δηλαδή κάθε δείγμα ανήκει εκ των προτέρων σε μια γνωστή κατηγορία [173]. Ένας αλγόριθμος ταξινόμησης εικόνων, για παράδειγμα, μπορεί να λάβει ως είσοδο χιλιάδες εικόνες γατών και σκύλων, όπου κάθε εικόνα έχει την αντίστοιχη ετικέτα («γάτα» ή «σκύλος»). Με την πάροδο του χρόνου, το μοντέλο μαθαίνει να συσχετίζει χαρακτηριστικά όπως το σχήμα, το χρώμα και η υφή με τη σωστή κατηγορία, με αποτέλεσμα να μπορεί να αποδόσει σε μια νέα άγνωστη εικόνα, την κατηγορία «γάτα» ή «σκύλος». Στην μη επιβλεπόμενη μάθηση, ο αλγόριθμος δεν λαμβάνει ετικέτες αλλά προσπαθεί να ανακαλύψει μοτίβα και κρυφές σχέσεις μέσα στα δεδομένα, ενώ η προσέγγιση αυτή είναι ιδιαίτερα χρήσιμη για τη δημιουργία ομάδων (*clustering*) και την εύρεση κρυφών δομών σε δεδομένα [180]. Άλλες προσεγγίσεις, χρησιμοποιούν ένα σύστημα ανταμοιβών και ποινών για τον αλγόριθμο, ανάλογα με τις ενέργειες που εκτελεί [263], ενισχύοντας έτσι την καλή συμπεριφορά του. Με αυτό τον τρόπο, προσομοιάζεται η ανθρώπινη μάθηση μέσω δοκιμής και σφάλματος. Ένα κλασικό παράδειγμα αυτής της προσέγγισης, είναι το AlphaGo, το οποίο κατάφερε να ξεπεράσει τους κορυφαίους ανθρώπινους παίκτες του παιχνιδιού Go, μαθαίνοντας μέσα από εκατομμύρια παρτίδες και βελτιστοποιώντας τη στρατηγική του μέσω ανταμοιβών [254].

Ανεξάρτητα από τη συγκεκριμένη μέθοδο που χρησιμοποιείται, όσον αφορά τους υπολογιστές, η διαδικασία της μάθησης εξαρτάται σε μεγάλο βαθμό από την ποιότητα των δεδομένων, τον σχεδιασμό των αλγορίθμων και την ικανότητα γενίκευσης του μοντέλου. Εάν ένα μοντέλο έχει εκπαιδευτεί αποκλειστικά σε ένα περιορισμένο ή προκατειλημμένο σύνολο δεδομένων, ενδέχεται να μην μπορεί να αποδώσει καλά σε νέα, άγνωστα δεδομένα. Για τον λόγο αυτό, οι ερευνητές αναπτύσσουν τεχνικές όπως η διασταυρούμενη επικύρωση (cross-validation) και η κανονικοποίηση (regularization), ώστε να διασφαλίσουν ότι ένα μοντέλο δεν προσαρμόζεται υπερβολικά σε συγκεκριμένα δεδομένα εκπαίδευσης αλλά μπορεί να εφαρμόσει τη γνώση του σε διαφορετικές συνθήκες [112].

Επιπλέον, ακόμη ένας λόγος για τον οποίο οι υπολογιστές δεν μαθαίνουν με τον ίδιο τρόπο που μαθαίνει ο άνθρωπος, είναι ότι στερούνται εννοιών όπως η διαίσθηση, η δημιουργικότητα και η κοινή λογική [235]. Ενώ οι αλγόριθμοι μπορούν να εξάγουν πολύπλοκα μοτίβα από δεδομένα και να πραγματοποιούν ακριβείς προβλέψεις, εξακολουθούν να έχουν περιορισμούς στην κατανόηση και στη λήψη αποφάσεων σε ασαφή ή απρόβλεπτα περιβάλλοντα. Αυτός είναι και ο λόγος που η ανάπτυξη υβριδικών μεθόδων, οι οποίες συνδυάζουν τη στατιστική ανάλυση με την ανθρώπινη επίβλεψη και παρέμβαση, καθίσταται απαραίτητη σε κρίσιμες εφαρμογές, όπως η ιατρική διάγνωση και η αυτόνομη οδήγηση.

Καθώς η Μηχανική Μάθηση συνεχίζει να εξελίσσεται, νέες τεχνικές, όπως τα μετα-μαθησιακά (meta-learning) συστήματα [85] και τα γενικευμένα μοντέλα τεχνητής νοημοσύνης (Artificial general intelligence - AGI), υπόσχονται να γεφυρώσουν το χάσμα μεταξύ της ανθρώπινης και της μηχανικής νοημοσύνης [97]. Παρ΄ όλα αυτά, οι υπολογιστές εξακολουθούν να βασίζονται αποκλειστικά σε αλγοριθμικές δομές και δεν μπορούν να αναπτύξουν συνείδηση ή πραγματική κατανόηση του κόσμου όπως οι άνθρωποι. Το πως θα εξελιχθεί αυτή η σχέση μεταξύ ανθρώπου και τεχνητής νοημοσύνης στο μέλλον παραμένει ένα από τα πιο ενδιαφέροντα και ανοιχτά ερωτήματα

της σύγχρονης επιστήμης.

2. Κατηγορίες Μηχανικής Μάθησης

Η Μηχανική Μάθηση περιλαμβάνει ένα ευρύ φάσμα τεχνικών και μεθόδων που επιτρέπουν στους υπολογιστές να αναλύουν δεδομένα, να αναγνωρίζουν μοτίβα και να λαμβάνουν αποφάσεις χωρίς ρητό προγραμματισμό. Οι αλγόριθμοι της Μηχανικής Μάθησης διακρίνονται σε διαφορετικές κατηγορίες, ανάλογα με τον τρόπο εκπαίδευσης και τον τρόπο με τον οποίο χρησιμοποιούν τα δεδομένα. Οι τρεις βασικές κατηγορίες είναι η επιβλεπόμενη μάθηση (supervised learning), η μη επιβλεπόμενη μάθηση (unsupervised learning) και η ενισχυτική μάθηση (reinforcement learning), ενώ σε ορισμένες περιπτώσεις αναγνωρίζεται και μια τέταρτη κατηγορία, η ημι-επιβλεπόμενη μάθηση (semi-supervised learning). Στη συνέχεια, θα παρουσιαστούν οι παραπάνω κατηγορίες, δίνοντας έμαφηση στην πρώτη, που απασχολέι και το μεγαλύτερο μέρος της παρούσας εργασίας.

2.1 Επιβλεπόμενη Μηχανική Μάθηση

Ως επιβλεπόμενη μάθηση (supervised learning) ορίζεται η υποκατηγορία της Μηχανικής Μάθησης όπου ο εκμαθητής (learner) εκπαιδεύεται σε ένα σύνολο επισημασμένων ή αλλιώς ετικετοποιημένων παραδειγμάτων - δηλαδή παραδειγμάτων των οποίων η ετικέτα (label) είναι γνωστή - με στόχο να μπορεί να προβλέψει επιτυχώς την ετικέτα μη ετικετοποιημένων παραδειγμάτων [173, 29]. Συγκεκριμένα, στην περίπτωση της επιβλεπόμενης μάθησης σκοπός είναι η εκμάθηση μιας συνάρτησης:

$$f : X \to Y \tag{2.1}$$

όπου: X ο χώρος εισόδου ή αλλιώς χώρος χαρακτηριστικών (input/feature space) και Y ο χώρος εξόδου, με βάση ένα σύνολο επισημασμένων παραδειγμάτων, το λεγόμενο σύνολο δεδομένων:

$$\{(x_i, y_i)\}_{i=1}^{n}, \tag{2.2}$$

όπου $x_i \in X \subseteq R^d$ και $y_i \in Y$ και n το πλήθος των παραδειγμάτων.

Στόχος είναι η συνάρτηση να αποδίδει αξιόπιστες προβλέψεις για νέα δείγματα των οποίων η ετικέτα είναι άγνωστη, μεγιστοποιώντας έτσι την ικανότητα γενίκευσης [11].

Η επιβλεπόμενη μάθηση αποτελεί τομέα που παρέχει ισχυρά εργαλεία που επιτρέπουν ακριβείς προβλέψεις και ενισχύουν τη βέλτιστη λήψη αποφάσεων με βάση τα δεδομένα [180]. Παράλληλα, δίνει τη δυνατότητα κατασκευής σύνθετων μοντέλων που διαχειρίζονται με αποτελεσματικότητα μεγάλα ή απαιτητικά σύνολα δεδομένων, εφόσον όμως διασφαλίζεται η ποιότητα και η αντιπροσωπευτικότητα των επισημασμένων δεδομένων. Ωστόσο, υπάρχουν και μειονεκτήματα, το σημαντικότερο ίσως είναι ο κίνδυνος υπερπροσαρμογής (overfitting), όπου το μοντέλο μαθαίνει υπέρμετρα τις ιδιαιτερότητες του συνόλου εκπαίδευσης και παρουσιάζει μειωμένη απόδοση σε νέα δεδομένα[69, 227].

Επιπλέον, η διαδικασία εξαγωγής κατάλληλων χαρακτηριστικών απαιτεί συχνά σημαντικούς πόρους και εξειδικευμένη γνώση, ενώ τυχόν μεροληψία (bias) στα δεδομένα μπορεί να οδηγήσει σε λανθασμένες προβλέψεις [235, 287, 42]. Τέλος, η ετικετοποίηση (labeling) δεδομένων συνδέεται συχνά με αυξημένο κόστος και χρόνο, γεγονός που μπορεί να περιορίσει την κλίμακα και το εύρος των εφαρμογών της επιβλεπόμενης μάθησης [178].

Στη συνέχεια, θα ασχοληθούμε με τα βασικότερα είδη προβλημάτων που ανήκουν στον τομέα της επιβλεπόμενης μάθησης: την ταξινόμηση (classification) και την παλινδρόμηση (regression).

2.1.1 Ταξινόμηση

2.1.1.1 Ορισμός προβλήματος ταξινόμησης και εφαρμογές

Η ταξινόμηση αποτελεί μια από τις θεμελιώδεις κατηγορίες προβλημάτων επιβλεπόμενης μάθησης, όπου το μοντέλο καλείται να προβλέψει σε ποια κατηγορία - η οποια ονομάζεται κλάση - ανήκει ένα δεδομένο δείγμα, βασιζόμενο στα χαρακτηριστικά του. Ένας περισσότερο φορμαλιστικός ορισμός της ταξινόμησης είναι ο παρακάτω:

Δεδομένου ενός συνόλου δεδομένων εκπαίδευσης

$$D = \{(x_i, y_i)\}_{i=1}^{n} \tag{2.3}$$

όπου:

- $x_i \in X \subseteq R^d$ είναι ένα διάνυσμα χαρακτηριστικών (feature vector) που περιγράφει το δείγμα i,
- $y_i \in C$, είναι η αντίστοιχη κατηγορία κλάσης (label) του δείγματος i,
- C είναι το σύνολο των πιθανών κλάσεων ($C \geq 2$).

Ζητείται η εκμάθηση μιας συνάρτησης ταξινόμησης:

$$f : R^d \to C \tag{2.4}$$

η οποία, δεδομένου ενός άγνωστου δείγματος x^*, εκτιμά την πιθανότερη κλάση στην οποία αυτό ανήκει:

$$\hat{y} = f(x^*) \tag{2.5}$$

με στόχο την ελαχιστοποίηση ενός κριτηρίου απώλειας (Loss Function), όπως η εμπειρική πιθανότητα σφάλματος:

$$L(f) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{J}\big(\varphi(f(x_i) \neq y_i)\big) \tag{2.6}$$

όπου $\mathcal{J}(\cdot)$ είναι η χαρακτηριστική συνάρτηση που επιστρέφει 1 αν η πρόβλεψη είναι λανθασμένη και 0 αν είναι σωστή.

Η ταξινόμηση καλύπτει ένα μεγάλο εύρος εφαρμογών, καθώς αποτελεί έναν από τους πιο θεμελιώδεις τομείς της Μηχανικής Μάθησης με σημαντική συνεισφορά σε πολλούς κλάδους. Μια από τις πιο διαδεδομένες εφαρμογές της είναι η Ανάλυση Συναισθήματος [28], με την οποία θα ασχοληθούμε εκτενέστερα στη συνέχεια της παρούσας

εργασίας. Η συγκεκριμένη εφαρμογή αφορά την κατηγοριοποίηση κειμένου με βάση την πολικότητά του, η οποία μπορεί να είναι θετική, αρνητική ή ουδέτερη, ενώ επεκτείνεται και σε πιο λεπτομερείς συναισθηματικές καταστάσεις, όπως ο θυμός, η χαρά ή ο φόβος.

Εκτός από την ανάλυση συναισθήματος, η ταξινόμηση χρησιμοποιείται ευρέως στην ανίχνευση ανεπιθύμητης αλληλογραφίας (spam detection) [217], όπου τα εισερχόμενα email φιλτράρονται ως ανεπιθύμητα ή μη, συμβάλλοντας στην ασφάλεια και την αποτελεσματικότητα των ηλεκτρονικών επικοινωνιών. Επιπλέον, η ταξινόμηση εφαρμόζεται ευρέως σε προβλήματα που αφορούν εικόνα [45, 153] και ήχο [290, 304], όπου μοντέλα Μηχανικής Μάθησης καλούνται να ταξινομήσουν αντικείμενα, πρόσωπα ή ήχους σε συγκεκριμένες κατηγορίες. Σε τέτοια μοντέλα βασίζονται προηγμένες τεχνολογίες, όπως η αυτόνομη οδήγηση [20, 92], η αναγνώριση προσώπου [126, 12] και οι έξυπνοι ψηφιακοί βοηθοί [165, 302].

Επιπλέον, στον τομέα της ιατρικής διάγνωσης, η ταξινόμηση χρησιμοποιείται για την ανίχνευση παθολογικών καταστάσεων από ιατρικές απεικονίσεις (π.χ., ανάλυση ακτινογραφιών ή μαγνητικών τομογραφιών [13, 176]), την ταξινόμηση των τύπων καρκίνου [7] και την πρόβλεψη του κινδύνου εμφάνισης ασθενειών [252]. Παράλληλα, η ταξινόμηση βρίσκει εφαρμογή και στη χρηματοοικονομική ανάλυση, όπου χρησιμοποιείται για την πρόβλεψη κινήσεων της αγοράς [149], την εκτίμηση της φερεγγυότητας δανειοληπτών [140] και τον εντοπισμό ύποπτων συναλλαγών μέσω συστημάτων ανίχνευσης απάτης [6].

2.1.1.2 Είδη προβλημάτων ταξινόμησησς

Τα προβλήματα ταξινόμησης μπορούν να διακριθούν σε τρεις κατηγορίες, σύμφωνα με το πλήθος των κλάσεων στις οποίες καλούνται να κατηγοριοποιηθούν τα παραδείγματα του προβλήματος. Οι κατηγορίες αυτές είναι:

- Δυαδική ταξινόμηση (Binary Classification) [40]: Στη δυαδική ταξινόμηση, υπάρχουν μόνο δύο κατηγορίες στις οποίες καλούνται να ταξινομηθούν τα δείγματα, δηλαδή σύμφωνα με τον ορισμό στην αρχή αυτής της ενότητας $C = \{0, 1\}$, οπότε $|C| = 2$. Αποτελεί την πιο βασική μορφή ταξινόμησης και χρησιμοποιείται εκτενώς σε πληθώρα εφαρμογών.

- Πολυκατηγορική (Multi-Class Classification) [115]: Στην πολυκατηγορική ταξινόμηση, κάθε δείγμα καλείται να ταξινομηθεί σε μία από περισσότερες από δύο κλάσεις, δηλαδή $|C| > 2$.

- Πολυετικετική (Multi-Label Classification) [269]: Στην πολυετικετική ταξινόμηση, κάθε δείγμα μπορεί να ανήκει ταυτόχρονα σε περισσότερες από μία κατηγορίες. Σε αντίθεση με την πολυκατηγορική ταξινόμηση όπου κάθε δείγμα κατατάσσεται σε μία μόνο κατηγορία, στην πολυετικετική ταξινόμηση ένα δεδομένο μπορεί να φέρει πολλαπλές ετικέτες ταυτόχρονα. Ως παράδειγμα μπορούμε να αναφέρουμε κατηγοριοποίηση ταινιών, όπου μια ταινία μπορεί να ανήκει ταυτόχρονα στις κατηγορίες «Δράση» και «Επιστημονική Φαντασία», ενώ στην πολυκατηγορική ταξινόμηση θα μπορούσε να καταταχθεί μόνο σε μία από αυτές. Αντίστοιχα, σε προβλήματα ταξινόμησης κειμένου, ένα άρθρο μπορεί να σχετίζεται ταυτόχρονα τόσο με την κατηγορία «Τεχνολογία» όσο και με την κατηγορία «Οικονομία».

2.1.1.3 Προκλήσεις των προβλημάτων ταξινόμησης

Στο σημείο αυτό, όπου ήδη έχει τονιστεί η σημασία των προβλημάτων ταξινόμησης στη Μηχανική Μάθηση, λόγω του μεγάλου εύρους εφαρμογών τους, πρέπει να τονιστούν και οι προκλήσεις που σχετίζονται με αυτά και θα πρέπει να λαμβάνονται σοβαρά υπόψιν κατά την ενασχόληση μαζί τους. Παρακάτω, θα γίνει μια σύντομη αναφορά σε αυτές.

- **Ανισορροπία Δεδομένων (Class Imbalance)**: Ως ανισορροπία δεδομένων, αναφέρουμε το φαινόμενο που εμφανίζεται σε πολλά προβλήματα όπου τα παραδείγματα του συνόλου δεδομένων δεν κατανέμονται ομοιόμορφα στις κλάσεις, με αποτέλεσμα ορισμένες από αυτές να περιέχουν πολύ περισσότερα δείγματα από άλλες [127, 1]. Βασική συνέπεια είναι τα μοντέλα να προτιμούν την πρόβλεψη της κυρίαρχης κλάσης, εκείνης δηλαδή που υπερέχει αριθμητικά, παραβλέποντας τις λιγότερο συχνές αλλά συχνά σημαντικές κλάσεις. Το φαινόμενο αυτό, μπορεί να αποτελέσει βασικό πρόβλημα σε προβλήματα όπου μπορεί η κλάση με τα λιγότερα δείγματα να είναι ιδιαίτερα σημαντική, όπως για παράδειγμα σε προβλήματα ιατρικής φύσης, οπου για παράδειγμα σπάνιες ασθένειες έχουν πολύ λιγότερα δείγματα σε σύγκριση με την υγιή κατάσταση, αλλά ο μη εντοπισμός τους από το μοντέλο, μπορεί να έχει κόστος ανθρώπινων ζωών. Για την αντιμετώπιση αυτού του φαινομένου, χρησιμοποιούνται συχνά τεχνικές oversampling [102] ή undersampling [64]. Οι πρώτες χρησιμοποιούνται για τη δημιουργία συνθετικών δειγμάτων από τις υποεκπροσωπούμενες κατηγορίες, ενώ οι δεύτερες στοχεύουν στη μείωση του αριθμού των δειγμάτων της κυρίαρχης κατηγορίας. Επίσης συχνή είναι και η χρήση συναρτήσεων σταθμισμένων απωλειών [219, 218], όπου σύμφωνα με τις οποίες τα σφάλματα στις λιγότερο συχνές κατηγορίες έχουν μεγαλύτερη βαρύτητα, ώστε το μοντέλο να τις λαμβάνει περισσότερο υπόψη κατά τη διαδικασία εκπαίδευσης.

- **Υπερπροσαρμογή (Overfitting) και Υποπροσαρμογή (Underfitting)**: Ως υπερπροσαμογή, ορίζεται το φαινόμενο κατά το οποίο το μοντέλο μαθαίνει πολύ καλά τις λεπτομέρειες των δεδομένων εκπαίδευσης, αλλά αποτυγχάνει να γενικεύσει σε νέα δεδομένα, οδηγώντας σε χαμηλή απόδοση. Στον αντίποδα, υποπροσαρμογή σημαίνει ότι το μοντέλο δεν έχει μάθει αρκετές σημαντικές σχέσεις, με αποτέλεσμα να έχει χαμηλή ακρίβεια ακόμα και στα δεδομένα εκπαίδευσης [9]. Οι κυριότερες αιτίες της υπερπροσαμογής είναι η χρήση ιδιαίτερα πολύπλοκων μοντέλων σε περιπτώσεις όπου δεν είναι διαθέσιμα αρκετά δεδομένα, αλλά και η χρήση μεγάλου πλήθους χαρακτηριστικών που προκαλεί απομνημόνευση των δεδομένων και δυσχεραίνει την ικανότητα του μοντέλου για γενίκευση. Αντίθετα, η υποπροσαρμογή συνήθως οφείλεται σε χρήση υπερβολικά απλών μοντέλων που δεν είναι ικανά και κατάλληλα να συλλάβουν την πολυπλοκότητα του προβλήματος, αλλά και η χρήση δεδομένων που είτε είναι ελλιπή είτε δεν έχουν προεπεξεργαστεί ορθά. Για την αντιμετώπιση των δύο αυτών φαινομένων, υιοθετούνται συνήθως τεχνικές κανονικοποίησης των δεδομένων [242], αλλά και επιλογής χαρακτηριστικών [225] για την επιλογή των πιο «χρήσιμων» χαρακτηριστικών. Η χρήση διασταυρωμένης επικύρωσης [182] επίσης ενδείκνυται, καθώς επιτρέπει την αξιολόγηση της απόδοσης του μοντέλου σε διαφορετικά υποσύνολα των δεδομένων εκπαίδευσης, μειώνοντας τον κίνδυνο υπερπροσαρμογής. Τέλος, η χρήση πρόωρου τερματισμού [158] κατά την εκπαίδευση μπορεί να αποτρέψει την υπερπροσαρμογή, διακόπτοντας την εκπαίδευση όταν η απόδοση στα δεδομένα επικύρωσης αρχίσει να μειώνεται.

- **Προκατάληψη Δεδομένων (Bias in Data)**: Τα μοντέλα Μηχανικής Μάθησης είναι τόσο καλά όσο τα δεδομένα με τα οποία εκπαιδεύονται. Αν τα δεδομένα εκπαίδευσης είναι μεροληπτικά ή μη αντιπροσωπευτικά, τότε το μοντέλο θα μάθει και συνεπώς θα αναπαράγει αυτές τις προκαταλήψεις. Για τον περιορισμό του φαινομένου αυτού, συστήνεται η χρήση διαφορετικών πηγών δεδομένων με στόχο τη διασφάλιση της ποικιλομορφίας, ενώ συχνός θα πρέπει να είναι και ο έλεγχος των προβλέψεων του μοντέλου για σημάδια συστηματικής προκατάληψης [281]. Η παραδοχή πως τα δημιουργούμενα μοντέλα τείνουν να αναπαραγάγουν τις προκαταλήψεις, οδήγησε στην ανάπτυξη μιας νέας κατηγορίας Μηχανικής Μάθησης, της Μηχανικής Μάθησης με επίγνωση της δικαιοσύνης (Fairness-Aware Machine Learning) [73, 312], ο οποίος επικεντρώνεται στην ανάπτυξη αλγορίθμων και μοντέλων που μειώνουν ή εξαλείφουν μεροληψίες, εξασφαλίζοντας ότι οι προβλέψεις τους δεν εισάγουν κοινωνικές ή ηθικές ανισότητες.

- **Ερμηνευσιμότητα Μοντέλων (Model Interpretability)**: Πολλά από τα σύγχρονα μοντέλα ταξινόμησης, όπως τα Νευρωνικά Δίκτυα και οι Μετασχηματιστές (Transformers, BERT, RoBERTa), έχουν υψηλή απόδοση αλλά είναι «μαύρα κουτιά», γεγονός που δυσκολεύει την κατανόηση των προβλέψεων τους [146, 236]. Το γεγονός αυτό, όχι μόνο εγείρει αίσθημα δυσπιστίας για τη χρήση τέτοιων μοντέλων σε κρίσιμες εφαρμογές όπως για παράδειγμα η ιατρική διάγνωση και η χρηματοοικονομική ανάλυση, αλλά επιπλέον επιφέρει σημαντικές δυσκολίες στον εντοπισμό σφαλμάτων και στην προσπάθεια για βελτίωσή τους. Άμεση συνέπεια αποτελεί η χρήση απλούστερων - και συνεπώς λιγότερο ισχυρών - μοντέλων σε περιπτώσεις όπου η διαφάνεια κρίνεται αδιαπραγμάτευτη. Στην προσπάθεια για αύξηση της ερμηνευσιμότητας των μοντέλων, αναπτύχθηκαν τεχνικές όπως οι SHAP (SHapley Additive Explanations) [14] και LIME (Local Interpretable Model-agnostic Explanations) [306] οι οποίες χρησιμοποιούνται για την ερμηνεία των αποφάσεων μοντέλων Μηχανικής Μάθησης, ιδιαίτερα σε περιπτώσεις όπου οι αλγόριθμοι είναι πολύπλοκοι και δύσκολοι στην κατανόηση. Η SHAP, η ανάπτυξη της οποίας βασίζεται στη θεωρία παιγνίων [164], αποδίδει σε κάθε χαρακτηριστικό μια ποσοτική τιμή που δείχνει πόσο συμβάλλει στην πρόβλεψη του μοντέλου, τόσο σε ατομικό όσο και σε συνολικό επίπεδο. Από την άλλη, η LIME λειτουργεί με τοπικές προσεγγίσεις, εκπαιδεύοντας απλά μοντέλα (π.χ. γραμμική παλινδρόμηση) κοντά σε μια συγκεκριμένη πρόβλεψη, ώστε να εξηγηθεί ο ρόλος κάθε χαρακτηριστικού στην έξοδο του μοντέλου [74].

2.1.1.4 Μετρικές Αξιολόγησης

Η αξιολόγηση της απόδοσης των ταξινομητών πραγματοποιείται μέσω συγκεκριμένων μετρικών που βασίζονται στον πίνακα σύγχυσης (confusion matrix). Ο πίνακας σύγχυσης παρουσιάζει τις πραγματικές και τις προβλεπόμενες κλάσεις ενός μοντέλου ταξινόμησης. Τα βασικά στοιχεία του είναι:

- **True Positives (TP)**: Ταξινομημένες ως θετικές περιπτώσεις που είναι πράγματι θετικές.

- **False Positives (FP)**: Ταξινομημένες ως θετικές περιπτώσεις που είναι στην πραγματικότητα αρνητικές (Σφάλμα Τύπου I).

- **True Negatives (TN)**: Ταξινομημένες ως αρνητικές περιπτώσεις που είναι πράγματι αρνητικές.
- **False Negatives (FN)**: Ταξινομημένες ως αρνητικές περιπτώσεις που είναι στην πραγματικότητα θετικές (Σφάλμα Τύπου II).

Στο Σχήμα 2.1, παρουσιάζεται η μορφή του πίνακα σύγχυσης για ένα δυαδικό πρόβλημα ταξινόμησης.

Η αξιολόγηση ενός ταξινομητή απαιτεί τη χρήση διαφορετικών μετρικών, καθώς κάθε μέτρική προσφέρει διαφορετική πληροφόρηση για την ποιότητα των προβλέψεων. Οι μετρικές αυτές είναι ιδιαίτερα σημαντικές όταν τα δεδομένα είναι μη ισορροπημένα, όπου η χρήση μόνο της ακρίβειας μπορεί να δώσει παραπλανητικά αποτελέσματα.

Ακρίβεια (Accuracy): Η ακρίβεια μετρά τη συνολική ορθότητα του μοντέλου. Ορίζεται ως ο λόγος των ορθώς ταξινομημένων δειγμάτων προς το συνολικό αριθμό των δειγμάτων και εκφράζεται ως ποσοστό. Υπολογίζεται ως:

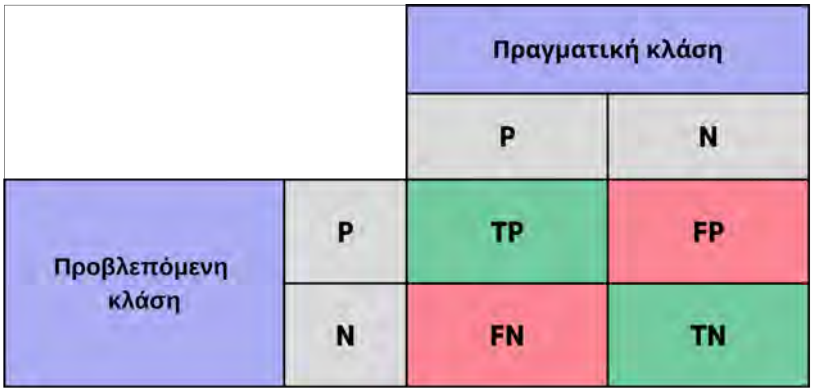
$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.7}$$

Παρ΄ όλο που είναι μια απλή και εύκολα κατανοητή μετρική, δεν είναι πάντα κατάλληλη για μη ισορροπημένα δεδομένα.

Ανάκληση (Recall): Η ανάκληση, γνωστή και ως ευαισθησία (Sensitivity) ή True Positive Rate (TPR), είναι ένα βασικό μέτρο απόδοσης που αξιολογεί την ικανότητα του μοντέλου να αναγνωρίζει σωστά τα θετικά δείγματα. Ορίζεται ως ο λόγος των σωστά ταξινομημένων θετικών δειγμάτων (TP) προς το σύνολο των πραγματικών θετικών δειγμάτων (TP + FN):

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2.8}$$

Η ανάκληση είναι ιδιαίτερα σημαντική σε εφαρμογές όπου το να χάσουμε ένα θετικό δείγμα μπορεί να έχει σοβαρές συνέπειες, όπως στη διάγνωση ασθενειών ή στην ανίχνευση απάτης. Ωστόσο, η υψηλή ανάκληση μπορεί να επιτευχθεί εις βάρος της



Σχήμα 2.1: Ο πίνακας σύγχυσης σε ένα δυαδικό πρόβλημα ταξινόμησης

ακρίβειας των θετικών προβλέψεων (Precision), καθώς ένα μοντέλο που χαρακτηρίζει σχεδόν όλα τα δείγματα ως θετικά θα έχει υψηλή ανάκληση αλλά χαμηλή εξειδίκευση (Specificity). Για μια ισορροπημένη αξιολόγηση του ταξινομητή, χρησιμοποιείται συχνά η F1-score, η οποία συνδυάζει την ανάκληση και την ακρίβεια θετικών προβλέψεων.

Ακρίβεια Θετικών Προβλέψεων (Precision): Η ακρίβεια θετικών προβλέψεων Precision είναι ένα βασικό μέτρο απόδοσης που δείχνει πόσο αξιόπιστες είναι οι θετικές προβλέψεις του μοντέλου. Ορίζεται ως ο λόγος των σωστά ταξινομημένων θετικών δειγμάτων (TP) προς το σύνολο των δειγμάτων που ταξινομήθηκαν ως θετικά (TP + FP):

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2.9}$$

Η ακρίβεια θετικών προβλέψεων είναι ιδιαίτερα σημαντική σε περιπτώσεις όπου οι ψευδώς θετικές προβλέψεις (FP) έχουν μεγάλο κόστος. Για παράδειγμα, σε ένα σύστημα ανίχνευσης ανεπιθύμητων μηνυμάτων (spam filter), ένα υψηλό ποσοστό ψευδώς θετικών σημαίνει ότι σημαντικά μηνύματα μπορεί να χαρακτηριστούν λανθασμένα ως ανεπιθύμητα. Από την άλλη πλευρά, η υψηλή ακρίβεια θετικών προβλέψεων δεν εγγυάται απαραίτητα καλή συνολική απόδοση, καθώς μπορεί να επιτευχθεί εις βάρος της ανάκλησης.

F1-score: Η F1-score είναι ένα μέτρο απόδοσης που συνδυάζει την ακρίβεια θετικών προβλέψεων και την ανάκληση σε μία μετρική, δίνοντας μια πιο ισορροπημένη εκτίμηση της απόδοσης ενός ταξινομητή. Ορίζεται ως ο αρμονικός μέσος των δύο αυτών μετρικών:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.10}$$

Η F1-score είναι ιδιαίτερα χρήσιμη σε περιπτώσεις όπου υπάρχει ανισορροπία μεταξύ των τάξεων, καθώς η απλή ακρίβεια μπορεί να είναι παραπλανητική. Για παράδειγμα, αν ένα μοντέλο ταξινομεί μια σπάνια κατηγορία με υψηλή ανάκληση αλλά χαμηλή ακρίβεια θετικών προβλέψεων, η F1-score θα αντικατοπτρίζει αυτήν την ανισορροπία και δεν θα εμφανίσει υπερβολικά αισιόδοξα αποτελέσματα. Έτσι, χρησιμοποιείται ευρέως σε προβλήματα όπως η ανίχνευση απάτης, η ιατρική διάγνωση και η αναγνώριση ανεπιθύμητων μηνυμάτων.

Συντελεστής Συσχέτισης του Matthews (Matthews Correlation Coefficient - MCC): Ο Συντελεστής Συσχέτισης του Matthews είναι ένα μέτρο απόδοσης που χρησιμοποιείται για την αξιολόγηση της ποιότητας κυρίως δυαδικών, αλλά και multiclass ταξινομητών μετά από κατάλληλη τροποποίηση. Λαμβάνει υπόψη όλες τις τιμές του πίνακα σύγχυσης και θεωρείται ένα από τα πιο ισορροπημένα μέτρα, ειδικά σε προβλήματα με άνισες κατανομές κλάσεων. Ορίζεται ως:

$$\text{MCC} = \frac{(\text{TP} \times \text{TN}) - (\text{FP} \times \text{FN})}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \tag{2.11}$$

Ο MCC παίρνει τιμές από -1 έως 1. Μια τιμή κοντά στο 1 δείχνει εξαιρετική ταξινόμηση, 0 σημαίνει ότι το μοντέλο δεν έχει καλύτερη απόδοση από την τυχαία ταξινόμηση, ενώ -1 υποδηλώνει εντελώς λανθασμένες προβλέψεις. Σε αντίθεση με την

ακρίβεια, που μπορεί να είναι παραπλανητική σε ανισομερή σύνολα δεδομένων, ο MCC προσφέρει μια πιο αξιόπιστη εκτίμηση της συνολικής απόδοσης του μοντέλου.

Συντελεστής Συμφωνίας του Cohen (Cohen's Kappa Coefficient): Ο Cohen's Kappa Coefficient (k) είναι ένα μέτρο που χρησιμοποιείται για την αξιολόγηση της συμφωνίας μεταξύ δύο ανεξάρτητων ταξινομητών, λαμβάνοντας υπόψη τη συμφωνία που μπορεί να προκύψει τυχαία. Ορίζεται ως:

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{2.12}$$

όπου:

- P_o είναι η παρατηρούμενη συμφωνία μεταξύ των δύο ταξινομητών, δηλαδή το ποσοστό των περιπτώσεων όπου έδωσαν την ίδια κατηγορία,
- P_e είναι η αναμενόμενη συμφωνία αν οι ταξινομητές έκαναν ταξινόμηση τυχαία.

Ο Cohen's Kappa κυμαίνεται από -1 έως 1: Η τιμή 1 σημαίνει απόλυτη συμφωνία μεταξύ των δύο ταξινομητών, ενώ η τιμή 0 σημαίνει ότι η συμφωνία τους δεν είναι καλύτερη από την τυχαία. Τέλος, τιμή μικρότερη του 0 υποδηλώνει χειρότερη από την τυχαία συμφωνία.

Η μέτρηση αυτή χρησιμοποιείται ευρέως στην ανάλυση διαγνώσεων, στην επεξεργασία φυσικής γλώσσας και σε εφαρμογές όπου είναι κρίσιμο να εκτιμηθεί η συνοχή μεταξύ διαφορετικών ταξινομητών. Σε περιπτώσεις με περισσότερους από δύο ταξινομητές, ο Cohen's Kappa μπορεί να επεκταθεί στον Fleiss' Kappa, που υπολογίζει τη συμφωνία μεταξύ πολλών ταξινομητών.

Εμβαδόν Κάτω από την Καμπύλη ROC - (Receiver Operating Characteristic - ROC AUC) Η καμπύλη ROC είναι ένα διαγνωστικό εργαλείο που χρησιμοποιείται για την αξιολόγηση της απόδοσης ενός ταξινομητή, ειδικά σε προβλήματα δυαδικής ταξινόμησης. Προκύπτει απεικονίζοντας τον True Positive Rate (TPR) (ή Recall) σε σχέση με τον False Positive Rate (FPR) για διαφορετικά κατώφλια απόφασης.

Το εμβαδόν κάτω από την καμπύλη (Area Under Curve - AUC είναι ένα αριθμητικό μέτρο που αντιπροσωπεύει το συνολικό εμβαδόν κάτω από την καμπύλη ROC και δίνει μια συνοπτική εκτίμηση της απόδοσης του μοντέλου.

Η μετρική ROC AUC παίρνει τιμές μεταξύ 0 και 1. Η τιμή 1 σημαίνει τέλεια ταξινόμηση, τιμές κοντά στο 0.5 σημαίνουν ότι το μοντέλο δεν είναι καλύτερο από την τυχαία ταξινόμηση, ενώ τιμή μικρότερη από 0.5 υποδηλώνει ότι το μοντέλο ταξινομεί τις τάξεις αντίστροφα. Είναι ιδιαίτερα χρήσιμη σε προβλήματα με μη ισορροπημένα σύνολα δεδομένων, καθώς δεν επηρεάζεται από την κατανομή των κλάσεων. Ωστόσο, μπορεί να είναι λιγότερο αποτελεσματική σε περιπτώσεις όπου βασικός στόχος είναι η μείωση των ψευδώς θετικών ή ψευδώς αρνητικών προβλέψεων, οπότε μπορεί να προτιμηθεί η Precision-Recall AUC.

2.1.2 Παλινδρόμηση

Η παλινδρόμηση (regression) είναι η δεύτερη θεμελιώδης κατηγορία προβλημάτων της επιβλεπόμενης μάθησης που στοχεύει στη μοντελοποίηση της σχέσης μεταξύ μιας εξαρτημένης μεταβλητής (γνωστής ως response variable ή target variable) και μίας ή

περισσότερων ανεξάρτητων μεταβλητών (predictors ή features). Ο βασικός στόχος της παλινδρόμησης είναι η εκμάθηση μιας συνάρτησης που μπορεί να προβλέψει συνεχείς τιμές με βάση ένα νέο σύνολο εισόδων [237, 166].

Μαθηματικά, ένα μοντέλο παλινδρόμσησης μπορεί πολύ απλοϊκά να διατυπωθεί ως εξής:

$$y = f(x) + \epsilon \tag{2.13}$$

όπου:

- y είναι η εξαρτημένη μεταβλητή,
- x είναι το διάνυσμα των ανεξάρτητων μεταβλητών,
- $f(x)$ είναι μια άγνωστη συνάρτηση που περιγράφει τη σχέση μεταξύ x και y,
- ϵ είναι ένας στοχαστικός όρος σφάλματος που μοντελοποιεί τον θόρυβο ή την ανεξήγητη μεταβλητότητα στα δεδομένα.

Η συνάρτηση $f(x)$ μπορεί να προσεγγιστεί μέσω παραμετρικών ή μη παραμετρικών μοντέλων, ανάλογα με τις υποθέσεις που γίνονται για την κατανομή και τη δομή των δεδομένων. Στην πιο απλή περίπτωση, η γραμμική παλινδρόμηση (linear regression) υποθέτει μια γραμμική σχέση μεταξύ των χαρακτηριστικών και της μεταβλητής στόχου, ενώ πιο πολύπλοκα μοντέλα - όπως η πολυωνυμική παλινδρόμηση (polynomial regression), οι μέθοδοι πυρήνων (kernel methods) και τα νευρωνικά δίκτυα - επιτρέπουν τη μοντελοποίηση μη γραμμικών σχέσεων.

Μια περισσότερο φορμαλιστική περιγραφή του προβλήματος, θα μπορούσε να είναι η εξής:

Δεδομένου ενός συνόλου δεδομένων:

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N} \tag{2.14}$$

όπου $x_i \in R^d$ αντιπροσωπεύει το διάνυσμα χαρακτηριστικών της i-οστής παρατήρησης και $y_i \in R$ την αντίστοιχη εξαρτημένη μεταβλητή, ο στόχος ενός μοντέλου παλινδρόμησης είναι να μάθει μια συνάρτηση $\hat{f}$ τέτοια ώστε:

$$\hat{y} = \hat{f}(x) \tag{2.15}$$

Η απόδοση του μοντέλου αξιολογείται συνήθως μέσω μιας συνάρτησης κόστους (loss function) [286], με τις πιο διαδεδομένες να είναι το Μέσο Τετραγωνικό Σφάλμα (Mean Squared Error - MSE), το Μέσο Απόλυτο Σφάλμα (Mean Absolute Error - MAE) - τα οποία θα συζητηθούν στη συνέχεια - και η συνάρτηση Huber,που συνδυάζει τα πλεονεκτήματα του MSE και του MAE, μειώνοντας την ευαισθησία σε ακραίες τιμές [125].

Η παλινδρόμηση αποτελεί μία από τις παλαιότερες και πιο ευρέως χρησιμοποιούμενες στατιστικές τεχνικές, με τις ρίζες της να ανάγονται στο έργο των Gauss και Legendre σχετικά με την εκτίμηση ελαχίστων τετραγώνων [259]. Στη σύγχρονη Μηχανική Μάθηση, η παλινδρόμηση αποτελεί το θεμέλιο πολλών προβλεπτικών μοντέλων και έχει επεκταθεί σε πιο σύνθετες περιοχές, όπως η πρόβλεψη χρονοσειρών [261, 177, 135], η ανάλυση επιβίωσης [94] και η επεξεργασία δεδομένων υψηλών διαστάσεων [128, 83]. Τα κύρια πλεονεκτήματα των μοντέλων παλινδρόμησης περιλαμβάνουν την ερμηνευσιμότητα των σχέσεων, καθώς μοντέλα όπως η γραμμική παλινδρόμηση και

τα γενικευμένα προσθετικά μοντέλα παρέχουν συντελεστές που ποσοτικοποιούν την επίδραση των χαρακτηριστικών στη μεταβλητή - στόχο, την εκτίμηση της αβεβαιότητας μέσω της υπολογιστικής ανάλυσης διαστημάτων εμπιστοσύνης και προβλεπτικών κατανομών και τη δυνατότητα κλιμάκωσης, καθώς μέθοδοι που βασίζονται σε κυρτή βελτιστοποίηση, όπως η Ridge Regression και η LASSO, μπορούν να εφαρμοστούν αποτελεσματικά σε μεγάλα σύνολα δεδομένων. Επιπλέον, μπορούν να ενσωματωθούν σε σχήματα Βαθιάς Μάθησης, όπου συναρτήσεις απώλειας όπως το μέσο τετραγωνικό σφάλμα χρησιμοποιούνται ευρέως στην εκπαίδευση νευρωνικών δικτύων [111].

2.1.2.1 Μέθοδοι παλινδρόμησης

Γραμμική παλινδρόμηση: Η γραμμική παλινδρόμηση [291] είναι ένας από τους πιο δημοφιλείς αλγορίθμους παλινδρόμησης και πιθανώς από τους βασικότερους. Στην περίπτωση που υπάρχει μία μόνο ανεξάρτητη μεταβλητή, η παλινδρόμηση ονομάζεται απλή γραμμική παλινδρόμηση (simple linear regression) [162], ενώ αν υπάρχουν πολλές ανεξάρτητες μεταβλητές, τότε ονομάζεται πολλαπλή γραμμική παλινδρόμηση (multiple linear regression) [274].

Η γραμμική παλινδρόμηση χρησιμοποιεί γραμμικές συναρτήσεις πρόβλεψης, των οποίων οι τιμές εκτιμώνται από τα δεδομένα του μοντέλου. Τα γραμμικά μοντέλα αποτελούν τις απλούστερες παραμετρικές μεθόδους και αξίζουν πάντα προσοχής, διότι πολλά προβλήματα, ακόμα και μη γραμμικά στη φύση τους, μπορούν να προσεγγιστούν αποτελεσματικά από αυτά.

Όπως αναφέρθηκε προηγουμένως, η παλινδρόμηση αφορά την πρόβλεψη μιας συνεχούς μεταβλητής-στόχου, με πολλές εφαρμογές. Είναι επομένως κρίσιμο να μπορούμε να διακρίνουμε πώς ένα γραμμικό μοντέλο μπορεί να προσαρμοστεί στα δεδομένα, ποια είναι τα πλεονεκτήματα και τα μειονεκτήματά του και πότε είναι προτιμότερο να επιλεγεί ένα εναλλακτικό μοντέλο.

Απλή γραμμική παλινδρόμηση: Ο όρος απλή αναφέρεται στο γεγονός ότι υπάρχει μόνο μία ανεξάρτητη μεταβλητή, ενώ η λέξη γραμμική δεν σημαίνει απαραίτητα ότι η σχέση μεταξύ των δύο μεταβλητών είναι οπτικά μια ευθεία γραμμή, αλλά ότι το μοντέλο είναι γραμμικό ως προς τις παραμέτρους.

Το βασικό μοντέλο της απλής γραμμικής παλινδρόμησης ορίζεται ως:

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{2.16}$$

όπου:

- Y είναι η εξαρτημένη μεταβλητή (dependent variable)
- X είναι η ανεξάρτητη μεταβλητή (independent variable)
- β_0 είναι σταθερά (intercept)
- β_1 είναι ο συντελεστής κλίσης (slope)
- ϵ είναι ο όρος σφάλματος, ο οποίος εκφράζει τη διαφορά μεταξύ της πραγματικής και της προβλεπόμενης τιμής.

Ο όρος ϵ δεν υποδηλώνει κάποιο λάθος αλλά απλά το γεγονός ότι δεν υπάρχει τέλεια συσχέτιση μεταξύ X και Y. Η εξίσωση (2.16) είναι η μαθηματική αναπαράσταση μιας ευθείας γραμμής, όπου το β_1 καθορίζει την κλίση της ευθείας, ενώ το β_0 είναι η τιμή του Y όταν $X = 0$.

Η εκτίμηση των παραμέτρων β_0 και β_1 γίνεται με τη μέθοδο των ελαχίστων τετραγώνων, η οποία ελαχιστοποιεί το άθροισμα των τετραγώνων των διαφορών μεταξύ των πραγματικών και των προβλεπόμενων τιμών:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{N}(X_i - \bar{X})^2} \tag{2.17}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \tag{2.18}$$

όπου $\bar{X}$ και $\bar{Y}$ είναι οι μέσοι όροι των X και Y αντίστοιχα.

Το τελικό μοντέλο χρησιμοποιείται για την πρόβλεψη νέων τιμών Y με βάση δεδομένες τιμές X. Η ποιότητα της προσαρμογής (goodness of fit) [108, 98] μπορεί να αξιολογεί μετρικές όπως ο συντελεστής προσδιορισμού R^2, ο οποίος μετρά το ποσοστό της διακύμανσης του Y που εξηγείται από το μοντέλο:

$$R^2 = 1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2} \tag{2.19}$$

Όσο πιο κοντά είναι το R^2 στη μονάδα, τόσο καλύτερη είναι η προσαρμογή του μοντέλου στα δεδομένα.

Πολλαπλή γραμμική παλινδρόμηση: Στις περισσότερες εφαρμογές παλινδρόμησης, χρησιμοποιούνται περισσότερες από μία ανεξάρτητες μεταβλητές για την πρόβλεψη της εξαρτημένης μεταβλητής. Σε αυτή την περίπτωση, το μοντέλο γενικεύεται σε αυτό που ονομάζεται πολλαπλή γραμμική παλινδρόμηση.

Η γενική μορφή του μαθηματικού μοντέλου της πολλαπλής γραμμικής παλινδρόμησης είναι:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m + \epsilon \tag{2.20}$$

με αντίστοιχη εξίσωση πρόβλεψης:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_m X_m \tag{2.21}$$

όπου:

- Y είναι η εξαρτημένη μεταβλητή,
- $X_1, X_2, \ldots, X_m$ είναι οι ανεξάρτητες μεταβλητές,
- β_0 είναι η σταθερά του μοντέλου,
- $\beta_1, \beta_2, \ldots, \beta_m$ είναι οι συντελεστές παλινδρόμησης που αντιπροσωπεύουν τη συμβολή κάθε ανεξάρτητης μεταβλητής στη διαμόρφωση του Y,
- ϵ είναι ο όρος σφάλματος που εκφράζει τη διαφορά μεταξύ της πραγματικής και της προβλεπόμενης τιμής του Y.

Το πλήθος των ανεξάρτητων μεταβλητών m καθορίζει το πόσο πολύπλοκο γίνεται το μοντέλο. Η πολλαπλή γραμμική παλινδρόμηση αποτελεί γενίκευση της απλής γραμμικής παλινδρόμησης και επιτρέπει τη μοντελοποίηση περιπτώσεων όπου η εξαρτημένη μεταβλητή επηρεάζεται από περισσότερους από έναν παράγοντες.

Ωστόσο, η αύξηση του αριθμού των ανεξάρτητων μεταβλητών ενέχει ορισμένες προκλήσεις:

- Επιλογή μεταβλητών: Δεν είναι πάντα σαφές ποιες μεταβλητές πρέπει να συμπεριληφθούν στο μοντέλο. Η χρήση υπερβολικά πολλών μεταβλητών μπορεί να οδηγήσει σε υπερπροσαρμογή (overfitting), ενώ η παράλειψη σημαντικών μεταβλητών μπορεί να οδηγήσει σε προκατάληψη (bias).

- Συσχέτιση μεταξύ ανεξάρτητων μεταβλητών: Όταν δύο ή περισσότερες ανεξάρτητες μεταβλητές έχουν ισχυρή συσχέτιση (multicollinearity) [57], η στατιστική ερμηνεία των συντελεστών μπορεί να γίνει προβληματική, καθώς καθίσταται δύσκολο να διακριθεί η επίδραση κάθε μεταβλητής ξεχωριστά.

- Απαιτήσεις σε δεδομένα: Για να εκπαιδευτεί ένα σταθερό μοντέλο, απαιτείται αρκετά μεγάλο δείγμα δεδομένων, ώστε να αποφευχθεί η προσαρμογή του μοντέλου σε τυχαίους θορύβους του συνόλου εκπαίδευσης.

Οι συντελεστές $\beta_0, \beta_1, ..., \beta_m$ εκτιμώνται χρησιμοποιώντας τη μέθοδο των ελαχίστων τετραγώνων, που στοχεύει στην ελαχιστοποίηση του συνολικού τετραγωνικού σφάλματος:

$$\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2 \tag{2.22}$$

όπου N είναι το πλήθος των δειγμάτων εκπαίδευσης. Η ελαχιστοποίηση αυτής της συνάρτησης επιτυγχάνεται με τη χρήση διανυσματικής μορφής:

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{2.23}$$

όπου:

- X είναι ο πίνακας των χαρακτηριστικών
- Y είναι το διάνυσμα των πραγματικών τιμών της εξαρτημένης μεταβλητής
- $\hat{\beta}$ είναι το διάνυσμα των εκτιμώμενων παραμέτρων

Η πολλαπλή γραμμική παλινδρόμηση παραμένει ένα από τα πιο χρήσιμα και ερμηνεύσιμα μοντέλα της Στατιστικής και της Μηχανικής Μάθησης. Ωστόσο, απαιτεί προσεκτική επιλογή χαρακτηριστικών και αξιολόγηση της απόδοσης, ώστε να αποφευχθούν φαινόμενα υπερπροσαρμογής και πολυσυγγραμμικότητας.

Πολυωνυμική παλινδρόμηση: Η Πολυωνυμική παλινδρόμηση [116] είναι μια τεχνική που επιτρέπει τη χρήση γραμμικών μοντέλων ακόμη και σε περιπτώσεις όπου τα δεδομένα παρουσιάζουν έντονες μη γραμμικές σχέσεις. Η βασική ιδέα είναι να μετασχηματίσουμε τις ανεξάρτητες μεταβλητές έτσι ώστε το μοντέλο να μπορεί να προσαρμοστεί καλύτερα στα δεδομένα, χρησιμοποιώντας πολυωνυμικούς όρους.

Το μοντέλο της πολυωνυμικής παλινδρόμησης μπορεί να οριστεί ως:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_d X^d + \epsilon \tag{2.24}$$

όπου:

- Y είναι η εξαρτημένη μεταβλητή
- X είναι η ανεξάρτητη μεταβλητή
- d είναι ο βαθμός του πολυωνύμου
- $\beta_0, \beta_1, ..., \beta_d$ είναι οι συντελεστές παλινδρόμησης
- ϵ είναι ο όρος σφάλματος

Αντί να χρησιμοποιούμε μόνο την ανεξάρτητη μεταβλητή X, εισάγουμε νέες μεταβλητές που αντιστοιχούν στις πολυωνυμικές δυνάμεις του X, ώστε να προσαρμόσουμε το μοντέλο σε δεδομένα που δεν ακολουθούν μια απλή γραμμική τάση.

Για να γίνουν περισσότερο κατανοητά τα παραπάνω, περιγράφεται εν τάχει ο Γενικός Πολυωνυμικός Μετασχηματισμός. Αν έχουμε m αρχικές μεταβλητές $X_1, X_2, \ldots, X_m$, μπορούμε να ορίσουμε ένα πιο σύνθετο πολυωνυμικό μοντέλο με όρους πολλαπλών μεταβλητών:

$$\tilde{Y_k} = \alpha_0 + \sum_{i=1}^{m} \alpha_i \tilde{X}_k^{(i)} + \sum_{j=m+1}^{k} \alpha_j f_p(\tilde{X}_k^{(1)}, \tilde{X}_k^{(2)}, ..., \tilde{X}_k^{(m)}) \tag{2.25}$$

όπου κάθε $f_p(\cdot)$ είναι πολυωνυμική συνάρτηση των χαρακτηριστικών. Για παράδειγμα, αν έχουμε δύο ανεξάρτητες μεταβλητές X_1, X_2, μπορούμε να επεκτείνουμε το μοντέλο δεύτερου βαθμού ως εξής:

$$\tilde{X} = (X_1, X_2) \rightarrow \tilde{X}_t = (X_1, X_2, X_1^2, X_2^2, X_1 X_2) \tag{2.26}$$

Δηλαδή, το διανυσματικό μοντέλο περιλαμβάνει τόσο τις αρχικές μεταβλητές όσο και τις τετραγωνικές και διασταυρούμενες τιμές τους.

Η πολυωνυμική παλινδρόμηση μπορεί να προσαρμόζεται καλύτερα σε δεδομένα που εμφανίζουν μη γραμμικές τάσεις, όμως χαρακτηρίζεται και αυτή από τα προβλήματα της υπερπροσαρμογής και της πολυσυγγραμμικότητας που αναφέρθηκαν προηγουμένως, ενώ συνοδεύεται επίσης από αυξημένη ανάγκη σε υπολογιστικούς πόρους, καθώς η εισαγωγή πολλών πολυωνυμικών όρων αυξάνει τον αριθμό των χαρακτηριστικών.

Συνεπώς, αν και η πολυωνυμική παλινδρόμηση αποτελεί μια ισχυρή τεχνική για την προσέγγιση μη γραμμικών σχέσεων, απαιτεί προσεκτική επιλογή του βαθμού του πολυωνύμου για να επιτευχθεί η βέλτιστη ισορροπία μεταξύ προσαρμογής και γενίκευσης. Επιπλέον, σε πολλές περιπτώσεις, η κανονικοποίηση είναι απαραίτητη για τη βελτίωση της σταθερότητας του μοντέλου και την αποφυγή της υπερπροσαρμογής. Περισσότερα για την κανονικοποίηση θα δούμε στη συνέχεια.

2.1.2.2 Μείωση της Υπερπροσαρμογής - Κανονικοποίηση

Η κανονικοποίηση (regularization) αποτελεί μια ευρύτερη τεχνική που περιλαμβάνει μεθόδους που επιβάλλουν περιορισμούς στην πολυπλοκότητα του μοντέλου, ώστε να μειωθεί το φαινόμενο της υπερπροσαρμογής. Στην πράξη, η κανονικοποίηση μπορεί να

οδηγήσει σε μικρή αύξηση της προκατάληψης αλλά σημαντική μείωση της διακύμανσης εξισορροπώντας την ικανότητα γενίκευσης του μοντέλου.

Η ανάγκη για κανονικοποίηση προκύπτει όταν το μοντέλο διαθέτει υπερβολικά πολλές μεταβλητές ή όταν ο αριθμός των παραμέτρων είναι συγκρίσιμος ή μεγαλύτερος από το μέγεθος των δεδομένων. Σε αυτές τις περιπτώσεις, οι παράμετροι μπορεί να προσαρμόζονται σε θόρυβο αντί για τη βασική δομή των δεδομένων, οδηγώντας σε ασταθείς εκτιμήσεις.

Τρεις από τις πιο δημοφιλείς μέθοδοι κανονικοποίησης στη Μηχανική Μάθηση είναι:

- **Ridge Regression**: Προσθέτει έναν όρο ποινής L_2, ώστε να αποφεύγονται μεγάλες τιμές των συντελεστών.
- **Lasso Regression**: Προσθέτει έναν όρο ποινής L_1, ο οποίος μηδενίζει κάποιους συντελεστές, επιτρέποντας έτσι την επιλογή μεταβλητών.
- **Elastic Net**: Συνδυάζει τις δύο παραπάνω μεθόδους, επιτυγχάνοντας καλύτερη ισορροπία μεταξύ συρρίκνωσης των συντελεστών και επιλογής μεταβλητών.

Ridge Regression: Η μέθοδος παλινδρόμισης Ridge (Ridge regression) [169] είναι μια τεχνική που εφαρμόζεται όταν το πρόβλημα παλινδρόμησης παρουσιάζει υψηλή πολυσυγγραμμικότητα, δηλαδή, όπως προαναφέρθηκε, όταν οι ανεξάρτητες μεταβλητές είναι ισχυρά συσχετισμένες μεταξύ τους. Σε τέτοιες περιπτώσεις, η ύπαρξη υψηλής συσχέτισης οδηγεί σε μεγάλες διακυμάνσεις στις εκτιμήσεις των συντελεστών της παλινδρόμησης, καθιστώντας το μοντέλο ασταθές.

Για να αντιμετωπιστεί αυτό το ζήτημα, η μέθοδος αυτή εισάγει έναν όρο ποινής στο κόστος εκπαίδευσης του μοντέλου, προσθέτοντας το τετράγωνο των συντελεστών (L_2 κανονικοποίηση):

$$L = ||Y - X\beta||^2 + \lambda||\beta||^2 \tag{2.27}$$

όπου:

- λ είναι μια παράμετρος που ελέγχει το μέγεθος της ποινής,
- $||\beta||^2 = \sum_{j=1}^{m} \beta_j^2$ είναι το άθροισμα των τετραγώνων των συντελεστών.

Όταν $\lambda = 0$, η μέθοδος ισοδυναμεί με τη συνήθη γραμμική παλινδρόμηση. Όσο αυξάνεται η τιμή του λ, οι συντελεστές συρρικνώνονται προς το μηδέν, μειώνοντας τη διακύμανση του μοντέλου και βελτιώνοντας την ικανότητα γενίκευσης.

Lasso Regression: Η μέθοδος παλινδρόμησης Lasso (Least Absolute Shrinkage and Selection Operator) [222] είναι μια μέθοδος κανονικοποίησης που, εκτός από τη συρρίκνωση των συντελεστών, έχει την ιδιότητα να μηδενίζει κάποιους από αυτούς, επιτρέποντας έτσι την επιλογή των σημαντικότερων χαρακτηριστικών.

Η λειτουργία της βασίζεται στην προσθήκη ενός όρου ποινής L_1, ο οποίος περιλαμβάνει το άθροισμα των απόλυτων τιμών των συντελεστών:

$$L = ||Y - X\beta||^2 + \lambda||\beta||_1 \tag{2.28}$$

όπου:

- $||\beta||_1 = \sum_{j=1}^{m} |\beta_j|$ είναι το άθροισμα των απόλυτων τιμών των συντελεστών.

Λόγω της φύσης της L_1 κανονικοποίησης, η παλινδρόμηση lasso οδηγεί σε αραιά διανύσματα συντελεστών, δηλαδή ορισμένοι συντελεστές μπορεί να μηδενιστούν πλήρως. Αυτό καθιστά τη lasso ιδιαίτερα χρήσιμη για προβλήματα όπου απαιτείται επιλογή χαρακτηριστικών, μειώνοντας έτσι τη διάσταση των δεδομένων.

Elastic Net: Η μέθοδος elastic net [314] συνδυάζει τις ποινές των ridge και lasso, επιτυγχάνοντας ταυτόχρονα τη συρρίκνωση των συντελεστών και την επιλογή μεταβλητών. Αυτό επιτυγχάνεται με την προσθήκη δύο όρων ποινής:

$$L = ||Y - X\beta||^2 + \alpha\beta||\beta||_1 + \frac{\alpha(1-\beta)}{2}||\beta||^2 \tag{2.29}$$

όπου:

- α είναι η παράμετρος που ελέγχει τη συνολική επίδραση της κανονικοποίησης
- β είναι ένας συντελεστής που ελέγχει τη σχέση του L_1 και του L_2 όρου

Η elastic net προσφέρει μια πιο ισορροπημένη προσέγγιση σε σχέση με τις μεθόδους ridge και lasso, ειδικά σε περιπτώσεις όπου υπάρχουν ισχυρές συσχετίσεις μεταξύ των ανεξάρτητων μεταβλητών.

Οι τεχνικές κανονικοποίησης είναι απαραίτητες στη μηχανική μάθηση για να αντιμετωπιστεί το πρόβλημα της υπερπροσαρμογής και της πολυσυγγραμμικότητας. Η ridge regression είναι ιδιαίτερα χρήσιμη όταν οι ανεξάρτητες μεταβλητές είναι ισχυρά συσχετισμένες, η lasso regression επιτρέπει την επιλογή μεταβλητών, ενώ η elastic net παρέχει έναν συνδυασμό των δύο. Η σωστή επιλογή της μεθόδου εξαρτάται από τη φύση των δεδομένων και τις απαιτήσεις της εφαρμογής.

2.1.2.3 Μετρικές αξιολόγησης της παλινδρόμησης

Η αξιολόγηση της απόδοσης ενός μοντέλου παλινδρόμησης είναι απαραίτητη για την κατανόηση της ικανότητάς του να γενικεύει σε νέα δεδομένα. Δεδομένου ότι κάθε πρόβλημα έχει διαφορετικές απαιτήσεις και χαρακτηριστικά, είναι κρίσιμο να χρησιμοποιούνται πολλαπλές μετρικές για την ολοκληρωμένη αποτίμηση της ποιότητας των προβλέψεων. Οι πιο διαδεδομένες μετρικές παλινδρόμησης περιλαμβάνουν το Μέσο Τετραγωνικό Σφάλμα (Mean Squared Error - MSE), τη Ρίζα του Μέσου Τετραγωνικού Σφάλματος (Root Mean Squared Error - RMSE), η Ρίζα του Μέσου Τετραγωνικού Λογαριθμικού Σφάλματος (Root Mean Squared Logarithmic Error - RMSLE), το Μέσο Απόλυτο Σφάλμα (Mean Absolute Error - MAE), το Ποσοστό του Μέσου Απόλυτου Σφάλματος (Mean Absolute Percentage Error - MAPE), και την μετρική R^2 (Coefficient of determination).

Στις παρακάτω εξισώσεις, το y_i αναπαριστά την πραγματική τιμή, το $\hat{y}_i$ την προβλεπόμενη τιμή και N τον αριθμό των δειγμάτων.

Mean Squared Error (MSE): Η μετρική MSE υπολογίζει το μέσο τετράγωνο της διαφοράς μεταξύ των προβλεπόμενων και πραγματικών τιμών:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 \tag{2.30}$$

Αποτελεί μία από τις πιο διαδεδομένες μετρικές, καθώς τιμωρεί μεγάλα σφάλματα λόγω της ύψωσης στο τετράγωνο. Ωστόσο, είναι ευαίσθητη στην ύπαρξη ακραίων τιμών (outliers).

Root Mean Squared Error (RMSE): Η RMSE είναι η τετραγωνική ρίζα του MSE και παρέχει μια πιο άμεσα κατανοητή μονάδα μέτρησης, καθώς βρίσκεται στην ίδια κλίμακα με τα δεδομένα:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2} \tag{2.31}$$

Σε αντίθεση με το MSE, η RMSE μπορεί να ερμηνευθεί άμεσα ως η μέση απόσταση των προβλέψεων από τις πραγματικές τιμές.

Root Mean Squared Logarithmic Error (RMSLE): Η RMSLE είναι μια τροποποιημένη έκδοση του MSE που χρησιμοποιείται όταν οι προβλέψεις διαφέρουν σημαντικά από τις πραγματικές τιμές, καθώς εφαρμόζει λογαριθμικό μετασχηματισμό:

$$\text{RMSLE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\log(\hat{y}_i + 1) - \log(y_i + 1))^2} \tag{2.32}$$

Αυτή η μετρική μειώνει την επίδραση των ακραίων τιμών και είναι χρήσιμη σε προβλήματα όπου τα σφάλματα των μικρότερων τιμών είναι πιο σημαντικά από εκείνα των μεγαλύτερων τιμών.

Mean Absolute Error (MAE): Η MAE μετρά τη μέση απόλυτη διαφορά μεταξύ των πραγματικών και προβλεπόμενων τιμών:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i| \tag{2.33}$$

Σε αντίθεση με το MSE και το RMSE, η MAE δεν δίνει μεγαλύτερο βάρος στα μεγάλα σφάλματα και είναι πιο ανθεκτική στα ακραία δεδομένα.

Mean Absolute Percentage Error (MAPE): Η MAPE εκφράζει το σφάλμα ως ποσοστό της πραγματικής τιμής, καθιστώντας την χρήσιμη για σύγκριση μοντέλων σε διαφορετικές κλίμακες:

$$\text{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \times 100 \tag{2.34}$$

Παρόλο που είναι εύχρηστη, η MAPE μπορεί να έχει αστάθειες όταν οι πραγματικές τιμές είναι κοντά στο μηδέν.

Coefficient of Determination ($\mathbf{R^2}$): Ο συντελεστής προσδιορισμού R^2 μετρά το ποσοστό της διακύμανσης των δεδομένων που εξηγείται από το μοντέλο παλινδρόμησης:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{2.35}$$

όπου $\bar{y}$ είναι η μέση τιμή των πραγματικών τιμών.

Το R^2 κυμαίνεται από 0 έως 1, όπου υψηλότερες τιμές υποδηλώνουν καλύτερη προσαρμογή του μοντέλου. Σε ορισμένες περιπτώσεις μπορεί να πάρει αρνητικές τιμές, κάτι που σημαίνει ότι το μοντέλο δεν αποδίδει καλύτερα από μια απλή μέση πρόβλεψη.

Η επιλογή της κατάλληλης μετρικής παλινδρόμησης εξαρτάται από το συγκεκριμένο πρόβλημα και τις ανάγκες του μοντέλου. Για γενική αξιολόγηση, η MSE και η RMSE χρησιμοποιούνται συχνά λόγω της βελτιστοποίησης που προσφέρουν. Η MAE είναι πιο εύκολα ερμηνεύσιμη και λιγότερο ευαίσθητη στα ακραία σφάλματα, ενώ η MAPE είναι χρήσιμη για αναλύσεις όπου το σφάλμα σε ποσοστά είναι πιο σημαντικό. Τέλος, το R^2 είναι απαραίτητο για την εκτίμηση της συνολικής ποιότητας του μοντέλου.

2.2 Μη - Επιβλεπόμενη Μηχανική Μάθηση

Στην παράγραφο [2.1] αναφέρθηκε ότι στο σενάριο της επιβλεπόμενης μάθησης, είναι διαθέσιμο ένα σύνολο δεδομένων D που αποτελείται από ζεύγη εισόδων - εξόδων:

$$D = \{(x_i, y_i)\}_{i=1}^{N} \tag{2.36}$$

όπου: x_i είναι ένα διάνυσμα χαρακτηριστικών και y_i η αντίστοιχη επιθυμητή έξοδος. Ο στόχος είναι να βρεθεί μια συνάρτηση $f : X \to Y$, η οποία να ελαχιστοποιεί το σφάλμα $L(y, \hat{y})$.

Αντίθετα, στη μη επιβλεπόμενη μάθηση, ο αλγόριθμος λαμβάνει μόνο δεδομένα εισόδου:

$$D = \{x_i\}_{i=1}^{N} \tag{2.37}$$

χωρίς να υπάρχουν επισημημασμένες τιμές y_i. Ο στόχος είναι η ανακάλυψη κρυφών δομών στα δεδομένα, χρησιμοποιώντας πιθανοτικά μοντέλα που εκτιμούν την κατανομή $P(x)$.

Υπάρχουν τρία βασικά είδη μη επιβλεπόμενης μάθησης. Η ομαδοποίηση, που ίσως είναι το πιο διάσημο, οι κανόνες συσχέτισης και οι μέθοδοι τον εντοπισμό ανωμαλιών. Παρακάτω, θα δοθούν πληροφορίες εν τάχει για τα τρια αυτά ήδη, που αντίστοιχα οδηγούν σε τρεις κατηγορίες προβλημάτων και κατ΄ επέκταση αλγορίθμων.

2.2.1 Τύποι μη επιβλεπόμενης μάθησης

Όπως ήδη αναφέρθηκε παραπάνω, υπάρχουν τρία βασικά είδη μη επιβλεπόμενης μάθησης: η ομαδοποίηση (clustering), οι κανόνες συσχέτισης (association rules) και η ανίσχνευση ανωμαλιών anomaly detection. Παρακάτω περιγράφονται αναλυτικά αυτά τα τρία είδη.

Ομαδοποίηση (Clustering): Η ομαδοποίηση είναι μια από τις πιο διαδεδομένες τεχνικές της μη επιβλεπόμενης μάθησης [301]. Αφορά τη διαίρεση ενός συνόλου δεδομένων σε ομάδες (clusters) με βάση την ομοιότητά τους, χωρίς να υπάρχει εκ των

προτέρων γνώση για τις ετικέτες των δεδομένων. Η διαδικασία αυτή στηρίζεται σε μετρικές απόστασης όπως η Ευκλείδεια απόσταση ή η απόσταση Manhattan [266].

Υπάρχουν διάφοροι τύποι αλγορίθμων ομαδοποίησης:

- **K-Means Clustering:** Ένας από τους πιο γνωστούς αλγορίθμους που βασίζεται στην εύρεση K κέντρων και στην ελαχιστοποίηση της απόστασης κάθε σημείου από το πλησιέστερο κέντρο [156].

- **Ιεραρχική Ομαδοποίηση (Hierarchical Clustering):** Δημιουργεί μια δενδρική δομή ομαδοποίησης, είτε με συγχώνευση ομάδων (συσσωρευτικοί αλγόριθμοι bottom - up) είτε με διαχωρισμό ομάδων (διαιρετικοί αλγόριθμοι top - down) [289]. Οι συσσωρευτικοί αλγόριθμοι, ξεκινούν με κάθε στοιχείο να αποτελεί μια ομάδα και δρουν διαδοχικά, συγχωνεύοντας, ώσπου να μείνει τελικά, μια καθολική ομάδα. Αντίθετα, οι διαιρετικοί αλγόριθμοι ξεκινούν με όλα τα στοιχεία να ανήκουν σε μια, αρχική ομάδα και δρουν διαιρετικά διασπώντας την.

- **Ομαδοποίηση με Βασισμένη σε Πυκνότητα (DBSCAN):** Βασίζεται στην έννοια της πυκνότητας και μπορεί να εντοπίσει ομάδες διαφορετικού σχήματος, καθώς και ανωμαλίες [77].

- **Γκαουσιανή Μείξη (Gaussian Mixture Models - GMM):** Μια πιο ευέλικτη προσέγγιση που αναπαριστά τις ομάδες ως μείγματα κατανομών Gauss [29].

Η ομαδοποίηση εφαρμόζεται σε διάφορα πεδία, όπως η τμηματοποίηση πελατών, όπου βοηθά τις επιχειρήσεις να κατηγοριοποιούν τους πελάτες τους με βάση τη συμπεριφορά ή τα δημογραφικά χαρακτηριστικά τους, επιτρέποντας στοχευμένο μάρκετινγκ και εξατομικευμένες προσφορές. Στην αναγνώριση προτύπων, η ομαδοποίηση χρησιμοποιείται για τον εντοπισμό κοινών χαρακτηριστικών σε δεδομένα, διευκολύνοντας εφαρμογές όπως η κατηγοριοποίηση κειμένων. Επιπλέον, στην ανάλυση εικόνων, η ομαδοποίηση βοηθά στον διαχωρισμό αντικειμένων ή περιοχών μέσα σε μια εικόνα, διευκολύνοντας διαδικασίες όπως η τμηματοποίηση εικόνας και η ανάλυση ιατρικών δεδομένων. Συνεπώς, η ομαδοποίηση είναι ένα ισχυρό εργαλείο για την αποκάλυψη κρυφών σχέσεων στα δεδομένα και χρησιμοποιείται ευρέως σε εφαρμογές που απαιτούν αυτοματοποιημένη εξερεύνηση και ανάλυση μεγάλων όγκων πληροφοριών.

Κανόνες Συσχέτισης (Association Rules): Οι κανόνες συσχέτισης αποσκοπούν στην ανίχνευση σχέσεων μεταξύ διαφορετικών μεταβλητών μέσα σε ένα σύνολο δεδομένων [2]. Η προσέγγιση αυτή εφαρμόζεται ιδιαίτερα σε συστήματα προτάσεων και σε εμπορικές εφαρμογές για την κατανόηση της συμπεριφοράς των καταναλωτών.

Δύο βασικές τεχνικές που χρησιμοποιούνται για τη δημιουργία κανόνων συσχέτισης είναι:

- **Ο αλγόριθμος Apriori:** Βασίζεται στην εύρεση συχνά επαναλαμβανόμενων στοιχείων και στη δημιουργία κανόνων της μορφής ¨A $\rightarrow$ B' μετρώντας δείκτες όπως η υποστήριξη, η εμπιστοσύνη και ο δείκτης βελτίωσης [2].

- **Ο αλγόριθμος FP-Growth:** Αποτελέι μια πιο αποδοτική μέθοδος εξόρυξης συχνών μοτίβων που χρησιμοποιεί δομές δέντρων (FP-Trees) [110].

Η ανάλυση κανόνων συσχέτισης εφαρμόζεται σε διάφορους τομείς, όπως η κατηγοριοποίηση αγοραστικών προτύπων, όπου βοηθά τις επιχειρήσεις να εντοπίσουν συνδυασμούς προϊόντων που αγοράζονται συχνά μαζί, οδηγώντας στη βελτίωση της στρατηγικής πωλήσεων και στην αποτελεσματικότερη τοποθέτηση προϊόντων. Στη βελτιστοποίηση αποθεμάτων, η ανάλυση κανόνων συσχέτισης επιτρέπει την πρόβλεψη της ζήτησης για συγκεκριμένα προϊόντα, μειώνοντας τα αποθέματα και εξασφαλίζοντας αποτελεσματική διαχείριση της εφοδιαστικής αλυσίδας. Επιπλέον, στη δημιουργία εξατομικευμένων συστάσεων, η τεχνική αυτή χρησιμοποιείται από συστήματα προτάσεων, όπως οι πλατφόρμες ηλεκτρονικού εμπορίου και οι υπηρεσίες streaming, για την προσαρμογή προτάσεων βάσει των προτιμήσεων των χρηστών.

Ανίχνευση Ανωμαλιών (Anomaly Detection): Η ανίχνευση ανωμαλιών αποτελεί μια διαδικασία εντοπισμού παρατηρήσεων που αποκλίνουν σημαντικά από τη γενική συμπεριφορά των δεδομένων [43]. Αυτές οι ανωμαλίες μπορεί να υποδηλώνουν σφάλματα ή απροσδόκητες αλλαγές.

Υπάρχουν διάφορες προσεγγίσεις για την ανίχνευση ανωμαλιών:

- **Μέθοδοι βασισμένες σε στατιστικά μοντέλα:** Αξιοποιούν τις κατανομές των δεδομένων για να καθορίσουν ποια παρατήρηση αποκλίνει από την κανονική συμπεριφορά [43].
- **Μέθοδοι βασισμένες σε αποστάσεις:** Χρησιμοποιούν μετρικές απόστασης για να εντοπίσουν σημεία δεδομένων που βρίσκονται απομονωμένα από το υπόλοιπο σύνολο [37].
- **Μέθοδοι Μηχανικής Μάθησης:** Εφαρμόζουν αλγορίθμους για την εκπαίδευση μοντέλων που διακρίνουν μεταξύ φυσιολογικών και ανώμαλων μοτίβων.

Η ανίχνευση ανωμαλιών έχει εφαρμογές σε τομείς όπως η ασφάλεια δικτύων, η χρηματοοικονομική απάτη και η προληπτική συντήρηση σε βιομηχανικές εφαρμογές, συμβάλλοντας στην έγκαιρη αναγνώριση και αντιμετώπιση πιθανών προβλημάτων.

2.3 Ημι - Επιβλεπόμενη Μάθηση

Η ημι - επιβλεπόμενη μάθηση (Semi - Supervised Learning) αποτελέι μια υβριδική κατηγορία Μηχανικής Μάθησης που συνδυάζει στοιχεία της επιβλεπόμενης μάθησης και της μη επιβλεπόμενης μάθησης. Σε ένα τυπικό πρόβλημα ημί-επιβλεπόμενης μάθησης, το σύνολο δεδομένων περιέχει έναν μικρό αριθμό δειγμάτων με ετικέτες και έναν πολύ μεγαλύτερο αριθμό μη επισημασμένων δειγμάτων. Τα μοντέλα ημι-επιβλεπόμενης μάθησης εκμεταλλεύονται την πληροφορία από τα μη επισημασμένα δεδομένα για να βελτιώσουν την συνολική την απόδοση του μοντέλου. Συνεπώς, η ημι-επιβλεπόμενη μάθηση είναι ιδιαίτερα χρήσιμη σε περιπτώσεις όπου η απόκτηση επισημασμένων δεδομένων είναι δαπανηρή ή χρονοβόρα, ενώ τα μη επισημασμένα δεδομένα είναι άφθονα.

Υπάρχουν διάφορες μέθοδοι που χρησιμοποιούνται στην ημι-επιβλεπόμενη μάθηση, όπως:

- **Αλγόριθμοι βασισμένοι στη συνέπεια (Consistency-based methods)**: Χρησιμοποιούν μη επισημασμένα δεδομένα για να μειώσουν την αβεβαιότητα των ταξινομητών μέσω τεχνικών όπως η ψευδο-ετικετοποίηση (pseudo-labeling),

όπου το μοντέλο δημιουργεί τις δικές του ετικέτες για τα μη επισημασμένα δεδομένα.

- **Μοντέλα πολλαπλών απόψεων (Multi-view learning)**: Συνδυάζουν διαφορετικές αναπαραστάσεις των δεδομένων για να βελτιώσουν την ταξινόμηση.
- **Αυτοεπιβλεπόμενη μάθηση (Self-supervised learning)**: Δημιουργεί συνθετικές ετικέτες από τα μη επισημασμένα δεδομένα για να εκπαιδεύσει το μοντέλο.
- **Μέθοδοι γραφημάτων (Graph-based methods)**: Μοντελοποιούν τη σχέση μεταξύ των δειγμάτων ως γράφο και χρησιμοποιούν αλγορίθμους διάδοσης ετικετών.

Παρ΄ όλο που η ημι-επιβλεπόμενη μάθηση προσφέρει σημαντικά πλεονεκτήματα, υπάρχουν και προκλήσεις που πρέπει να αντιμετωπιστούν. Ένα από τα βασικά ζητήματα είναι η ποιότητα των μη επισημασμένων δεδομένων, καθώς αν περιέχουν θόρυβο ή δεν είναι αντιπροσωπευτικά, μπορεί να επηρεάσουν αρνητικά την ακρίβεια και τη γενίκευση του μοντέλου. Επιπλέον, η επιλογή των κατάλληλων υπερπαραμέτρων αποτελεί κρίσιμο παράγοντα, καθώς ορισμένοι αλγόριθμοι απαιτούν λεπτομερή ρύθμιση για να λειτουργήσουν αποτελεσματικά. Τέλος, μια ακόμη σημαντική πρόκληση είναι η γενίκευση σε νέα δεδομένα, καθώς ενώ τα μη επισημασμένα δεδομένα μπορούν να βελτιώσουν την απόδοση του μοντέλου, η ακατάλληλη χρήση τους ενδέχεται να οδηγήσει σε υπερπροσαρμογή (overfitting).

2.4 Ενισχυτική Μάθηση

Η ενισχυτική μάθηση (Reinforcement Learning - RL) είναι ένα υποπεδίο της Μηχανικής Μάθησης που επικεντρώνεται στη διαδικασία λήψης αποφάσεων ενός πράκτορα (agent) σε ένα δυναμικό περιβάλλον [264]. Αντί να διαθέτει ένα στατικό σύνολο δεδομένων εκπαίδευσης, όπως συμβαίνει στην επιβλεπόμενη μάθηση, ο πράκτορας αλληλεπιδρά με το περιβάλλον και μαθαίνει μέσω ενός μηχανισμού ανταμοιβών, με στόχο τη μεγιστοποίηση της συνολικής ανταμοιβής μέσω εξερεύνησης και εκμετάλλευσης [264].

Η ενισχυτική μάθηση έχει πολλές εφαρμογές σε τομείς όπου η λήψη αποφάσεων σε αλληλεπιδρώντα συστήματα είναι κρίσιμη. Στη ρομποτική, επιτρέπει σε αυτόνομα ρομπότ να μάθουν να εκτελούν σύνθετες εργασίες μέσω πειραματισμού [137]. Στα αυτόνομα οχήματα, χρησιμοποιείται για τη λήψη βέλτιστων αποφάσεων πλοήγησης σε πραγματικό χρόνο. Στα παιχνίδια στρατηγικής, όπως το Go και το σκάκι, έχει οδηγήσει στην ανάπτυξη υπερ-ανθρώπινων πρακτόρων που ξεπερνούν την ανθρώπινη ικανότητα [254]. Επιπλέον, εφαρμόζεται στη βελτιστοποίηση συστημάτων, όπως η διαχείριση πόρων σε κέντρα δεδομένων και η ρύθμιση χρηματοοικονομικών χαρτοφυλακίων.

Παρ΄ όλο που η ενισχυτική μάθηση έχει σημειώσει σημαντική πρόοδο, υπάρχουν αρκετές προκλήσεις που πρέπει να αντιμετωπιστούν. Ένα από τα κύρια ζητήματα είναι η ανάγκη για μεγάλο αριθμό αλληλεπιδράσεων με το περιβάλλον, καθώς η εκπαίδευση των πρακτόρων απαιτεί συχνά χιλιάδες ή εκατομμύρια δοκιμές για την εύρεση βέλτιστων στρατηγικών. Επιπλέον, η διαδικασία εκπαίδευσης είναι συχνά ασταθής και υπολογιστικά απαιτητική, δεδομένου ότι η εξερεύνηση του περιβάλλοντος μπορεί να οδηγήσει σε μη επιθυμητές ή καταστροφικές ενέργειες. Ένα άλλο σημαντικό πρόβλημα είναι η γενίκευση, καθώς οι πράκτορες που εκπαιδεύονται σε ένα συγκεκριμένο περιβάλλον

συχνά δυσκολεύονται να προσαρμοστούν σε διαφορετικά περιβάλλοντα. Παρά τις προκλήσεις, η συνεχιζόμενη έρευνα στην εκμάθηση με ελάχιστη επίβλεψη, στις μεθόδους μεταφοράς μάθησης (transfer learning) και στην εκμάθηση με ανθρώπινη καθοδήγηση (human-in-the-loop learning) αναμένεται να βελτιώσει την αποδοτικότητα και τη γενίκευση των αλγόριθμων ενισχυτικής μάθησης, διευρύνοντας την εφαρμογή τους σε πραγματικά προβλήματα.

3. Ενεργή Μάθηση

Η Ενεργή Μάθηση (Active Learning) είναι μια προσέγγιση στη Μηχανική Μάθηση που επιτρέπει σε ένα σύστημα να επιλέγει τα δεδομένα από τα οποία θα μάθει, αντί να βασίζεται σε ένα τυχαίο ή προεπιλεγμένο σύνολο δεδομένων. Σε αντίθεση με την κλασσική Μηχανική Μάθηση, όπου το μοντέλο εκπαιδεύεται με ένα μεγάλο σύνολο ετικετοποιημένων δεδομένων, η Ενεργή Μάθηση δίνει τη δυνατότητα στο σύστημα να διατυπώνει ερωτήσεις και να ζητά ετικέτες μόνο για τα πιο κρίσιμα δείγματα. Είναι σύνηθες το φαινόμενο της ύπαρξης τεράστιου όγκου δεδομένων, χωρίς όμως αυτά να συνοδεύονται με την ετικέτα τους. Τις περισσότερες φορές μάλιστα, η απόκτηση των αντίστοιχων ετικετών αποτελεί δαπανηρή διαδικασία, τόσο σε κόστος όσο και σε χρόνο. Συνήθως η διαδικασία ετικετοποίησης (annotation) απαιτεί την εμπλοκή ενός ειδικού του πεδίου - στη βιβλιογραφία συναντάμε συχνά τη λέξη μαντείο (oracle) για να περιγράψει την πηγή προέλευσης της πληροφορίας αυτής.

Σαν έννοια, η Ενεργή Μάθηση εμφανίζεται πρώτη φορά περί τη δεκαετία του 1990, με μια από τις πρώτες και πιο επιδραστικές δημοσιεύσεις να είναι το άρθρο των D. Cohn, L. Atlas και R. Ladner το 1994 [50]. Σε αυτό το έργο, οι συγγραφείς παρουσίασαν έναν αλγόριθμο Ενεργής Μάθησης που επιλέγει δείγματα προς ετικετοποίηση με βάση την αβεβαιότητα του μοντέλου, θέτοντας τα θεμέλια για μελλοντικές έρευνες και εφαρμογές στον τομέα. Λίγο αργότερα, εισάγεται στη βιβλιογραφία το πλαίσιο της στρατηγικής Query by Committee [87] , η οποία αναδεικνύει τη σημασία της επιλογής δείγματος με βάση την πληροφοριακή του αξία. Η προσέγγιση αυτή έδειξε ότι η Ενεργή Μάθηση μπορεί να επιταχύνει τη διαδικασία εκπαίδευσης ενός ταξινομητή, μειώνοντας σημαντικά την ανάγκη για ετικετοποιημένα δεδομένα.

Στη συνέχεια, οι Tong και Koller [273] συνδύασαν τεχνικές Ενεργής Μάθησης με Μηχανές Διανισματικής Υποστήριξης. Η εργασία τους εισήγαγε την έννοια της τεχνικής της δειγματοληψίας περιθωρίου margin-based sampling, όπου το μοντέλο επιλέγει δείγματα κοντά στο όριο απόφασης, βελτιώνοντας έτσι την απόδοσή του με ελάχιστες ετικέτες, όπως θα δούμε στη συνέχεια. Μια σημαντική επέκταση έγινε από τους Roy και McCallum [232], οι οποίοι εισήγαγαν τη στρατηγική μείωσης του αναμενόμενου σφάλματος Expected Error Reduction, αποδεικνύοντας θεωρητικά ότι η στρατηγική επιλογής δειγμάτων μπορεί να μειώσει δραματικά το συνολικό σφάλμα του μοντέλου. Το 2003, οι McCallum και Nigam [167] εφάρμοσαν Ενεργή Μάθηση σε μοντέλα πιθανοτήτων, όπως σε Αφελείς Μπευζιανούς ταξινομητές, προτείνοντας τη χρήση πληροφοριακής εντροπίας για την επιλογή δειγμάτων. Αυτή η μέθοδος αποτέλεσε τη βάση για μεταγενέστερες Μπευζιανές προσεγγίσεις.

Λίγα χρόνια αργότερα, οι Settles και Craven [249] εισήγαγαν ημι-επιβλεπόμενες μεθόδους Ενεργής Μάθησης, οι οποίες συνδυάζουν μικρό αριθμό ετικετοποιημένων δειγμάτων με μεγάλα σύνολα μη ετικετοποιημένων δεδομένων. Στη συνέχεια, ο Dasgupta [58] ανέπτυσει θεωρητικές εγγυήσεις για την αποδοτικότητα της Ενεργής Μάθησης, θέτοντας τα θεμέλια για τη σύγκριση μεταξύ Ενεργής και κλασικής μάθησης. Το 2012, ο Settles [248] δημοσίευσε μια επισκόπηση των τεχνικών Ενεργής Μάθησης, καλύπτο-

ντας τόσο θεωρητικά όσο και πρακτικά ζητήματα, ενώ οι Hoi et al. [120] παρουσίασαν τις πρώτες εφαρμογές της ενεργής μάθησης σε χρηματοοικονομικά δεδομένα.

Το 2015, οι Wang και Yeung [285] συνδύασαν την Ενεργή Μάθηση με τεχνικές Βαθιάς Μάθησης χρησιμοποιώντας νευρωνικά δίκτυα για την επιλογή δειγμάτων, ανοίγοντας τον δρόμο για σύγχρονες εφαρμογές. Δύο χρόνια αργότερα, το 2017, οι Gal et al. [90] ανέπτυξαν και αξιοποίησαν την Ενεργή Μάθηση με αβεβαιότητα σε συνδυασμό με στοχαστικά βαθιά δίκτυα. Το επόμενο έτος, οι Sener και Savarese [247] πρότειναν μια προσέγγιση που βασίζεται σε Ενεργή Μάθηση βασισμένη σε μεθόδους επιλογής αντιπροσωπευτικών συνόλων (Core-Set Selection), όπου τα δείγματα επιλέγονται με γνώμονα την καλύτερη αναπαράσταση της κατανομής δεδομένων.

Το 2020, οι Ash et al. [17] πρότειναν το BADGE (Batch Active Learning by Diverse Gradient Embeddings), το οποίο εκμεταλλεύεται πληροφορίες από τις κλίσεις των νευρωνικών δικτύων για επιλογή δειγμάτων. Τέλος, το 2022, οι Fujii και Matsuo [89] συνέδεσαν την Ενεργή Μάθηση με την ενισχυτική μάθηση (Reinforcement Learning), ανοίγοντας νέες προοπτικές για την αυτοματοποιημένη επιλογή δεδομένων.

Όπως ήδη σχολιάστηκε, η βασική ιδέα πίσω από την προσέγγιση είναι ότι ένα μοντέλο μπορεί να βελτιώσει την ακρίβειά του με λιγότερα δεδομένα και συνεπώς να μειωθεί το κόστος της όλης διαδικασίας, εάν επιλέξει έξυπνα τις πιο χρήσιμες παρατηρήσεις για εκμάθηση. Έτσι, η κεντρική διαφορά του κλασσικού σεναρίου σε αντίθεση με αυτό της Ενεργής Μηχανικής Μάθησης, συνοψίζεται ως εξής:

- **Κλασική Μηχανική Μάθηση**: Το μοντέλο εκπαιδεύεται με ένα προεπιλεγμένο σύνολο δεδομένων, το οποίο μπορεί να είναι μεγάλο και ενδεχομένως να περιέχει περιττές ή μη αντιπροσωπευτικές παρατηρήσεις.
- **Ενεργή Μηχανική Μάθηση**: Το μοντέλο επιλέγει τα δεδομένα που θεωρεί πιο χρήσιμα για εκμάθηση, μειώνοντας την ανάγκη για μεγάλο αριθμό ετικετών και βελτιώνοντας την αποδοτικότητα της εκπαίδευσης.

Η Ενεργή Μάθηση βασίζεται στη διατύπωση ερωτημάτων (queries) προς μια πηγή πληροφοριών, όπως ένας ανθρώπινος ειδικός, η οποία παρέχει τις απαραίτητες ετικέτες για τα δεδομένα. Με αυτόν τον τρόπο, το σύστημα συγκεντρώνει μόνο τις πιο κρίσιμες πληροφορίες, εξοικονομώντας χρόνο και κόστος. Για τις στρατηγικές διατύπωσης ερωτημάτων, θα γίνει εκτενέστερη αναφορά παρακάτω.

3.1 Διαδικασία Ενεργής Μάθησης

Ένας αλγόριθμος Ενεργής Μάθησης λειτουργεί αρχικά με ένα μικρό σύνολο ετικετοποιημένων δεδομένων ως σύνολο εκπαίδευσης, το οποίο χρησιμοποιείται για την εκπαίδευση του αρχικού μοντέλου. Στη συνέχεια, το μοντέλο αναλύει ένα μεγαλύτερο σύνολο μη ετικετοποιημένων δεδομένων και αξιολογεί την αβεβαιότητα ή την πληροφοριακή αξία κάθε δείγματος. Με βάση αυτό το κριτήριο, επιλέγει τα πιο αβέβαια ή χρήσιμα δείγματα και ζητά από ένα «μαντείο» (oracle), όπως έναν ανθρώπινο αναλυτή ή αυτοματοποιημένο σύστημα, να τους αποδώσει τις σωστές ετικέτες. Αυτά τα νέα δεδομένα προστίθενται στο σύνολο εκπαίδευσης και το μοντέλο επανεκπαιδεύεται στο νέο εμπλουτισμένο σύνολο δεδομένων. Η διαδικασία επαναλαμβάνεται έως ότου το μοντέλο επιτύχει την επιθυμητή ακρίβεια ή εξαντληθούν οι διαθέσιμοι πόροι, βελτιώνοντας την απόδοσή του με λιγότερα δεδομένα εκπαίδευσης σε σύγκριση με μια παραδοσιακή προσέγγιση. Η παραπάνω διαδικασία περιγράφεται στον Αλγόριθμο 1.

Αλγόριθμος 1 Διαδικασία Ενεργής Μάθησης

1: **Αρχικοποίηση**: Ορισμός αρχικού συνόλου ετικετοποιημένων δεδομένων L και μη ετικετοποιημένων δεδομένων U.
2: Εκπαίδευση του μοντέλου M με το σύνολο L.
3: **Όσο** δεν επιτευχθεί η επιθυμητή ακρίβεια ή δεν εξαντληθούν οι πόροι **δο**
4: Υπολογισμός μέτρου αβεβαιότητας $u(x)$ για κάθε $x \in U$.
5: Επιλογή του δείγματος:

$$x^* = \arg\max_{x \in U} u(x)$$

6: Απόκτηση ετικέτας y^* από τον oracle.
7: Προσθήκη του (x^*, y^*) στο L και αφαίρεση από το U.
8: Επανεκπαίδευση του μοντέλου M με το ενημερωμένο L.
9: **Τέλος Όσο**

3.2 Σενάρια Ενεργής Μάθησης

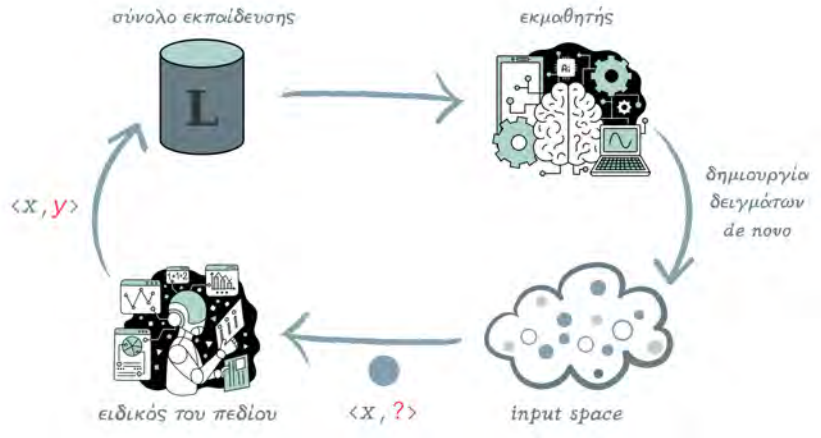
Πριν εξετάσουμε τα βασικά σενάρια χρήσης της Ενεργής Μάθησης, είναι σημαντικό να κατανοήσουμε πότε η χρήση της συνίσταται και πότε όχι. Σε ορισμένες εφαρμογές, η απόκτηση ετικετών για τα δεδομένα είναι φθηνή ή ακόμα και δωρεάν, όπως όταν οι χρήστες σημειώνουν ως ανεπιθύμητα emails της αλληλογραφίας τους ή βαθμολογούν ταινίες σε μια πλατφόρμα. Σε τέτοιες περιπτώσεις, τα συστήματα μπορούν να συλλέγουν φυσικά πολλές ετικέτες χωρίς επιπλέον κόστος, καθιστώντας την Ενεργή Μάθηση λιγότερο απαραίτητη. Επιπλέον, όταν απαιτούνται μόνο λίγες δεκάδες ή εκατοντάδες ετικετοποιημένα παραδείγματα για την εκπαίδευση ενός ακριβούς μοντέλου, η πολυπλοκότητα της Ενεργής Μάθησης μπορεί να είναι αποτρεπτική. Αντίθετα, η Ενεργή Μάθηση είναι χρήσιμη όταν υπάρχουν μεγάλες ποσότητες μη ετικετοποιημένων δεδομένων, τα οποία μπορούν να συλλεχθούν εύκολα, αλλά η ετικετοποίησή τους είναι δαπανηρή ή χρονοβόρα. Σε αυτές τις περιπτώσεις, η επιλογή των πιο χρήσιμων δειγμάτων προς ετικετοποίηση μπορεί να οδηγήσει σε σημαντική εξοικονόμηση πόρων και βελτίωση της ακρίβειας του μοντέλου. Για τους σκοπούς αυτής της ανάλυσης, υποθέτουμε ότι το «μαντείο» μπορεί να παρέχει απαντήσεις για τις ετικέτες των δεδομένων και ότι η κατάλληλη κατηγορία κλάσεων είναι ήδη γνωστή. Δεδομένου ότι η Ενεργή Μάθηση είναι κατάλληλη, υπάρχουν διάφοροι τρόποι με τους οποίους ο αλγόριθμος μπορεί να διατυπώσει ερωτήματα. Οι τρεις βασικές προσεγγίσεις που έχουν προταθεί στη βιβλιογραφία είναι οι εξής [248]:

- **Query Synthesis**: Το σενάριο αυτό, αποτελεί μία από τις πρώτες μορφές Ενεργής Μάθησης που μελετήθηκαν. Σε αυτό το σενάριο, ο εκμαθητής μπορεί να ζητήσει την επισήμανση οποιουδήποτε μη επισημασμένου δείγματος δεδομένων στον χώρο εισόδου, ακόμα και να κατασκευάσει δείγματα προς ετικετοποίηση de novo (Σχήμα 3.1). Αν και η μεθοδολογία αυτή είναι ιδιαίτερα αποδοτική για πεπερασμένους χώρους προβλημάτων, παρουσιάζει προκλήσεις σε πιο περίπλοκες εφαρμογές, όπως η αναγνώριση χειρόγραφων χαρακτήρων, όπου η αυτόματη δημιουργία ερωτημάτων συχνά καταλήγει σε ακατανόητα ή μη φυσικά δείγματα. Αυτό μπορεί να οδηγήσει σε δυσκολία επισήμανσης από ανθρώπινους ειδικούς ή να παράγει άχρηστα δεδομένα για την εκπαίδευση. Παρόμοιες προκλήσεις εμφανίζονται και σε προβλήματα φυσικής γλώσσας, όπου η αυτόματη σύνθεση

ερωτημάτων μπορεί να δημιουργήσει κείμενα άνευ νοήματος ή ακατανόητο ηχητικό υλικό. Για να ξεπεραστούν αυτές οι αδυναμίες, έχουν προταθεί εναλλακτικά μοντέλα, όπως τα σενάρια που βασίζονται σε ροές δεδομένων (stream-based) και δεξαμενές δεδομένων (pool-based), τα οποία λαμβάνουν περισσότερο υπόψη την κατανομή των δεδομένων και αποτρέπουν την παραγωγή άσκοπων ή μη χρήσιμων ερωτημάτων.

- **Stream-Based Selective Sampling**: Το σενάριο αυτό, που στα ελληνικά μπορεί να αποδοθεί ως «επιλεκτική δειγματοληψία βασισμένη σε ροές δεδομένων», αποτελεί μια εναλλακτική προσέγγιση στην ενεργή μάθηση, όπου τα μη επισημασμένα δείγματα αντλούνται από την πραγματική κατανομή δεδομένων και στη συνέχεια ο εκμαθητής αποφασίζει αν θα ζητήσει την ετικέτα τους ή θα τα απορρίψει 3.2. Σε αντίθεση με τη σύνθεση ερωτημάτων, αυτή η μέθοδος διασφαλίζει ότι τα επιλεγμένα δείγματα έχουν νόημα, καθώς προέρχονται από μια πραγματική κατανομή δεδομένων. Η απόφαση για το αν θα γίνει ερώτημα μπορεί να βασίζεται σε μέτρα που αξιολογούν τη χρησιμότητα ή το περιεχόμενο πληροφορίας του δείγματος, ή να καθορίζεται από την περιοχή αβεβαιότητας του εκπαιδευόμενου μοντέλου. Παρ΄ όλο που ο πλήρης υπολογισμός αυτής της περιοχής είναι υπολογιστικά απαιτητικός, προσεγγίσεις όπως η διατήρηση ενός version space επιτρέπουν αποδοτικότερη επιλογή δειγμάτων. Η μέθοδος αυτή έχει εφαρμοστεί σε διάφορα πεδία, όπως η ταξινόμηση μερών του λόγου, ο προγραμματισμός αισθητήρων και η διάκριση σημασιολογικών εννοιών σε φυσική γλώσσα. Αν και ορισμένοι ερευνητές χρησιμοποιούν τον όρο «selective sampling» για να περιγράψουν και το pool-based σενάριο, στις περισσότερες περιπτώσεις αναφέρεται στη ροή δεδομένων, όπου η επιλογή των δειγμάτων γίνεται δυναμικά, ένα-ένα, από την πηγή δεδομένων.

- **Pool-Based Sampling**: Αυτό το είδος της δειγματοληψίας χρησιμοποιείται ευρέως σε πραγματικά προβλήματα, όπου υπάρχει ένα μικρό σύνολο επισημασμένων δεδομένων και ένας μεγάλος στατικός ή ημι-στατικός χώρος μη επισημασμένων δεδομένων. Ο εκμαθητής αξιολογεί όλα τα διαθέσιμα δείγματα (ή ένα υποσύνολό τους αν το σύνολο δεδομένων είναι πολύ μεγάλο) χρησιμοποιώντας ένα κριτήριο



Σχήμα 3.1: Query synthesis: Ο εκμαθητής μπορεί να ζητά την ετικέτα οποιουδήποτε δείγματος - αυτό σημαίνει ότι είναι δυνατή η δημιουργία δυγμάτων de novo

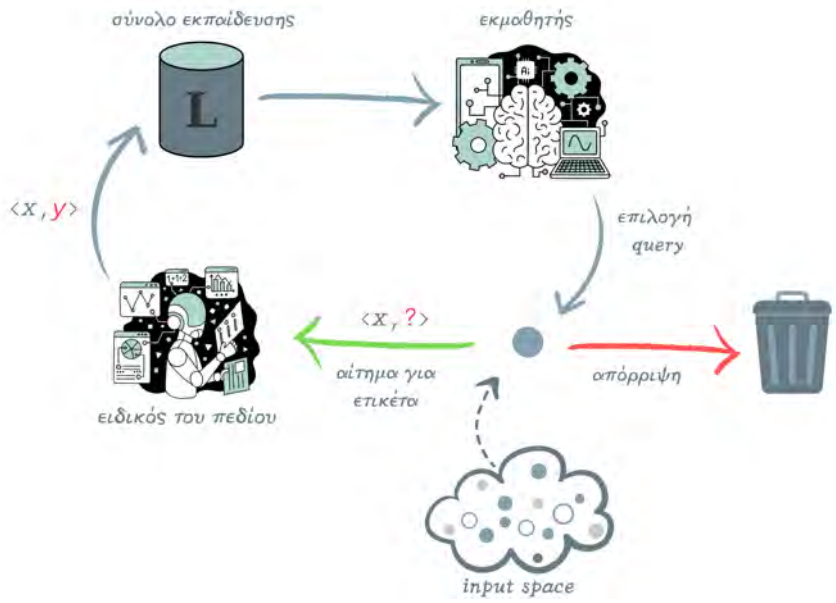
.

βασισμένο στην πληροφορία και επιλέγει εκείνα με την περισσότερη πληροφορία για επισήμανση. Αυτή η μέθοδος έχει εφαρμοστεί σε διάφορα πεδία, όπως η ταξινόμηση κειμένου, η εξαγωγή πληροφορίας, η αναγνώριση εικόνων και βίντεο, η αναγνώριση ομιλίας και η διάγνωση ασθενειών, καθιστώντας την ιδιαίτερα δημοφιλή στην εφαρμοσμένη έρευνα. Η βασική διαφορά μεταξύ του pool-based και του stream-based σεναρίου είναι ότι η πρώτη αξιολογεί ολόκληρο το σύνολο των μη επισημασμένων δεδομένων πριν επιλέξει τα πιο κατάλληλα ερωτήματα, ενώ η δεύτερη διαχειρίζεται τις αποφάσεις ερωτημάτων διαδοχικά, επεξεργαζόμενη τα δεδομένα ένα-ένα καθώς εμφανίζονται. Παρόλο που το pool-based σενάριο είναι πιο διαδεδομένο, το stream-based είναι προτιμότερο όταν οι διαθέσιμοι υπολογιστικοί πόροι είναι περιορισμένοι ή όταν το σύνολο δεδομένων είναι πολύ μεγάλο για να αποθηκευτεί στη μνήμη και πρέπει να προσπελαστεί σταδιακά. Στην πράξη, η επιλογή της κατάλληλης μεθόδου εξαρτάται από τις απαιτήσεις του εκάστοτε προβλήματος και των διαθέσιμων πόρων.

3.3 Στρατηγικές επιλογής ερωτημάτων

Παραπάνω παρουσιάστηκαν τα βασικά σενάρια της Ενεργής Μάθησης που αφορούν την επιλογή των προς ετικετοποίηση δειγμάτων. Στη συνέχεια, παρουσιάζονται οι βασικές στρατηγικές, με τις οποίες αποφασίζεται αν η ετικέτα ενός δείγματος θα ζητηθεί ή όχι και εν γένει, οι τεχνικές που καθορίζουν ποια δείγματα θα ετικετοποιηθούν.

Δειγματοληψία Αβεβαιότητας (Uncertainty Sampling): Η δειγματοληψία αβεβαιότητας είναι μία από τις πιο απλές και διαδεδομένες στρατηγικές για την επιλογή δειγμάτων προς ετικετοποίηση στα σενάρια εφαρμογής Ενεργής Μάθησης. Η βασική ιδέα της είναι να επιλέγει για ετικετοποίηση τα δείγματα για τα οποία το μοντέλο είναι πιο αβέβαιο σχετικά με την πρόβλεψή τους. Η λογική πίσω από αυτή τη μέθοδο είναι



Σχήμα 3.2: Stream-Based Selective Sampling: Ο εκμαθητής αποφασίζει αν θα αιτηθεί την ετικέτα ενός δείγματος ή αν θα το απορρίψει για να περάσει στο επόμενο.

ότι τα πιο αβέβαια παραδείγματα παρέχουν τη μέγιστη πληροφορία για τη βελτίωση του μοντέλου.

Υπάρχουν διάφοροι τρόποι μέτρησης της αβεβαιότητας ενός δείγματος x, με τις πιο κοινές τεχνικές να περιλαμβάνουν:

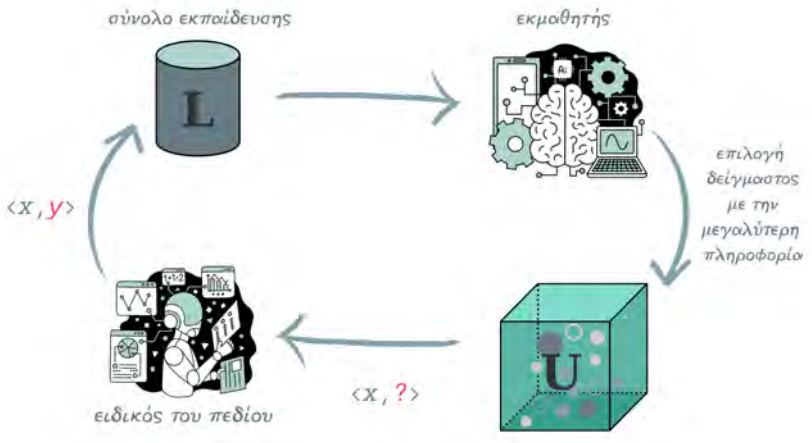
- **Δειγματοληψία με βάση τη Λιγότερη Εμπιστοσύνη (Least Confident Sampling):** Στη μέθοδο αυτή, επιλέγεται το δείγμα για το οποίο το μοντέλο έχει τη χαμηλότερη πιθανότητα πρόβλεψης για την επικρατέστερη κλάση. Η αβεβαιότητα για ένα δείγμα x υπολογίζεται ως:

$$\begin{aligned} U(x) &= \arg\min_x P_\theta(\hat{y}|x) \\ &= \arg\max_x (1 - P_\theta(\hat{y}|x)) \end{aligned} \tag{3.1}$$

  όπου $\hat{y} = \arg\max_y P_\theta(y|x)$, δηλαδή η κλάση με τη μεγαλύτερη εκ των προτέρων πιθανότητα για το δείγμα x βάσει του μοντέλου θ. Με άλλα λόγια, σύμφωνα με αυτή τη στρατηγική επιλέγεται το δείγμα για το οποίο το μοντέλο εμφανίζει τη μικρότερη βεβαιότητα για την προβλεπόμενη κλάση. Αν και αυτή η μέθοδος είναι ιδιαίτερα απλή και εύκολη στην υλοποίηση, ένα βασικό της μειονέκτημα είναι ότι αγνοεί πληροφορία σχετικά με την υπόλοιπη κατανομή πιθανοτήτων στις κλάσεις, λαμβάνοντας υπόψη μόνο την πιο πιθανή πρόβλεψη.

- **Δειγματοληψία Με Βάση Το Περιθώριο (Margin Sampling)**: Η μέθοδος αυτή επιλέγει το δείγμα για το οποίο η διαφορά μεταξύ των δύο πιο πιθανών κλάσεων είναι η μικρότερη, δηλαδή το μοντέλο έχει δυσκολία να ξεχωρίσει ανάμεσα σε δύο πιθανές κλάσεις:

$$\begin{aligned} x_M^* &= \arg\min_x [P_\theta(\hat{y}_1|x) - P_\theta(\hat{y}_2|x)] \\ &= \arg\max_x [P_\theta(\hat{y}_2|x) - P_\theta(\hat{y}_1|x)] \end{aligned} \tag{3.2}$$



Σχήμα 3.3: Pool-based selective sampling: Ο εκμαθητής αξιολογεί όλα τα διαθέσιμα δείγματα (ή ένα υποσύνολό τους) χρησιμοποιώντας ένα κριτήριο βασισμένο στην πληροφορία και επιλέγει εκείνο με την περισσότερη πληροφορία για επισήμανση.

όπου $\hat{y}_1$ και $\hat{y}_2$ είναι οι δύο πιο πιθανές προβλέψεις του μοντέλου για το δείγμα x. Η προσέγγιση του περιθωρίου βελτιώνει τη μέθοδο που βασίζεται στη λιγότερη εμπιστοσύνη. καθώς ενσωματώνει πληροφορία όχι μόνο για την πιο πιθανή, αλλά και για τη δεύτερη πιθανότερη κλάση. Ωστόσο, σε περιπτώσεις με μεγάλο πλήθος κλάσεων, η μέθοδος αυτή συνεχίζει να αγνοεί σημαντικό μέρος της πληροφορίας που περιέχεται στην πλήρη κατανομή των πιθανοτήτων.

- **Δειγματοληψία Με Βάση Την Εντροπία** (Entropy-Based Sampling): Η εντροπία μετράει την αβεβαιότητα ενός δείγματος σε όλες τις πιθανές κλάσεις. Έτσι, ως μέτρο χρησιμότητας επιλέγεται το δείγμα για το οποίο η εντροπία της εκτιμώμενης κατανομής πιθανοτήτων είναι μέγιστη:

$$\begin{aligned} x_H^* &= \arg\max_x \; H_\theta(Y|x) \\ &= \arg\max_x - \sum_y P_\theta(y|x) \log P_\theta(y|x) \end{aligned} \tag{3.3}$$

όπου y διατρέχει όλες τις δυνατές ετικέτες για το x. Η εντροπία εκφράζει το μέσο πληροφοριακό περιεχόμενο μιας μεταβλητής και, στο πλαίσιο της ενεργής μάθησης, αποτελεί έναν από τους πιο αποτελεσματικούς τρόπους μέτρησης της αβεβαιότητας του μοντέλου ως προς την πρόβλεψή του.

Τα πλεονεκτήματα και τα μειονεκτήματα της δειγματοληψίας με αβεβαιότητα συνοψίζονται στον Πίνακα 3.1. Ανάλογα με το πρόβλημα, μπορεί να επιλεγεί ο κατάλληλος τρόπος αποτίμησης της αβεβαιότητας, ώστε να μεγιστοποιηθεί η απόδοση του αλγορίθμου ενεργής μάθησης.

Δειγματοληψία με βάση την αβεβαιότητα	
Πλεονεκτήματα	**Μειονεκτήματα**
Απλή στην υλοποίηση και εύκολα εφαρμόσιμη	Μπορεί να επιλέγει παραδείγματα εκτός της κατανομής των δεδομένων
Μειώνει την ανάγκη για πολλές ετικέτες	Συχνά εστιάζει σε «ακραία» δείγματα που δεν είναι τόσο αντιπροσωπευτικά
Αποδοτική σε πολλά πρακτικά προβλήματα	Ενδέχεται να μην καλύψει όλο το χώρο των δεδομένων

Πίνακας 3.1: Σύγκριση πλεονεκτημάτων και μειονεκτημάτων της μεθόδου δειγματοληψίας με βάση την αβεβαιότητα.

Επιλογή ερωτημάτων με ψηφοφορία επιτροπής (Query by Committee): Η επιλογή ερωτημάτων με ψηφοφορία επιτροπής βασίζεται στην ιδέα ότι αντί να έχουμε ένα μόνο μοντέλο που επιλέγει τα πιο αβέβαια δείγματα, μπορούμε να χρησιμοποιήσουμε μια επιτροπή από μοντέλα, τα οποία εκπαιδεύονται στο ίδιο σύνολο δεδομένων, αλλά με διαφορετικούς τρόπους. Η βασική υπόθεση είναι ότι όταν τα μοντέλα της επιτροπής

διαφωνούν έντονα για την ετικέτα ενός δείγματος, τότε αυτό το δείγμα είναι ιδιαίτερα χρήσιμο για εκμάθηση. Η επιτροπή αποτελείται συνήθως από μοντέλα που προέρχονται από διαφορετικές τυχαίες εκδοχές του ίδιου συνόλου δεδομένων εκπαίδευσης ή από διαφορετικούς αρχικούς όρους εκπαίδευσης.

Υπάρχουν διάφοροι τρόποι για να ποσοτικοποιηθεί η διαφωνία μεταξύ των μοντέλων της επιτροπής. Οι πιο συνηθισμένοι είναι:

- **Μείωση Διακύμανσης (Variance Reduction)**: Επιλογή των δειγμάτων που, όταν προστεθούν στο σύνολο εκπαίδευσης, αναμένεται να μειώσουν τη διακύμανση μεταξύ των μοντέλων της επιτροπής.

- **Ψηφοφορία Εντροπίας (Vote Entropy)**: Η αβεβαιότητα μετριέται υπολογίζοντας την κατανομή των ψήφων της επιτροπής για κάθε πιθανή ετικέτα ενός δείγματος. Αν όλα τα μοντέλα συμφωνήσουν σε μία ετικέτα, τότε η αβεβαιότητα είναι χαμηλή, ενώ αν η ψηφοφορία κατανέμεται ομοιόμορφα μεταξύ πολλών ετικετών, η αβεβαιότητα είναι υψηλή:

$$U(x) = -\sum_{y} \frac{\text{vote}_{\mathcal{C}}(y,x)}{|\mathcal{C}|} \log \frac{\text{vote}_{\mathcal{C}}(y,x)}{|\mathcal{C}|} \tag{3.4}$$

  όπου $\text{vote}_{\mathcal{C}}(y,x)$ είναι ο αριθμός των μοντέλων της επιτροπής που επέλεξαν την ετικέτα y για το δείγμα x, και $|\mathcal{C}|$ είναι το πλήθος των μοντέλων στην επιτροπή.

  Προαιρετικά, όταν κάθε μοντέλο της επιτροπής παρέχει προβλεπόμενη κατανομή πιθανοτήτων, η ήπια ψηφοφορία επιτροπής ορίζεται ως:

$$U(x) = -\sum_{y} P_{\mathcal{C}}(y|x) \log P_{\mathcal{C}}(y|x) \tag{3.5}$$

  όπου:

$$P_{\mathcal{C}}(y|x) = \frac{1}{|\mathcal{C}|} \sum_{\theta \in \mathcal{C}} P_{\theta}(y|x) \tag{3.6}$$

- **Απόκλιση Kullback-Leibler (KL Divergence)**: Σε αυτή την περίπτωση, η διαφωνία μεταξύ των μοντέλων της επιτροπής μετριέται ως ο μέσος όρος της KL απόκλισης κάθε μοντέλου από την «συνεπή» κατανομή της επιτροπής:

$$U(x) = \frac{1}{|\mathcal{C}|} \sum_{\theta \in \mathcal{C}} KL\left(P_{\theta}(Y|x) \parallel P_{\mathcal{C}}(Y|x)\right) \tag{3.7}$$

  όπου η ΚΛ απόκλιση ορίζεται ως:

$$KL\left(P_{\theta}(Y|x) \parallel P_{\mathcal{C}}(Y|x)\right) = \sum_{y} P_{\theta}(y|x) \log \frac{P_{\theta}(y|x)}{P_{\mathcal{C}}(y|x)} \tag{3.8}$$

Η μέθοδος επιλογής ερωτημάτων με βάση την ψηφοφορία επιτροπής είναι ιδιαίτερα χρήσιμη σε περιπτώσεις όπου υπάρχει μεγάλη αβεβαιότητα σχετικά με την κατανομή των δεδομένων και η ύπαρξη διαφορετικών εκτιμήσεων βοηθά στη βελτίωση της συνολικής απόδοσης του συστήματος. Συγκεκριμένες πληροφορίες για τα πλεονεκτήματα και τα μειονεκτήματα των παραπάνω πρακτικών βρίσκονται στον Πίνακα 3.2.

Επιλογή ερωτημάτων με ψηφοφορία επιτροπής	
Πλεονεκτήματα	**Μειονεκτήματα**
Αξιοποιεί τη γνώμη πολλών μοντέλων, βελτιώνοντας την επιλογή δειγμάτων	Απαιτεί την εκπαίδευση πολλαπλών μοντέλων, αυξάνοντας τον υπολογιστικό φόρτο
Αντιμετωπίζει περιπτώσεις όπου ένα μεμονωμένο μοντέλο έχει συστηματικές προκαταλήψεις	Η απόδοση εξαρτάται από τη σύνθεση και την ποικιλομορφία της επιτροπής
Χρησιμοποιείται σε διάφορες εφαρμογές μηχανικής μάθησης, ειδικά όταν υπάρχουν ετερογενείς δομές δεδομένων	Μπορεί να απαιτεί σύνθετους μηχανισμούς συνδυασμού των αποτελεσμάτων των μοντέλων

Πίνακας 3.2: Σύγκριση πλεονεκτημάτων και μειονεκτημάτων της μεθόδου επιλογής ερωτήσεων με βάση την ψηφοφορία επιτροπής

Επιλογή ερωτημάτων με βάση την Αναμενόμενη Αλλαγή Μοντέλου (Expected Model Change): Η μέθοδος επιλογής ερωτημάτων με βάση την Αναμενόμενη Αλλαγή Μοντέλουβασίζεται στην αρχή ότι τα πιο χρήσιμα δείγματα για επικετοποίηση είναι εκείνα που, όταν προστεθούν στο σύνολο εκπαίδευσης, προκαλούν τη μεγαλύτερη αλλαγή στις παραμέτρους του μοντέλου. Η ιδέα πίσω από αυτή τη μέθοδο είναι ότι αν ένα νέο δείγμα έχει τη δυνατότητα να επηρεάσει σημαντικά το μοντέλο, τότε είναι πιθανό να βελτιώσει τη συνολική του απόδοση.

Για να ποσοτικοποιηθεί η αλλαγή του μοντέλου μετά την εκπαίδευση με ένα νέο δείγμα x, υπολογίζεται η μεταβολή στις παραμέτρους θ του μοντέλου:

$$\Delta\theta = ||\theta' - \theta|| \tag{3.9}$$

Όπου θ και θ' είναι το σύνολο των παραμέτρων πριν και αντίστοιχα μετά την ενημέρωση του και $||\cdot||$ είναι κάποιος μετρικός κανόνας (συνήθως η Ευκλείδεια απόσταση ή κάποιος άλλος μετρικός χώρος).

Η στρατηγική επιλογής δείγματος βασίζεται στο να επιλέγεται το x^* που οδηγεί στη μέγιστη μεταβολή των παραμέτρων του μοντέλου:

$$x^* = \arg\max_{x \in U} E[\Delta\theta|x] \tag{3.10}$$

Όπου U το σύνολο μη επισημασμένων δεδομένων.

Υπάρχουν διάφοροι τρόποι με τους οποίους μπορεί να εκτιμηθεί η αλλαγή του μοντέλου, όπως:

- **Gradient-Based Model Change**: Υπολογισμός της αναμενόμενης αλλαγής των παραμέτρων με βάση το μέτρο της απόκλισης στη διαδικασία ενημέρωσης των βαρών. Η τεχνική αυτή είναι ιδιαίτερα χρήσιμη σε νευρωνικά δίκτυα.
- **Fisher Information Matrix-Based Model Change**: Χρήση του Fisher Information Matrix για να προσδιοριστεί η επίδραση ενός δείγματος στην εκπαίδευση του μοντέλου.

- **Influence Function Analysis**: Ανάλυση της επιρροής ενός δεδομένου στην τελική απόδοση του μοντέλου, μέσω του υπολογισμού της συνεισφοράς του στην τροποποίηση των παραμέτρων.

Στον Πίνακα 3.3 συγκεντρώνονται περιγραφικά τα πλεονεκτήματα και τα μειονεκτήματα της στρατηγικής αυτής.

Επιλογή ερωτημάτων με βάση την Αναμενόμενη Αλλαγή Μοντέλου	
Πλεονεκτήματα	**Μειονεκτήματα**
Εξασφαλίζει ότι τα επιλεγμένα δείγματα οδηγούν σε πραγματική βελτίωση του μοντέλου	Υπολογιστικά ακριβή διαδικασία, ιδιαίτερα σε μεγάλα μοντέλα
Λειτουργεί καλά για πολύπλοκα μοντέλα όπως τα νευρωνικά δίκτυα	Απαιτεί επανεκπαίδευση ή προσομοίωση της ενημέρωσης των παραμέτρων
Δίνει προτεραιότητα σε δείγματα που προκαλούν σημαντικές αλλαγές	Μπορεί να αγνοήσει δείγματα που είναι κρίσιμα για γενίκευση σε άλλες περιπτώσεις

Πίνακας 3.3: Σύγκριση πλεονεκτημάτων και μειονεκτημάτων της μεθόδου επιλογής ερωτημάτων με βάση την Αναμενόμενη Αλλαγή Μοντέλου

Επιλογή ερωτημάτων με βάση την Αναμενόμενη Μείωση Σφάλματος Expected Error Reduction): Η μέθοδος επιλογής ερωτημάτων με βάση της Αναμενόμενη Μείωση Σφάλματος επιλέγει τα δείγματα προς ετικετοποίηση με βάση τη μέγιστη αναμενόμενη μείωση του συνολικού σφάλματος πρόβλεψης μετά την ενημέρωση του μοντέλου. Η διαδικασία ξεκινά με την εκπαίδευση του μοντέλου στα ήδη διαθέσιμα δεδομένα και τον υπολογισμό του τρέχοντος σφάλματος. Στη συνέχεια, για κάθε μη ετικετοποιημένο δείγμα, γίνεται προσομοίωση της εκπαίδευσης με το συγκεκριμένο δείγμα και εκτίμηση του νέου σφάλματος του μοντέλου. Η διαφορά μεταξύ του αρχικού και του νέου σφάλματος καταγράφεται και το δείγμα που οδηγεί στη μεγαλύτερη μείωση του σφάλματος επιλέγεται για ετικετοποίηση.

$$
\begin{aligned}
x^*_{ER} = \arg\min_x E_{Y|\theta,x} \left[\sum_{x' \in \mathcal{U}} E_{Y'|\theta^+,x'} \left[y \neq \hat{y} \right] \right] \; = \\
\arg\min_x \sum_y P_\theta(y|x) \left[\sum_{x' \in \mathcal{U}} (1 - P_{\theta^+}(\hat{y}|x')) \right]
\end{aligned}
\tag{3.11}
$$

Όπου $E_{Y|\theta,x}$ είναι η αναμενόμενη τιμή ως προς τις πιθανές ετικέτες Y που προβλέπει το τρέχον μοντέλο θ για το δείγμα x, ενώ $P_{\theta^+}(\hat{y}|x')$ είναι η πιθανότητα που προβλέπει το μοντέλο μετά την ενημέρωση με το νέο δείγμα, για το x'.

Ο υπολογισμός της αναμενόμενης μείωσης του σφάλματος μπορεί να γίνει με διάφορες προσεγγίσεις:

- **Μείωση Διακύμανσης Σφάλματος (Variance Reduction)**: Επιλογή των δειγμάτων που μειώνουν περισσότερο τη διακύμανση της απόκλισης του μοντέλου.
- **Μείωση Μπευζιανού Σφάλματος (Bayesian Error Reduction)**: Υπολογισμός της αναμενόμενης επίδρασης ενός δείγματος στο σφάλμα βάσει Μπευζιανών προσεγγίσεων.
- **Εκτίμηση Σφάλματος Γενίκευσης (Generalization Error Estimation)**: Πρόβλεψη της μεταβολής του γενικού σφάλματος μέσω της επίδρασης κάθε δείγματος στα όρια απόφασης του μοντέλου.

Συγκεντρωτικά τα πλεονεκτήματα και τα μειονεκτήματα της στρατηγικής βρίσκονται στον Πίνακα 3.4.

Επιλογή ερωτημάτων με βάση την Αναμενόμενη Μείωση Σφάλματος	
Πλεονεκτήματα	**Μειονεκτήματα**
Επιλέγει δείγματα που πραγματικά μειώνουν το σφάλμα του μοντέλου	Υπολογιστικά απαιτητική μέθοδος, ιδιαίτερα για πολύπλοκα μοντέλα
Μπορεί να εφαρμοστεί σε διάφορα μοντέλα και κατανομές δεδομένων	Χρειάζεται προσεγμένες προσεγγίσεις για εκτίμηση της μείωσης σφάλματος
Περιορίζει την πιθανότητα επιλογής περιττών ή μη χρήσιμων δειγμάτων	Η ακριβής εκτίμηση της μείωσης σφάλματος μπορεί να είναι δύσκολη σε μεγάλα δατασετς

Πίνακας 3.4: Σύγκριση πλεονεκτημάτων και μειονεκτημάτων της μεθόδου επιλογής ερωτημάτων με βάση την Αναμενόμενη Μείωση Σφάλματος

Επιλογή ερωτημάτων με βάση τη Μείωση Διακύμανσης Σφάλματος (Variance Reduction): Η μέθοδος επιλογής ερωτημάτων με βάση τη Μείωση Διακύμανσης Σφάλματος επιδιώκει να μειώσει τη συνολική αβεβαιότητα του μοντέλου επιλέγοντας δείγματα που μειώνουν τη διακύμανση των προβλέψεών του. Η βασική υπόθεση αυτής της μεθόδου είναι ότι οι υψηλές διακυμάνσεις στις προβλέψεις υποδεικνύουν περιοχές του χώρου χαρακτηριστικών όπου το μοντέλο δεν είναι καλά εκπαιδευμένο, και επομένως αυτά τα δείγματα είναι πιο χρήσιμα για την εκπαίδευση. Η διακύμανση ενός μοντέλου σε σχέση με ένα δείγμα x μπορεί να υπολογιστεί ως:

$$Var[f(x)] = E[(f(x) - E[f(x)])^2] \tag{3.12}$$

όπου $f(x)$ είναι η συνάρτηση πρόβλεψης του μοντέλου. Με αυτό τον τρόπο επιλέγονται τα δείγματα x^* που μειώνουν το συνολικό μέτρο διακύμανσης, δηλαδή:

$$x^* = \arg\min_{x \in U} E[Var(f(x))] \tag{3.13}$$

Αυτή η στρατηγική είναι ιδιαίτερα χρήσιμη σε εφαρμογές όπου η σταθερότητα του μοντέλου είναι κρίσιμη, όπως στην πρόβλεψη καιρού, την οικονομική ανάλυση και τα

ιατρικά συστήματα διάγνωσης. Παρά την αποτελεσματικότητά της, απαιτεί μεγαλύτερη υπολογιστική ισχύ για την εκτίμηση της διακύμανσης και την επιλογή των πιο κρίσιμων δειγμάτων για εκμάθηση. Τα πλεονεκτήματα και τα μειονεκτήματα της παρέχονται συνοπτικά στον Πίνακα 3.5.

Πίνακας 3.5: Σύγκριση πλεονεκτημάτων και μειονεκτημάτων της μεθόδου επιλογής ερωτημάτων με βάση τη Μείωση Διακύμανσης Σφάλματος

Επιλογή ερωτημάτων με βάση τη Μείωση Διακύμανσης Σφάλματος	
Πλεονεκτήματα	**Μειονεκτήματα**
Μειώνει την αβεβαιότητα του μοντέλου σε περιοχές υψηλής διακύμανσης	Υπολογιστικά δαπανηρή διαδικασία
Παρέχει πιο σταθερές προβλέψεις	Χρειάζεται επανειλημμένες μετρήσεις για την εκτίμηση της διακύμανσης
Συμβάλλει στη δημιουργία μοντέλων ικανότερων για γενίκευση	Μπορεί να παραβλέψει χρήσιμες πληροφορίες εκτός των περιοχών υψηλής διακύμανσης

3.4 Πρακτικές παρατηρήσεις και προβλήματα

Η υλοποίηση της Ενεργής Μάθησης στην πράξη συνοδεύεται από διάφορες προκλήσεις και πρακτικά ζητήματα, τα οποία πρέπει να ληφθούν υπόψη για την αποτελεσματική εφαρμογή της. Οι παράγοντες που επηρεάζουν την επιτυχία ενός συστήματος Ενεργής Μάθησης περιλαμβάνουν τη διαχείριση των μη ετικετοποιημένων δεδομένων, την επιλογή του κατάλληλου oracle, την πολυπλοκότητα του υπολογιστικού κόστους και την αξιολόγηση της απόδοσης του συστήματος.

Ένας από τους σημαντικότερους περιορισμούς της Ενεργής Μάθησης είναι η ανάγκη ύπαρξης ενός oracle, δηλαδή μιας πηγής που παρέχει τις σωστές ετικέτες για τα δεδομένα. Σε πολλές περιπτώσεις, αυτός είναι ένας ανθρώπινος ειδικός, γεγονός που καθιστά τη διαδικασία χρονοβόρα και δαπανηρή. Επιπλέον, η ποιότητα των ετικετών μπορεί να επηρεαστεί από την υποκειμενικότητα του ετικετοποιητή, την κούραση και τα λάθη που μπορεί να προκύψουν, καθώς και από το κόστος και τον χρόνο που απαιτείται, ειδικά σε τομείς όπως η ιατρική διάγνωση. Για την αντιμετώπιση αυτών των προβλημάτων, προσεγγίσεις όπως η συναίνεση πολλών oracles ή η χρήση ημι-επιβλεπόμενων μεθόδων μπορούν να βελτιώσουν την ποιότητα των ετικετών.

Η Ενεργή Μάθηση, αν και μειώνει την ανάγκη για μεγάλο αριθμό ετικετοποιημένων δεδομένων, μπορεί να έχει υψηλές υπολογιστικές απαιτήσεις. Οι λόγοι για αυτό περιλαμβάνουν την ανάλυση μεγάλου όγκου μη ετικετοποιημένων δεδομένων για την επιλογή των πιο χρήσιμων παραδειγμάτων, την επαναλαμβανόμενη εκπαίδευση του μοντέλου κάθε φορά που προστίθενται νέα δείγματα, καθώς και την ανάγκη για πολυπλοκότερες μεθόδους επιλογής δειγμάτων, όπως Μπευζιανές προσεγγίσεις ή τεχνικές Βαθιάς Μάθησης. Για την αποδοτική εφαρμογή της Ενεργής Μάθησης σε μεγάλα σύνολα δεδομένων, χρησιμοποιούνται τεχνικές όπως η δειγματοληψία υπό περιορισμο-

ύς, η κατανομή της διαδικασίας σε παράλληλα συστήματα και η χρήση προσεγγιστικών μοντέλων για τον εντοπισμό των πιο χρήσιμων δειγμάτων.

Ένα ακόμα πρόβλημα είναι η μέτρηση της απόδοσης των συστημάτων Ενεργής Μάθησης. Δεδομένου ότι η μέθοδος επικεντρώνεται στη βελτιστοποίηση των ερωτημάτων, οι κλασικές μετρικές αξιολόγησης, όπως η ακρίβεια, μπορεί να μην αντικατοπτρίζουν πλήρως την αποδοτικότητα της προσέγγισης. Οι πιο κατάλληλες μετρικές περιλαμβάνουν καμπύλες μάθησης, που παρακολουθούν την απόδοση του μοντέλου καθώς προστίθενται νέες ετικέτες, ανάλυση κόστους και οφέλους, που αναλύει το κόστος ετικετοποίησης σε σχέση με τη βελτίωση της ακρίβειας, και ανάλυση της αποδοτικότητας των ερωτημάτων, που μετρά τον αριθμό των ερωτημάτων που απαιτούνται για να φτάσει το μοντέλο σε ένα συγκεκριμένο επίπεδο απόδοσης.

Επιπλέον, συχνό πρόβλημα στην Ενεργή Μάθηση αποτελεί η επιλογή δεδομένων που δεν είναι αντιπροσωπευτικά της συνολικής κατανομής, γεγονός που μπορεί να οδηγήσει σε προκαταλήψεις του μοντέλου. Αυτό συμβαίνει όταν η στρατηγική επιλογής δειγμάτων επικεντρώνεται υπερβολικά σε αβέβαια δείγματα και παραβλέπει σημαντικά μοτίβα στα δεδομένα. Για να αντιμετωπιστεί αυτό, χρησιμοποιούνται μέθοδοι που συνδυάζουν αβεβαιότητα και πυκνότητα δεδομένων και εφαρμόζονται προσεγγίσεις που λαμβάνουν υπόψη την ποικιλομορφία των δειγμάτων.

Συμπερασματικά, η Ενεργή Μάθηση παρέχει ισχυρές τεχνικές που μπορεί να βελτιώσουν την αποδοτικότητα της εκπαίδευσης των μοντέλων, αλλά συνοδεύεται από πλήθος πρακτικών προκλήσεων. Η επιλογή του κατάλληλου oracle, η διαχείριση του υπολογιστικού κόστους, η σωστή αξιολόγηση της απόδοσης και η αποφυγή προκαταλήψεων στα δεδομένα είναι παράγοντες που πρέπει να ληφθούν υπόψη για την επιτυχή εφαρμογή της. Παρά τις δυσκολίες, η Ενεργή Μάθηση παραμένει μια από τις πιο αποτελεσματικές προσεγγίσεις για εφαρμογές όπου η απόκτηση ετικετοποιημένων δεδομένων είναι δαπανηρή ή δύσκολη.

4. Ανάλυση Συναισθήματος

Η Ανάλυση Συναισθήματος (Sentiment Analysis) ή, εναλλακτικά, η Εξόρυξη Γνώμης (Opinion Mining) αναφέρεται στη διαδικασία αναγνώρισης, εξαγωγής και ποσοτικοποίησης των στάσεων, των απόψεων και των συναισθημάτων που εκφράζονται σε ένα κείμενο [148]. Ο κεντρικός στόχος της συγκεκριμένης προσέγγισης είναι η αυτόματη ταξινόμηση των εκφραζόμενων συναισθημάτων σε θετικά, αρνητικά ή ουδέτερα, αν και σε αρκετές περιπτώσεις εξετάζονται πιο λεπτοφυείς συναισθηματικές κατηγοριοποιήσεις (π.χ. χαρά, λύπη, θυμός) [39].

Σε πρώτο επίπεδο, η Ανάλυση Συναισθήματος αποσκοπεί στην εξαγωγή πολικότητας (polarity) από κείμενα, δηλαδή στην απλή διάκριση μεταξύ θετικού και αρνητικού συναισθήματος [205]. Εντούτοις, ανάλογα με τους στόχους της εφαρμογής, το πεδίο δύναται να επεκταθεί και σε πιο σύνθετες μορφές ανάλυσης, όπως η ανάλυση σε επίπεδο εγγράφου (document-level), πρότασης (sentence-level), ή και χαρακτηριστικού (feature-level) η οποία εστιάζει σε συγκεκριμένες οντότητες ή χαρακτηριστικά ενός παραδείγματος [170].

Η πρακτική σημασία της Ανάλυσης Συναισθήματος διαφαίνεται από τις πολλαπλές εφαρμογές της σε τομείς όπως το μάρκετινγκ, τα κοινωνικά δίκτυα, η πολιτική επιστήμη και η δημοσιογραφία. Για παράδειγμα, οι εταιρείες μπορούν να παρακολουθούν τις αντιδράσεις των καταναλωτών για τα προϊόντα τους, ενώ οι πολιτικοί αναλυτές εκτιμούν τη δημόσια γνώμη σε ευαίσθητα γεγονότα [258]. Παράλληλα, η ραγδαία αύξηση περιεχομένου που δημιουργείται από χρήστες καθιστά επιτακτική την ανάγκη για αυτοματοποιημένες μεθόδους επεξεργασίας κειμένου μεγάλης κλίμακας, σε συνδυασμό με τεχνικές Επεξεργασίας Φυσικής Γλώσσας. [213].

Η ιδέα της αυτόματης κατανόησης και αξιολόγησης του ύφους και του συναισθήματος ενός κειμένου έχει τις ρίζες της σε πρώιμες έρευνες στον χώρο της Τεχνητής Νοημοσύνης. Ωστόσο, το ενδιαφέρον για τη Ανάλυση Συναισθήματος ενισχύθηκε σημαντικά στις αρχές της δεκαετίας του 2000, με τις πρώτες συστηματικές μελέτες να εστιάζουν κυρίως στην ταξινόμηση κριτικών προϊόντων και ταινιών [206].

Στα μέσα της δεκαετίας του 2000, η ανάγκη αυτοματοποίησης της αποτίμησης κριτικών (π.χ. σε ηλεκτρονικά καταστήματα και πλατφόρμες όπως το IMDB) κατέστησε τα κλασικά μοντέλα Μηχανικής Μάθησης, όπως τον Αφελή Μπευζιανό Ταξινομητή και τις Μηχανές Διανυσματικής Υποστήριξης, τους πρώτους πρωταγωνιστές στον χώρο. Την ίδια περίοδο αναπτύχθηκαν τα πρώτα λεξικά συναισθήματος (sentiment lexicons), τα οποία παρείχαν λίστες λέξεων και σχετικών πολικοτήτων [122]. Μετά το 2005, οι ερευνητές εστίασαν σε πιο σύνθετες πτυχές, όπως η απόχρωση (intensity) του συναισθήματος και η υποκειμενικότητα (subjectivity) [205]. Παράλληλα, η ραγδαία άνοδος των κοινωνικών δικτύων (Facebook, Twitter κ.λπ.) εισήγαγε σύντομα και συχνά ατελή κείμενα (short text), τα οποία περιείχαν emoticons ή ιδιωματισμούς [23].

Η ανάπτυξη των μεθόδων Βαθιάς Μάθησης, όπως τα Convolutional Neural Networks (CNNs) και τα Long Short-Term Memory (LSTM) Recurrent Neural Networks, οδήγησε σε νέα άλματα στην Ανάλυση Συναισθήματος [267]. Η εισαγωγή προεκπαιδευ-

μένων μοντέλων, όπως το BERT [66], επέτρεψε την αξιοποίηση τεράστιων συλλογών κειμένων για γενικευμένες γλωσσικές αναπαραστάσεις (embeddings). Παράλληλα, εμφανίστηκαν και υβριδικές προσεγγίσεις, οι οποίες συνδυάζουν λεξικά ή κανόνες με μοντέλα Μηχανικής Μάθησης [62].

Σε πιο πρόσφατη βιβλιογραφία παρατηρείται στροφή προς την αναγνώριση φαινομένων όπως ο σαρκασμός και η ειρωνεία [96], καθώς και προς την πολυτροπική ανάλυση (multimodal sentiment analysis), όπου συνδυάζεται κείμενο, ήχος και εικόνα [213]. Επιπλέον, η ανάγκη για επεξεργασία μεγάλου όγκου δεδομένων σε πραγματικό χρόνο ενισχύει το ερευνητικό ενδιαφέρον προς συστήματα Ανάλυσης Συναισθήματος με βελτιστοποιημένη απόδοση [258].

4.1 Είδη και μέθοδοι Ανάλυσης Συναισθήματος

Η Ανάλυση Συναισθήματος βασίζεται σε ένα σύνολο από θεμελιώδεις έννοιες που αφορούν στην κατανόηση της γλωσσικής έκφρασης συναισθημάτων και απόψεων μέσα σε κείμενα. Μία από τις κεντρικές διακρίσεις είναι αυτή ανάμεσα στο sentiment (π.χ. θετικό, αρνητικό ή ουδέτερο) και στο emotion (π.χ. χαρά, λύπη, θυμός), με το πρώτο να επικεντρώνεται κυρίως στη συνολική στάση του συντάκτη, ενώ το δεύτερο αναφέρεται σε επιμέρους συναισθηματικές κατηγοριοποιήσεις [39]. Περαιτέρω, γίνεται διαχωρισμός ανάμεσα στην υποκειμενικότητα (subjectivity) και την αντικειμενικότητα (objectivity), καθώς συχνά ένα κείμενο περιλαμβάνει τόσο γνώμες όσο και πραγματολογικές πληροφορίες [205]. Στο πλαίσιο αυτό, η πολυσημία (ambiguity) και η σημασία του συμφραζομένου (context) αποτελούν ιδιαίτερες προκλήσεις, δεδομένου ότι η ίδια λέξη μπορεί να φέρει διαφορετικό συναισθηματικό φορτίο ανάλογα με τη χρήση της μέσα στην πρόταση.

Η Ανάλυση Συναισθήματος χωρίζεται σε τρία κυρίως επίπεδα ανάλυσης όπως είδαμε παραπάνω[148]:

- **Επίπεδο κειμένου (Document-level)**: Η αποτίμηση του συναισθήματος γίνεται στο σύνολο ενός κειμένου (π.χ. μιας κριτικής). Η κεντρική ιδέα είναι να εξαχθεί η συνολική στάση που εκφράζεται.

- **Επίπεδο πρότασης (Sentence-level)**: Η πολικότητα προσδιορίζεται σε επίπεδο πρότασης, δίνοντας μεγαλύτερη λεπτομέρεια στην ανάλυση.

- **Επίπεδο Χαρακτηριστικού (Aspect-level ή feature-level)**: Εστιάζει σε συγκεκριμένα χαρακτηριστικά ενός αντικειμένου ή συμβάντος (π.χ. ποιότητα ήχου σε μια κριτική κινητού), ώστε να διαχωριστεί το συναισθηματικό φορτίο που αφορά μόνο σε εκείνα τα χαρακτηριστικά.

Η αξιολόγηση των αλγορίθμων Ανάλυσης Συναισθήματος συνήθως γίνεται με τις κλασικές μετρικές (Accuracy, Precision, Recall, F1-score) κ.λπ. [245], οι οποίες έχουν ήδη παρουσιαστεί αναλυτικά στο προηγούμενο κεφάλαιο. Επομένως, εδώ αρκεί να επισημανθεί ότι αποτελούν τις βασικές μεθόδους για την αποτίμηση της απόδοσης στην ταξινόμηση συναισθήματος.

Η Ανάλυση Συναισθήματος έχει εξελιχθεί σε ένα πολυδιάστατο πεδίο, όπου εφαρμόζονται τόσο παραδοσιακές μέθοδοι Μηχανικής Μάθησης όσο και σύγχρονες τεχνικές Βαθιάς Μάθησης. Στις επόμενες υποενότητες παρουσιάζονται οι κυριότερες

κατηγορίες μεθόδων, συμπεριλαμβανομένων και των εργαλείων που χρησιμοποιήθηκαν στις παρούσες εφαρμογές, όπως VADER, TextBlob, FinBERT και λοιποί Transformers.

4.2 Μέθοδοι βασισμένες σε λεξικά (Lexicon-based Methods)

Στις μεθόδους βασιζμένες σε λεξικά lexicon-based, η βασική υπόθεση είναι ότι κάθε λέξη ή φράση φέρει από τη φύση της ένα συγκεκριμένο συναισθηματικό φορτίο, το οποίο μπορεί να χαρακτηριστεί ως θετικό, αρνητικό ή ουδέτερο [122, 170]. Ένα λεξικό (lexicon) στο πλαίσιο της Ανάλυσης Συναισθήματος, αποτελεί ουσιαστικά μία συλλογή τέτοιων λέξεων/εκφράσεων, όπου κάθε στοιχείο έχει συσχετιστεί με έναν βαθμό πολικότητας (polarity score). Για παράδειγμα, λέξεις όπως *«θαυμάσιος»*, *«υπέροχος»* κ.λπ. λαμβάνουν υψηλό θετικό σκορ, ενώ λέξεις όπως *«κακός»*, *«φρικτός»* φέρουν υψηλό αρνητικό σκορ. Η ύπαρξη ενός τέτοιου λεξικού επιτρέπει στο σύστημα Ανάλυσης Συναισθήματος να εκτιμά, σε πρώτο επίπεδο, το συνολικό συναίσθημα ενός κειμένου μετρώντας τη συχνότητα ή τη βαρύτητα (π.χ. με σταθμισμένους συντελεστές) των λέξεων αυτών [205].

Ωστόσο, το κατά πόσο ένα λεξικό καλύπτει επαρκώς τα γλωσσικά φαινόμενα ενός κειμένου εξαρτάται από πολλούς παράγοντες, όπως η πολυσημία των λέξεων, η ύπαρξη ειρωνείας και σαρκασμού, καθώς και η πιθανή εξειδίκευση του λεξιλογίου σε κάποιο θεματικό πεδίο [148]. Επομένως, η κατασκευή ή επιλογή κατάλληλου λεξικού θεωρείται κρίσιμη για την απόδοση των lexicon-based συστημάτων Ανάλυσης Συναισθήματος. Στις επόμενες παραγράφους παρουσιάζονται οι δύο κύριες προσεγγίσεις για τη δημιουργία και τη συντήρηση τέτοιων λεξικών: Dictionary-based Method και η Corpus-based Method.

4.2.1 Dictionary-based Method:

Οι dictionary-based μέθοδοι ξεκινούν συνήθως με μια αρχική λίστα λέξεων (seed list), στην οποία κάθε λήμμα συνοδεύεται από μια δηλωμένη πολικότητα (π.χ. «θετικό», «αρνητικό»). Στη συνέχεια, η λίστα αυτή εμπλουτίζεται χρησιμοποιώντας λεξικά αναφοράς, όπως το WordNet ή το SentiWordNet, όπου αναζητούνται συνώνυμα και αντώνυμα [122]. Η βασική παραδοχή είναι ότι τα συνώνυμα διατηρούν την ίδια πολικότητα, ενώ τα αντώνυμα συνήθως έχουν αντίθετη πολικότητα. Για παράδειγμα, αν γνωρίζουμε ότι η λέξη «όμορφος» έχει θετική πολικότητα, τότε η λέξη «εντυπωσιακός» (ως συνώνυμο) κληρονομεί την ίδια θετική πολικότητα. Αντίθετα, η λέξη «άσχημος» (ως αντώνυμο) θα λάβει αρνητική πολικότητα.

Στην πρακτική εφαρμογή, οι dictionary-based μέθοδοι συχνά ενσωματώνουν επιπλέον κανόνες για την αντιμετώπιση ενισχυτών, όπως «πολύ», «εξαιρετικά»), αρνητών, όπως «μη», «δεν») ή συνδέσμων που ανατρέπουν το εκφραζόμενο συναίσθημα σε μια πρόταση [170]. Έτσι, από έναν απλό κατάλογο λέξεων καταλήγουμε σε ένα πιο σύνθετο σύστημα που λαμβάνει υπόψη του διάφορες συντακτικές και σημασιολογικές πτυχές του κειμένου.

Επιπρόσθετα, υπάρχουν εξειδικευμένα λεξικά για συγκεκριμένους τομείς (domain-specific lexicons), τα οποία δημιουργούνται μέσω της ίδιας λογικής. Για παράδειγμα, όροι που θεωρούνται θετικοί σε ένα γενικό λεξικό, μπορεί να είναι ουδέτεροι ή αρνητι-

κοί στο πεδίο της ιατρικής ή των χρηματοοικονομικών. Αυτή η ανάγκη προσαρμογής καθιστά σημαντική την περιοδική ανανέωση του λεξικού, ώστε να αντικατοπτρίζει τυχόν αλλαγές στη γλώσσα ή την εμφάνιση νεολογισμών [205].

4.2.2 Corpus-based Method:

Η corpus-based προσέγγιση διαφοροποιείται από τη dictionary-based κυρίως ως προς τον τρόπο δημιουργίας του λεξικού. Αντί να στηριζόμαστε αποκλειστικά σε ένα λεξικό γενικής χρήσης, συλλέγουμε μεγάλο σύνολο κειμένων από το πεδίο ενδιαφέροντος και κάνουμε στατιστική ανάλυση των συσχετίσεων μεταξύ λέξεων. Η διαδικασία συνήθως ακολουθεί τα εξής διαδικασία: Αρχικά, ξεκινάμε με μια μικρή ομάδα λέξεων γνωστής πολικότητας (θετικής ή αρνητικής). Για κάθε υποψήφια λέξη, ελέγχουμε πόσο συχνά και με ποιες λέξεις συνυπάρχει στο corpus. Με βάση στατιστικά κριτήρια, καθορίζουμε αν η λέξη τείνει να συναναστρέφεται συχνότερα με θετικούς ή αρνητικούς όρους. Αν ανιχνεύεται ισχυρή συσχέτιση με θετικές λέξεις, τότε η υπό εξέταση λέξη χαρακτηρίζεται θετική, και αντιστρόφως.

Μέσω της corpus-based μεθόδου, το λεξικό που προκύπτει είναι προσαρμοσμένο στη συγκεκριμένη θεματική ενότητα, καταγράφοντας το πραγματικό γλωσσικό περιβάλλον των όρων [148]. Για παράδειγμα, η λέξη «κέρδος» σε γενικό λεξικό μπορεί να έχει θετική χροιά, αλλά σε ένα σύνολο κειμένων που αναφέρεται σε περιβαλλοντικές επιπτώσεις θα μπορούσε να κριθεί ουδέτερη ή ακόμα και αρνητική, αν το οικονομικό κέρδος συσχετίζεται συστηματικά με αρνητικές συνέπειες. Επομένως, η corpus-based μέθοδος επιτρέπει μεγαλύτερη ευελιξία και καλύτερη προσαρμογή, ιδίως σε κλάδους όπου οι όροι έχουν εξειδικευμένη σημασιολογία.

Παρά τις δυνατότητές της, η επιτυχία της corpus-based προσέγγισης εξαρτάται από την ποιότητα και το μέγεθος του διαθέσιμου corpus, καθώς απαιτείται ικανός όγκος δεδομένων προκειμένου να προκύψουν στατιστικά αξιόπιστες συσχετίσεις. Επιπλέον, η παρουσία πολυσημικών λέξεων (π.χ. όρος που μπορεί να είναι θετικός σε ένα συμφραζόμενο, μα αρνητικός σε άλλο) αποτελεί διαρκή πρόκληση, καθιστώντας αναγκαία την ανάπτυξη πιο σύνθετων αλγορίθμων που λαμβάνουν υπόψη το γλωσσικό πλαίσιο [205].

Σε γενικές γραμμές, οι μέθοδοι που βασίζονται σε λεξικά διακρίνονται για την ερμηνευσιμότητά τους και την απλή λογική λειτουργίας. Η διαφοροποίηση μεταξύ dictionary-based και corpus-based έγκειται κυρίως στο πώς εμπλουτίζεται το λεξικό: είτε μέσω συνωνυμιών/αντωνυμιών σε έτοιμα λεξικά, είτε μέσα από στατιστικές αναλύσεις ενός συγκεκριμένου συνόλου κειμένων. Αμφότερες οι προσεγγίσεις μπορούν να προσφέρουν σημαντικά πλεονεκτήματα στην Ανάλυση Συναισθήματος, αρκεί να χρησιμοποιούνται σε συνάρτηση με το πεδίο εφαρμογής, το διαθέσιμο λεξιλόγιο και το μέγεθος των δεδομένων [170, 148].

4.2.3 TextBlob και VADER

TextBlob: Η TextBlob είναι μία βιβλιοθήκη γραμμένη σε Python, η οποία προσφέρει ένα φιλικό προς τον χρήστη περιβάλλον για βασικές λειτουργίες Επεξεργασίας Φυσικής Γλώσσας, όπως ταξινόμηση κειμένου, διόρθωση ορθογραφίας, και φυσικά, Ανάλυση Συναισθήματος [152]. Για την Ανάλυση Συναισθήματος, αξιοποιεί εσωτερικά ένα λεξικό πολικότητας σε συνδυασμό με κανόνες, ώστε να εντοπίσει λέξεις με θετική, αρνητική ή ουδέτερη φόρτιση. Παράλληλα, υποστηρίζει και πιο σύνθετες τεχνικές Μη-

χανικής Μάθησης, επιτρέποντας την εκπαίδευση ταξινομητών πάνω σε επισημασμένα σύνολα δεδομένων, γεγονός που ενισχύει την ευελιξία της. Έτσι, ο χρήστης μπορεί είτε να βασιστεί στις διαθέσιμες συναρτήσεις που παρέχει η βιβλιοθήκη, είτε να δημιουργήσει το δικό του μοντέλο ταξινόμησης.

Το βασικό πλεονέκτημα της TextBlob έγκειται στο ότι προσφέρει ένα API, με περιορισμένες απαιτήσεις σε γνώση Επεξεργασίας Φυσικής Γλώσσας. Επιπλέον, σε περιπτώσεις όπου το προεπιλεγμένο λεξικό πολικότητας δεν ανταποκρίνεται σε εξειδικευμένη ορολογία (π.χ. ιατρικά ή οικονομικά κείμενα), ο χρήστης μπορεί να προσαρμόσει χειροκίνητα τη λίστα λέξεων. Ωστόσο, όπως συμβαίνει με τις περισσότερες lexicon-based μεθόδους, η TextBlob μπορεί να αντιμετωπίσει δυσκολίες σε κείμενα με εκτεταμένη χρήση ιδιωματισμών, σαρκασμού ή νεολογισμών, απαιτώντας συχνά κάποια επιπλέον ρύθμιση ή βελτιστοποίηση [27].

VADER: Η μέθοδος VADER (Valence Aware Dictionary and Sentiment Reasoner) είναι ένα εξειδικευμένο εργαλείο Ανάλυσης Συναισθήματος, σχεδιασμένο κυρίως για σύντομα κείμενα και αναρτήσεις σε κοινωνικά δίκτυα [124]. Διαθέτει ένα εκτεταμένο, χειροκίνητα επιμελημένο λεξικό που βαθμονομεί λέξεις, ακρωνύμια, συντομογραφίες, καθώς και emoticons βάσει της συναισθηματικής τους φόρτισης. Η καινοτομία της μεθόδου έγκειται στην ικανότητά του να λαμβάνει υπόψη συμφραζόμενα χαρακτηριστικά όπως τα κεφαλαία γράμματα, τα σημεία στίξης (π.χ. πολλαπλά «!» ή «;»), τους ενισχυτές και τους αρνητές, επιτρέποντας έτσι ακριβέστερη βαθμολόγηση της πολικότητας.

Όταν εφαρμόζεται σε πλατφόρμες όπως το Twitter ή το Reddit, η VADER επιτυγχάνει γενικά υψηλά επίπεδα ακρίβειας, καθώς έχει σχεδιαστεί με γνώμονα τα χαρακτηριστικά της ανεπίσημης, συχνά ιδιωματικής, γλώσσας που συναντάται στα κοινωνικά δίκτυα. Παρότι κατατάσσεται στις lexicon-based μεθόδους, χρησιμοποιεί έναν συνδυασμό κανόνων (heuristics) και σταθμισμένων σκορ που αναπτύχθηκαν εμπειρικά από ομάδα ανθρώπων-αξιολογητών. Επομένως, διατηρεί την ερμηνευσιμότητα των παραδοσιακών λεξικών, ενώ παράλληλα βελτιώνει την ευστοχία του χάρη στους προκαθορισμένους κανόνες χειρισμού της γλώσσας των μέσων κοινωνικής δικτύωσης.

4.3 Μέθοδοι από το χώρο της Μηχανικής Μάθησης

Οι μέθοδοι Μηχανικής Μάθησης βασίζονται συνήθως στην επιβλεπόμενη μάθηση, όπου τα δείγματα ένος συνόλου δεδομένων (π.χ. κριτικές ταινιών, προϊόντων) έχει ήδη επισημανθεί ως θετικά, αρνητικά ή ουδέτερα. Έπειτα, εκπαιδεύονται κλασικοί ταξινομητές, όπως ο Απλός Μπευζιανός Ταξινομητής ή οι Μηχανές Διανυσματικής Υποστήριξης, οι οποίοι μαθαίνουν να συσχετίζουν συγκεκριμένα χαρακτηριστικά του κειμένου με αντίστοιχες ετικέτες πολικότητας [206].

Σημαντικό βήμα σε αυτές τις προσεγγίσεις αποτελεί η επεξεργασία των χαρακτηριστικών του συνόλου δεδομένων, που συχνά περιλαμβάνει την εξαγωγή n-grams, τον υπολογισμό συχνοτήτων εμφάνισης (μέθοδος TF-IDF), τη χρήση γλωσσικών αναπαραστάσεων όπως η bag-of-words ή και πιο σύνθετων γλωσσολογικών χαρακτηριστικών, όπως μορφολογικών ετικετών ή συντακτικών συναρτήσεων). Το βασικό μειονέκτημα της συγκεκριμένης κατηγορίας είναι ότι απαιτεί συχνά μεγάλο όγκο ετικετοποιημένων δεδομένων, ενώ η απόδοση μπορεί να υποβαθμιστεί σημαντικά σε διαφορετικά πεδία.

4.4 Μέθοδοι από το χωρο της Βαθιάς Μάθησης

Με την πρόοδο λοιπόν της Τεχνητής Νοημοσύνης, οι μέθοδοι Ανάλυσης Συναισθήματος έχουν εξελιχθεί από απλές μεθόδους βασισμένες σε κανόνες και παραδοσιακές τεχνικές Μηχανικής Μάθησης σε πιο προηγμένες προσεγγίσεις, όπως τα Συνελικτικά Νευρωνικά Δίκτυα (Convolutional Neural Networks - CNNs) και τα Αναδρομικά Νευρωνικά Δίκτυα (Recurrent Neural Networks - RNNs) (π.χ. LSTM, GRU) [267] και τελευταία τους Transformers. Οι Transformers [280] έχουν μεταμορφώσει το πεδίο της Επεξεργασίας Φυσικής Γλώσσας, εισάγοντας μια αρχιτεκτονική που βασίζεται σε συγκεκριμένους μηχανισμούς αυτοπροσοχής (self-attention), εξαλείφοντας την ανάγκη για επαναλαμβανόμενες ή συνελικτικές δομές.

Η αυτοπροσοχή εν ολίγοις επιτρέπει στο μοντέλο να συσχετίζει κάθε λέξη με άλλες λέξεις της ίδιας ακολουθίας, ανεξαρτήτως απόστασης μεταξύ τους, γεγονός που καθιστά τους Transformers ιδιαίτερα αποτελεσματικούς στην κατανόηση συμφραζομένων. Η βασική δομή ενός Transformer αποτελείται από διαδοχικά μπλοκς κωδικοποιήσης και αποκωδικοποίησης. Στην περίπτωση της ανάλυσης συναισθήματος, συχνά χρησιμοποιούνται μόνο τα τμήματα κωδικοποίησης, όπως το BERT (Bidirectional Encoder Representations from Transformers), που αξιοποιούν τη διπλή κατεύθυνση της πληροφορίας για πληρέστερη κατανόηση του κειμένου.

Η χρήση των Transformers στην ανάλυση συναισθήματος επιτρέπει την ακριβή εξαγωγή συναισθημάτων από κείμενα, ξεπερνώντας προγενέστερες μεθόδους όπως τα Recurrent Neural Networks (RNNs) και τα Long Short-Term Memory (LSTM) δίκτυα. Τα εν λόγω σχήματα είναι επί της ουσίας προεκπαιδευμένες δομές οι οποίες δύνανται να χρησιμοποιηθούν σε μια σειρά από επιμέρους προβλήματα - όπως εκείνα της Ανάλυσης Συναισθήματος. Ένα από τα πιο διαδεδομένα προεκπαιδευμένα μοντέλα για την ανάλυση συναισθήματος είναι το BERT. Το BERT επιτρέπει τη διπλής κατεύθυνσης κατανόηση των συμφραζομένων και έχει αποδειχθεί εξαιρετικά αποδοτικό στην Ανάλυση Συναισθήματος. Προτάθηκε από την Google AI το 2018 [67] και αποτελεί ένα από τα πιο επιτυχημένα προεκπαιδευμένα μοντέλα Επεξεργασίας Φυσικής Γλώσσας. Το κύριο πλεονέκτημά του είναι η ικανότητά του να κατανοεί το πλαίσιο μιας λέξης μέσω αμφίδρομης εκπαίδευσης, σε αντίθεση με παλαιότερα μοντέλα που διάβαζαν το κείμενο μόνο προς μία κατεύθυνση.

Σε αδρές γραμμές, η δομή του αποτελείται αποκλειστικά από το κομμάτι των κωδικοποιητών ενός Transformer, με πολλαπλά επίπεδα αυτοπροσοχής. Οι βασικές εκδόσεις του είναι:

- **BERT Base:** Περιλαμβάνει 12 encoder layers, 768 hidden units, 12 attention heads και 110 εκατομμύρια παραμέτρους.
- **BERT Large:** Περιλαμβάνει 24 encoder layers, 1024 hidden units, 16 attention heads και 340 εκατομμύρια παραμέτρους.

Για την διαδικασία της προεκπαίδευση χρησιμοποιεί δύο βασικές τεχνικές:

- Masked Language Model (MLM): Ένα ποσοστό των λέξεων στο εισαγόμενο κείμενο «κρύβεται», και το μοντέλο εκπαιδεύεται να «ανακαλύψει» τις κρυμμένες λέξεις.

- Next Sentence Prediction (NSP): Το μοντέλο λαμβάνει δύο προτάσεις και πρέπει να προσδιορίσει εάν η δεύτερη πρόταση έπεται λογικά της πρώτης.

Σε αυτή τη βάση, στο πλαίσιο της Ανάλυσης Συναισθήματος, το BERT μπορεί να χρησιμοποιηθεί είτε ως εργαλείο εξαγωγής χαρακτηριστικών (feature extractor) είτε μέσω λεπτομερούς προσαρμογής (fine-tuning) σε συγκεκριμένα δεδομένα.

Παράδειγμα εξειδικευμένης βαριάντας του BERT είναι το FinBERT, το οποίο έχει προεκπαιδευτεί σε δεδομένα χρηματοοικονομικού περιεχομένου. FinBERT προσαρμόζει τη βασική λογική του BERT σε κείμενα όπως εταιρικές αναφορές, ειδησεογραφικά άρθρα και άλλα έγγραφα του χρηματοπιστωτικού κλάδου, επιτυγχάνοντας έτσι υψηλή ακρίβεια στην ανάλυση πολικότητας για οικοομικές ειδήσεις.

Περισσότερα όμως για τους τρανσφορμερς και το BEPT, ιδιατέρως επί της πρακτικής εφαρμογής τους, θα δούμε στο Κεφάλαιο ;;.

4.5 Υβριδικές Προσεγγίσεις

Οι λεγόμενες υβριδικές μέθοδοι (hybrid approaches) συνδυάζουν τα πλεονεκτήματα των λεξικών συναισθήματος και των αλγορίθμων Μηχανικής και Βαθιάς Μάθησης. Για παράδειγμα, ένα υβριδικό σύστημα μπορεί αρχικά να βασίζεται σε ένα λεξικό για τον εντοπισμό πιθανών ενδείξεων πολικότητας, αξιοποιώντας ωστόσο στην επόμενη φάση ένα νευρωνικό δίκτυο για την τελική ταξινόμηση, λαμβάνοντας υπόψη τα συμφραζόμενα ή ειδικά γλωσσικά φαινόμενα (π.χ. ειρωνεία, εξειδικευμένη ορολογία) [62].

Η πρακτική εφαρμογή των παραπάνω μεθόδων και εργαλείων εξαρτάται από το είδος των δεδομένων, τη διαθεσιμότητα ετικετοποιημένων συνόλων και τους περιορισμούς σε χρόνο επεξεργασίας. Οι σύγχρονες τάσεις εστιάζουν ιδιαίτερα σε προεκπαιδευμένα μοντέλα όπως το BERT και τα παράγωγά του (FinBERT, SciBERT κ.λ.π.), αλλά και στη χρήση πολυτροπικών δεδομένων (multimodal data), όπου το συναίσθημα εξάγεται από συνδυασμό κειμένου, ήχου και εικόνας [213].

Μέρος ΙΙ

Πειραματικά Αποτελέσματα σε Πρακτικές εφαρμογές

5. Πρόβλεψη χρονοσειρών με χρήση ανάλυσης συναισθήματος

5.1 Βασικές έννοιες χρονοσειρών και σύντομη ιστορική αναδρομή

Η ιστορία της ανάγκης πρόβλεψης γεγονότων που επηρεάζουν τις ανθρώπινες αποφάσεις και εν γένει την ανθρώπινη δραστηριότητα, έχει τις ρίζες της στην αρχαιότητα. Την εποχή εκείνη, η ικανότητα πρόβλεψης του μέλλοντος θεωρούνταν άλλες φορές σημάδι θεϊκής φώτισης και άλλες κατοχής ανίερων δυνάμεων. Η ανώτερη ιέρεια του Απόλλωνα στο μαντείο των Δελφών, η Πυθία, αποτελεί ένα από τα πιο γνωστά παραδείγματα αρχαίων προφητών και πολυάριθμες αποφάσεις κατά την κλασική περίοδο συνδέονται με τους – συνήθως αμφιλεγόμενους - χρησμούς της. Αν και ο κατάλογος των φωτισμένων ιερέων, προφητών και μάντεων μέσα στους αιώνες είναι μακρύς, η επιστημονική επανάσταση είναι αυτή που μάλλον τελικά έφερε τις προβλέψεις ξανά στο προσκήνιο, και αυτή τη φορά εκδεδυμένες το μανδύα του μυστικισμού και του μυστηρίου.

Στις μέρες μας, οι προβλέψεις δεν βασίζονται στις βουλήσεις πανίσχυρων θεών ή στις κινήσεις των άστρων στο στερέωμα. Ο κλάδος της Μηχανικής Μάθησης που ασχολείται με την πρόβλεψη, έχει προσφέρει πληθώρα μεθόδων που μπορούν να βρουν χρήση σε ένα μεγάλο εύρος προβλημάτων. Μπορεί το μέσο να άλλαξε ριζικά, όμως ο στόχος παραμένει ο ίδιος: αν μπορούμε να προβλέψουμε πώς θα εξελιχθούν διάφορες παράμετροι ενός φαινομένου, τότε μπορούμε να λάβουμε καλύτερες αποφάσεις γι' αυτό, μειώνοντας πιθανούς κινδύνους. Συνεπώς, ο τομέας της πρόβλεψης χρονοσειρών γνωρίζει ταχεία ανάπτυξη τα τελευταία χρόνια, καθώς η ανάγκη για ακριβή μοντέλα πρόβλεψης αυξάνεται συνεχώς.

Η ανάγκη αυτή, και η βαθιά συνειδητοποίηση του πως οι ακριβείς προβλέψεις έχουν άμεσο επακόλουθο ορθότερες αποφάσεις, οδήγησε στην άνθηση του πεδίου της πρόβλεψης χρονοσειρών υπό το πρίσμα αρχικά της Στατιστικής και πλέον της Μηχανικής Μάθησης. Τα πρώτα θεμέλια μπήκαν περίπου 200 χρόνια πριν – όταν ο Adrien-Marie Legendre εισήγαγε τη Μέθοδο των Ελάχιστων Τετραγώνων [144], που υπήρξε θεμελιώδης για τη στατιστική πρόβλεψη. Το Αυτοπαλίνδρομο Μοντέλο για ανάλυση χρονοσειρών προτάθηκε περίπου 100 χρόνια αργότερα από τον Yule [305], ενώ την ίδια δεκαετία αναπτύχθηκε η θεωρία του Andrei Kolmogorov πάνω στις στοχαστικές διαδικασίες [138]. Λίγο πριν το 1950 παρουσιάζονται και οι πρώτες τεχνικές για το φιλτράρισμα των χρονοσειρών, από τους Wiener και Hopf [292]. Το 1950 αποτέλεσε σταθμό για την επιστήμη των υπολογιστών, καθώς ο Alan Turing δημοσίευσε το άρθρο του για την Τεχνητή Νοημοσύνη και τις «σκεπτόμενες μηχανές» [277]. Το 1957 ο Frank Rosenblatt ανέπτυξε το Perceptron, ένα από τα πρώτα μοντέλα νευρωνικών δικτύων [229], και τρία χρόνια αργότερα τα μοντέλα ARIMA αναπτύχθηκαν από τους

Box & Jenkins [31]. Μια δεκαετία αργότερα εμφανίζονται τα πρώτα Κρυφά Μοντέλα Markov (HMM) για πρόβλεψη σειρών δεδομένων [26], ενώ το 1980 παρουσιάστηκε ο γνωστός αλγόριθμος Backpropagation, θεμελιώδης για τα σύγχρονα νευρωνικά δίκτυα [59].

Μπορεί οι δεκαετίες αυτές να φαίνονται ήδη πλούσιες σε ανακαλύψεις στον τομέα των χρονοσειρών, αλλά δε θα ήταν υπερβολή να πούμε ότι οι μέθοδοι αυτές μοιάζουν πλέον ξεπερασμένες. Το διάστημα 1990 – 2010 υπήρξε παραγωγικότατο, καθώς πλέον η πρόβλεψη χρονοσειρών αλλάζει σκοπιά προσέγγισης και πλέον η Μηχανική Μάθηση κυριαρχεί. Το 1994 προτείνονται τα Long Short-Term Memory (LSTM) δίκτυα, ειδικά για χρονοσειρές, από τον Hochreiter [119], ενώ έως το 2010 χρησιμοποιείται πληθώρα μεθόδων Μηχανικής Μάθησης όπως οι Μηχανές Διανυσματικές Υποστήριξης [53], τα Μπευζιανά Δίκτυα [208] και οι Γκαουσιανές Διεργασίες [223] για την πρόβλεψη χρονοσειρών. Από το 2010 και έπειτα, οι εξελίξεις είναι πραγματικά καταιγιστικές, με πολλά δημοφιλή μοντέλα να αναπτύσσονται αυτή τη δεκαετία. Η χρήση μοντέλων Βαθιάς Μάθησης, όπως τα Long Short-Term Memory (LSTM) και τα Gated Recurrent Units (GRU), επέτρεψε την αποδοτικότερη ανάλυση μεγάλων και πολύπλοκων δεδομένων. Το 2017, η ανάπτυξη του Transformer από την Google έφερε επανάσταση στην επεξεργασία ακολουθιακών δεδομένων, επιτρέποντας μεγαλύτερη παράλληλη επεξεργασία και βελτιωμένη πρόβλεψη χρονοσειρών [280]. Τα τελευταία χρόνια, εργαλεία όπως το Facebook Prophet [270], το DeepAR της Amazon [238], το N-BEATS [198] και το NeuralProphet [159] βελτίωσαν την ακρίβεια και την ευκολία εφαρμογής της πρόβλεψης χρονοσειρών. Παράλληλα, οι Graph Neural Networks (GNNs) [296] και τα Self-Supervised Learning μοντέλα [130] ανοίγουν νέους δρόμους, ενσωματώνοντας πληροφορίες από αλληλοσχετιζόμενα δεδομένα. Σήμερα, η έμφαση δίνεται στη χρήση Generative AI [100], Explainable AI [105] και προσαρμοστικών τεχνικών για δυναμικά μεταβαλλόμενες χρονοσειρές, επιτρέποντας πιο αξιόπιστες και ευέλικτες προβλέψεις σε τομείς όπως τα χρηματοοικονομικά, η κλιματική αλλαγή και η βιομηχανική παραγωγή.

Είναι αναμφισβήτητο πως τα δεδομένα αποτελούν τον ακρογωνιαίο λίθο σε κάθε διαδικασία που εμπλέκεται η Μηχανική Μάθηση. Πριν συνεχίσουμε λοιπόν στο κυρίως μέρος αυτού του κεφαλαίου, είναι απαραίτητο να περιγράψουμε τη μορφή των δεδομένων στα προβλήματα πρόβλεψης χρονοσειρών, ή, με άλλα λόγια, να ορίσουμε τι είναι μια χρονοσειρά, καθώς και να αναφέρουμε ορισμένες βασικές έννοιες που σχετίζονται με αυτήν.

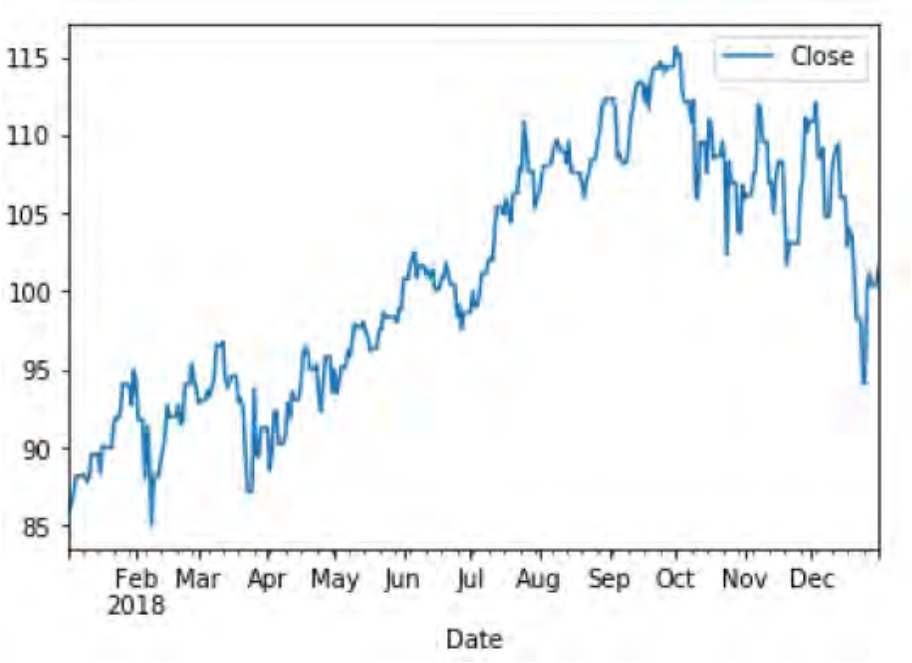
Μια χρονοσειρά είναι μια ακολουθία από παρατηρήσεις που λαμβάνονται σε διαδοχικές χρονικές στιγμές. Τυπικά, ορίζεται ως:

$$X_t, \quad t = 1, 2, 3, \ldots, T$$

Όπου:

- X_t είναι η τιμή της χρονοσειράς την χρονική στιγμή t
- T το πλήθος των παρατηρήσεων.

Μια χρονοσειρά λοιπόν, είναι ένα σύνολο αριθμητικών παρατηρήσεων, καθεμία από τις οποίες καταγράφεται σε μια συγκεκριμένη χρονική στιγμή, η οποία χρησιμοποιείται ως δείκτης. Όταν οι παρατηρήσεις καταγράφονται σε σταθερά χρονικά διαστήματα, τότε η χρονοσειρά αναφέρεται ως διακριτού χρόνου (discrete-time time series). Αντίθετα, όταν οι παρατηρήσεις καταγράφονται συνεχώς σε ένα δεδομένο χρονικό διάστημα, τότε αναφέρεται ως συνεχούς χρόνου (continuous-time).



Σχήμα 5.1: Γράφημα που απεικονίζει την τιμή κλεισίματος της μετοχής της Microsoft Corporation από τον Ιανουάριο 2018 έως τον Δεκέμβριο 2018

Ανάλογα με τον αριθμό των μεταβλητών που μεταβάλλονται συναρτήσει του χρόνου, οι χρονοσειρές χωρίζονται σε δύο κατηγορίες: μονομεταβλητές (univariate), όπου μία μεταβλητή μεταβάλλεται σε συνάρτηση με το χρόνο, και πολυμεταβλητές (multivariate), όπου περισσότερες από μία μεταβλητές μεταβάλλονται σε συνάρτηση με το χρόνο.

Οι χρονοσειρές απεικονίζονται συχνά σε απλά δισδιάστατα (2-D) γραφήματα, όπου κάθε παρατήρηση σημειώνεται με μια κουκκίδα. Ο οριζόντιος άξονας αναπαριστά το χρόνο, ενώ ο κατακόρυφος άξονας αναπαριστά το μέγεθος του οποίου η συμπεριφορά μελετάται με την πάροδο του χρόνου. Στο Σχήμα 5.1 , παρουσιάζεται η τιμή κλεισίματος (σε δολάρια ΗΠΑ) της μετοχής της Microsoft Corporation από τον Ιανουάριο 2018 έως τον Δεκέμβριο 2018, σχηματίζοντας μια χρονοσειρά που αποτελείται από 365 παρατηρήσεις.

Οι χρονοσειρές αποτελούνται από τέσσερα διαφορετικά συστατικά, τα οποία και μπορούν να εξαχθούν από αυτές. Αυτά είναι η τάση (trend), οι εποχιακές διακυμάνσεις (seasonal variations), οι κυκλικές διακυμάνσεις (cyclic variations) και οι τυχαίες μεταβολές (irregular movements), τα οποία και περιγράφονται εν συντομία στη συνέχεια:

- **Τάση**: Η τάση είναι ένα μακροχρόνιο συστατικό που αποτυπώνει τη γενική τάση μιας χρονοσειράς να αυξάνεται (ανοδική τάση) ή να μειώνεται (καθοδική τάση) συνεχώς. Το μοτίβο της μπορεί να είναι γραμμικό ή μη γραμμικό και ενδέχεται να παραμένει σταθερό ή να μεταβάλλεται με την πάροδο του χρόνου.

- **Εποχικότητα**: Οι εποχιακές διακυμάνσεις είναι κανονικές, σχετικά βραχυπρόθεσμες επαναλαμβανόμενες αυξομειώσεις που, σε αντίθεση με τους κύκλους, συμβαίνουν εντός μιας συγκεκριμένης χρονικής περιόδου, για παράδειγμα εβδομαδιαία, μηνιαία ή ετήσια.

- **Κύκλοι**: Οι κυκλικές διακυμάνσεις είναι μακροπρόθεσμες, ανοδικές και καθο-

δικές διακυμάνσεις, οι οποίες μπορεί να είναι ακανόνιστες και προκαλούνται από συνθήκες που εμφανίζονται κυκλικά, αλλά χωρίς σταθερή διάρκεια.

- **Τυχαίες διακυμάνσεις**: Οι τυχαίες διακυμάνσεις δεν ακολουθούν ένα προβλέψιμο μοτίβο και προκαλούνται από απρόβλεπτα γεγονότα. Σε αυτήν την κατηγορία περιλαμβάνονται όλες οι μεταβολές που δεν μπορούν να ταξινομηθούν ως τάση, κυκλικές ή εποχιακές.

Τα τέσσερα συστατικά που αναφέρθηκαν μπορούν να αξιοποιηθούν από δύο τύπους μοντέλων: τα προσθετικά (additive) [70] και τα πολλαπλασιαστικά (multiplicative) [303]. Στην πρώτη περίπτωση, θεωρείται ότι κάθε παρατήρηση προκύπτει ως άθροισμα των τεσσάρων στοιχείων, υποδηλώνοντας πως είναι ανεξάρτητα μεταξύ τους. Αντιθέτως, στα πολλαπλασιαστικά μοντέλα, οι τιμές υπολογίζονται βάσει του γινομένου των εν λόγω συστατικών, υποθέτοντας έτσι ότι το καθένα μπορεί να ενισχύει ή να αλληλεπιδρά με τα υπόλοιπα. Εάν το μέγεθος των στοιχείων τάσης, κυκλικών και εποχιακών διακυμάνσεων παραμένει σχετικά σταθερό στον χρόνο, τότε προτιμάται ένα προσθετικό μοντέλο. Όταν όμως αυτά τα συστατικά μεταβάλλονται ανάλογα με την εκάστοτε τάση, ένα πολλαπλασιαστικό μοντέλο τείνει να είναι πιο κατάλληλο.

Στην πράξη, μια χρονοσειρά δεν είναι απαραίτητο να ενσωματώνει και τα τέσσερα συστατικά ταυτόχρονα. Ο ρόλος της ανάλυσης χρονοσειρών είναι να ξεχωρίσει και να απομονώσει αυτά τα στοιχεία, ώστε να αποκαλυφθεί με ποιον τρόπο συμβάλλουν στις τιμές που παρατηρούμε. Στην ουσία, μόνο η τάση, οι κύκλοι και η εποχικότητα δίνουν συνεπή μοτίβα, ενώ οι τυχαίες μεταβολές αντιμετωπίζονται ως σφάλματα ή απρόβλεπτοι παράγοντες. Η πρόβλεψη, συνεπώς, βασίζεται στην ιδέα ότι τα αναγνωρίσιμα μοτίβα του παρελθόντος μπορούν να προεκταθούν στο μέλλον. Με άλλα λόγια, επεκτείνουμε τα ήδη διαπιστωμένα μοτίβα, προϋποθέτοντας ότι όσα συνέβησαν ξανά, έχουν πιθανότητες να συμβούν εκ νέου.

Η διάρθρωση αυτού του κεφαλαίου είναι η ακόλουθη: Αρχικά, εν είδει εισαγωγής, παρουσιάζεται μια συγκριτική ανάλυση σύγχρονων μεθόδων πρόβλεψης μονοδιάστατων χρονοσειρών η οποία περιλαμβάνει κλασικές στατιστικές μεθόδους πρόβλεψης όπως η ARIMA και μέθοδοι εκθετικές εξομάλυνσης, υβριδικές μεθόδους που συνδυάζουν στατιστικές τεχνικές και διαχειρίζονται πολλαπλές εποχικότητες και μεθόδους Μηχανικής Μάθησης που βασίζονται στα νευρωνικά δίκτυα. Στη συνέχεια, εισάγεται η χρήση της Ανάλυσης Συναισθήματος στην πρόβλεψη χρονοσειρών με στόχο το να διερευνηθεί το κατά πόσο η εισαγωγή πληροφορίας που αφορά την άποψη της κοινής γνώμης όπως αυτή αποτυπώνεται στα μέσα κοινωνικής δικτύωσης, επηρεάζει θετικά τις εξαγόμενες προβλέψεις. Παρουσιάζεται μια ενδελεχής σύγκριση, με διττό στόχο. Πρώτον, τη σύγκριση μεθόδων πρόβλεψης και δεύτερον τη σύγκριση διαφορετικών διαμορφώσεων χαρακτηριστικών εισόδου. Στο πλαίσιο μιας εκτεταμένης πειραματικής διαδικασίας, δοκιμάστηκαν 22 διαφορετικές διαμορφώσεις χαρακτηριστικών, συνδυάζοντας τιμές κλεισίματος μετοχών και δείκτες συναισθήματος, σε πολλαπλά σύνολα δεδομένων και μετρικές αξιολόγησης, κάνοντας χρήση 27 αλγορίθμων Μηχανικής Μάθησης.

Τέλος, στο τρίτο τμήμα, διερευνάται η χρήση μεθόδων Ανάλυσης Συναισθήματος σε δεδομένα που εξάγονται από κοινωνικά δίκτυα και την ενσωμάτωσή τους σε προβλήματα πρόβλεψης χρονοσειρών με χρήση αλγορίθμων Βαθιάς Μάθησης. Μέσω μιας εκτεταμένης πειραματικής διαδικασίας, δοκιμάστηκαν 67 διαφορετικές διαμορφώσεις εισόδου σε 16 διαφορετικά σύνολα δεδομένων, αξιοποιώντας 30 διαφορετικούς αλγορίθμους. Η σύγκριση των μεθόδων πραγματοποιήθηκε σε δύο επίπεδα: την διερεύνηση της πιθανής

βελτίωσης της ακρίβειας των προβλέψεων μέσω της χρήσης Ανάλυσης Συναισθήματος και την επιλογή των μεθόδων που επιτυγχάνουν την καλύτερη απόδοση.

5.2 Συγκριτική μελέτη μεθόδων για προβλέψεις σε χρηματοοικονομικά δεδομένα

Στην πρώτη υποενότητα αυτού του κεφαλαίου, πραγματοποιείται μια εκτενής συγκριτική ανάλυση σύγχρονων προβλεπτικών μεθόδων στο πρόβλημα της πρόβλεψης της τιμής κλεισίματος μετοχών του χρηματιστηρίου. Η ανάλυση αυτή έχει εισαγωγικό χαρακτήρα, καθώς στοχεύει στην παροχή μιας συνολικής εικόνας των διαθέσιμων τεχνικών, καλύπτοντας διαφορετικές οικογένειες μεθόδων, όπως στατιστικές προσεγγίσεις, υβριδικά μοντέλα και αλγορίθμους Μηχανικής Μάθησης.

Στη σύγκριση αυτή περιλαμβάνονται στατιστικά μοντέλα όπως το ARIMA (Auto-Regressive Integrated Moving Average), το οποίο χρησιμοποιείται ευρέως για την πρόβλεψη χρονοσειρών λόγω της ικανότητάς του να συλλαμβάνει αυτοπαλίνδρομες σχέσεις και κινούμενους μέσους όρους [31]. Επιπλέον, εξετάζεται το μοντέλο εκθετικής εξομάλυνσης του Holt-Winters, το οποίο λαμβάνει υπόψη τις τάσεις και την εποχικότητα της χρονοσειράς [121, 293].

Στη σύγκριση επίσης συμπεριλαμβάνονται υβριδικές προσεγγίσεις, όπως οι μέθοδοι TBATS και TATS, που ενσωματώνουν τριγωνομετρικές συναρτήσεις για μοντελοποίηση εποχικότητας [61], καθώς και μετασχηματισμούς Box-Cox [32] για τη σταθεροποίηση της διακύμανσης. Τα μοντέλα αυτά είναι γνωστά για την καλή προσαρμογή τους σε δεδομένα με περίπλοκες εποχικότητες και τάσεις μεταβλητές στο χρόνο.

Τέλος, συμπεριλαμβάνονται σύγχρονες μέθοδοι Μηχανικής Μάθησης που βασίζονται σε βαθιά νευρωνικά δίκτυα, όπως η μέθοδος N-BEATS (Neural Basis Expansion Analysis for Time Series), που αποτελείται από ένα πλήρως συνδεδεμένο νευρωνικό δίκτυο [198], αλλά και η DeepAR, που υλοποιείται μέσω ενός αυτοπαλίνδρομου αναδρομικού νευρωνικού δικτύου βασισμένου σε αρχιτεκτονική LSTM, το οποίο επιτρέπει την παραγωγή πιθανοτικών προβλέψεων και την εκμάθηση εποχικών μοτίβων [238]. Επίσης, στη σύγκριση συμπεριλαμβάνεται και ο Prophet, ένα μοντέλο βασισμένο σε γραμμική και μη γραμμική παλινδρόμηση, το οποίο επιτρέπει διαχείριση εποχικότητας και εξωτερικών παραγόντων, όπως αργίες και γεγονότα, και είναι ιδιαίτερα γνωστό, καθώς αναπτύχθηκε από την ομάδα αναλυτών του κοινωνικού δικτύου Facebook [270].

Δεδομένης της ποικιλομορφίας αυτών των μεθόδων και των χωρικών περιορισμών της παρούσας εργασίας, η παρουσίαση εστιάζει σε μια συνοπτική αλλά ουσιαστική εισαγωγή στο θεωρητικό τους υπόβαθρο. Ο σκοπός δεν είναι μια εξαντλητική ανάλυση των ιδιαίτερων χαρακτηριστικών κάθε μεθόδου, αλλά η κατανόηση των βασικών διαφορών τους, ώστε να διευκολυνθεί η αξιολόγηση της καταλληλότητάς τους για διαφορετικές εφαρμογές πρόβλεψης χρηματοοικονομικών χρονοσειρών.

5.3 Οικογένειες προβλεπτικών μοντέλων

5.3.1 Μοντέλα ARIMA

Τα μοντέλα ARIMA (AutoRegressive Integrated Moving Average) [33] αποτελούν μια από τις πλέον δημοφιλείς και ευρέως χρησιμοποιούμενες μεθόδους για την πρόβλεψη χρονοσειρών. Πρόκειται για γενίκευση των μοντέλων ARMA (AutoRegressive

Moving Average), τα οποία εφαρμόζονται σε στάσιμες χρονοσειρές, ενώ τα ARIMA μοντέλα είναι κατάλληλα για μη στάσιμα δεδομένα, αφού ενσωματώνουν ένα επιπλέον στάδιο ολοκλήρωσης (Integrated), που μετατρέπει μια μη στάσιμη χρονοσειρά σε στάσιμη μέσω διαφορών.

Ένα μοντέλο της οικογένειας ARIMA αποτελείται από τρία βασικά στοιχεία:

1. **Αυτοπαλινδρόμηση (AutoRegressive - AR) με παράμετρο p**: Το μοντέλο χρησιμοποιεί τις παρελθοντικές τιμές της χρονοσειράς για να προβλέψει τις μελλοντικές. Η σχέση αυτή μπορεί να περιγραφεί από την εξίσωση:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t, \tag{5.1}$$

   Όπου:

   - y_t είναι η τρέχουσα τιμή της χρονοσειράς,
   - c είναι μια σταθερά,
   - ϕ_i είναι οι συντελεστές της αυτοπαλινδρόμησης,
   - ϵ_t είναι ένας τυχαίος όρος σφάλματος (λευκός θόρυβος).

2. **Ολοκλήρωση (Integrated - I) με παράμετρο d**: Στην περίπτωση μη στάσιμων χρονοσειρών, δηλαδή σε χρονοσειρές που έχουν αυξανόμενη ή φθίνουσα τάση ή μεταβαλλόμενη διακύμανση, χρησιμοποιείται η διαφοροποίηση ώστε να μετατραπούν σε στάσιμες. Το επίπεδο διαφοροποίησης d καθορίζει πόσες φορές πρέπει να υπολογίσουμε τη διαφορά μεταξύ διαδοχικών παρατηρήσεων για να επιτευχθεί στασιμότητα. Ισχύει πως:

$$y'_t = y_t - y_{t-1} \tag{5.2}$$

3. **Κινούμενος Μέσος (Moving Average - MA) με παράμετρο q**: Αντί να βασίζεται στις προηγούμενες τιμές της χρονοσειράς, το MA μέρος του μοντέλου χρησιμοποιεί παρελθοντικά σφάλματα πρόβλεψης για να βελτιώσει την ακρίβεια της πρόβλεψης. Το τμήμα αυτό του μοντέλου επιτρέπει τη διόρθωση των τυχαίων διακυμάνσεων και των μη αναμενόμενων αποκλίσεων στη χρονοσειρά. Η εξίσωση που περιγράφει τη διαδικασία αυτή είναι:

$$y_t = c + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q}, \tag{5.3}$$

   Όπου:

   - θ_1 είναι οι συντελεστές του κινούμενου μέσου,
   - ϵ_t είναι ο όρος τυχαίου σφάλματος.

Σε συνέπεια των παραπάνω, κάθε μοντέλο της οικογένειας ARIMA χαρακτηρίζεται από τις τρεις παραπάνω τιμές, p, d και q και στη βιβλιογραφία απαντάται ως ARIMA(p,d,q). Συνδυάζοντας τα τρία αυτά στοιχεία, η εξίσωση που τελικά περιγράφει ένα τέτοιο μοντέλο παίρνει τη μορφή:

$$y'_t = c + \phi_1 y'_{t-1} + \cdots + \phi_p y'_{t-p} + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q} + \epsilon_t, \tag{5.4}$$

Όπου:

- y'_t η διαφοροποιημένη σειρά
- p η τάξη του αυτοπαλίνδρομου μέρους
- d η τάξη των πρώτων διαφορών
- q η τάξη του μέρους του κινητού μέσου.

.

Μερικές χαρακτηριστικές περιπτώσεις που προκύπτυον για διάφορες τιμές των παραμέτρων αυτών, φαίνονται στον Πίνακα 5.1.

βοοκταβς αρραψ [ταβλε]ξςολορ

Πίνακας 5.1: Ειδικές περιπτώσεις του μοντέλου ARIMA

Περίπτωση	**Μορφή**
Λευκός θόρυβος	ARIMA(0,0,0) χωρίς σταθερά
Τυχαίος περίπατος	ARIMA(0,1,0) χωρίς σταθερά
Τυχαίος περίπατος με περιπλάνηση	ARIMA(0,1,0) με σταθερά
Αυτοπαλινδρόμηση	ARIMA(p,0,0)
Κινητός μέσος	ARIMA(0,0,q)

5.3.2 Μοντέλο Prophet

Το μοντέλο Prophet αποτελεί μια πρόσφατη προσθήκη στο χώρο των προβλεπτικών μοντέλων. Αναπτύχθηκε από το Facebook [270] και αρχικά σχεδιάστηκε για την πρόβλεψη ημερήσιων δεδομένων, λαμβάνοντας υπόψη εβδομαδιαία και ετήσια εποχικότητα, καθώς και επιδράσεις από αργίες. Στη συνέχεια, επεκτάθηκε ώστε να μπορεί να διαχειρίζεται περισσότερους τύπους εποχικών δεδομένων. Το μοντέλο αυτό, αποδίδει καλύτερα σε χρονοσειρές με έντονη εποχικότητα και επαρκές ιστορικό δεδομένων που καλύπτει πολλαπλές εποχικές περιόδους.

Μαθηματικά μπορεί να περιγραφεί ως εξής:

$$y(t) = g(t) + s(t) + h(t) + e_t \tag{5.5}$$

Όπου:

- ο όρος $g(t)$ μοντελοποιεί τις μη περιοδικές μεταβολές,
- ο όρος $s(t)$ εκφράζει περιοδικές διακυμάνσεις (όπως ημερήσια, εβδομαδιαία ή ετήσια εποχικότητα),
- ο όρος $h(t)$ αναπαριστά τις επιδράσεις γεγονότων που συμβαίνουν σε ακανόνιστα χρονικά διαστήματα (όπως αργίες)
- και ο παράγοντας e_t περιλαμβάνει όλες τις υπόλοιπες διακυμάνσεις που δεν εξηγούνται από τα κύρια συστατικά του μοντέλου.

Ο Prophet χρησιμοποιεί δύο διαφορετικά μοντέλα τάσης. Το πρώτο είναι ένα μοντέλο κορεσμένης ανάπτυξης, όπου η αύξηση είναι μη γραμμική και φτάνει σε ένα όριο φέρουσας. Αυτό το είδος ανάπτυξης περιγράφεται από το λογιστικό μοντέλο ανάπτυξης, το οποίο έχει τροποποιηθεί ώστε να προσαρμόζεται σε περιπτώσεις όπου τόσο ο ρυθμός ανάπτυξης όσο και η φέρουσα ικανότητα δεν είναι σταθερά. Το δεύτερο μοντέλο είναι ένα τμηματικά γραμμικό μοντέλο, κατάλληλο για περιπτώσεις όπου δεν παρατηρείται κορεσμένη ανάπτυξη.

Για την προσαρμογή της εποχικότητας, ο Prophet βασίζεται στη χρήση σειρών Fourier, καθιστώντας τον αλγόριθμο ευέλικτο, ενώ ο αριθμός των όρων Fourier καθορίζει το επίπεδο λεπτομέρειας της εποχικότητας (ετήσια, εβδομαδιαία κ.λπ.). Δεδομένου ότι γεγονότα όπως αργίες μπορούν να επιφέρουν απότομες μεταβολές στη χρονοσειρά (κάτι που είναι ιδιαίτερα σημαντικό σε επιχειρησιακά δεδομένα), ο Prophet διαθέτει μια λίστα με τέτοια γεγονότα, ενσωματώνοντάς τα αυτόματα στην πρόβλεψη.

Το γεγονός ότι η απόδοση του Prophet είναι ιδιαίτερα καλή σε χρονοσειρές με έντονη εποχικότητα και δεδομένα που καλύπτουν πολλαπλές εποχικές περιόδους, τον καθιστά κατάλληλο για εφαρμογές όπως η ζήτηση προϊόντων, η ανάλυση επισκεψιμότητας ιστοσελίδων και η πρόβλεψη πωλήσεων.

5.3.3 Εποχικά Μοντέλα Holt - Winters

Το μοντέλο Holt-Winters είναι ένα ευρέως διαδεδομένο μοντέλο εξομάλυνσης για την πρόβλεψη χρονοσειρών, δημοφιλές σε περιπτώσεις που υπάρχουν εποχικές διακυμάνσεις [121, 293]. Βασίζεται στην εκθετική εξομάλυνση και αποτελεί επέκταση των μεθόδων του Holt, επιτρέποντας την πρόβλεψη με τη χρήση τάσης και εποχικότητας. Υπάρχουν δύο βασικές εκδοχές του μοντέλου: η προσθετική (Holt-Winters Additive, HWAAS) και η πολλαπλασιαστική (Holt-Winters Multiplicative, HWAMS), με την επιλογή να εξαρτάται από τη φύση της εποχικότητας – αν αυτή παραμένει σταθερή (προσθετική) ή αν μεταβάλλεται ανάλογα με το επίπεδο της χρονοσειράς (πολλαπλασιαστική). Η διαδικασία μοντελοποίησης περιλαμβάνει τον υπολογισμό τριών βασικών συνιστωσών: του επιπέδου (level), της τάσης (trend) και της εποχικότητας (seasonality), με τη χρήση τριών αντίστοιχων παραμέτρων εξομάλυνσης. Το μοντέλο αυτό έχει αποδειχθεί ιδιαίτερα αποτελεσματικό για προβλέψεις βραχυπρόθεσμου και μεσοπρόθεσμου ορίζοντα, ειδικά σε δεδομένα όπου οι εποχικές τάσεις είναι εμφανείς και επαναλαμβανόμενες [93]. Χρησιμοποιείται συχνά σε χρηματοοικονομικά δεδομένα, αποδεικνύοντας τη χρησιμότητά του στην πρόβλεψη τιμών μετοχών και άλλων οικονομικών μεγεθών [44].

Ακολουθεί η φορμαλιστική περιγραφή των δύο βαριάντων του:

5.3.4 Προσθετικό Μοντέλο Holt-Winters (HWAAS)

Στην προσθετική εκδοχή, η εποχικότητα προστίθεται στο βασικό επίπεδο της χρονοσειράς. Οι συναρτήσεις εξομάλυνσης δίνονται από τους τύπους:

$$l_t = \alpha(y_t - s_{t-m}) + (1-\alpha)(l_{t-1} + b_{t-1}) \tag{5.6}$$

$$b_t = \beta^*(l_t - l_{t-1}) + (1-\beta^*)b_{t-1} \tag{5.7}$$

$$s_t = \gamma(y_t - l_t) + (1-\gamma)s_{t-m} \tag{5.8}$$

ενώ η συνάρτηση πρόβλεψης παίρνει τη μορφή:

$$\hat{y}_{t+h} = l_t + hb_t + s_{t+h-m(k+1)} \tag{5.9}$$

5.3.5 Πολλαπλασιαστικό Μοντέλο Holt-Winters (HWAMS)

Στην πολλαπλασιαστική εκδοχή, η εποχικότητα πολλαπλασιάζεται στο βασικό επίπεδο της χρονοσειράς. Οι συναρτήσεις εξομάλυνσης δίνονταια αντίστοιχα από τους τύπους:

$$l_t = \alpha \frac{y_t}{s_{t-m}} + (1-\alpha)(l_{t-1} + b_{t-1}) \tag{5.10}$$

$$b_t = \beta^*(l_t - l_{t-1}) + (1-\beta^*)b_{t-1} \tag{5.11}$$

$$s_t = \gamma \frac{y_t}{l_t} + (1-\gamma)s_{t-m} \tag{5.12}$$

ενώ η αντίστοιχη συνάρτηση πρόβλεψης είναι:

$$\hat{y}_{t+h} = (l_t + hb_t)s_{t+h-m(k+1)} \tag{5.13}$$

Και στα δύο παραπάνω μοντέλα:

- y_t: Η πραγματική τιμή της χρονοσειράς στη χρονική στιγμή t
- l_t: Το εκτιμώμενο επίπεδο στη χρονική στιγμή t
- b_t: Η εκτιμώμενη τάση στη χρονική στιγμή t
- s_t: Η εκτιμώμενη εποχικότητα στη χρονική στιγμή t
- m: Η περίοδος της εποχικότητας (π.χ., 12 για μηνιαία δεδομένα με ετήσια εποχικότητα)
- k: Ο αριθμός των πλήρων περιόδων που χωράνε μέσα στο βήμα πρόβλεψης h
- α, β^*, γ: Παράμετροι εξομάλυνσης με τιμές που κυμαίνονται από 0 έως 1 και καθορίζουν τον βαθμό προσαρμογής στις νέες τιμές δεδομένων.

5.3.6 Το μοντέλο NBEATS

Το N-BEATS (Neural Basis Expansion Analysis for Time Series) [199] είναι ένα βαθύ νευρωνικό δίκτυο σχεδιασμένο ειδικά για την πρόβλεψη χρονοσειρών. Αποτελείται από στοίβες, οι οποίες περιλαμβάνουν πολλαπλά βασικά μπλοκ, με κάθε μπλοκ να περιέχει τέσσερα πλήρως συνδεδεμένα επίπεδα. Η έξοδος κάθε μπλοκ διαχωρίζεται σε δύο μέρη: την ανακατασκευή εισόδου (backcast) και την πρόβλεψη εξόδου (forecast). Τα δεδομένα περνούν από διαδοχικές στοίβες, κάθε μία από τις οποίες εξειδικεύεται σε διαφορετικές συνιστώσες της χρονοσειράς, όπως η τάση, η εποχικότητα και τα υπολειμματικά μοτίβα.

Μαθηματικά, το N-BEATS εκφράζεται ως εξής:

$$\mathbf{x} = f_\theta(\mathbf{B}, \mathbf{F}) \tag{5.14}$$

όπου:

- $\mathbf{B}$: Η backcast συνιστώσα (ανακατασκευή εισόδου).

- $\mathbf{F}$: Η forecast συνιστώσα (πρόβλεψη εξόδου).
- f_θ: Η μη γραμμική χαρτογράφηση που πραγματοποιείται από τα πλήρως συνδεδεμένα επίπεδα.

Η αρχιτεκτονική περιλαμβάνει ένα μηχανισμό Residual Stacking, όπου κάθε νέο επίπεδο ενημερώνει την προηγούμενη αναπαράσταση, επιτρέποντας έτσι στο μοντέλο να συσσωρεύει πληροφορία και να βελτιώνει την πρόβλεψη:

$$\mathbf{B}^{(i)} = g_\theta^{(i)}(\mathbf{B}^{(i-1)}) \tag{5.15}$$

$$\mathbf{F} = \sum_i \mathbf{F}^{(i)} \tag{5.16}$$

όπου:

- $\mathbf{B}^{(i)}$: Η ανακατασκευή εισόδου στο μπλοκ i.
- $\mathbf{F}^{(i)}$: Η πρόβλεψη εξόδου από το μπλοκ i.
- $g_\theta^{(i)}$: Μη γραμμική συνάρτηση μετασχηματισμού στο μπλοκ i.
- θ: Παράμετροι του νευρωνικού δικτύου που προσαρμόζονται κατά την εκπαίδευση.

Το N-BEATS έχει αποδειχθεί ιδιαίτερα αποτελεσματικό σε προβλήματα πρόβλεψης όπως τα χρηματοοικονομικά δεδομένα, οι μετεωρολογικές μετρήσεις και οι πωλήσεις προϊόντων. Η συνολική πρόβλεψη του μοντέλου προκύπτει ως το άθροισμα των επιμέρους προβλέψεων από κάθε στοίβα. Η χρήση ενός συνδυασμού τάσης, εποχικότητας και μη γραμμικών μοτίβων καθιστά το N-BEATS εξαιρετικά ευέλικτο και ικανό να αποδίδει υψηλής ποιότητας προβλέψεις χωρίς να απαιτεί προεπεξεργασία των δεδομένων.

5.3.7 Το μοντέλο DeepAR

Το DeepAR [239] είναι ένα βαθύ νευρωνικό δίκτυο βασισμένο σε αναδρομικές αρχιτεκτονικές RNNs που χρησιμοποιείται για την πρόβλεψη πιθανοτικών χρονοσειρών. Σχεδιάστηκε για να αξιοποιεί μεγάλο όγκο σχετικών χρονοσειρών, εκπαιδεύοντας ένα μοντέλο που μπορεί να μάθει κοινές δομές μεταξύ τους. Χρησιμοποιεί ένα μηχανισμό βασισμένο σε LSTMs ώστε να διαχειρίζεται εξαρτήσεις μεγάλης διάρκειας στη χρονοσειρά και να επιτυγχάνει ακριβείς προβλέψεις.

Στη γλώσσα των μαθηματικών, το DeepAR εκφράζεται ως εξής:

$$p(\mathbf{y}_{t+1:T}|\mathbf{y}_{1:t}, \mathbf{x}) = \prod_{i=t+1}^{T} p(y_i|\mathbf{y}_{1:i-1}, \mathbf{x}; \theta) \tag{5.17}$$

όπου:

- $\mathbf{y}_t$: Η παρατηρούμενη τιμή της χρονοσειράς στη χρονική στιγμή t.
- $\mathbf{x}$: Επιπρόσθετες εξωγενείς μεταβλητές.
- θ: Παράμετροι του νευρωνικού δικτύου που εκτιμώνται κατά την εκπαίδευση.

- $p(y_i|\mathbf{y}_{1:i-1}, \mathbf{x}; \theta)$: Η κατανομή πιθανότητας της μελλοντικής τιμής y_i με βάση τις προηγούμενες τιμές και τις εξωγενείς μεταβλητές.

Το μοντέλο χρησιμοποιεί δειγματοληψία Monte Carlo για την παραγωγή πιθανοτικών προβλέψεων, επιτρέποντας την εκτίμηση της αβεβαιότητας στις προβλέψεις. Επίσης, υποστηρίζει μοντελοποίηση εποχικών και τάσεων μέσω της χρήσης επιπρόσθετων χαρακτηριστικών και δυνατοτήτων μάθησης από πολλές χρονοσειρές ταυτόχρονα.

Το DeepAR έχει χρησιμοποιηθεί με επιτυχία σε εφαρμογές όπως η πρόβλεψη πωλήσεων, η διαχείριση αποθεμάτων και η εκτίμηση της ζήτησης, καθώς μπορεί να αξιοποιήσει μεγάλο όγκο δεδομένων και να αποδώσει προβλέψεις υψηλής ακρίβειας.

5.3.8 Trigonometric BAT

Το Trigonometric BATS (Box-Cox transformation, ARMA errors, Trend, and Seasonal components) είναι ένα στατιστικό μοντέλο για την πρόβλεψη χρονοσειρών, το οποίο ενσωματώνει τριγωνομετρικές συναρτήσεις για να χειρίζεται πολύπλοκα εποχικά μοτίβα. Χρησιμοποιείται σε δεδομένα με μη ακέραιες ή πολλαπλές εποχικές περιόδους, επιτρέποντας μεγαλύτερη ευελιξία στη μοντελοποίηση της χρονοσειράς.

Μαθηματικά, το μοντέλο εκφράζεται ως:

$$y_t = \lambda^{-1} \left[\sum_{i=1}^{k} \alpha_i \cos(\omega_i t) + \beta_i \sin(\omega_i t) + X_t \right] + \varepsilon_t \tag{5.18}$$

όπου:

- y_t: Η προβλεπόμενη τιμή της χρονοσειράς στη χρονική στιγμή t.
- λ: Παράμετρος Box-Cox μετασχηματισμού.
- α_i, β_i: Συντελεστές των τριγωνομετρικών όρων.
- ω_i: Συχνότητα των εποχικών συνιστωσών.
- X_t: Στοιχείο που περιλαμβάνει τάση και συνιστώσες.
- ε_t: Τυχαίο σφάλμα.

Το μοντέλο BATS ενσωματώνει το μετασχηματισμό Box-Cox, επιτρέποντας την καλύτερη σταθεροποίηση της διακύμανσης. Τα ARMA σφάλματα βοηθούν στην αποτύπωση αυτοσυσχετιζόμενων δομών, ενώ η χρήση τριγωνομετρικών όρων διευκολύνει την πρόβλεψη δεδομένων με περίπλοκη εποχικότητα.

Το BATS είναι ιδιαίτερα χρήσιμο σε προβλέψεις όπου η εποχικότητα δεν είναι σταθερή και δεν μπορεί να περιγραφεί από απλές προσθετικές ή πολλαπλασιαστικές μεθόδους, όπως στην κατανάλωση ενέργειας, τη ζήτηση προϊόντων και τα κλιματολογικά δεδομένα.

5.4 Πειραματική διαδικασία

Στη συνέχεια παρουσιάζεται η περιγραφή της πειραματικής διαδικασίας. Αρχικά, παρουσιάζονται πληροφορίες για τα δεδομένα και τις μεθόδους που χρησιμοποιήθηκαν, ενώ στην συνέχεια περιγράφονται τα αποτελέσματα και τα ευρήματα που εξήχθησαν.

5.4.1 Δεδομένα

Στο πλαίσιο της πειραματικής διαδικασίας χρησιμοποιήθηκαν 40 σύνολα δεδομένων που αντλήθηκαν από την ιστοσελίδα Yahoo! Finance. Η επιλογή αυτή έγινε με σκοπό να διασφαλιστεί το ότι τα δεδομένα θα είναι δημόσια διαθέσιμα, επιτρέποντας την εύκολη αναπαραγωγή των αποτελεσμάτων. Τα τελικά σύνολα δεδομένων σχηματίστηκαν επιλέγοντας 40 από τις πιο ενεργές μετοχές. Για λόγους ευκολίας από εδώ και στο εξής κάθε μια από αυτές θα αναφέρεται με τη συντομογραφία του ονόματός της, όπως γίνεται και στην εν λόγω σελίδα από όπου τα δεδομένα αντλήθηκαν. Στον Πίνακα 5.2 παρουσιάζονται οι πλήρεις ονομασίες των μετοχών μαζί με τις αντίστοιχες συντομογραφίες τους. Για κάθε μετοχή, το σύνολο δεδομένων καλύπτει περίοδο ενός έτους και συγκεκριμένα το διάστημα από την 1η Απριλίου 2018 έως την 31η Μαρτίου 2019. Δεδομένου ότι η μελέτη επικεντρώνεται αποκλειστικά στην πρόβλεψη μονομεταβλητών χρονοσειρών, κάθε σύνολο δεδομένων περιλαμβάνει την ημερήσια τιμή κλεισίματος της μετοχής, περιέχοντας περίπου 365 τιμές. Για να αντιμετωπιστεί το ζήτημα των ελλιπών δεδομένων που προκύπτουν τις ημέρες που η χρηματιστηριακή αγορά έμεινε κλειστή, χρησιμοποιήθηκε η τιμή κλεισίματος της προηγούμενης ημέρας. Η μετατροπή αυτή διευκολύνει την ομαλή διεξαγωγή της πειραματικής διαδικασίας, καθώς κάποιες από τις υλοποιήσεις των μεθόδων που χρησιμοποιήθηκαν δε μπορούν να διαχειριστούν ελλιπή δεδομένα.

5.4.2 Αλγόριθμοι

Όπως έχει ήδη τονιστεί, ο βασικός στόχος αυτής της μελέτης είναι η σύγκριση διαφόρων ευρέως χρησιμοποιούμενων και ισχυρών αλγορίθμων για την πρόβλεψη μονομεταβλητών χρονοσειρών, οι οποίοι θεωρούνται state-of-the-art. Για την υλοποίηση των αλγορίθμων, χρησιμοποιήθηκε το πλαίσιο AtsPy [257] το οποίο είναι ένα πρόσφατα ανεπτυγμένο αποθετήριο αυτοματοποιημένων μοντέλων χρονοσειρών στη γλώσσα Python.

Το AtsPy επιλέχθηκε επειδή περιλαμβάνει μια ποικιλία αυτοματοποιημένων αλγορίθμων πρόβλεψης χρονοσειρών, ενώ ταυτόχρονα μειώνει το σφάλμα των δομικών των μοντέλων κατά 30%-50% χρησιμοποιώντας ένα Gradient Boosting Model (GBM) με χαρακτηριστικά που εξάγονται από τη χρονοσειρά. Το αποθετήριο αυτό αξιοποιεί τις ήδη υπάρχουσες βιβλιοθήκες Python, όπως οι statsmodels, fbprophet, gluonts, pmdarima και tbats, ώστε να χρησιμοποιήσει όλους τους αλγορίθμους που παρουσιάζονται στην Ενότητα 5.3, με έναν ενιαίο τρόπο, καθιστώντας το ένα εύχρηστο αλλά ταυτόχρονα αποτελεσματικό εργαλείο για την πρόβλεψη χρονοσειρών. Επιπλέον, το AtsPy εντοπίζει αυτόματα εποχικότητες στα δεδομένα χρησιμοποιώντας ανάλυση φάσματος και ανάλυση αιχμών (peak analysis).

Οι αλγόριθμοι που χρησιμοποιήθηκαν ανήκουν στις έξι οικογένειες μεθόδων που παρουσιάστηκαν στην Ενότητα 5.3. Συγκεκριμένα χρησιμοποιήθηκαν οι:

- ARIMA(p,d,q): Πρόκειται για το μοντέλο Auto-Regressive Integrated Moving Average, το οποίο είναι μια εξέλιξη του AutoRegressive Moving Average (ARMA) με p την παράμετρο που ορίζει τον αριθμό των χρονικών καθυστερήσεων (lags) του αυτοπαλινδρομούμενου μοντέλου, d την τάξη διαφοροποίησης και q την τάξη του μοντέλου κινούμενου μέσου.

- Prophet: Πρόκειται για τη μέθοδο πρόβλεψης που ανέπτυξε το Facebook, η οποία

Πίνακας 5.2: Οι μετοχές και οι αντίστοιχες συντομογραφίες τους

No	Συντ.	Όνομασία μετοχής	No	Συντ.	Όνομασία μετοχής
1	AAL	American Airlines Group Inc.	21	INO	Inovio Pharmaceuticals, Inc.
2	AAPL	Apple Inc.	22	INTC	Intel Corporation
3	AMD	Advanced Micro Devices, Inc.	23	LUV	Southwest Airlines Co.
4	APA	Apache Corporation	24	MSFT	Microsoft Corporation
5	AUY	Yamana Gold Inc.	25	MU	Micron Technology, Inc.
6	BA	The Boeing Company	26	NCLH	Norwegian Cruise Line Holdings Ltd.
7	BABA	Alibaba Group Holding Limited	27	ORCL	Oracle Corporation
8	BAC	Bank of America Corporation	28	PE	Parsley Energy, Inc.
9	BB	BlackBerry Limited	29	PG	The Procter & Gamble Company
10	CCL	Carnival Corporation & Plc	30	PYPL	PayPal Holdings, Inc.
11	CSCO	Cisco Systems, Inc.	31	SBUX	Starbucks Corporation
12	DAL	Delta Air Lines, Inc.	32	SIRI	Sirius XM Holdings Inc.
13	DIS	The Walt Disney Company	33	SNAP	Snap Inc.
14	EBAY	eBay Inc.	34	SPOT	Spotify Technology S.A.
15	ERI	Eldorado Resorts, Inc.	35	TGS	Transportadora de Gas del Sur S.A.
16	F	Ford Motor Company	36	TSLA	Tesla, Inc.
17	FB	Facebook, Inc.	37	TWTR	Twitter, Inc.
18	GM	General Motors Company	38	UAL	United Airlines Holdings, Inc.
19	HPE	Hewlett Packard Enterprise Company	39	WFC	Wells Fargo & Company
20	HPQ	HP Inc.	40	ZNGA	Zynga Inc.

βασίζεται σε ένα προσθετικό μοντέλο όπου οι μη γραμμικές τάσεις υπολογίζονται μέσω ημερήσιας εποχικότητας.

- HWAAS: Πρόκειται για εκθετική εξομάλυνση με προσθετική τάση και προσθετική εποχικότητα.
- HWAMS: Παρόμοιο με το HWAAS, αλλά εφαρμόζει πολλαπλασιαστική εποχικότητα αντί για προσθετική.
- NBEATS: Ένα βαθύ νευρωνικό δίκτυο βασισμένο σε πλήρως συνδεδεμένα επίπεδα, που χρησιμοποιεί μια στρατηγική υπολογισμού της τάσης και εποχικότητας μέσω backcast και forecast residual stacking.
- GluonTS [DeepAR]: Πρόκειται για μια υλοποίηση RNN μέσω της βιβλιοθήκης GluonTS.

- TBAT: Μέθοδος της οικογένειας Trigonometric BATS που αξιοποιεί τριγωνομετρική εποχικότητα, μετασχηματισμούς Box-Cox, ARMA σφάλματα και πληροφορία που αφορά την εποχικότητα αλλά και τις τάσεις, επιτρέποντας τον χειρισμό πολύπλοκων εποχικών μοτίβων.

- TATS: Παραλλαγή του TBAT, που αξιοποιεί τριγωνομετρική εποχικότητα χωρίς τη χρήση Box-Cox μετασχηματισμών.

- TBATS1: Παρόμοιο με το TBAT, αλλά χρησιμοποιεί μία εποχική περίοδο.

- TBATS2: Παρόμοιο με το TBATS1, αλλά χρησιμοποιεί δύο εποχικές περιόδους.

- TBATP1: Παρόμοιο με το TBATS1, όπου η εποχικότητα καθορίζεται αυστηρά από την περιοδικότητα.

Κατά τη διάρκεια της πειραματικής διαδικασίας, κάθε αλγόριθμος δοκιμάστηκε σε καθένα από τα 40 σύνολα δεδομένων. Για το ARIMA(p,d,q), οι παράμετροι τέθηκαν ως (0,1,0), ενώ για τα DeepAR και NBEATS, ο αριθμός εποχών (epochs) ορίστηκε στις 20.

Σχετικά με τα μοντέλα HWAAS και HWAMS, το μοντέλο εκθετικής εξομάλυνσης HWAAS ρυθμίστηκε ώστε να βασίζεται στις προσθετικές τάσεις θέτοντας την παράμετρο trend ίση με add, ενώ το HWAMS πολλαπλασιαστικές θέτοντας την παράμετρο trend ίση με mul. Για το TBATS και τις παραλλαγές του (TATS, TBATS1, TBATP1, TBATS2), οι αντίστοιχες παράμετροι seasonal_period και using_box_cox ρυθμίστηκαν κατάλληλα, σύμφωνα με τα όσα ορίστηκαν παραπάνω.

Κατά την εκτέλεση των αλγορίθμων, πραγματοποιήθηκαν δύο διαφορετικοί τύποι προβλέψεων: η πρόβλεψη εντός δείγματος (in-sample prediction), όπου το σύνολο δεδομένων χωρίστηκε σε εκπαίδευση (70%) και έλεγχο (30%), με το πρώτο τμήμα να χρησιμοποιείται για την εκπαίδευση του μοντέλου και το δεύτερο για την αξιολόγηση των προβλέψεων, και η πρόβλεψη εκτός δείγματος (out-sample prediction), όπου χρησιμοποιήθηκε ολόκληρο το σύνολο δεδομένων για εκπαίδευση και το μοντέλο προέβλεψε τιμές για τις επόμενες δύο εβδομάδες.

Αν και και οι δύο τύποι προβλέψεων παρέχουν χρήσιμες πληροφορίες, σε αυτήν τη μελέτη εστιάσαμε στα αποτελέσματα της in-sample πρόβλεψης, καθώς τείνουν να είναι πιο αξιόπιστα.

5.4.3 Αποτελέσματα

Σε αυτή την ενότητα παρουσιάζονται τα αποτελέσματα της πειραματικής διαδικασίας. Όπως προαναφέρθηκε, η απόδοση 11 διαφορετικών μεθόδων συγκρίθηκε σε 40 διακριτά σύνολα δεδομένων του χρηματιστηρίου. Για την αξιολόγηση των μεθόδων, χρησιμοποιήθηκαν τρεις μετρικές αξιολόγησης οι: MAE, RMSE και MAPE, πληροφορίες για τις οποίες έχουν παρουσιαστεί στην υποενότητα 2.1.2.3 Επιπροσθέτως, για την αύξηση της αξιοπιστίας των αποτελεσμάτων, εκτός από τα αριθμητικά αποτελέσματα της απόδοσης των μεθόδων στις παραπάνω μετρικές, χρησιμοποιήθηκαν και στατιστικοί έλεγχοι, των οποίων τα αποτελέσματα επίσης παρουσιάζονται. Λόγω του μεγάλου αριθμού μεθόδων και συνόλων δεδομένων, τα αναλυτικά αποτελέσματα της πειραματικής διαδικασίας απεικονίζονται σε εκτενείς πίνακες. Ωστόσο, δεδομένων των περιορισμών χώρου και του γεγονότος ότι οι μεγάλες αριθμητικές αναπαραστάσεις μπορεί να είναι περισσότερο αποτροσανατολιστικές παρά επεξηγηματικές, δεν παρατίθενται εδώ οι

πλήρεις πίνακες αποτελεσμάτων. Αντ' αυτού, επιλέχθηκαν πιο αποδοτικοί τρόποι παρουσίασης των αποτελεσμάτων, όπως οι πίνακες νικών/ισοπαλιών/ηττών (w/t/l), τα διαγράμματα κρίσιμων διαφορών (cd diagrams) και τα θηκογράμματα. Οι πλήρεις πίνακες αποτελεσμάτων των τριών μετρικών είναι διαθέσιμοι στον ακόλουθο σύνδεσμο: https://rb.gy/7lqayb.

Όπως έχει ήδη αναφερθεί, για να διασφαλιστεί η αξιοπιστία των αποτελεσμάτων και να διερευνηθεί η στατιστική ανεξαρτησία των συγκρινόμενων μεθόδων πρόβλεψης, η πειραματική διαδικασία συνοδεύτηκε από στατιστικούς ελέγχους. Ειδικότερα, εφαρμόστηκε ο μη παραμετρικός έλεγχος Friedman (all vs. all), ακολουθούμενος από τον post-hoc έλεγχο Bonferroni-Dunn, με την παράμετρο α να έχει καθοριστεί στο 0.05. Ο έλεγχος Friedman εξετάζει αν ισχύει η μηδενική υπόθεση H_0, η οποία υποστηρίζει ότι οι μέσες τιμές των αποτελεσμάτων δύο ή περισσότερων αλγορίθμων δεν διαφέρουν. Εάν η μηδενική υπόθεση απορριφθεί, σημαίνει ότι υπάρχουν διαφορές μεταξύ των μεθόδων. Επιπρόσθετα, ο έλεγχος Bonferroni-Dunn εντοπίζει συγκεκριμένες ομάδες μεθόδων που εμφανίζουν στατιστικά σημαντικές αποκλίσεις. Οι στατιστικές αναλύσεις δεν αποσκοπούν μόνο στον καθορισμό του βαθμού ομοιότητας ή διαφοροποίησης μεταξύ των μεθόδων, αλλά επιπλέον, ο έλεγχος Friedman ταξινομεί τις εξεταζόμενες μεθόδους, αναδεικνύοντας εκείνες που υπερέχουν συγκριτικά με τις υπόλοιπες.

Δεδομένου ότι χρησιμοποιήθηκαν τρεις διαφορετικές μετρικές αξιολόγησης, η συνέχεια της παρούσας ενότητας χωρίζεται αναλόγως, ώστε κάθε μια από τις επόμενες υποενότητες να επικεντρώνεται σε μια από τις μετρικές MAE, RMSE και MAPE, αντίστοιχα. Για κάθε μία από αυτές τις μετρικές, παρουσιάζονται τα ακόλουθα:

1. Ένας πίνακας νικών/ισοπαλιών/ηττών (w/t/l), ο οποίος καταγράφει πόσες φορές μία μέθοδος υπερτερεί, ισοφαρίζει ή χάνει έναντι μιας άλλης.

2. Ένας πίνακας που περιέχει την κατάταξη Friedman.

3. Ένα διάγραμμα κρίσιμων διαφορών (cd-diagram), το οποίο απεικονίζει τις μεθόδους που θεωρούνται στατιστικά εξαρτημένες.

4. Ένα θηκόγραμμα, που παρέχει μια γραφική σύνοψη των αποτελεσμάτων της εκάστοτε μετρικής.

Οι πίνακες w/t/l παρέχουν πληροφορίες για το πόσες φορές ένας αλγόριθμος ξεπερνάει έναν άλλο κατά μία κρίσιμη διαφορά, υποδεικνύοντας εάν οι μέθοδοι είναι στατιστικά εξαρτημένες. Κάθε κελί (i, j) του πίνακα περιλαμβάνει μια τριάδα αριθμών, που δηλώνουν τις νίκες, τις ισοπαλίες και τις ήττες του αλγορίθμου της γραμμής i έναντι του αλγορίθμου της στήλης j. Ως αποτέλεσμα, οι πίνακες w/t/l είναι πάντα συμμετρικοί ως προς τη διαγώνιο, ενώ όλες οι τιμές της διαγωνίου είναι μηδενικές.

Στην παρούσα μελέτη, μια νίκη ή μια ήττα θεωρείται ότι συμβαίνει όταν η διαφορά στην απόδοση των αλγορίθμων, με βάση τη συγκεκριμένη μετρική, υπερβαίνει το 0.01; διαφορετικά, καταγράφεται ως ισοπαλία. Στους πίνακες κατάταξης Friedman, οι αλγόριθμοι ταξινομούνται από τον καλύτερο προς τον χειρότερο, σύμφωνα με την υπό εξέταση μετρική. Όσο μικρότερη είναι η τιμή κατάταξης που αποδίδει ο έλεγχος Friedman, τόσο υψηλότερη είναι η απόδοση της μεθόδου.

Στα διαγράμματα κρίσιμων διαφορών, η βαθμολογία κατάταξης Friedman χρησιμοποιείται για τον υπολογισμό της κρίσιμης διαφοράς. Όταν η αριθμητική διαφορά μεταξύ των κατατάξεων δύο μεθόδων υπερβαίνει την κρίσιμη διαφορά, τότε οι μέθοδοι θεωρούνται στατιστικά διακριτές.

Πίνακας 5.3: N/I/H για τη μετρική αξιολόγησης MAE

	ARIMA	Gluonts	Prophet	HWAAS	HWAMS	NBEATS	TBAT	TATS	TBATS1	TBATP1	TBATS2
ARIMA	0/0/0	19/1/20	10/0/38	17/0/23	18/0/22	27/1/3	11/2/27	23/3/14	17/1/22	13/3/34	10/0/30
Gluonts	20/1/19	0/0/0	15/0/25	21/0/19	22/0/18	22/0/18	17/1/22	22/1/17	19/1/20	19/1/20	19/1/20
Prophet	30/0/10	25/0/15	0/0/0	29/0/11	31/0/9	31/0/9	22/1/17	29/0/11	25/1/14	24/1/15	22/0/18
HWAAS	22/0/18	19/0/21	11/0/29	0/0/0	15/5/20	27/0/13	16/0/24	25/0/15	17/0/23	17/0/23	19/0/21
HWAMS	23/1/17	18/0/22	9/0/31	13/0/27	0/0/0	14/0/26	0/0/0	12/0/28	12/0/28	12/0/28	11/0/29
NBEATS	13/0/27	18/0/22	9/0/31	13/0/27	14/0/26	0/0/0	12/0/28	12/0/28	12/0/28	12/0/28	11/0/29
TBAT	27/2/11	22/1/17	17/1/22	24/0/16	25/0/15	28/0/12	0/0/0	25/9/6	9/2/23	8/2/26	8/2/26
TATS	14/3/23	17/3/20	12/1/27	19/0/21	20/0/20	26/0/14	25/9/6	0/0/0	8/9/23	8/9/23	8/9/23
TBATS1	17/1/22	21/0/19	14/1/25	21/0/19	21/0/19	29/0/11	5/24/11	22/8/10	0/0/0	19/19/12	7/18/15
TBATP1	24/3/13	20/1/19	15/1/24	23/0/17	23/0/17	28/0/12	3/28/9	23/9/8	12/9/17	0/0/0	7/18/15
TBATS2	30/0/10	21/0/19	18/0/22	19/0/21	23/0/17	29/0/11	6/26/8	27/8/5	15/18/7	15/18/7	0/0/0

5.4.3.1 Αποτελέσματα της μετρικής MAE

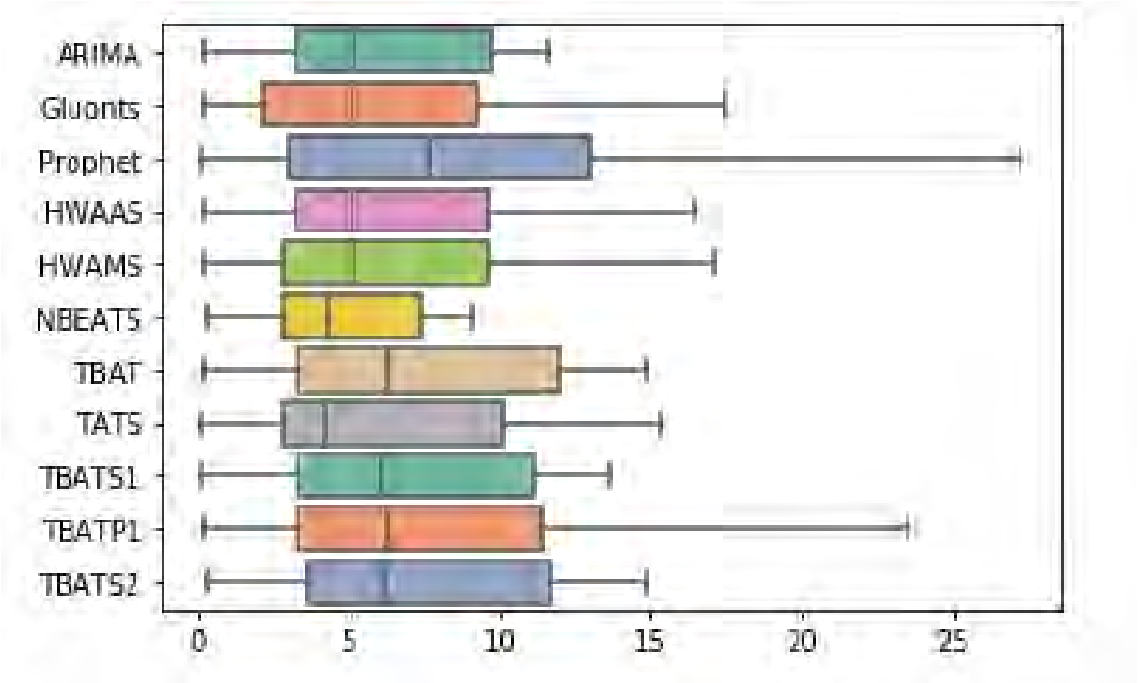
Σχετικά με τη μετρική MAE, και σύμφωνα με τα αποτελέσματα της κατάταξης που προκύπτει από τον έλεγχο Friedman, όπως αυτά αποτυπώνονται στον Πίνακα 5.4, ο αλγόριθμος NBEATS επιτυγχάνει τις χαμηλότερες τιμές MAE, κάτι που υποδηλώνει την καλύτερη επίδοση σε σύγκριση με τους ανταγωνιστές του. Ο αλγόριθμος TATS ακολουθεί με ελαφρώς υψηλότερες τιμές MAE, ενώ ο Prophet εμφανίζει τις υψηλότερες τιμές MAE, υποδηλώνοντας τη χαμηλότερη ακρίβεια προβλέψεων μεταξύ των εξεταζόμενων μεθόδων.

Πίνακας 5.4: Κατάταξη Friedman για τις μετρικές MAE, RMSE, MAPE

	MAE		RMSE		MAPE	
ΑΑ	Μέθοδος	Τιμή	Μέθοδος	Τιμή	Μέθοδος	Τιμή
1	NBEATS	4.225	TATS	4.375	NBEATS	3.15
2	TATS	4.525	NBEATS	4.6	Gluonts	4.7
3	ARIMA	5.2	ARIMA	5.15	TATS	5.3
4	HWAAS	5.8	HWAAS	5.85	HWAAS	5.475
5	HWAMS	5.9	Gluonts	5.875	ARIMA	5.55
6	Gluonts	5.975	HWAMS	5.9	HWAMS	5.625
7	TBATS1	6.2125	TBATS1	6.1625	TBATS1	6.4875
8	TBATP1	6.3875	TBATP1	6.4625	TBATP1	6.6125
9	TBATS2	7.0375	TBATS2	6.9625	TBATS2	7.3875
10	TBAT	7.0375	TBAT	7.0125	TBAT	7.5625
11	Prophet	7.7	Prophet	7.65	Prophet	8.15

Συγκεκριμένα, ο NBEATS κατατάσσεται πρώτος με τη χαμηλότερη μέση τιμή MAE (4.225), επιβεβαιώνοντας ότι έχει την καλύτερη απόδοση. Ο TATS βρίσκεται στη δεύτερη θέση με MAE 4.525. Οι αλγόριθμοι ARIMA, HWAA και HWAMS ακολουθούν με λίγο υψηλότερες τιμές (5.2 - 5.9), ενώ ο Prophet κατατάσσεται τελευταίος, με τη μεγαλύτερη τιμή MAE (7.7), υποδεικνύοντας τη χαμηλότερη αποτελεσματικότητα.

Τα αποτελέσματα αυτά είναι εμφανή και στα θηκογράμματα στο Σχήμα 5.2, όπου παρατηρείται ότι οι υπόλοιποι αλγόριθμοι (ARIMA, HWAA, HWAMS, TBATS κ.λπ.) παρουσιάζουν παρόμοια απόδοση μεταξύ τους, με σχετικά κοντινές τιμές σφάλματος. Επιπλέον, στον πίνακα N/I/H (Πίνακας 5.3) παρατηρούμε ότι ο NBEATS παρουσιάζει



Σχήμα 5.2: Θηκογράμματα της απόδοσης των μεθόδων στην μετρική MAE

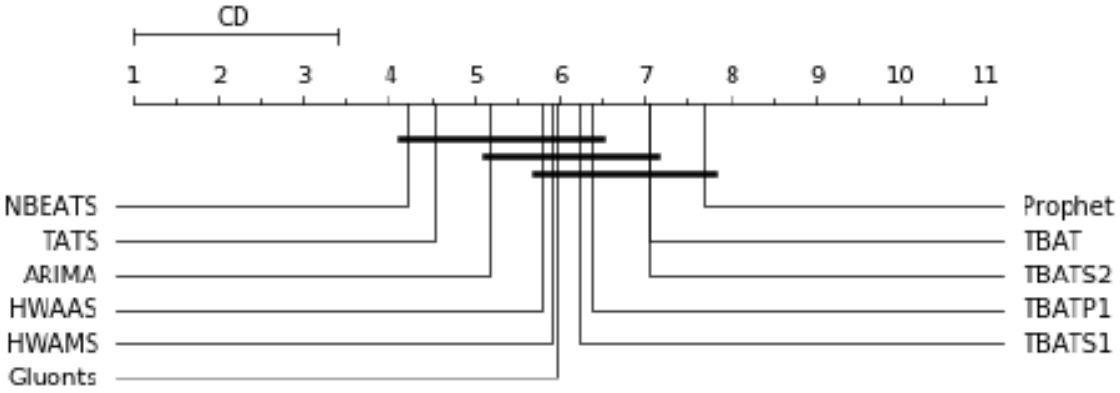
Πίνακας 5.5: N/I/H για τη μετρική αξιολόγησης RMSE

	ARIMA	Gluonts	Prophet	HWAAS	HWAMS	NBEATS	TBAT	TATS	TBATS1	TBATP1	TBATS2
ARIMA	0/0/0	19/1/20	12/0/28	16/2/22	16/2/22	22/8/18	11/2/27	24/2/14	18/1/21	13/3/24	10/1/29
Gluonts	20/1/19	0/0/0	14/0/26	21/0/19	22/0/18	22/0/18	18/2/22	19/0/21	18/0/22	18/0/22	18/0/22
Prophet	28/0/12	26/0/14	0/0/0	29/0/11	27/1/12	30/0/10	23/0/17	29/0/11	27/0/13	25/0/15	22/0/18
HWAAS	22/2/16	19/0/21	11/0/29	0/0/0	17/4/19	27/0/13	20/4/16	26/0/14	18/0/22	18/0/22	18/0/22
HWAMS	24/0/16	18/0/22	12/1/27	19/4/17	0/0/0	25/1/15	17/0/23	19/1/20	18/0/22	18/0/22	18/0/22
NBEATS	18/0/22	18/0/22	10/0/30	15/0/25	10/0/30	0/0/0	13/0/27	19/0/21	13/0/27	13/0/27	12/0/28
TBAT	27/2/11	22/0/18	17/0/23	24/0/16	24/1/15	27/0/13	0/0/0	26/8/6	12/1/25	9/2/28	9/26/5
TATS	14/2/24	19/0/21	11/0/29	20/0/22	19/1/20	19/0/21	26/8/6	0/0/0	24/8/8	14/8/7	6/7/27
TBATS1	21/1/18	21/0/19	13/0/27	21/0/19	21/0/19	13/0/27	5/24/11	24/8/8	0/0/0	9/19/12	7/19/14
TBATP1	24/3/13	22/0/18	15/0/25	22/0/18	22/0/18	13/0/27	3/28/9	14/8/7	9/19/12	0/0/0	7/19/14
TBATS2	29/1/10	22/0/18	18/0/22	18/0/22	24/0/16	28/0/12	5/26/9	27/7/6	14/18/7	14/19/7	0/0/0

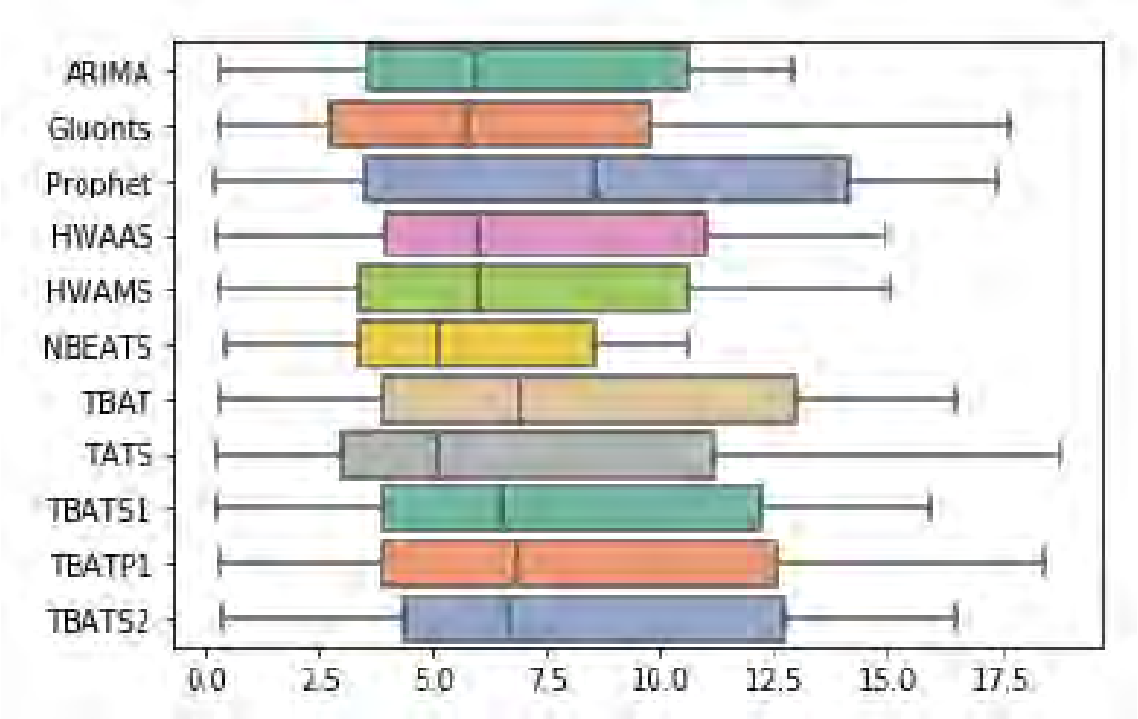
τις περισσότερες νίκες, ενώ ο TBAT τις λιγότερες. Τέλος, ο στατιστικός έλεγχος που πραγματοποιήθηκε απέρριψε τη μηδενική υπόθεση H_0, καθώς η τιμή p ήταν ίση με 0.00001, επιβεβαιώνοντας ότι δεν έχουν όλοι οι αλγόριθμοι την ίδια απόδοση. Επομένως, οι διαφορές που παρατηρήθηκαν στα αποτελέσματα είναι στατιστικά σημαντικές και δεν οφείλονται σε τυχαιότητα. Οι μέθοδοι με παραπλήσια συμπεριφορά μπο- ρούν να παρατηρηθούν στο διάγραμμα των κρίσιμων διαφορών του Σχήματος 5.3, όπου ενώνονται με σκούρες οριζόντιες γραμμές.

5.4.3.2 Αποτελέσματα της μετρικής RMSE

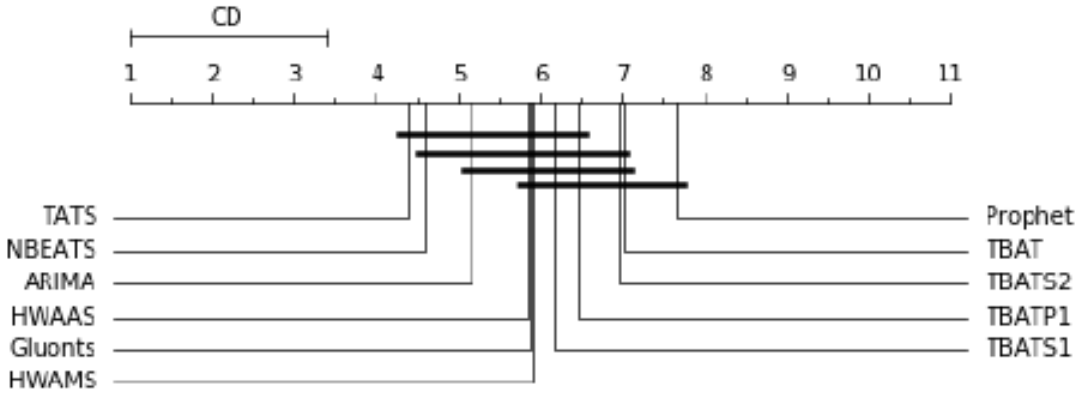
Σύμφωνα με την κατάταξη του Friedman όπως αυτή παρουσιάζεται στον 5.4, βλέπουμε ότι ο TATS καταλαμβάνει την πρώτη θέση με μέση τιμή 4.375, ακολουθούμενος από τον NBEATS με 4.6, επιβεβαιώνοντας ότι αυτοί οι δύο αλγόριθμοι ξεπερνούν τους ανταγωνιστές τους. Στον αντίποδα, ο Prophet συγκεντρώνει τη χειρότερη κατάταξη με 7.65, υποδηλώνοντας τη χαμηλότερη ακρίβεια στις προβλέψεις του με βάση τη μετρική RMSE.



Σχήμα 5.3: Διάγραμμα κρίσιμων διαφορών της μετρικής MAE



Σχήμα 5.4: Θηκογράμματα της απόδοσης των μεθόδων στην μετρική RMSE



Σχήμα 5.5: Διάγραμμα κρίσιμων διαφορών της μετρικής RMSE

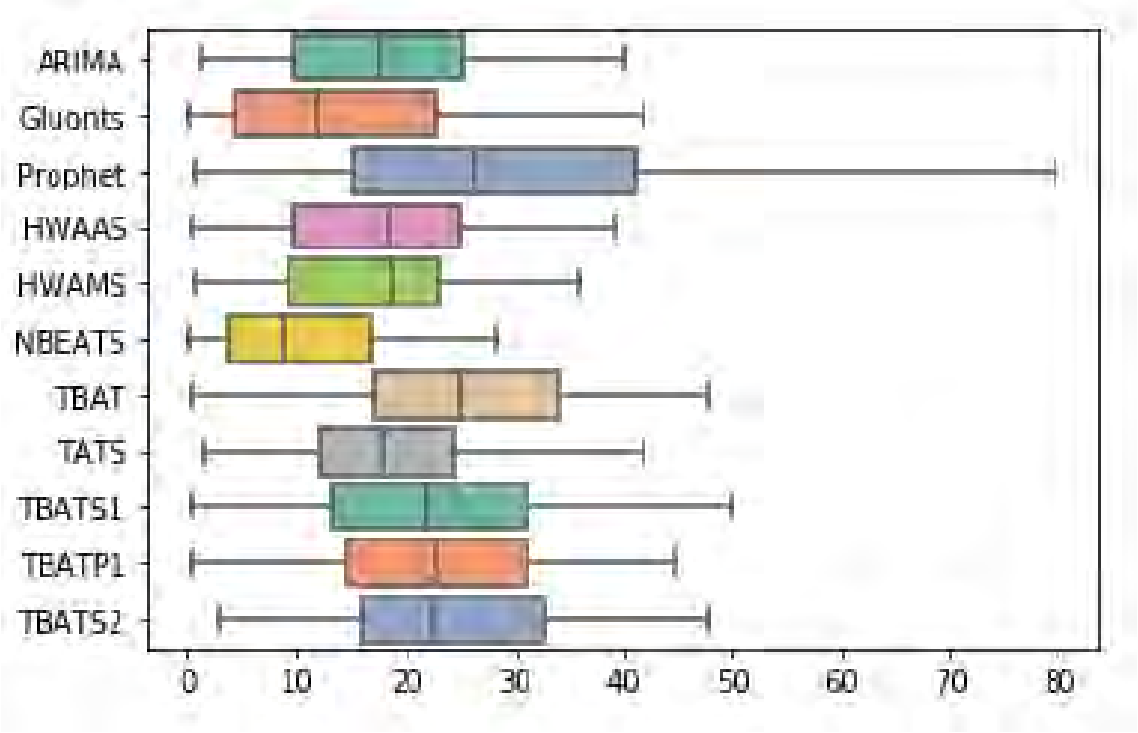
Πίνακας 5.6: N/I/H για τη μετρική αξιολόγησης MAPE

	ARIMA	Gluonts	Prophet	HWAAS	HWAMS	NBEATS	TBAT	TATS	TBATS1	TBATP1	TBATS2
ARIMA	0/0/0	24/0/16	11/0/29	18/1/21	20/0/20	28/0/12	12/0/28	23/0/17	16/1/23	11/0/29	11/0/29
Gluonts	16/0/24	0/0/0	9/0/31	18/0/22	20/0/20	22/0/18	12/0/29	16/0/24	13/0/27	12/0/28	11/0/29
Prophet	29/0/11	31/0/9	0/0/0	33/0/7	33/0/7	34/0/6	29/0/11	29/0/11	26/0/14	26/0/14	26/0/14
HWAAS	21/1/18	22/0/18	7/0/33	0/0/0	17/3/30	30/0/10	30/0/10	24/0/16	15/0/25	17/0/23	17/0/23
HWAMS	20/0/20	20/0/20	7/0/33	0/3/17	0/0/0	33/0/7	30/0/10	21/1/18	16/2/22	15/0/25	15/0/25
NBEATS	12/0/28	18/0/22	6/0/34	10/0/30	7/0/33	0/0/0	12/0/28	30/0/10	36/0/32	12/0/28	11/0/29
TBAT	28/0/12	29/0/11	16/0/24	30/0/10	28/2/10	35/0/5	0/0/0	23/8/9	11/2/25	10/2/27	9/26/5
TATS	17/0/23	24/0/16	11/0/29	16/0/24	18/1/21	33/0/7	9/8/23	0/0/0	11/1/22	12/8/20	8/7/25
TBATS1	23/1/16	28/0/12	16/0/24	25/0/15	22/0/16	34/0/6	21/1/19	11/19/10	0/0/0	17/0/15	7/18/15
TBATP1	23/1/16	28/0/12	16/0/24	25/0/15	22/0/16	34/0/6	21/1/19	11/19/10	17/0/15	0/0/0	7/18/15
TBATS2	29/0/11	29/0/11	19/0/21	23/0/17	25/0/15	35/0/5	6/25/9	25/8/7	15/18/7	15/18/7	0/0/0

Οι παραπάνω παρατηρήσεις επιβεβαιώνονται και στο θηκόγραμμα που παρουσιάζεται στο Σχήμα 5.4. Οι αλγόριθμοι NBEATS και TATS επιτυγχάνουν τις χαμηλότερες τιμές, γεγονός που υποδηλώνει καλύτερη απόδοση. Αντίθετα, ο Prophet παρουσιάζει τη χειρότερη απόδοση, ακολουθούμενος από τον TBAT, ο οποίος σημειώνει επίσης υψηλές τιμές RMSE. Επιπλέον, στον πίνακα N/I/H, ο οποίος παρουσιάζεται στον Πίνακα 5.4, ο TATS καταγράφει τις περισσότερες νίκες, επιβεβαιώνοντας την ανωτερότητα της απόδοσής του, ενώ οι Prophet και TBAT συγκεντρώνουν τις περισσότερες ήττες. Σχετικά με τη στατιστική σημαντικότητα των αποτελεσμάτων, ο στατιστικός έλεγχος που διεξήχθη για να εξετάσει αν οι αλγόριθμοι έχουν παρόμοιες επιδόσεις οδήγησε στην απόρριψη της H_0, καθώς η τιμή του p ήταν ίση με 0.00003, επιβεβαιώνοντας ότι όλες οι μέθοδοι δεν είναι στατιστικά ισοδύναμες. Στο αντίστοιχο διάγραμμα κρίσιμων διαφορών στο Σχήμα 5.5, απεικονίζονται οι ομάδες αλγορίθμων που εμφανίζουν στατιστικά παρόμοιες επιδόσεις. Παρατηρούμε ότι οι κορυφαίοι αλγόριθμοι NBEATS, TATS ξεχωρίζουν από τους υπόλοιπους αν και παρουσιάζουν παραπλήσια συμπεριφορά.

5.4.3.3 Αποτελέσματα της μετρικής MAPE

Σύμφωνα με την κατάταξη του στατιστικού ελέγχου Friedman στον Πίνακα 5.4 σε σχέση με τη μετρική MAPE, ο NBEATS καταλαμβάνει την πρώτη θέση με μέση τιμή 3.15, επιβεβαιώνοντας ότι υπερέχει έναντι των άλλων μεθόδων. Ο Gluonts ακολουθεί στη δεύτερη θέση με 4.7, ενώ ο TATS καταλαμβάνει την τρίτη θέση με 5.3. Από την άλλη, ο Prophet εμφανίζει την υψηλότερη τιμή MAPE (8.15), γεγονός που υποδηλώνει τη χαμηλότερη ακρίβεια στις προβλέψεις του.



Σχήμα 5.6: Θηκογράμματα της απόδοσης των μεθόδων στην μετρική MAPE

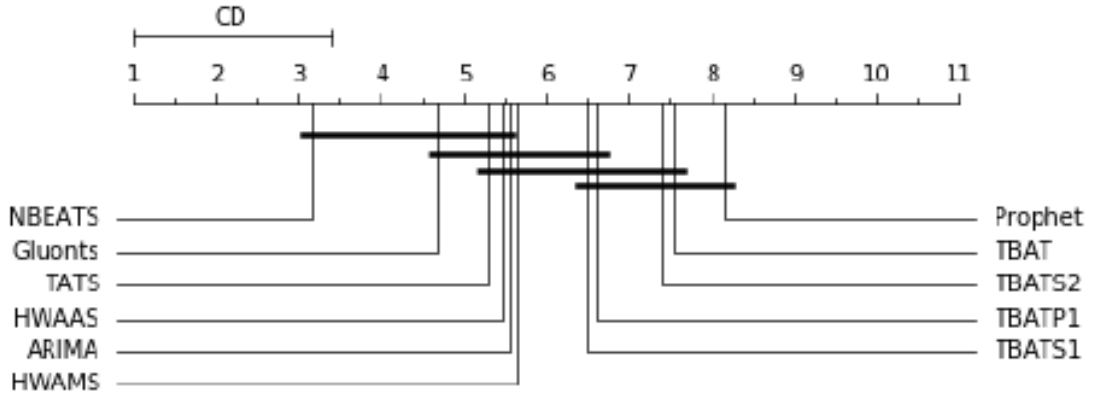
Τα παραπάνω αποτελέσματα επιβεβαιώνονται και από το θηκόγραμμα που παρουσιάζεται στο Σχήμα 5.6, το οποίο επίσης απεικονίζει την απόδοση των αλγορίθμων με βάση τη μετρική MAPE, όπου διακρίνεται ξεκάθαρα ότι ο NBEATS επιτυγχάνει το χαμηλότερο ποσοστό σφάλματος, ενώ ακολουθεί ο Gluonts. Αντίθετα, ο Prophet εμφανίζει το υψηλότερο ποσοστό σφάλματος, γεγονός που υποδηλώνει την κατώτερη απόδοσή του σε σχέση με τις υπόλοιπες μεθόδους.

Όπως είναι αναμενόμενο, και στον πίνακα N/I/H (Πίνακας 5.6), οι αλγόριθμοι NBEATS και Gluonts καταγράφουν τις περισσότερες νίκες. Αντίθετα, οι Prophet και TBAT συγκεντρώνουν τις περισσότερες ήττες, επιβεβαιώνοντας τη χαμηλότερη ακρίβεια των προβλέψεών τους. Επιπλέον, ο στατιστικός έλεγχος που πραγματοποιήθηκε απέρριψε τη μηδενική υπόθεση (H_0), καθώς η τιμή του p ήταν ίση με 0.00001, γεγονός που υποδηλώνει ότι υπάρχουν στατιστικά σημαντικές διαφορές μεταξύ των αλγορίθμων, ενώ τα αποτελέσματα του post-hoc Bonferroni–Dunn ελέγχου αποτυπώνονται στα διαγράμματα των κρίσιμων διαφορών του Σχήματος 5.7.

5.4.4 Σύνοψη Αποτελεσμάτων και Συμπεράσματα

Λαμβάνοντας υπόψιν τα διαφορετικά πλεονεκτήματα και μειονεκτήματα των διάφορων μετρικών αξιολόγησης, η χρήση τριών διαφορετικών μετρικών MAE, RMSE και MAPE επιτρέπει μια πιο ολοκληρωμένη αξιολόγηση των συγκρινόμενων μεθόδων. Μια από τις βασικότερες διαφορές τους είναι ότι οι μετρικές MAE και RMSE είναι εξαρτώμενες από την κλίμακα των δεδομένων, ενώ η MAPE προσφέρει ποσοστιαία εκτίμηση του σφάλματος και για αυτό το λόγο κρίθηκε αναγκαία η ενσωμάτωσή της στην αξιολόγηση των αποτελεσμάτων.

Και οι τρεις μετρικές αναδεικνύουν την υπεροχή της μεθόδου NBEATS, καθώς βρίσκεται στην πρώτη θέση στις μετρικές MAE και MAPE και στη δεύτερη στη με-



Σχήμα 5.7: Διάγραμμα κρίσιμων διαφορών της μετρικής MAPE

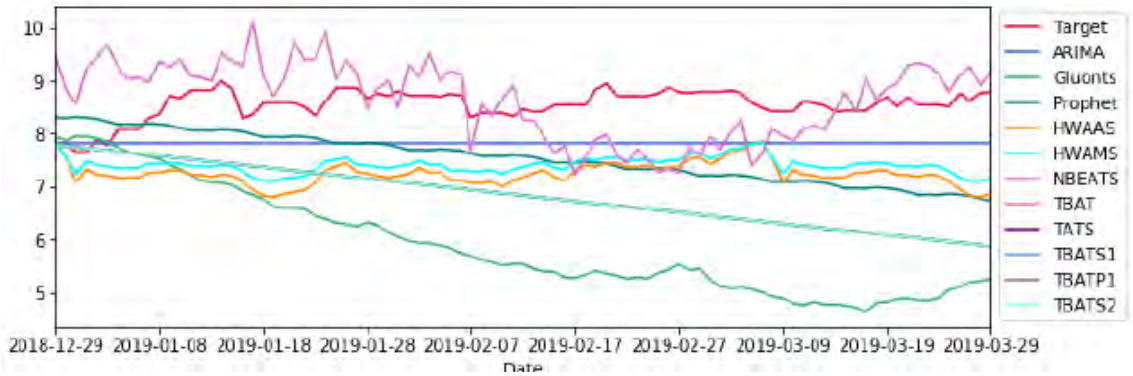
τρική RMSE. Ιδιαίτερα υποσχόμενος αναδεικνύεται και ο TATS, καταλαμβάνοντας την πρώτη θέση στη μετρική RMSE, τη δεύτερη στη MAE και την τρίτη στη MAPE. Στον αντίποδα, η μέθοδος Prophet καταλαμβάνει την τελευταία θέση στην κατάταξη του Friedman και στις τρεις μετρικές, εμφανίζοντας τις υψηλότερες τιμές σφάλματος. Η μεγάλη διαφορά του Prophet από τους υπόλοιπους αλγορίθμους υποδηλώνει ότι δεν είναι ιδανικός για προβλέψεις σε σύντομες χρονοσειρές. Το γεγονός αυτό ενισχύει την εμπειρική αρχή ότι δεν υπάρχει μία μέθοδος που να λύνει ικανοποιητικά όλα τα προβλήματα πρόβλεψης. Ο Prophet, αν και θεωρείται state-of-the-art προβλεπτικός αλγόριθμος, ικανός να παράγει ακριβείς προβλέψεις όταν εφαρμόζεται σε δεδομένα μακράς διάρκειας, στην περίπτωσή μας εμφανίζει χαμηλή απόδοση λόγω του περιορισμένου χρονικού εύρους των δεδομένων.

Ένα αξιοσημείωτο γεγονός είναι ότι η υλοποίηση του DeepAR μέσω του Gluonts παρουσίασε σημαντική βελτίωση στην κατάταξη με βάση τη MAPE, σε σύγκριση με τις MAE και RMSE, καθώς κατατάχθηκε έκτη και πέμπτη στις MAE και RMSE αντίστοιχα, αλλά βρέθηκε στη δεύτερη θέση στη μετρική MAPE. Αυτό μπορεί να υποδηλώνει καλή απόδοση, δεδομένου ότι η MAPE εκφράζει ποσοστιαία σφάλματα, σε αντίθεση με τις MAE και RMSE, που εξαρτώνται από την κλίμακα των δεδομένων. Παράλληλα, αυτό το εύρημα υπογραμμίζει την ανάγκη χρήσης πολλαπλών μετρικών αξιολόγησης για την καλύτερη κατανόηση των αποτελεσμάτων.

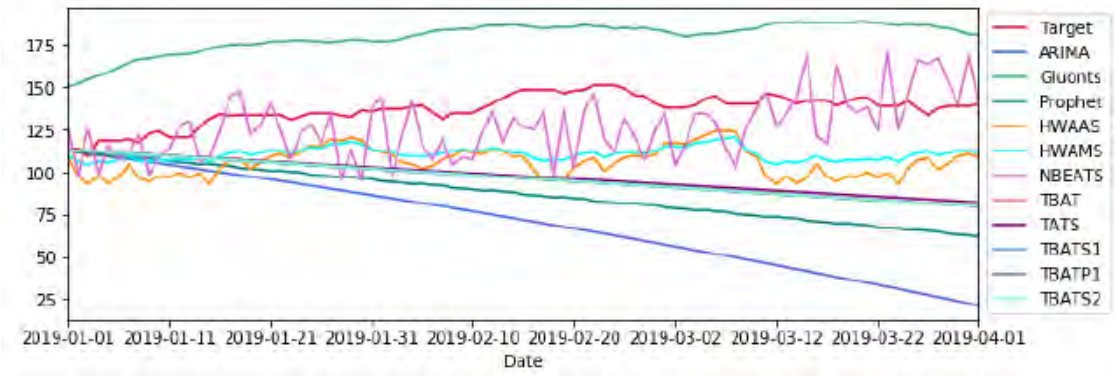
Στις ενδιάμεσες θέσεις της κατάταξης συναντάμε τις ίδιες μεθόδους σε παραπλήσιες θέσεις και στις τρεις μετρικές. Οι ARIMA και HWAAS παρουσίασαν αρκετά καλή απόδοση, καθώς και στις τρεις μετρικές βρέθηκαν μεταξύ της τρίτης και πέμπτης θέσης. Αντίθετα, ο TBAT κατατάχθηκε προτελευταίος σε όλες τις περιπτώσεις, ενώ οι υπόλοιπες μέθοδοι της ίδιας οικογένειας (TBATS1, TBATP1 και TBATS2) κατατάχθηκαν αντίστοιχα στην έβδομη, όγδοη και ένατη θέση.

Ενδεικτικά διαγράμματα που απεικονίζουν την απόδοση πρόβλεψης των μεθόδων σε in-sample δεδομένα παρουσιάζονται στα Σχήματα 5.8, 5.9 και 5.10.

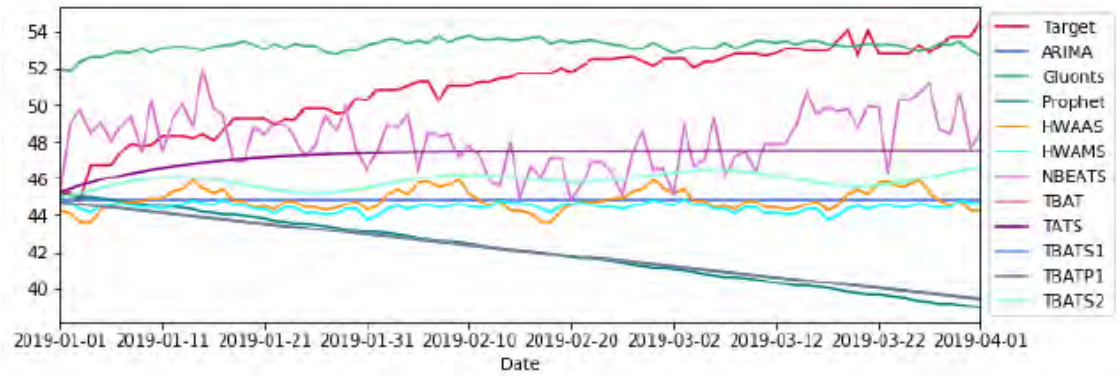
Πρέπει να τονίσουμε ότι, σε όλες τις περιπτώσεις, οι μέθοδοι με την καλύτερη και τη χειρότερη απόδοση ήταν στατιστικά ανεξάρτητες, γεγονός που επιβεβαιώνει τη συγκριτική εγκυρότητα των αποτελεσμάτων. Επιπλέον, τα διαγράμματα κρίσιμων διαφορών παρουσιάζουν ομάδες αλγορίθμων που εμφανίζουν στατιστικά παρόμοια απόδοση, επιβεβαιώνοντας και ενισχύοντας περαιτέρω το συμπέρασμα ότι δεν υπάρχει μία μοναδική, ιδανική μέθοδος – είτε στατιστική είτε βασισμένη στη Μηχανική Μάθηση – που να είναι κατάλληλη για όλα τα προβλήματα πρόβλεψης.



Σχήμα 5.8: In-sample προβλέψεις της μετοχής F



Σχήμα 5.9: In-sample προβλέψεις της μετοχής SPOT



Σχήμα 5.10: In-sample προβλέψεις της μετοχής ORCL

5.5 Πρόβλεψη μονομεταβλητών χρηματοοικονομικών χρονοσειρών με χρήση Ανάλυσης Συναισθήματος

Στο δεύτερο μέρος αυτού του κεφαλαίου, η έρευνα που αφορά την πρόβλεψη χρονοσειρών επικεντρώνεται στη χρήση Ανάλυσης Συναισθήματος για τη βελτίωση των προβλέψεων σε μονομεταβλητές χρονοσειρές οικονομικών δεδομένων. Η ανάγκη για ακριβέστερες προβλέψεις έχει οδηγήσει στη χρήση νέων τεχνικών που στοχεύουν στη βελτίωση των εξαγόμενων προβλέψεων. Είναι εύλογο πως σε διαδικασίες πρόβλεψης που σχετίζονται με οικονομικά, κοινωνικά και τεχνολογικά φαινόμενα, όπου η ανθρώπινη ψυχολογία και οι απόψεις παίζουν καθοριστικό ρόλο, η γνώση της άποψης της κοινής γνώμης μπορεί να έχει ευεργετικό ρόλο στην ανάπτυξη των μοντέλων πρόβλεψης. Τέτοια χαρακτηριστικά παραδείγματα αποτελούν περιπτώσεις από τον τομέα του marketing, όπως η ανάλυση της καταναλωτικής συμπεριφοράς και της δημοφιλίας προϊόντων, αλλά και η πρόβλεψη επιχειρησιακών κρίσεων. Επιπλέον, ενδεικτικά είναι και τα παραδείγματα από τον τομέα της υγείας και της πολιτικής, όπως η πρόβλεψη της εξέλιξης πανδημιών και η πρόβλεψη της πολιτικής πόλωσης αντίστοιχα. Στα χρηματοοικονομικά, εκτός από το φαινόμενο πρόβλεψης της τιμής κλεισίματος του χρηματιστηρίου που μελετάται σε αυτό το κεφάλαιο, άλλα χαρακτηριστικά παραδείγματα είναι αυτά της ανίχνευσης χρηματοοικονομικών κρίσεων αλλά και των λεγόμενων «φουσκών», καθώς και της πρόβλεψης της μεταβλητότητας της αγοράς.

Όλα τα παραπάνω φαινόμενα έχουν έναν κοινό παρονομαστή. Επηρεάζουν σε κάποιο βαθμό την κοινή γνώμη, και με τη σειρά τους, επηρεάζονται από αυτήν σε μεγαλύτερο ή μικρότερο βαθμό. Φαινόμενα που για κοινωνικούς, οικονομικούς ή άλλους λόγους απασχολούν την επικαιρότητα, τείνουν να αποτελούν κύριο θέμα όχι μόνο στις πρόσωπο με πρόσωπο συζητήσεις, αλλά και στα κοινωνικά δίκτυα. Δεν θα ήταν λοιπόν υπερβολή να πούμε πως η άποψη της κοινής γνώμης, σε μεγάλο βαθμό, αποτυπώνεται στις συζητήσεις που λαμβάνουν χώρα στα κοινωνικά δίκτυα, και αποτελεί εύλογη συνέχεια η επιθυμία για αξιοποίηση αυτής της πληροφορίας.

Σε συνέχεια της προηγούμενης έρευνας, όπου διάφορα μοντέλα χρησιμοποιήθηκαν για την πρόβλεψη 40 μετοχών χρηματιστηρίου, στραφήκαμε στη χρήση της Ανάλυσης Συναισθήματος για την επίλυση του ίδιου προβλήματος. Συγκεκριμένα, ένας μεγάλος αριθμός προηγμένων μεθόδων τίθεται υπό ένα πειραματικό πλαίσιο που περιλαμβάνει πολλαπλές διαμορφώσεις χαρακτηριστικών εισόδου, ενσωματώνοντας ποσοτικοποιημένες τιμές συναισθηματικών στάσεων με τη μορφή χρονοσειρών. Αυτές οι χρονοσειρές αποτελούνται από συναισθηματικές βαθμολογίες που εξάγονται από το Twitter, χρησιμοποιώντας τρεις διαφορετικές μεθόδους Ανάλυσης Συναισθήματος.

Όσον αφορά τις μεθόδους πρόβλεψης, εξετάζονται τόσο στατιστικά μοντέλα όσο και μέθοδοι Μηχανικής Μάθησης. Στις μεθόδους Μηχανικής Μάθησης που χρησιμοποιήθηκαν, συμπεριλαμβάνονται επίσης σύγχρονες τεχνικές Βαθιάς Μάθησης και άλλες προηγμένες μέθοδοι, οι οποίες συγκρίνονται με βάση την απόδοσή τους. Επιπλέον, δοκιμάστηκαν διάφορες διαμορφώσεις χαρακτηριστικών εισόδου που σχετίζονται με την Ανάλυση Συναισθήματος. Διερευνήθηκαν έτσι δύο ξεχωριστές μελέτες περίπτωσης. Στην πρώτη μελέτη, οι αξιολογήσεις οργανώθηκαν σύμφωνα με τις χρησιμοποιούμενες μεθόδους πρόβλεψης και οι συγκρίσεις έγιναν με βάση αυτή την ταξινόμηση. Στη δεύτερη μελέτη, οι συγκρίσεις επικεντρώθηκαν στις διαφορετικές ρυθμίσεις των χαρακτηριστικών εισόδου, διερευνώντας αν υπάρχουν διαμορφώσεις χαρακτηριστικών που

βελτιώνουν την προβλεπτική ικανότητα των χρησιμοποιούμενων μοντέλων.

5.5.1 Πειραματική διαδικασία

5.5.1.1 Προεπεξεργασία δεδομένων και εξαγωγή συναισθηματικών πολικοτήτων

Έχει ήδη αναφερθεί ότι ο πυρήνας της παρούσας εργασίας αποτελείται από μια πειραματική διαδικασία που αποσκοπεί, αφενός, στην αξιολόγηση της αποδοτικότητας διαφόρων σύγχρονων αλγορίθμων και αφετέρου, στην ενσωμάτωση της ανάλυσης συναισθήματος σε προβλέψεις τιμών. Ως εκ τούτου, πραγματοποιήθηκαν συνολικά 28.512 πειράματα, λαμβάνοντας υπόψη:

- 16 σύνολα δεδομένων,
- 22 διαφορετικούς συνδυασμούς χαρακτηριστικών,
- 27 αλγορίθμους,
- 3 διαφορετικά χρονικά παράθυρα πρόβλεψης.

Τα δεδομένα που χρησιμοποιήθηκαν περιλάμβαναν χρονοσειρές με τις ημερήσιες τιμές κλεισίματος μετοχών από διάφορες εταιρείες, καλύπτοντας περίοδο τριών ετών (2 Ιανουαρίου 2018 - 24 Δεκεμβρίου 2020). Επιπλέον, χρησιμοποιήθηκαν δεδομένα Ανάλυσης Συναισθήματος, που προέκυψαν από σχετικές δημοσιεύσεις στο Twitter. Για την εξαγωγή των συναισθηματικών βαθμολογιών χρησιμοποιήθηκαν τρεις διαφορετικές μέθοδοι ανάλυσης συναισθήματος.

Οι χρονοσειρές των συναισθηματικών βαθμολογιών, μαζί με τις τιμές κλεισίματος, υποβλήθηκαν σε στρατηγικές κινητού μέσου 7 και 14 ημερών. Εκτός από αυτά, χρησιμοποιήθηκαν και συνδυασμοί τους, όπως φαίνεται στον Πίνακα 5.7.

Ξεκινώντας με την περιγραφή των δεδομένων, πρέπει να σημειωθεί πως οι μετοχές που χρησιμοποιήθηκαν αποτελούν ένα τυχαίο υποσύνολο αυτών που χρησιμοποιήθηκαν στην παράγραφο 5.2. Οι μετοχές που χρησιμοποιήθηκαν, καθώς και οι συντομογραφίες τους, βρίσκονται στον Πίνακα 5.8. Σημειώνεται ξανά ότι όλα τα σύνολα δεδομένων καλύπτουν περίοδο τριών ετών (2 Ιανουαρίου 2018 - 24 Δεκεμβρίου 2020).

Τα αρχικά χαρακτηριστικά περιλάμβαναν τιμές κλεισίματος και τρεις επιπλέον χρονοσειρές συναισθηματικών βαθμολογιών, κάθε μια από τις οποίες εξήχθη μέσω μιας διαφορετικής τεχνικής, όπως θα επεξηγηθεί στη συνέχεια. Μετά την εφαρμογή κινητών μέσων 7 και 14 ημερών, εξήχθησαν συνολικά 14 χαρακτηριστικά. Το τελικό σύνολο δεδομένων προέκυψε από επεξεργασία των αρχικών δεδομένων χρηματιστηριακών τιμών και από ανάλυση συναισθήματος ενός εκτεταμένου συνόλου αναρτήσεων στο Twitter.

Για την εξαγωγή των δεδομένων χρησιμοποιήθηκε το Twitter Intelligence Tool (TWINT), ένα ισχυρό εργαλείο εξαγωγής δεδομένων από το Twitter, υλοποιημένο σε Python. Για κάθε μετοχή, για την άντληση των δεδομένων από το Twitter, χρησιμοποιήθηκε ως λέξη κλειδί το όνομα της μετοχής, όπως αυτό εντοπίζεται στη σελίδα Yahoo! Finance και αντίστοιχα στον Πίνακα 5.8.

Στη συνέχεια, τα εξαγόμενα tweets ομαδοποιήθηκαν ανά ημέρα, αφού προηγήθηκε καθαρισμός τους. Συγκεκριμένα, αφαιρέθηκαν τυχόν υπερσύνδεσμοι και URLs χρησιμοποιώντας τη βιβλιοθήκη Re της Python. Στη συνέχεια, κάθε tweet μετατράπηκε

Πίνακας 5.7: Διαμορφώσεις συναισθηματικών αποτιμήσεων

ΑΑ	Συντομογραφία	Συναισθηματική διαμόρφωση
1	NS	No Sentiment
2	B	TextBlob
3	V	Vader
4	F	FinBERT
5	B7	Rolling Mean 7 TextBlob
6	V7	Rolling Mean 7 Vader
7	F7	Rolling Mean 7 FinBERT
8	BV	TextBlob and Vader
9	BF	TextBlob and FinBERT
10	BB7	TextBlob and Rolling Mean 7 TextBlob
11	BV7	TextBlob and Rolling Mean 7 Vader
12	BF7	TextBlob and Rolling Mean 7 FinBERT
13	VF	Vader and FinBERT
14	VB7	Vader and Rolling Mean 7 TextBlob
15	VV7	Vader and Rolling Mean 7 Vader
16	VF7	Vader and Rolling Mean 7 FinBERT
17	FB7	FinBERT and Rolling Mean 7 TextBlob
18	FV7	FinBERT and Rolling Mean 7 Vader
19	FF7	FinBERT and Rolling Mean 7 FinBERT
20	B7V7	Rolling Mean 7 TextBlob and Rolling Mean 7 Vader
21	B7F7	Rolling Mean 7 TextBlob and Rolling Mean 7 FinBERT
22	V7F7	Rolling Mean 7 Vader and Rolling Mean 7 FinBERT

σε πεζούς χαρακτήρες και διασπάστηκε σε λέξεις. Έπειτα, αφαιρέθηκαν ανεπιθύμητες φράσεις από μια χειροκίνητα παραγόμενη λίστα, καθώς και διάφορες αριθμητικές συμβολοσειρές.

Μετά την πραγματοποίηση των απαραίτητων συνδέσεων για την επαναφορά κάθε κειμένου στην αρχική του δομή, κάθε tweet υποβλήθηκε σε διαδικασία διαχωρισμού προτάσεων χρησιμοποιώντας τη βιβλιοθήκη NLTK. Τέλος, εφαρμόστηκε αφαίρεση σημείων στίξης με τη χρήση της βιβλιοθήκης String.

Στο Σχήμα 5.8, εμφανίζονται διαγραμματικά τα βήματα της παραπάνω διαδικασίας.

Με το πέρας της προεπεξεργασίας, τα συλλεγμένα δεδομένα είναι έτοιμα προς χρήση των μεθόδων Ανάλυσης Συναισθήματος, από τις οποίες εξάγεται και η αντίστοιχη πολικότητά τους (sentiment score), με στόχο την ποσοτικοποίηση της λεκτικής πληροφορίας. Για το σκοπό αυτό, χρησιμοποιήθηκαν τρεις διαφορετικές μέθοδοι ανάλυσης συναισθήματος. Συγκεκριμένα, η διαδικασία περιλάμβανε την εξαγωγή πολικοτήτων με τη χρήση των εργαλείων TextBlob, VADER και FinBERT. Ενώ τα δύο πρώτα αποτελούν εργαλεία γενικής χρήσης και χρησιμοποιούνται σε ένα μεγάλο εύρος εφαρμογών, η μέθοδος FinBERT αποτελεί εξειδικευμένη προσαρμογή του γλωσσικού μοντέλου BERT σε χρηματοοικονομικά δεδομένα.

Χρησιμοποιώντας κάθε μία από τις παραπάνω μεθόδους, εξήχθησαν ημερήσιες συ-

Πίνακας 5.8: Λίστα μετοχών

ΑΑ	Σύνολο Δεδομένων	Μετοχή
1	AAL	American Airlines Group
2	AMD	Advanced Micro Devices
3	AUY	Yamana Gold Inc.
4	BABA	Alibaba Group
5	BAC	Bank of America Corporation
6	ET	Energy Transfer L.P.
7	FCEL	FuelCell Energy Inc.
8	GE	General Electric
9	GM	General Motors
10	INTC	Intel Corporation
11	MRO	Marathon Oil Corporation
12	MSFT	Microsoft Corporation
13	OXY	Occidental Petroleum Corporation
14	RYCEY	Rolls-Royce Holdings
15	SQ	Square
16	VZ	Verizon Communications



Σχήμα 5.11: Διαγραμμάτική αναπαράσταση της προεπεξεργασίας

ναισθηματικές βαθμολογίες για κάθε μετοχή. Για να συμβεί αυτό, για κάθε μια από τις τρεις μεθόδους, έλαβε χώρα η εξής διαδικασία. Αρχικά, κάθε ένα από τα tweets έλαβε μέσω της εκάστοτε μεθόδου την αντίστοιχη πολικότητα. Στη συνέχεια, υπολογίστηκε ο ημερήσιος μέσος όρος των πολικοτήτων, αποδίδοντας έτσι σε κάθε ημέρα μια μοναδική πολικότητα, ανεξάρτητα από το πλήθος των tweets που συλλέχθηκαν. Έτσι, σχηματίστηκαν τρεις νέες χρονοσειρές – μια για κάθε μέθοδο – αποτελούμενες από πολικότητες συναισθημάτων.

Παρακάτω, παρέχονται συνοπτικά τεχνικές πληροφορίες για τα τρία αυτά μοντέλα, αλλά και πληροφορίες για τις διαφορές που παρουσίαζαν στο πώς οι τιμές των ημερήσιων πολικοτήτων προέκυψαν, ανάλογα με τις ιδιαιτερότητες του εκάστοτε μοντέλου.

- **TextBlob**: Η TextBlob είναι μια βιβλιοθήκη της Python που επιτρέπει την επεξεργασία φυσικής γλώσσας και περιλαμβάνει λειτουργίες όπως ανάλυση συναισθήματος, ανίχνευση γλώσσας και σημασιολογική ταξινόμηση. Η μέθοδος ανάλυσης συναισθήματος που χρησιμοποιείται από την TextBlob βασίζεται σε κανόνες βασισμένους σε λέξεις-κλειδιά. Δηλαδή, λειτουργεί εφαρμόζοντας προ-

καθορισμένους κανόνες βαθμολόγησης σε συγκεκριμένες λέξεις ή φράσεις. Με αυτό τον τρόπο, η μέθοδος επιστρέφει έναν αριθμό στο διάστημα [-1, 1], ο οποίος αναπαριστά την πολικότητα του συναισθήματος για κάθε tweet. Ο αριθμητικός μέσος των ημερήσιων συναισθηματικών βαθμολογιών εξάγεται για να δημιουργηθεί η τελική χρονοσειρά συναισθηματικών δεδομένων.

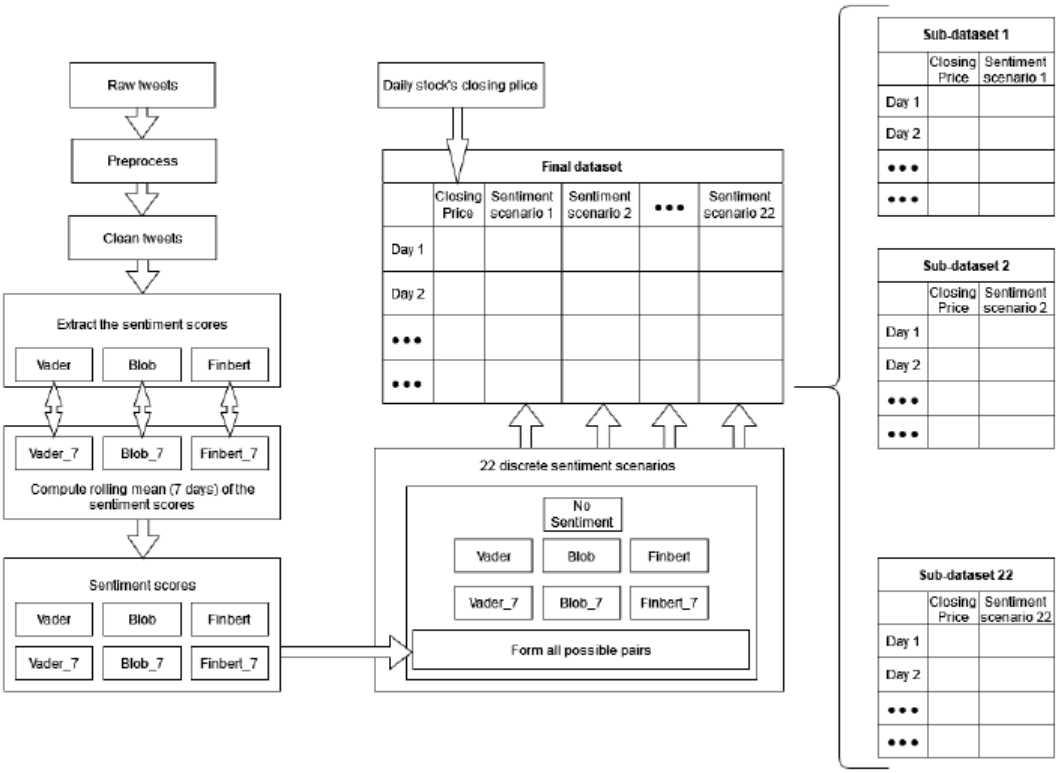
- **VADER**: Η VADER (Valence Aware Dictionary and sEntiment Reasoner) αποτελεί και αυτή με τη σειρά της μια βασισμένη σε κανόνες μέθοδο για ανάλυση συναισθήματος, η οποία έχει αναπτυχθεί ειδικά για την επεξεργασία συναισθημάτων που υπάρχουν σε δεδομένα κοινωνικών δικτύων. Στο πλαίσιο της παρούσας εργασίας, το εργαλείο VADER χρησιμοποιήθηκε για την εξαγωγή σύνθετων βαθμολογιών (compound scores) από τα tweets. Συγκεκριμένα, για κάθε κείμενο, ο αλγόριθμος επιστρέφει τέσσερις διαφορετικές τιμές: αρνητικό, ουδέτερο, θετικό και τη συνολική σύνθετη βαθμολογία. Στην παρούσα μελέτη χρησιμοποιήθηκε η σύνθετη βαθμολογία, η οποία υπολογίζεται με κανονικοποίηση των άλλων τριών μεγεθών. Οι ημερήσιες μέσες τιμές αυτών των βαθμολογιών χρησιμοποιήθηκαν για τη δημιουργία της αντίστοιχης χρονοσειράς συναισθήματος.

- **FinBERT**: Όσον αφορά το FinBERT, στην παρούσα μελέτη χρησιμοποιήθηκε η υλοποίηση που περιγράφεται στο X. Συγκεκριμένα, αξιοποιήθηκε το μοντέλο που έχει εκπαιδευτεί στα δεδομένα της PhraseBank. Η μέθοδος FinBERT αποτελεί μια εξειδικευμένη προσαρμογή του BERT, ενός transformer βασισμένου σε Βαθιά Μάθηση για Επεξεργασία Φυσικής Γλώσσας. Το BERT (Bidirectional Encoder Representations from Transformers) είναι ένα προεκπαιδευμένο μοντέλο που βασίζεται στις αρχές της μεταφοράς μάθησης. Η εκπαίδευση του FinBERT πραγματοποιείται σε ένα σύνολο δεδομένων που αποτελείται από οικονομικά κείμενα. Το αποτέλεσμα είναι ένα μοντέλο το οποίο είναι προσαρμοσμένο για την εξαγωγή συναισθημάτων από οικονομικές ειδήσεις και σχόλια. Το FinBERT επιστρέφει μια ταξινόμηση του συναισθήματος σε θετικό, αρνητικό ή ουδέτερο, με βάση τις πιθανότητες που υπολογίζονται από το μοντέλο. Στην παρούσα μελέτη, χρησιμοποιήθηκε η μέση ημερήσια βαθμολογία των ταξινομημένων tweets ως χρονοσειρά συναισθήματος.

Μετά τη δημιουργία των τριών αρχικών χρονοσειρών πολικοτήτων, εφαρμόστηκαν κινητοί μέσοι όροι 7 και 14 ημερών στις παραπάνω εξαγόμενες χρονοσειρές. Αυτό είχε ως αποτέλεσμα την παραγωγή εννέα διαφορετικών χρονοσειρών συναισθηματικών δεδομένων, οι οποίες, μαζί με τις αρχικές τιμές κλεισίματος των μετοχών και τις δικές τους αντίστοιχες κινητές μέσες τιμές, οδήγησαν σε 12 συνολικά χαρακτηριστικά. Οι διαφορετικοί συνδυασμοί των παραπάνω χαρακτηριστικών, καθώς και η εξέταση του σεναρίου όπου δεν λαμβάνεται υπόψιν η Ανάλυση Συναισθήματος, οδήγησαν στη διαμόρφωση των 22 διαφορετικών διαμορφώσεων εισόδου, οι οποίες χρησιμοποιήθηκαν στην πειραματική διαδικασία για κάθε έναν από τους συγκρινόμενους αλγορίθμους. Η διαδικασία εξαγωγής των διαφορετικών συναισθηματικών διαμορφώσεων, παρουσιάζεται στο Σχήμα 5.12.

5.5.2 Αλγόριθμοι

Σχετικά με τους αλγορίθμους, χρησιμοποιήθηκαν 27 Μέθοδοι Μηχανικής μάθησης, οι οποίες εντοπίζονται αλφαβητικά στον Πίνακα 5.9. Για την υλοποίηση των μεθόδων



Σχήμα 5.12: Διαγραμματική αναπαράσταση της διαδικασίας εξαγωγής των 22 συναισθηματικών διαμορφώσεων

στο πλαίσιο της πειραματικής διαδικασίας χρησιμοποιήθηκαν οι βιβλιοθήκες Keras και PyCaret. Η Keras είναι μια βιβλιοθήκη ανοικτού κώδικα υψηλού επιπέδου για Βαθιά Μάθηση, η οποία βασίζεται στο TensorFlow και επιτρέπει την εύκολη δημιουργία και εκπαίδευση νευρωνικών δικτύων. Το TensorFlow είναι ένα ισχυρό πλαίσιο Μηχανικής και Βαθιάς Μάθησης, ανοιχτού κώδικα, που αναπτύχθηκε από την Google Brain. Βασίζεται σε έναν υπολογιστικό γράφο, όπου οι κόμβοι αναπαριστούν μαθηματικές πράξεις και οι ακμές τις ροές δεδομένων, επιτρέποντας την αποδοτική εκτέλεση αλγορίθμων σε CPU, GPU και TPU. Η PyCaret, από την άλλη, είναι ένα low-code πλαίσιο Μηχανικής Μάθησης, το οποίο απλοποιεί τη διαδικασία προεπεξεργασίας δεδομένων, επιλογής μοντέλων, εκπαίδευσης και αξιολόγησης, επιταχύνοντας σημαντικά τη ροή εργασίας πειραμάτων Μηχανικής Μάθησης. Χρησιμοποιείται για την αυτοματοποίηση πολλών σταδίων της διαδικασίας, επιτρέποντας ταχύτερη ανάπτυξη και ανάλυση μοντέλων. Πρέπει να σημειωθεί, ότι στο πλαίσιο της υλοποίησής μας, όλοι οι αλγόριθμοι που δοκιμάστηκαν χρησιμοποιήθηκαν στη βασική τους διαμόρφωση, χωρίς να πραγματοποιηθεί κάποια διαδικασία βελτιστοποίησης.

Πίνακας 5.9: Λίστα αλγορίθμων

ΑΑ.	**Συντομογραφία**	**Αλγόριθμοι**
1	ABR	AdaBoost Regressor [71]
2	ARD	Automatic Relevance Determination [294]
3	BiLSTM (LSTM_2)	Bidirectional LSTM [104]
4	BiLSTM-LSTM (LSTM_3)	Bidirectional LSTM and LSTM Stacked [104], [118]
5	CBR	CatBoost Regressor [214]
6	DTR	Decision Tree Regressor [36]
7	ELN	Elastic Net [313]
8	ET	Extra Trees Regressor [95]
9	XGBoost	Extreme Gradient Boosting [46]
10	GB	Gradient Boosting Regressor [88]
11	HBR	Huber Regressor [109]
12	KNR	K-Neighbors Regressor [68]
13	KER	Kernel Ridge [282]
14	LSTM	LSTM [118]
15	LA-LAS	Lasso Least Angle Regression [75]
16	LAS	Lasso Regression [272]
17	LA	Least Angle Regression [75]
18	LGBM	Light Gradient Boosting Machine [78]
19	LNR	Linear Regression [246]
20	MLP	Multilevel Perceptron [181]
21	OMP	Orthogonal Matching Pursuit [233]
22	PAR	Passive Aggressive Regressor [54]
23	RF	Random Forest Regressor [35]
24	RSC	Random Sample Consensus [48]
25	RDG	Ridge Regression [163]
26	SVR	Support Vector Regression [256]
27	THS	Theil–Sen Regressor [56]

Στο παρόν πλαίσιο, η διαδικασία πρόβλεψης αντιμετωπίστηκε ως ένα πρόβλημα παλινδρόμησης. Κατά την εκτέλεση των πειραμάτων, οι αρχικές παράμετροι των αντίστοιχων μεθόδων από τη βιβλιοθήκη διατηρήθηκαν, διατηρώντας σταθερό το περιβάλλον υλοποίησης για όλα τα αλγοριθμικά σχήματα. Έτσι, όλοι οι αλγόριθμοι συγκρίθηκαν στη βασική τους διαμόρφωση, ενισχύοντας την εύκολη αναπαραγωγή των πειραμάτων, αλλά και τη σύγκριση με επιπλέον μεθόδους.

Επιπλέον, κατά την πειραματική διαδικασία, τα δεδομένα χωρίστηκαν σε αναλογία 80% - 20% σε σύνολα εκπαίδευσης και ελέγχου αντίστοιχα.

Όσον αφορά τους χρονικούς ορίζοντες πρόβλεψης, εφαρμόστηκαν τρεις διαφορετικές στρατηγικές: πρόβλεψη μιας ημέρας (single-step), 7 και 14 ημερών (multi-step).

Τα αποτελέσματα αξιολογήθηκαν με βάση τις μετρικές απόδοσης MAE, MAPE, MSE, RMSE, RMSLE και R^2.

Στο σημείο θα παρατεθούν εν τάχει πληροφορίες για τις μεθόδους που όπως διαφαινεται στην πειραματική διαδικασία υπερτερούν των υπολοίπων - αυτές είναι τα LSTMs και τα BiLSTMs. Καθώς το πλήθος των μεθόδων είναι απαγορευτικό για μια - έστω και σύντομη - περιγραφή τους, στον Πίνακα 5.9 υπάρχουν αναφορές που παραπέμπουν τον αναγνώστη σε αντίστοιχη βιβλιογραφία.

5.5.2.1 Long-Short Term Memory - (LSTM)

Τα Δίκτυα Μακράς Βραχύχρονης Μνήμης Long Short-Term Memory - LSTM αναπτύχθηκαν από τους Hochreiter & Schmidhuber (1997) ως λύση στα μειονεκτήματα των απλών RNNs, ιδιαίτερα στο πρόβλημα της εξαφάνισης ή/και της έκρηξης των βαθμίδων (vanishing/exploding gradients) σε μεγάλες ακολουθίες. Τα LSTMs χρησιμοποιούν μία ειδική εσωτερική δομή —την κατάσταση κελιού (cell state)— η οποία επιτρέπει τη μεταφορά πληροφορίας ανάμεσα σε χρονικά βήματα, διατηρώντας μακροπρόθεσμες εξαρτήσεις στα δεδομένα.

Σε αντίθεση με τα απλά RNNs, τα LSTMs εισάγουν τον μηχανισμό των πυλών, που ελέγχουν τη ροή πληροφορίας και τη διατήρησή της.

Κάθε LSTM «κελί» αποτελείται από τρεις βασικές πύλες:

- **Πύλη εισόδου (input gate)**: Ρυθμίζει ποια νέα πληροφορία θα εισέλθει στην κατάσταση κελιού.
- **Πύλη λήθης (forget gate)**: Ρυθμίζει ποια πληροφορία θα διατηρηθεί και ποια θα απορριφθεί από την προηγούμενη κατάσταση κελιού.
- **Πύλη εξόδου (output gate)**: Ρυθμίζει ποια πληροφορία θα χρησιμοποιηθεί για την παραγωγή της εξόδου (κρυφή κατάσταση).

Η ροή πληροφορίας σε ένα LSTM κελί περιγράφεται από τις ακόλουθες εξισώσεις:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \tag{5.19}$$

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \tag{5.20}$$

$$\tilde{c}_t = \tanh(W_c[h_{t-1}, x_t] + b_c) \tag{5.21}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \tag{5.22}$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{5.23}$$

$$h_t = o_t \odot \tanh(c_t) \tag{5.24}$$

Όπου:

- x_t το εισερχόμενο διάνυσμα στο χρονικό βήμα t
- h_{t-1} η κρυφή κατάσταση στο προηγούμενο χρονικό βήμα
- c_{t-1} η κατάσταση κελιού στο προηγούμενο χρονικό βήμα
- f_t, i_t, o_t οι πύλες λήθης, εισόδου και εξόδου αντίστοιχα
- $\tilde{c}_t$ η υποψήφια πληροφορία που θα προστεθεί στην κατάσταση κελιού
- c_t η νέα κατάσταση κελιού

- h_t η νέα κρυφή κατάσταση
- $\sigma(\cdot)$ η σιγμοειδής συνάρτηση ενεργοποίησης $\sigma(x) = \frac{1}{1+e^{-x}}$
- W_f, W_i, W_c, W_o τα βάρη που πολλαπλασιάζονται με το συζευγμένο διάνυσμα
- b_f, b_i, b_c, b_o οι αντίστοιχες μεταβλητές μετατόπισης

Οι παραπάνω εξισώσεις περιγράφουν τη λειτουργία ενός κελιού LSTM σε κάθε χρονικό βήμα t. Συγκεκριμένα, η πύλη λήθης (f_t) (5.19) αποφασίζει ποια πληροφορία από την προηγούμενη κατάσταση κελιού (c_{t-1}) θα διατηρηθεί ή θα απορριφθεί. Η πύλη εισόδου (i_t) (5.20), ελέγχει πόση νέα πληροφορία θα εισέλθει στην κατάσταση κελιού, ενώ η εξίσωση (5.21) δίνει την υποψήφια νέα πληροφορία ($\tilde{c}_t$) που μπορεί να προστεθεί στο κελί. Στη συνέχεια, η εξίσωση (5.22) ενημερώνει την ίδια την κατάσταση κελιού (c_t), συνδυάζοντας την προηγούμενη κατάσταση και τη νέα πληροφορία, ανάλογα με τις τιμές των πυλών. Η εξίσωση (5.23) καθορίζει την πύλη εξόδου (o_t), η οποία ρυθμίζει ποιο τμήμα της κατάστασης κελιού θα χρησιμοποιηθεί για την παραγωγή της εξόδου. Τέλος, η νέα κρυφή κατάσταση (h_t) (5.24), αποτελεί και την έξοδο του κελιού στο συγκεκριμένο χρονικό βήμα.

Τα LSTMs παρουσιάζουν σημαντικά πλεονεκτήματα σε σχέση με τα απλά επαναληπτικά νευρωνικά δίκτυα, κυρίως λόγω της ικανότητάς τους να διαχειρίζονται μακροπρόθεσμες εξαρτήσεις μέσα σε ακολουθιακά δεδομένα. Ο μηχανισμός των πυλών, που χαρακτηρίζει τα LSTMs, επιτρέπει στο δίκτυο να διατηρεί κρίσιμη πληροφορία για μεγάλο αριθμό χρονικών βημάτων, επιλύοντας έτσι το διαχρονικό πρόβλημα της εξαφάνισης (vanishing) ή της έκρηξης (exploding gradients) των βαθμίδων (κλίσεων) που συναντάται συχνά στα παραδοσιακά RNNs. Το πρόβλημα αυτό ανακύπτει κατά τον υπολογισμό των παραγώγων μέσω της μεθόδου της οπισθοδρόμησης στο χρόνο (Backpropagation Through Time, BPTT), που χρησιμοποιείται για τη ρύθμιση των βαρών του δικτύου. Σε ακολουθίες με μεγάλο μήκος, οι παραγόμενες κλίσεις προκύπτουν ως γινόμενο πολλών μικρών ή μεγάλων αριθμών. Εάν αυτές οι τιμές είναι μικρότερες της μονάδας, το γινόμενο τείνει γρήγορα στο μηδέν, ενώ αν είναι μεγαλύτερες, το γινόμενο μεγαλώνει εκθετικά. Χάρη σε αυτήν την αρχιτεκτονική, τα LSTMs μπορούν να μαθαίνουν σύνθετες χρονικές σχέσεις χωρίς να ξεχνιούνται σημαντικές πληροφορίες, κάτι που τα καθιστά ιδιαίτερα αποτελεσματικά στην επεξεργασία δεδομένων με πολύπλοκες ή μακροπρόθεσμες αλληλεξαρτήσεις.

5.5.2.2 Bidirectional Long Short-Term Memory (BiLSTM)

Τα Αμφίδρομα Δίκτυα Μακράς Βραχύχρονης Μνήμης (Bidirectional Long Short-Term Memory - BiLSTM)δίκτυα αποτελούν επέκταση των απλών LSTMs, προσφέροντας στο δίκτυο τη δυνατότητα να αξιοποιεί πληροφορία και από το παρελθόν και από το μέλλον της ακολουθίας. Συγκεκριμένα, κάθε BiLSTM αποτελείται από δύο ανεξάρτητες LSTM δομές: μία που επεξεργάζεται τα δεδομένα προς τα εμπρός (forward LSTM) και μία που τα επεξεργάζεται προς τα πίσω (backward LSTM). Η προσέγγιση αυτή επιτρέπει στο μοντέλο να έχει πλήρη εικόνα του συμφραζόμενου σε κάθε χρονικό βήμα, κάτι που είναι κρίσιμο σε πολλές εφαρμογές ανάλυσης ακολουθιών, όπως στην επεξεργασία φυσικής γλώσσας.

Μαθηματικά, οι καταστάσεις για το forward και backward LSTM υπολογίζονται ως εξής:

Forward LSTM:

$$f_t^F = \sigma(W_f^F[\overrightarrow{h}_{t-1}, x_t] + b_f^F) \tag{5.25}$$

$$i_t^F = \sigma(W_i^F[\overrightarrow{h}_{t-1}, x_t] + b_i^F) \tag{5.26}$$

$$\tilde{c}_t^F = \tanh(W_c^F[\overrightarrow{h}_{t-1}, x_t] + b_c^F) \tag{5.27}$$

$$c_t^F = f_t^F \odot c_{t-1}^F + i_t^F \odot \tilde{c}_t^F \tag{5.28}$$

$$o_t^F = \sigma(W_o^F[\overrightarrow{h}_{t-1}, x_t] + b_o^F) \tag{5.29}$$

$$\overrightarrow{h}_t = o_t^F \odot \tanh(c_t^F) \tag{5.30}$$

Backward LSTM:

$$f_t^B = \sigma(W_f^B[\overleftarrow{h}_{t+1}, x_t] + b_f^B) \tag{5.31}$$

$$i_t^B = \sigma(W_i^B[\overleftarrow{h}_{t+1}, x_t] + b_i^B) \tag{5.32}$$

$$\tilde{c}_t^B = \tanh(W_c^B[\overleftarrow{h}_{t+1}, x_t] + b_c^B) \tag{5.33}$$

$$c_t^B = f_t^B \odot c_{t+1}^B + i_t^B \odot \tilde{c}_t^B \tag{5.34}$$

$$o_t^B = \sigma(W_o^B[\overleftarrow{h}_{t+1}, x_t] + b_o^B) \tag{5.35}$$

$$\overleftarrow{h}_t = o_t^B \odot \tanh(c_t^B) \tag{5.36}$$

Η τελική έξοδος σε κάθε χρόνο t προκύπτει από τη συνένωση των δύο κρυφών καταστάσεων:

$$h_t = [\overrightarrow{h}_t ; \overleftarrow{h}_t] \tag{5.37}$$

Όπου x_t είναι το εισερχόμενο διάνυσμα στο χρονικό βήμα t, ενώ τα $\overrightarrow{h}_t$ και $\overleftarrow{h}_t$ είναι οι κρυφές καταστάσεις του forward και backward LSTM αντίστοιχα. Τα W και b με τοος αντίστοιχους εκθέτες F και B συμβολίζουν τους πίνακες βαρών και τις μεταβλητές μετατόπισης για κάθε κατεύθυνση.

Τα BiLSTM είναι εξαιρετικά αποτελεσματικά σε εφαρμογές όπου η πλήρης κατανόηση του συμφραζόμενου μιας ακολουθίας απαιτεί πληροφορία από όλο το εύρος της, όπως για παράδειγμα στη στην ανάλυση συντακτικής δομής κειμένου, στην αναγνώριση φωνής, αλλά και σε πολλές εφαρμογές βιοϊατρικής. Το μοναδικό επιπρόσθετο υπολογιστικό κόστος προέρχεται από το γεγονός ότι διατηρούνται δύο ανεξάρτητες LSTM δομές (forward & backward), όμως αυτό συνήθως αντισταθμίζεται από τη σημαντική βελτίωση στην ακρίβεια και στη συνολική απόδοση του μοντέλου.

5.5.3 Αποτελέσματα

Τα αποτελέσματα της εκτεταμένης πειραματικής διαδικασίας θα παρουσιαστούν με τρόπο τέτοιο που αντικατοπτρίζει το διττό στόχο της: αφενός, να αποτελέσει σύγκριση ενός μεγάλου αριθμού σύγχρονων αλγορίθμων πρόβλεψης χρονοσειρών, και αφετέρου να διερευνήσει το αν και κατά πόσο η γνώση της κοινής γνώμης, όπως ανταναακλάται στα μέσα κοινωνικής δικτύωσης και ποσοτικοποιείται μέσω τριών διαφορετικών μεθόδων ανάλυσης συναισθήματος, μπορεί να βελτιώσει τις παραγόμενες προβλέψεις. Αντίστοιχα, η παρουσίαση των αποτελεσμάτων της πειραματικής διαδικασίας χωρίζεται σε δύο διακριτά μέρη. Το πρώτο αφορά τη σύγκριση των μεθόδων, ενώ το δεύτερο τη σύγκριση των διαφορετικών σχημάτων Συναισθηματικής Ανάλυσης, με στόχο να διερευνηθεί η αποτελεσματικότητά τους.

Ωστόσο, πρέπει να σημειωθεί ότι ο μεγάλος αριθμός συγκρίσεων που πραγματοποιήθηκαν οδήγησε σε έναν τεράστιο όγκο αποτελεσμάτων. Συγκεκριμένα, σε κάθε περίπτωση, αξιολογήθηκε η απόδοση 27 διαφορετικών αλγοριθμικών σχημάτων και 22 διαφορετικών διαμορφώσεων χαρακτηριστικών για τρεις διαφορετικούς χρονικούς ορίζοντες πρόβλεψης (1, 7 και 14 ημερών). Κάθε ένα από αυτά τα σχήματα αξιολογήθηκε χρησιμοποιώντας έξι διαφορετικές μετρικές απόδοσης, ενώ η διαδικασία επαναλήφθηκε για κάθε ένα από τα σύνολα δεδομένων.

Ως εκ τούτου, είναι προφανές ότι η πλήρης παρουσίαση όλων των αριθμητικών αποτελεσμάτων δεν μπορεί να συμβάλει ουσιαστικά στην κατανόηση των συμπερασμάτων. Στη συνέχεια, ακολουθεί μία απαραίτητη σύνοψη της διαδικασίας, πριν την παρουσίαση των εξαγόμενων αποτελεσμάτων.

Όπως έχει ήδη αναφερθεί, κατά τη διαδικασία, για κάθε μία από τις μετοχές ακολουθήθηκε η εξής στρατηγική:

- Κάθε ένας από τους 27 αλγορίθμους πρόβλεψης εκτελέστηκε 22 φορές, κάθε φορά λαμβάνοντας ως είσοδο έναν από τους 22 διαφορετικούς συνδυασμούς χαρακτηριστικών.
- Αυτή η διαδικασία επαναλήφθηκε τρεις φορές, μία για κάθε χρονικό παράθυρο πρόβλεψης.
- Για κάθε εκτέλεση, υπολογίστηκαν οι έξι μετρικές απόδοσης.

Η σύγκριση των αλγορίθμων πραγματοποιήθηκε χρησιμοποιώντας στατιστικά τεστ Friedman, τα οποία εφαρμόστηκαν ξεχωριστά για κάθε σύνολο χαρακτηριστικών και κάθε χρονικό διάστημα πρόβλεψης. Έτσι, για την ανάλυση των αλγορίθμων, εκτελέστηκαν συνολικά:

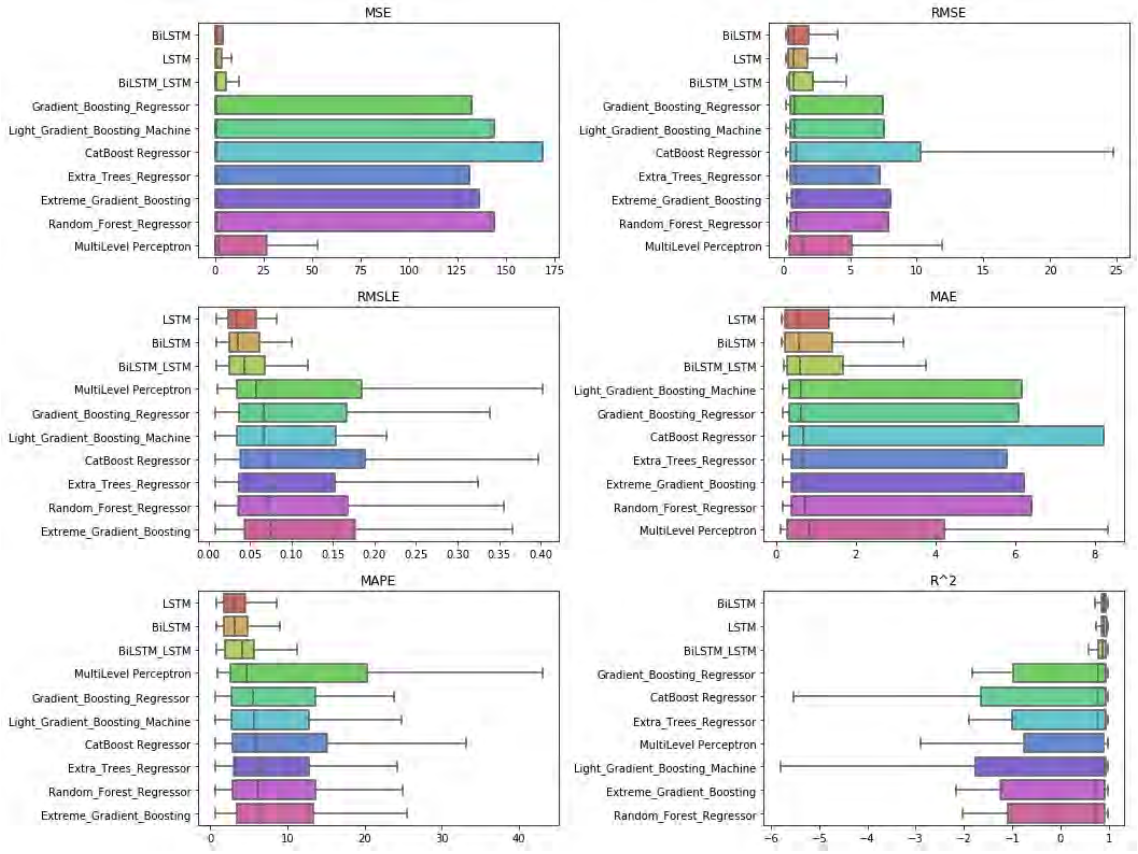
$$27 \times 6 \times 3 = 486 \ \text{ Friedman tests.} \tag{5.38}$$

Αντίστοιχα, η αξιολόγηση της αποδοτικότητας των συνδυασμών χαρακτηριστικών έγινε με βάση τη μέση απόδοση των 30 αλγορίθμων για κάθε έναν από τους 67 διαφορετικούς συνδυασμούς. Σε αυτή την περίπτωση, εκτελέστηκαν συνολικά:

$$22 \times 6 \times 3 = 396 \ \text{Friedman tests.} \tag{5.39}$$

Για να συνοψιστούν τα αποτελέσματα, υπολογίστηκε η μέση κατάταξη των μεθόδων και των συνόλων χαρακτηριστικών σε κάθε χρονικό διάστημα. Η κατάταξη βασίστηκε στην κατάταξη που παρήγαγε ο στατιστικός έλεγχος Friedman, ενώ στη συνέχεια, χρησιμοποιήθηκε το τεστ Nemenyi για να ελεγχθούν κατά ζεύγη οι σημαντικές διαφορές μεταξύ των μοντέλων. Τα αποτελέσματα του τεστ Nemenyi απεικονίζονται σε διαγράμματα κρίσιμων διαφορών, στα οποία οι μέθοδοι που δεν διαφέρουν σημαντικά μεταξύ τους συνδέονται με μαύρες οριζόντιες γραμμές. Κατά τα γνωστά, δύο μέθοδοι θεωρούνται μη σημαντικά διαφορετικές, εάν η διαφορά στις μέσες κατατάξεις τους είναι μικρότερη από την τιμή της κρίσιμης διαφοράς.

Στη συνέχεια, παρουσιάζονται τα αποτελέσματα που προέκυψαν από τις δύο μελέτες περίπτωσης: πρώτα η σύγκριση των αλγορίθμων πρόβλεψης (Υποενότητα 5.5.3.1) και στη συνέχεια η σύγκριση των συνδυασμών χαρακτηριστικών και Ανάλυσης Συναισθήματος (Υποενότητα 5.5.3.2).



Σχήμα 5.13: Θηκογράμματα για την απόδοση των αλγορίθμων - Χρονικός ορίζοντας μιας ημέρας

5.5.3.1 Από τη σκοπιά των αλγορίθμων

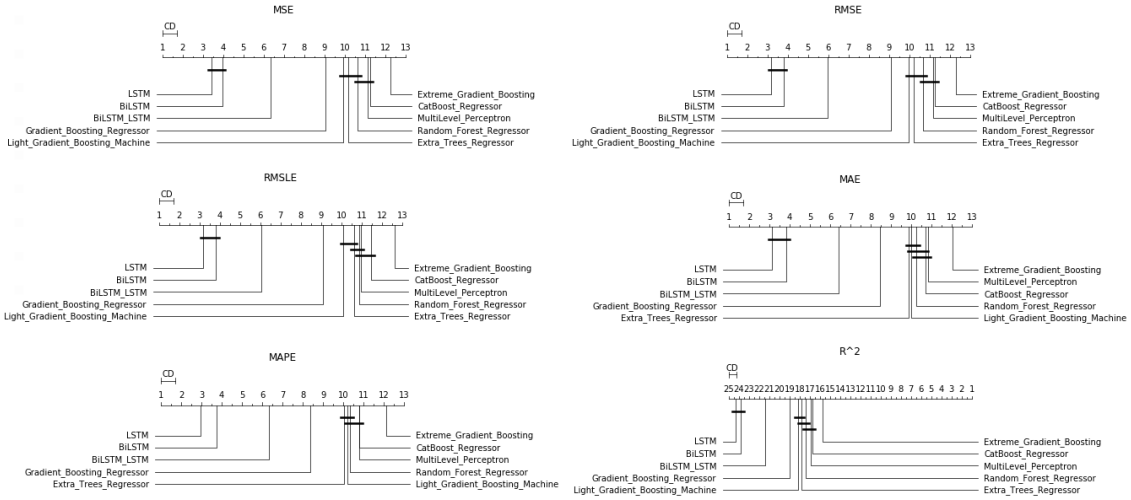
Για την ευκολότερη παρακολούθηση του κειμένου, τα αποτελέσματα θα παρουσιαστούν ανά χρονικό ορίζοντα, ξεκινώντας από αυτόν της μιας ημέρας και συνεχίζοντας στους μεγαλύτερους, δηλαδή αυτούς της μιας εβδομάδας και των δύο εβδομάδων.

Χρονικός ορίζοντας μιας ημέρας Ξεκινώντας με την απλή πρόβλεψη μίας ημέρας, από τα αποτελέσματα που παρουσιάζονται στον Πίνακα 5.10 και στα Σχήματα 5.13 και 5.14, μπορεί κανείς εύκολα να συμπεράνει ότι υπάρχει μια σχεδόν καθολική επικράτηση των μεθόδων LSTM.

Όσον αφορά τις τρεις κορυφαίες μεθόδους, τα διαγράμματα κρίσιμων διαφορών υποδεικνύουν στατιστική εξάρτηση μεταξύ των LSTM και Bi-LSTM, ενώ έπεται ο συνδυασμός που ενσωματώνει και τις δύο μεθόδους μέσω ενός stacked σχήματος τους, ο οποίος παρουσιάζεται ως στατιστικά ανεξάρτητος από όλες τις υπόλοιπες μεθόδους. Το γεγονός αυτό, σύμφωνα με όσα έχουν αναφερθεί για τον τρόπο εξαγωγής αυτών των διαγραμμάτων, μπορεί εύκολα να εξαχθεί και από τα αποτελέσματα της κατάταξης Friedman, όπου οι αποκλίσεις μεταξύ των μεθόδων είναι σημαντικές, αλλά και από τα αντίστοιχα θηκογράμματα, όπου η διασπορά όσο και οι τιμές των σφαλμάτων όπως αποτυπώνονται στις μετρικές αξιολόγησης των τριών κορυφαίων μεθόδων ξεχωρίζουν ξεκάθαρα από εκείνες όλων των άλλων.

	MSE		RMSE		RMSLE	
	Method	**F-Rank**	**Method**	**F-Rank**	**Method**	**F-Rank**
1	LSTM	3.4091	LSTM	3.1847	LSTM	3.1790
2	LSTM_2	3.9545	LSTM_2	3.7869	LSTM_2	3.8011
3	LSTM_3	6.3438	LSTM_3	5.9773	LSTM_3	6.0653
4	GB	9.0483	GB	9.0767	GB	9.1051
5	LGBM	9.9233	LGBM	9.9688	LGBM	10.1051
6	ET	10.1705	ET	10.2244	ET	10.6250
7	RF	10.6449	RF	10.6875	RF	10.8722
8	MLP	11.1364	MLP	11.1790	MLP	10.9574
9	CBR	11.2386	CBR	11.2585	CBR	11.4744
10	XGBoost	12.2642	XGBoost	12.3040	XGBoost	12.6335
11	ARD	13.1165	ARD	13.1676	ARD	13.5625
12	OMP	13.2116	OMP	13.2685	OMP	13.7060
13	LA	13.6662	LA	13.7116	LA	14.1832
14	RDG	13.7727	RDG	13.8182	RDG	14.2188
15	LNR	13.7827	LNR	13.8281	LNR	14.2983
16	ABR	14.6847	ABR	14.7102	ABR	15.1108
17	DTR	15.3409	DTR	15.3608	DTR	15.6477
18	KNR	16.1903	KNR	16.2102	KNR	15.9943
19	RSC	16.3168	RSC	16.3509	RSC	16.5057
20	HBR	16.9148	HBR	16.9602	HBR	17.3040
21	SVR	17.8551	SVR	17.8608	THS	17.5170
22	THS	18.1165	LAS	18.1307	SVR	17.6307
	MAE		**MAPE**		**R^2**	
	Method	**F-Rank**	**Method**	**F-Rank**	**Method**	**F-Rank**
1	LSTM	3.1136	LSTM	2.9602	LSTM	24.3438
2	LSTM_2	3.8352	LSTM_2	3.7472	LSTM_2	23.7955
3	LSTM_3	6.4233	LSTM_3	6.3466	LSTM_3	21.3693
4	GB	8.4489	GB	8.3608	GB	19.0057
5	ET	9.8892	ET	10.0199	LGBM	18.1534
6	LGBM	9.9972	LGBM	10.2188	ET	17.8494
7	RF	10.2443	RF	10.3494	RF	17.3864
8	CBR	10.6932	MLP	10.7841	MLP	16.8864
9	MLP	10.8182	CBR	10.8068	CBR	16.7699
10	XGBoost	12.0284	XGBoost	12.1108	XGBoost	15.7557
11	ARD	13.8352	ARD	13.7557	ARD	14.9460
12	OMP	13.9616	OMP	13.8764	OMP	14.8452
13	LA	14.2088	LA	14.1406	LA	14.4020
14	RDG	14.3125	RDG	14.2813	RDG	14.2955
15	LNR	14.3494	LNR	14.2912	LNR	14.2855
16	ABR	14.6761	ABR	14.6932	ABR	13.3381
17	DTR	14.9403	DTR	15.0511	DTR	12.6648
18	KNR	15.5369	KNR	15.3693	KNR	11.8523
19	SVR	15.7358	SVR	15.8438	RSC	11.7315
20	RSC	16.8153	RSC	16.7798	HBR	11.1335
21	HBR	17.4517	HBR	17.4403	SVR	10.1705
22	THS	18.6534	LAS	18.6222	THS	9.9091

Πίνακας 5.10: Κατάταξη Friedman (Μέθοδοι) - Ορίζοντας πρόβλεψης μιας ημέρας



Σχήμα 5.14: CD διαγράμματα για την απόδοση των αλγορίθμων - Χρονικός ορίζοντας μιας ημέρας

Χρονικός ορίζοντας μιας εβδομάδας Στην περίπτωση των εβδομαδιαίων προβλέψεων επανεμφανίζεται το μοτίβο που παρατηρήθηκε στον χρονικό ορίζοντα μιας ημέρας. Και πάλι, σε όλες τις μετρικές, οι τρεις καλύτερες μέθοδοι ήταν οι LSTM, Bi-LSTM και ο συνδυασμός τους. Τα σχετικά αποτελέσματα εμφανίζονται στον Πίνακα ;;, ενώ τα αντίστοιχα διαγράμματα κρίσιμων διαφορών στο Σχήμα 5.16.

Στα θηκογράμματα (Σχήμα 5.15), αν και οι παραλλαγές LSTM φαίνεται να σχηματίζουν μια ομάδα με παρόμοια απόδοση, οι βαθμολογίες Friedman δείχνουν ότι μόνο οι δύο κορυφαίες από αυτές διαφέρουν στατιστικά σημαντικά από όλες τις υπόλοιπες μεθόδους όσον αφορά την αξιολόγηση των αριθμητικών αποτελεσμάτων, ενώ ο συνδυασμός τους ακολουθεί, συμπληρώνοντας την τριάδα που ξεχωρίζει από τις υπόλοιπες.

Χρονικός ορίζοντας δύο εβδομάδων Τέλος, όσον αφορά την περίπτωση των προβλέψεων 14 ημερών, οι γενικές παρατηρήσεις που διατυπώθηκαν στην προηγούμενη ενότητα μπορούν να επεκταθούν και εδώ. Τα αποτελέσματα παρουσιάζονται παρακάτω, όπου στον Πίνακα ;; εμφανίζονται τα αποτελέσματα της κατάταξης του στατιστικού ελέγχου Friedman και στα Σχήματα 5.17 και 5.18 τα θηκογράμματα των αριθμητικών αποτελεσμάτων των έξι μετρικών και τα διαγράμματα κρίσιμων διαφορών αντίστοιχα. Ωστόσο, μια επιπλέον τελική παρατήρηση πρέπει να σημειωθεί είναι πως στα θηκογράμματα που αφορούν τα αποτελέσματα της μετρικής R^2, φαίνεται να υπάρχει μια διαφορά στην κατάταξη της διάμεσης τιμής. Ωστόσο, αυτή η διαφορά δεν εντοπίστηκε στην κατάταξη που παρέχει ο έλεγχος Friedman.

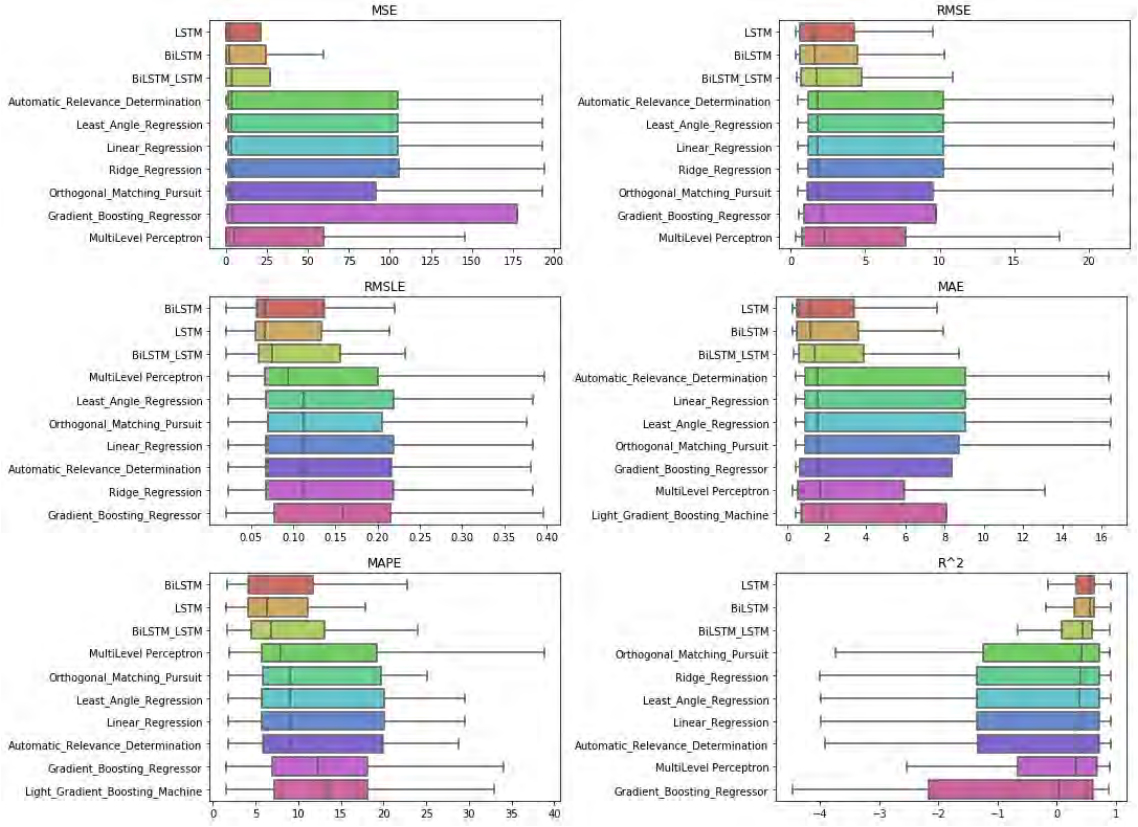
5.5.3.2 Από τη σκοπιά των συναισθηματικών διαμορφώσεων

Συνεχίζοντας με τα αποτελέσματα υπό το πρίσμα των διαφορετικών συναισθηματικών διαμορφώσεων, στη συνέχεια παρουσιάζονται τα αποτελέσματα και πάλι ανά χρονικό ορίζοντα, ξεκινώντας από των μικρότερο και συνεχίζοντας στους μεγαλύτερους.

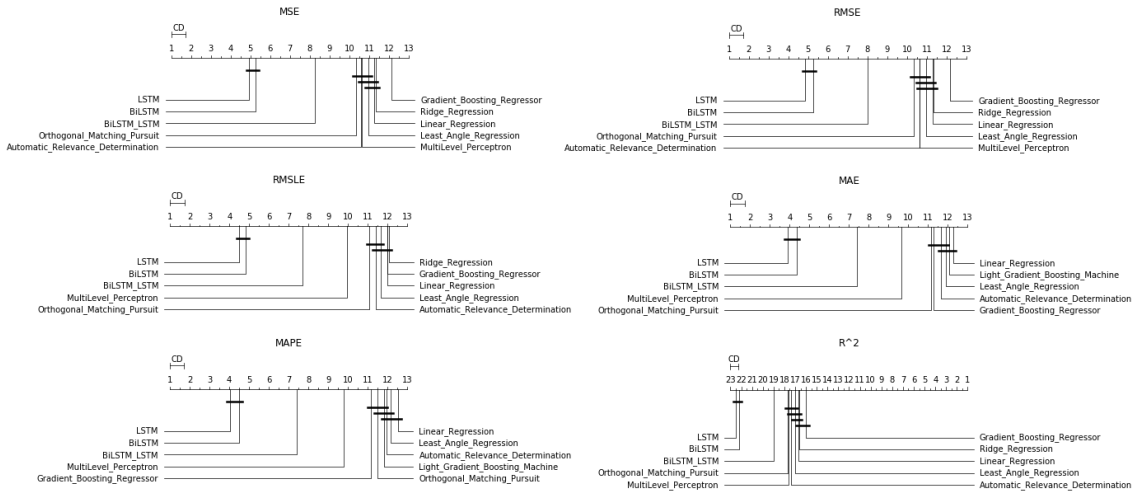
Χρονικός ορίζοντας μιας ημέρας Αρχικά, όσον αφορά την πρόβλεψη της επόμενης ημέρας, ο Πίνακας ;; καταδεικνύει τη γενική υπεροχή της περίπτωσης όπου η πρόβλεψη γίνεται μόνο με τη χρήση των ιστορικών δεδομένων της μετοχής. Τα

	MSE		RMSE		RMSLE	
	Method	**F-Rank**	**Method**	**F-Rank**	**Method**	**F-Rank**
1	LSTM	4.9403	LSTM	4.8523	LSTM	4.4915
2	LSTM_2	5.2727	LSTM_2	5.2358	LSTM_2	4.8239
3	LSTM_3	8.2415	LSTM_3	8.0057	LSTM_3	7.6960
4	OMP	10.3196	OMP	10.3366	MLP	9.9375
5	ARD	10.6051	ARD	10.6193	OMP	11.0753
6	MLP	10.6193	MLP	10.6449	ARD	11.3807
7	LA	10.9673	LA	10.9815	LA	11.6378
8	LNR	11.2699	LNR	11.2841	LNR	11.9830
9	RDG	11.3438	RDG	11.3580	GB	11.9886
10	GB	12.1477	GB	12.1733	RDG	12.0540
11	LGBM	12.7244	LGBM	12.7500	LGBM	12.9119
12	ET	13.6591	ET	13.6818	ABR	13.6903
13	HBR	13.6761	HBR	13.6932	ET	13.8892
14	CBR	13.9091	CBR	13.9091	CBR	13.8949
15	ABR	14.2528	ABR	14.2727	RF	14.4830
16	RF	14.3892	RF	14.4006	THS	14.5923
17	THS	14.4489	THS	14.4631	HBR	14.7045
18	RSC	14.9460	RSC	14.9716	LAS	15.4830
19	LAS	16.1080	LAS	16.1364	RSC	15.5611
20	KNR	16.1960	KNR	16.2159	KNR	16.0057
21	XGBoost	16.2528	XGBoost	16.2614	XGBoost	16.6307
22	DTR	17.7557	DTR	17.7699	SVR	17.0739
	MAE		**MAPE**		**R^2**	
	Method	**F-Rank**	**Method**	**F-Rank**	**Method**	**F-Rank**
1	LSTM	3.9006	LSTM	4.0227	LSTM	22.4744
2	LSTM_2	4.3409	LSTM_2	4.5142	LSTM_2	22.2045
3	LSTM_3	7.4148	LSTM_3	7.4006	LSTM_3	18.9972
4	MLP	9.6364	MLP	9.7699	OMP	17.6690
5	OMP	11.1662	GB	11.1477	MLP	17.5909
6	GB	11.2614	OMP	11.4986	ARD	17.3864
7	ARD	11.6676	LGBM	11.8466	LA	17.0327
8	LA	11.8821	ARD	11.9460	LNR	16.7301
9	LGBM	12.0682	LA	12.1491	RDG	16.6506
10	LNR	12.2557	LNR	12.5341	GB	15.9858
11	RDG	12.4460	RDG	12.6705	LGBM	15.4034
12	CBR	12.9972	CBR	12.9574	ET	14.5057
13	ET	13.2131	ET	12.9716	HBR	14.2955
14	ABR	13.6080	ABR	13.2670	CBR	14.2670
15	RF	13.7330	RF	13.3153	ABR	13.8494
16	HBR	14.9148	KNR	15.0085	RF	13.8295
17	KNR	15.3920	HBR	15.1847	THS	13.5114
18	THS	15.3920	XGBoost	15.3494	RSC	13.1307
19	XGBoost	15.5795	THS	15.5909	KNR	11.9659
20	RSC	15.9688	SVR	16.1364	XGBoost	11.9403
21	SVR	16.2642	RSC	16.2159	LAS	11.9063
22	LAS	17.4261	LAS	17.2273	SVR	10.3210

Πίνακας 5.11: Κατάταξη Friedman (Μέθοδοι) - Ορίζοντας πρόβλεψης μιας εβδομάδας



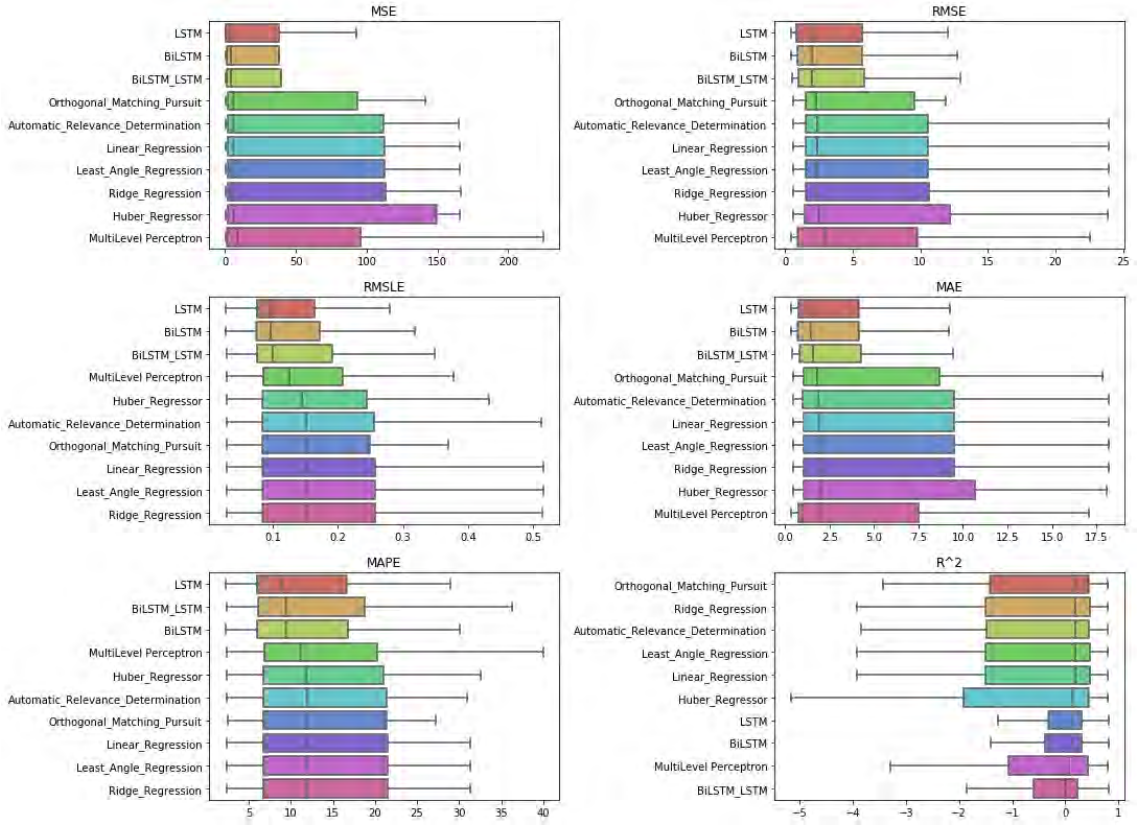
Σχήμα 5.15: Θηκογράμματα για την απόδοση των αλγορίθμων - Χρονικός ορίζοντας μιας εβδομαδας



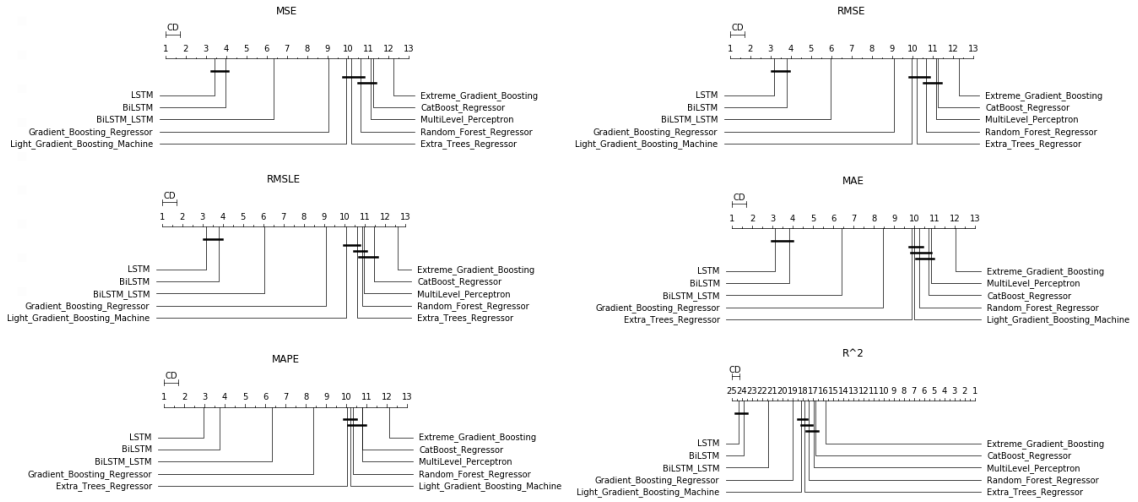
Σχήμα 5.16: CD διαγράμματα για την απόδοση των αλγορίθμων - Χρονικός ορίζοντας μιας εβδομαδας

	MSE		RMSE		RMSLE	
	Method	**F-Rank**	**Method**	**F-Rank**	**Method**	**F-Rank**
1	LSTM	5.4886	LSTM	5.3636	LSTM	4.9460
2	LSTM_2	5.7216	LSTM_2	5.5881	LSTM_2	5.1136
3	OMP	8.6733	LSTM_3	8.4432	LSTM_3	8.0483
4	LSTM_3	8.6790	OMP	8.7017	MLP	9.4886
5	ARD	9.1165	ARD	9.1563	OMP	9.5313
6	LA	10.0611	LA	10.0923	ARD	9.8409
7	MLP	10.1847	MLP	10.2017	LA	10.6491
8	LNR	10.2088	LNR	10.2401	LNR	10.8906
9	RDG	10.2983	RDG	10.3295	RDG	11.0199
10	HBR	11.8636	HBR	11.8892	HBR	12.5852
11	ABR	13.2642	ABR	13.2983	ABR	13.1591
12	RSC	13.4631	RSC	13.4858	THS	13.7230
13	THS	13.5369	THS	13.5511	LAS	13.7500
14	GB	13.9801	GB	13.9943	RSC	13.7827
15	LAS	14.4233	LAS	14.4744	GB	14.1676
16	LGBM	15.0398	LGBM	15.0653	LGBM	15.3608
17	RF	16.1847	RF	16.2017	ET	16.4602
18	ET	16.2017	CBR	16.2159	CBR	16.5426
19	CBR	16.2045	ET	16.2244	RF	16.5625
20	KNR	16.5455	KNR	16.5710	KNR	16.6108
21	SVR	17.4034	SVR	17.4176	SVR	17.0483
22	XGBoost	17.7301	XGBoost	17.7500	ELN	17.5426
	MAE		**MAPE**		**R^2**	
	Method	**F-Rank**	**Method**	**F-Rank**	**Method**	**F-Rank**
1	LSTM	5.5398	LSTM	5.1563	LSTM	20.6932
2	LSTM_2	5.7330	LSTM_2	5.3210	LSTM_2	20.4688
3	LSTM_3	8.7841	LSTM_3	8.4574	OMP	19.6420
4	OMP	8.8778	MLP	9.1705	ARD	19.1619
5	ARD	9.1761	OMP	9.3636	LA	18.2116
6	MLP	9.2955	ARD	9.6534	LNR	18.0639
7	LA	10.0866	LA	10.4389	MLP	18.0483
8	LNR	10.2741	LNR	10.6406	RDG	17.9716
9	RDG	10.4176	RDG	10.8182	LSTM_3	17.6960
10	HBR	11.9432	HBR	12.3949	HBR	16.4830
11	ABR	13.1847	ABR	12.8381	ABR	14.8722
12	THS	13.7386	GB	13.6193	RSC	14.7670
13	GB	13.7699	LGBM	14.1648	THS	14.5000
14	RSC	13.8210	RSC	14.2443	GB	14.2472
15	LGBM	14.3920	THS	14.2841	LAS	13.9773
16	CBR	15.6136	CBR	15.2955	LGBM	13.1903
17	ET	15.7670	ET	15.4261	RF	12.1136
18	RF	15.8409	RF	15.4489	ET	12.0767
19	LAS	15.9716	KNR	15.8153	CBR	12.0170
20	KNR	15.9744	LAS	16.2188	KNR	11.7102
21	SVR	16.4602	SVR	16.3579	SVR	10.8097
22	XGBoost	17.7216	XGBoost	17.4205	XGBoost	10.5540

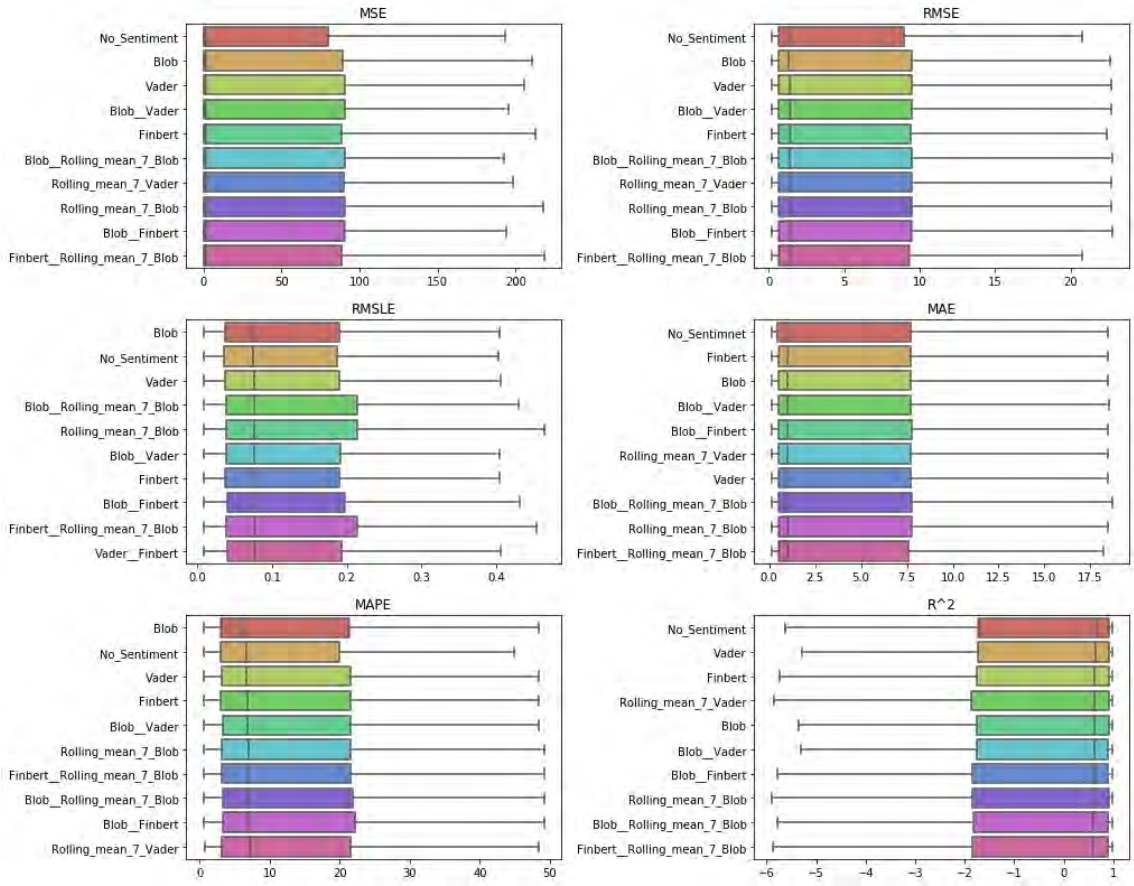
Πίνακας 5.12: Κατάταξη Friedman (Μέθοδοι) - Ορίζοντας πρόβλεψης δύο εβδομάδων



Σχήμα 5.17: Θηκογράμματα για την απόδοση των αλγορίθμων - Χρονικός ορίζοντας δύο εβδομάδων



Σχήμα 5.18: CD διαγράμματα για την απόδοση των αλγορίθμων - Χρονικός ορίζοντας δύο εβδομάδων



Σχήμα 5.19: Θηκογράμματα για την απόδοση των συναισθηματικών διαμορφώσεων - Χρονικός ορίζοντας μιας ημέρας

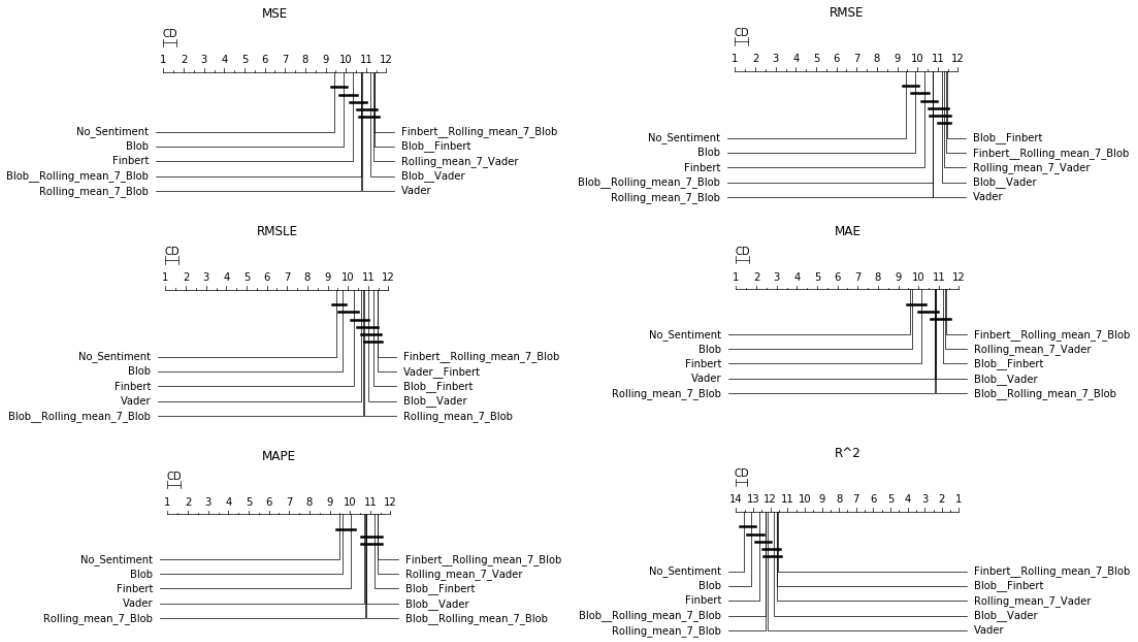
διαγράμματα κρίσιμων διαφορών (Σχήμα 5.20) δείχνουν πως η διαφορά αυτή δεν είναι στατιστικά σημαντική σε σχέση με την διαμόρφωση που καταλαμβάνει τη δεύτερη θέση, ενώ στην τρίτη θέση στην συντριπτική πλειοψηφία των περιπτώσεων εμφανίζεται η Finbert διαμόρφωση. Τα θηκογράμματα που παρουσιάζονται στο Σχήμα 5.19 και απεικονίζουν τους δέκα κορυφαίους συνδυασμούς συναισθήματος για κάθε μετρική επικυρώνουν την κυριαρχία της περίπτωσης όπου δεν χρησιμοποιήθηκε ανάλυση συναισθήματος, αν και στις μετρικές MAPE και RMSLE η περίπτωση χρήσης της τεχνικής Blob παρουσιάζει μικρότερη διάμεση τιμή. Πρέπει επίσης να σημειωθεί ότι η εμφανής ομοιότητα στις αποδόσεις των μεθόδων οφείλεται, αφενός, στην κλίμακα της αναπαράστασης, ενώ αφετέρου, αντικατοπτρίζει μια πιθανή ομοιομορφία.

5.5.3.3 Χρονικός ορίζοντας μιας εβδομάδας

Στο χρονικό ορίζοντα πρόβλεψης μιας εβδομάδας το τοπίο αλλάζει ριζικά. Όπως φαίνεται από την κατάταξη Friedman των εβδομαδιαίων προβλέψεων, μπορεί κανείς να παρατηρήσει από τον Πίνακα ;; αισθητή πτώση στην κατάταξη της διαμόρφωσης όπου δεν γίνεται χρήση ανάλυσης συναισθήματος, συνοδευόμενη από μια ταυτόχρονη βελτίωση των διαμορφώσεων που αξιοποιούν βαθμολογίες συναισθήματος. Συγκεκριμένα, σε τέσσερις από τις έξι μετρικές, η διαμόρφωση που βασίζεται μόνο στη μέθοδο FinBERT φαίνεται να υπερέχει, ενώ στις υπόλοιπες δύο, ο συνδυασμός των FinBERT και TextBlob κατέλαβε την πρώτη θέση στην κατάταξη. Πέρα από αυτό, οι VADER,

	MSE		RMSE		RMSLE	
	Method	**F-Rank**	**Method**	**F-Rank**	**Method**	**F-Rank**
1	NS	9.4456	NS	9.4410	NS	9.4201
2	B	10.3576	B	10.3576	B	9.7106
3	F	10.3576	F	10.3576	F	10.3009
4	BB7	10.7326	BB7	10.7303	V	10.6435
5	B7	10.7431	B7	10.7454	BB7	10.7755
6	V	10.8102	V	10.7755	B7	10.8137
7	BV	11.2257	BV	11.1956	BV	10.9977
8	V7	11.3542	V7	11.3356	BF	11.2836
9	BF	11.4074	FB7	11.4201	VF	11.4421
10	FB7	11.4236	BF	11.4444	FB7	11.4884
11	VB7	11.4363	VB7	11.4525	V7	11.5069
12	VF	11.4815	VF	11.5023	VB7	11.5706
13	VV7	11.6609	VV7	11.6655	F7	11.6678
14	F7	11.7697	F7	11.7720	FF7	11.7813
15	FF7	11.8449	FF7	11.8345	VV7	11.9201
16	BV7	12.0197	BV7	11.9780	BV7	12.1840
17	BF7	12.1574	BF7	12.2037	BF7	12.1887
18	VF7	12.2801	VF7	12.2662	VF7	12.2153
19	FV7	12.4676	FV7	12.5116	B7V7	12.6169
20	B7F7	12.7859	B7F7	12.7697	FV7	12.7338
21	V7F7	12.8553	B7V7	12.8576	B7F7	12.7801
22	B7V7	12.8669	V7F7	12.8623	V7F7	12.9583
	MAE		**MAPE**		**R^2**	
	Method	**F-Rank**	**Method**	**F-Rank**	**Method**	**F-Rank**
1	NS	9.5914	NS	9.5035	NS	13.5521
2	B	9.6887	B	9.6169	B	13.1262
3	F	10.1678	F	10.0567	F	12.6400
4	V	10.7905	V	10.7234	BB7	12.2674
5	B7	10.8218	B7	10.7384	B7	12.2581
6	BB7	10.8553	BB7	10.7836	V	12.1910
7	BV	10.8600	BV	10.7940	BV	11.7743
8	BF	11.2176	BF	11.1991	V7	11.6493
9	V7	11.3299	V7	11.3657	BF	11.5926
10	FB7	11.3877	FB7	11.3762	FB7	11.5764
11	VF	11.4861	VF	11.5000	VB7	11.5637
12	VB7	11.6400	VB7	11.6690	VF	11.5185
13	F7	11.6979	FF7	11.7257	VV7	11.3368
14	FF7	11.6979	F7	11.8032	F7	11.2292
15	VV7	11.8785	VV7	11.9248	FF7	11.1563
16	BV7	12.1169	BV7	12.1875	BV7	10.9803
17	BF7	12.1921	BF7	12.2824	BF7	10.8426
18	VF7	12.3403	VF7	12.2824	VF7	10.7211
19	FV7	12.3565	FV7	12.4201	FV7	10.5324
20	B7F7	12.7234	B7F7	12.6470	B7F7	10.2141
21	V7F7	13.0104	V7F7	13.0995	V7F7	10.1447
22	B7V7	13.1493	B7V7	13.1759	B7V7	10.1331

Πίνακας 5.13: Κατάταξη Friedman (Συναισθηματικές διαμορφώσεις) - Ορίζοντας πρόβλεψης μιας ημέρας



Σχήμα 5.20: CD διαγράμματα για την απόδοση των συναισθηματικών διαμορφώσεων - Χρονικός ορίζοντας μιας ημέρας

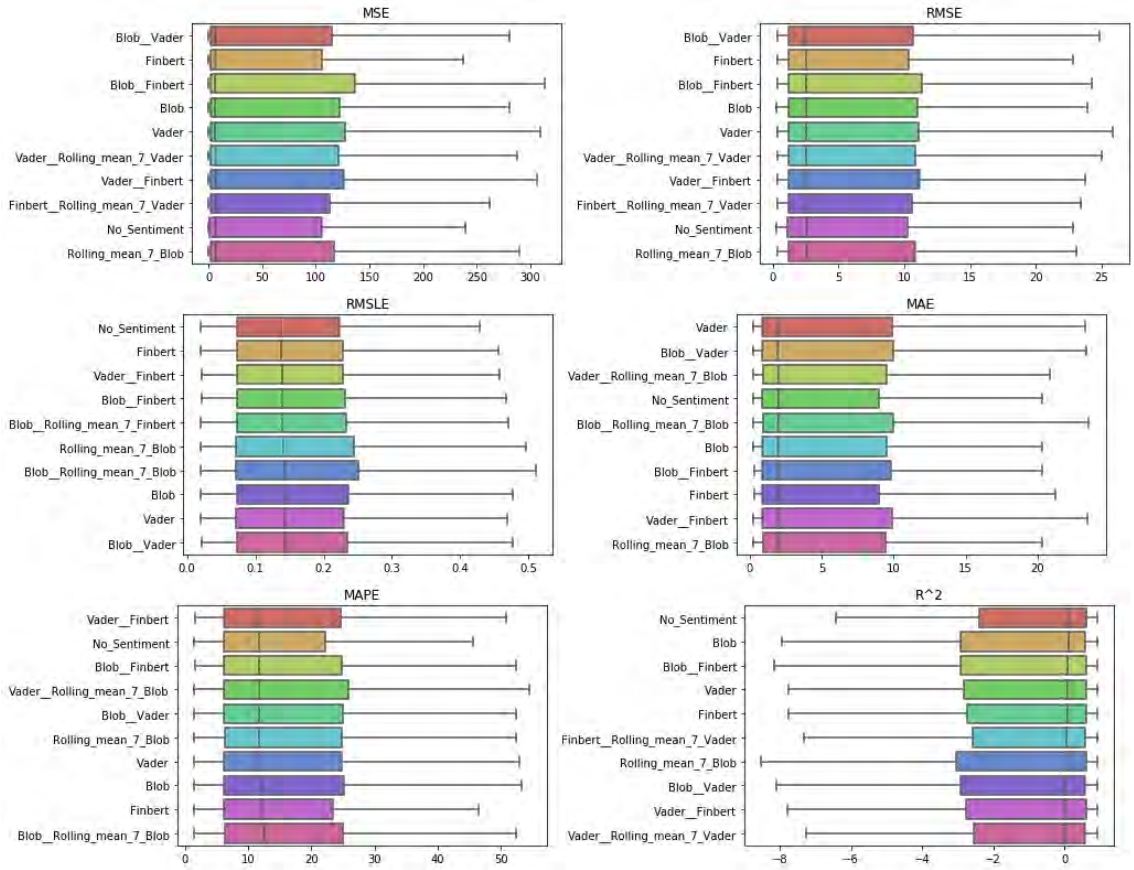
Blob και ο συνδυασμός των VADER και FinBERT παρουσιάζουν σχεδόν ισοδύναμη απόδοση με τις παραπάνω μεθόδους, καθώς οι διαφορές στις αντίστοιχες κατατάξεις ήταν ελάχιστες.

Επιπλέον, όσον αφορά τη χρήση κινητών μέσων όρων (rolling means), δεν παρατηρήθηκε ιδιαίτερη βελτίωση στο παρόν πλαίσιο, εκτός από σπάνιες περιπτώσεις, όπου εφαρμόστηκαν σε συνδυασμό με βαθμολογίες συναισθήματος όπου δεν έγινε χρήση κινητών μέσων. Παρέχονται επίσης και τα διαγράμματα κρίσιμων διαφορών (Σχήμα 5.22), όπου αποτυπώνεται ανά ζευγάρι διαμορφώσεων αν η διαφορά στην απόδοσή τους είναι ή όχι στατιστικά σημαντική.

Η μοναδική απεικόνιση των αποτελεσμάτων στην οποία η διαμόρφωση χωρίς χρήση συναισθήματος εμφανίζεται σε υψηλές θέσεις παρατηρείται μόνο μέσω των θηκογραμμάτων (Σχήμα 5.21) όπου απεικονίζεται η απόδοση των μετρικών. Υπενθυμίζεται, πως στην περίπτωση αυτή οι διαφορετικές συναισθηματικές διαμορφώσεις έχουν ταξινομηθεί με βάση τη διάμεση τιμή μέσα στο θηκόγραμμα, συνεπώς η πληροφορία αυτή υποδηλώνει τη χαμηλή διάμεση τιμή των αντίστοιχων σφαλμάτων. Ωστόσο, με βάση τις βαθμολογίες Friedman, στην καλύτερη περίπτωση, το σενάριο μη χρήσης Ανάλυσης Συναισθήματος κατατάχθηκε στην έκτη θέση.

	MSE		RMSE		RMSLE	
	Method	**F-Rank**	**Method**	**F-Rank**	**Method**	**F-Rank**
1	F	10.4549	F	10.4688	F	10.4444
2	BF	10.5162	BF	10.5301	BF	10.4641
3	V	10.6493	V	10.6285	V	10.7523
4	VF	10.9051	VF	10.9120	VF	10.7836
5	B	10.9225	B	10.9178	B	10.8495
6	NS	11.0683	NS	11.0938	NS	10.9606
7	BV	11.1991	BV	11.1852	BV	11.1458
8	B7	11.2326	B7	11.2396	B7	11.1609
9	FV7	11.3414	FV7	11.3669	BF7	11.3090
10	VV7	11.3900	VV7	11.4016	BB7	11.4248
11	FF7	11.5220	FF7	11.5208	FF7	11.4444
12	BF7	11.5440	BF7	11.5231	VB7	11.4769
13	BB7	11.5625	BB7	11.5440	FB7	11.5243
14	FB7	11.6088	FB7	11.6238	VV7	11.6771
15	BV7	11.7106	BV7	11.6991	VF7	11.7222
16	VB7	11.7396	VB7	11.7396	FV7	11.7442
17	V7	11.7697	V7	11.7650	BV7	11.8981
18	VF7	11.8785	VF7	11.8530	F7	12.1678
19	F7	12.1412	F7	12.1458	V7	12.2384
20	B7V7	12.5486	B7V7	12.5556	B7V7	12.5255
21	V7F7	12.6424	V7F7	12.6285	B7F7	12.5694
22	B7F7	12.6528	B7F7	12.6574	V7F7	12.7164
	MAE		**MAPE**		**R^2**	
	Method	**F-Rank**	**Method**	**F-Rank**	**Method**	**F-Rank**
1	BF	10.4595	BF	10.3808	F	12.5451
2	B	10.6794	B	10.6782	BF	12.4850
3	F	10.6863	V	10.7002	V	12.3507
4	V	10.7106	F	10.7396	VF	12.0949
5	B7	10.8553	B7	10.8530	B	12.0775
6	BB7	10.8785	BV	10.8692	NS	11.9317
7	BV	10.8831	VF	10.9016	BV	11.8009
8	VF	10.9850	BB7	10.9248	B7	11.7674
9	NS	11.0289	NS	11.0093	FV7	11.6597
10	VB7	11.1875	VB7	11.0347	VV7	11.6100
11	FB7	11.4201	FB7	11.4271	FF7	11.4780
12	BF7	11.6204	BF7	11.5231	BF7	11.4537
13	VV7	11.7419	VV7	11.7650	BB7	11.4387
14	FF7	11.7731	FF7	11.7801	FB7	11.3912
15	FV7	11.8530	VF7	11.9005	BV7	11.2859
16	VF7	11.9097	FV7	12.0475	VB7	11.2604
17	BV7	11.9132	BV7	12.0810	V7	11.2303
18	V7	12.1759	V7	12.1308	VF7	11.1215
19	F7	12.2639	F7	12.2269	F7	10.8588
20	B7F7	12.5000	B7F7	12.5394	B7V7	10.4525
21	B7V7	12.6331	B7V7	12.5417	V7F7	10.3588
22	V7F7	12.8414	V7F7	12.9456	B7F7	10.3472

Πίνακας 5.14: Κατάταξη Friedman (Συναισθηματικές διαμορφώσεις) - Ορίζοντας πρόβλεψης μιας εβδομάδας



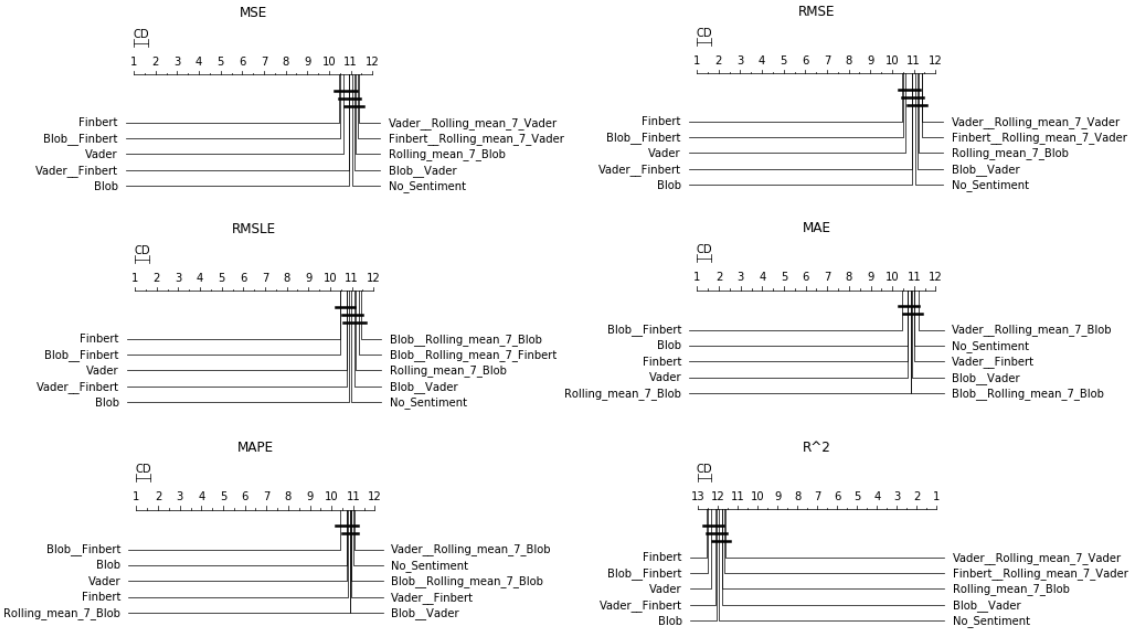
Σχήμα 5.21: Θηκογράμματα για την απόδοση των συναισθηματικών διαμορφώσεων - Χρονικός ορίζοντας μιας εβδομάδας

5.5.3.4 Χρονικός ορίζοντας δυο εβδομάδων

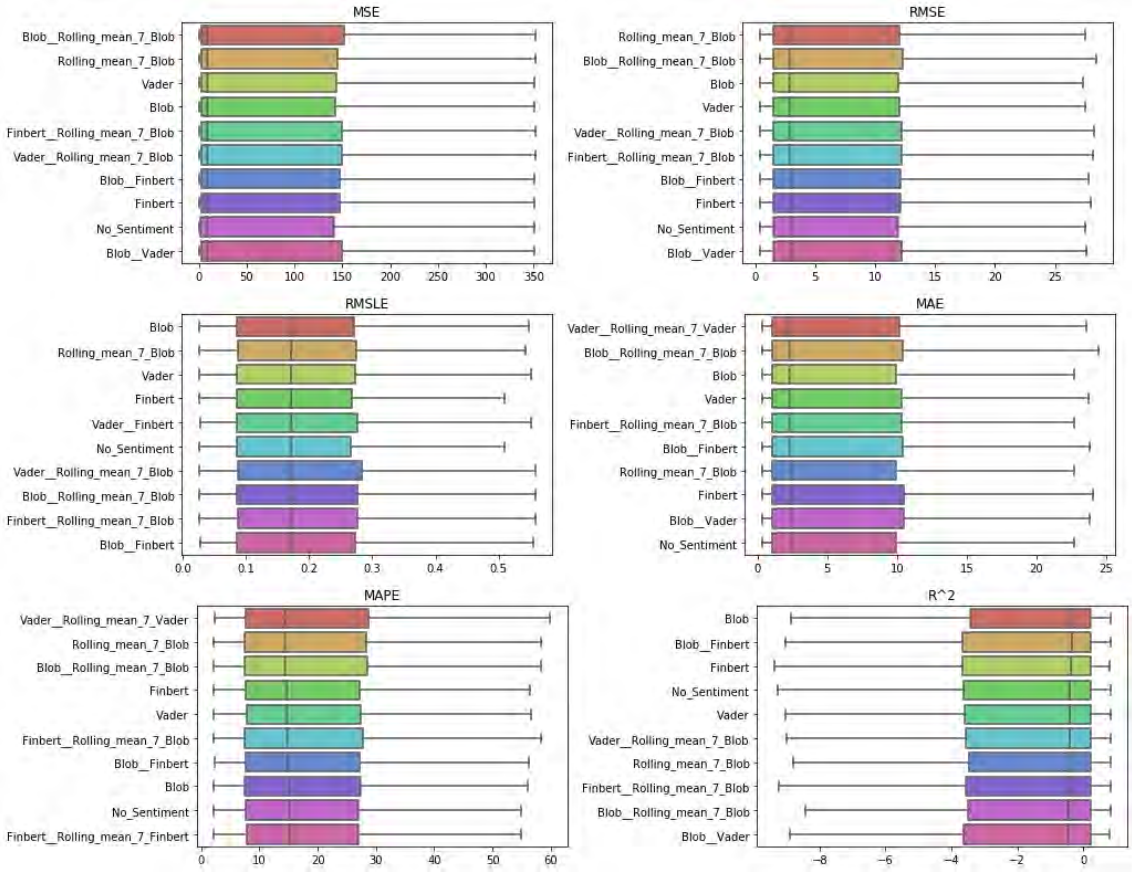
Τα αποτελέσματα των προβλέψεων δεκατεσσάρων ημερών παρουσίασαν παρόμοια συμπεριφορά με εκείνα των προβλέψεων επτά ημερών, με τη μοναδική εξαίρεση να αφορά την απόδοση των σχημάτων που περιλαμβάνουν κινητούς μέσους, ορισμένα από τα οποία κατέλαβαν υψηλότερες θέσεις στην κατάταξη. Πράγματι, και εδώ, η κατάταξη Friedman (Πίνακας ;;) σε όλες τις αξιολογήσεις υποδεικνύει ότι η αξιοποίηση πληροφοριών από τα μέσα κοινωνικής δικτύωσης είναι ευεργετική στο παρόν πλαίσιο πρόβλεψης.

Επιπροσθέτως, η σαφής βελτίωση των σχημάτων που χρησιμοποιούν κινητούς μέσους όρους γίνεται ιδιαίτερα εμφανής στα Σχήματα ;; και 5.23, τα οποία απεικονίζουν αντίστοιχα τα διαγράμματα κρίσιμων διαφορών και θηκογράμματα της απόδοσης των διαφορετικών συναισθηματικών διαμορφώσεων.

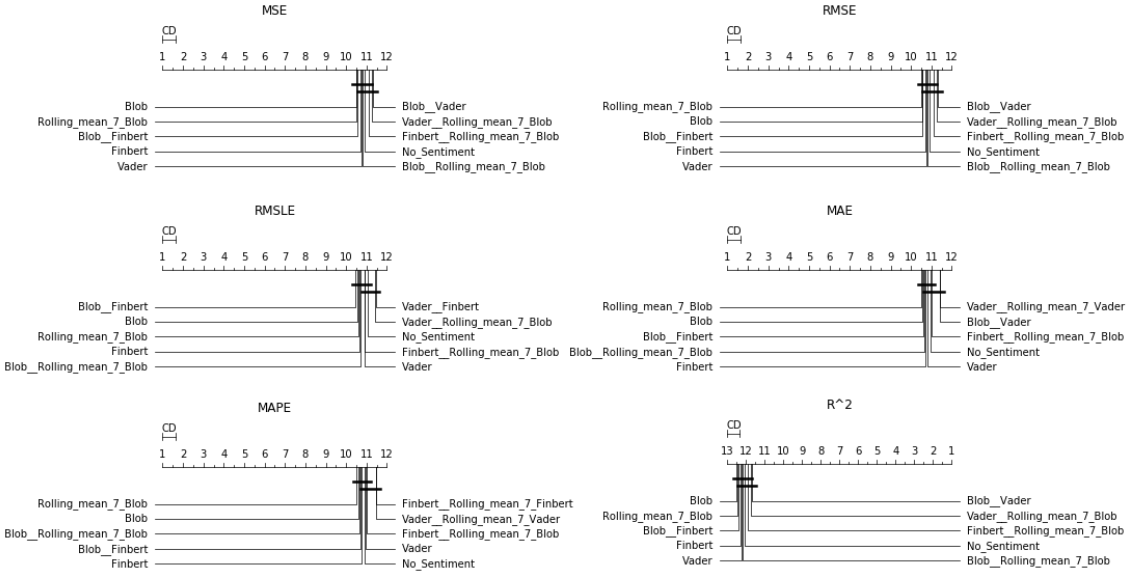
Εύκολα παρατηρείται ότι η διαμόρφωση του TextBlob που ενσωματώνει τον εβδομαδιαίο κινητό μέσο όρο καταλαμβάνει την πρώτη θέση στην κατάταξη Friedman για τρεις μετρικές απόδοσης (RMSE, MAE και MAPE). Έτσι, πέρα από τα συμπεράσματα που μπορούν να εξαχθούν από τη μελέτη των απεικονίσεων των αποτελεσμάτων—τα οποία ακολουθούν παρόμοιο μοτίβο με τις προηγούμενες περιπτώσεις—φαίνεται να προκύπτει ένα νέο εύρημα. Υπήρξε σταδιακή βελτίωση της απόδοσης των συνδυασμών που χρησιμοποιούν σταθμισμένες πληροφορίες. Επιπλέον, αυτή η βελτίωση φαίνεται να σχετίζεται άμεσα με το μεγαλύτερο χρονικό ορίζοντα πρόβλεψης.



Σχήμα 5.22: CD διαγράμματα για την απόδοση των συναισθηματικών διαμορφώσεων - Χρονικός ορίζοντας μιας εβδομάδας



Σχήμα 5.23: Θηκογράμματα για την απόδοση των συναισθηματικών διαμορφώσεων - Χρονικός ορίζοντας δύο εβδομάδων



Σχήμα 5.24: CD διαγράμματα για την απόδοση των συναισθηματικών διαμορφώσεων - Χρονικός ορίζοντας δύο εβδομάδων

5.5.3.5 Συμπεράσματα

Όσον αφορά τους αλγορίθμους, οι συγκρίσεις παρείχαν φαίνεται πως επιδέχονται άμεση και ξεκάθαρη ερμηνεία. Από τα αποτελέσματα, μπορεί να υποστηριχθεί με ασφάλεια ότι οι διαμορφώσεις που αξιοποιούν νευρωνικά δίκτυα—δηλαδή οι παραλλαγές LSTM—υπερείχαν σε απόδοση των κλασικών αλγορίθμων παλινδρόμησης. Μεταξύ αυτών, η LSTM ξεπέρασε την αρχιτεκτονική Bi-LSTM σε κάθε περίπτωση, ενώ ο stacking συνδυασμός των δύο ακολούθησε. Επιπλέον, η ανωτερότητα των δύο κυρίαρχων μεθόδων ήταν ξεκάθαρη, με την απόδοσή τους να σχηματίζει ένα κατώφλι, κάτω από το οποίο—και σε σημαντική απόσταση—βρέθηκαν όλες οι άλλες εξεταζόμενες μέθοδοι.

Συνεπώς, καταλήγοντας, αν ληφθεί υπόψη ότι οι χρησιμοποιούμενες αρχιτεκτονικές νευρωνικών δικτύων δεν περιείχαν σύνθετες διαμορφώσεις—για παράδειγμα, σε όρους βάθους—τότε, δεδομένου ότι οποιοδήποτε πρόσθετο υπολογιστικό κόστος γίνεται αμελητέο, η χρήση των LSTMs αποτελεί την ξεκάθαρη επιλογή.

Σχετικά με τη σύγκριση των διαφορετικών συναισθηματικών διαμορφώσεων, το κύριο συμπέρασμα που προκύπτει είναι ότι η χρήση ανάλυσης συναισθήματος φαίνεται να βελτιώνει τα μοντέλα όταν εφαρμόζεται σε μακροπρόθεσμες προβλέψεις. Ειδικότερα, ενώ η διαμόρφωση που βασίζεται μόνο στα ιστορικά δεδομένα χωρίς τη χρήση Ανάλυσης Συναισθήματος εμφανίζεται πιο αποδοτική στις προβλέψεις μίας ημέρας, όταν οι προβλέψεις επεκτείνονται στις επτά και δεκατέσσερις ημέρες, η χρήση διαμορφώσεων που αξιοποιούν Ανάλυση Συναισθήματος φαίνεται να βελτιώνει συνολικά τις προβλέψεις.

Συγκεκριμένα, στις εβδομαδιαίες προβλέψεις, όλες οι διαμορφώσεις που βασίζονται σε μία μέθοδο Ανάλυσης Συναισθήματος υπερέβησαν την προσέγγιση που δε χρησιμοποιεί ανάλυση συναισθήματος, με την FinBERT να αποδίδει καλύτερα στις MSE, RMSE, RMSLE και R^2, ενώ ο συνδυασμός Blob και FinBERT υπερείχε στις MAE και MAPE. Όταν το χρονικό πλαίσιο πρόβλεψης διπλασιάζεται σε 14 ημέρες, παρατηρείται ότι οι Blob και Rolling Mean 7 Blob κυριαρχούν έναντι των υπόλοιπων διαμορφώσεων συναισθήματος, ακολουθούμενοι από τον συνδυασμό Blob και FinBERT, καθώς και την FinBERT. Αντίθετα, η VADER κατέλαβε χαμηλότερες θέσεις σε όλες

	MSE		RMSE		RMSLE	
	Method	**F-Rank**	**Method**	**F-Rank**	**Method**	**F-Rank**
1	B	10.4873	B7	10.5081	BF	10.4653
2	B7	10.5046	B	10.5278	B	10.5313
3	BF	10.5660	BF	10.5521	B7	10.6030
4	F	10.7095	F	10.7199	F	10.6655
5	V	10.7593	V	10.7488	BB7	10.7072
6	BB7	10.7975	BB7	10.8021	V	10.9248
7	NS	10.9259	NS	10.8958	FB7	10.9294
8	FB7	11.1019	FB7	11.0903	NS	11.0498
9	VB7	11.2569	VB7	11.2824	VB7	11.4132
10	BV	11.3275	BV	11.3275	VF	11.4560
11	VF	11.3762	VF	11.3715	FF7	11.4630
12	VV7	11.5278	VV7	11.5208	BV	11.4699
13	V7	11.7002	V7	11.7106	VV7	11.7558
14	FF7	11.7593	FF7	11.7523	BF7	11.7824
15	BV7	11.8472	BV7	11.8414	BV7	11.8727
16	BF7	12.0012	BF7	11.9965	F7	11.9271
17	B7F7	12.0185	B7F7	12.0208	V7	11.9456
18	B7V7	12.0509	FV7	12.0567	FV7	12.1285
19	FV7	12.0625	B7V7	12.0660	B7F7	12.1551
20	F7	12.3056	F7	12.3067	VF7	12.2118
21	VF7	12.4711	VF7	12.4549	B7V7	12.2755
22	V7F7	13.4433	V7F7	13.4468	V7F7	13.2674
	MAE		**MAPE**		**R^2**	
	Method	**F-Rank**	**Method**	**F-Rank**	**Method**	**F-Rank**
1	B7	10.5116	B7	10.5069	B	12.5093
2	B	10.5313	B	10.6053	B7	12.4965
3	BF	10.5949	BB7	10.6435	BF	12.4317
4	BB7	10.6296	BF	10.7060	F	12.2917
5	F	10.6933	F	10.7419	V	12.2407
6	V	10.7951	NS	10.8843	BB7	12.2072
7	NS	10.9491	V	10.9363	NS	12.0741
8	FB7	10.9942	FB7	11.0336	FB7	11.8958
9	BV	11.4109	VV7	11.4421	VB7	11.7431
10	VV7	11.4282	FF7	11.4734	BV	11.6725
11	V7	11.4306	VB7	11.4954	VF	11.6238
12	VB7	11.4468	V7	11.5093	VV7	11.4722
13	VF	11.5938	VF	11.6632	V7	11.2998
14	BV7	11.6400	BV	11.6863	FF7	11.2384
15	FF7	11.6748	FV7	11.8021	BV7	11.1528
16	FV7	11.8090	BF7	11.8194	BF7	11.0035
17	BF7	12.0023	BV7	11.8646	B7F7	10.9792
18	B7F7	12.1701	B7F7	12.1076	B7V7	10.9491
19	F7	12.3935	F7	12.1458	FV7	10.9375
20	B7V7	12.4664	VF7	12.3669	F7	10.6956
21	VF7	12.5266	B7V7	12.5359	VF7	10.5289
22	V7F7	13.3079	V7F7	13.0301	V7F7	9.5567

Πίνακας 5.15: Κατάταξη Friedman (Συναισθηματικές διαμορφώσεις) - Ορίζοντας πρόβλεψης δύο εβδομάδων

τις μετρικές, παρουσιάζοντας ασθενέστερη απόδοση σε σύγκριση με τις δύο προηγούμενες περιπτώσεις.

Ωστόσο, πρέπει να τονιστεί πως η χρήση της Ανάλυσης Συναισθήματος εξετάζεται ανεξάρτητα από την απόδοση της εκάστοτε μεθόδου παλινδρόμησης και αυτό γιατί σε ένα πρακτικό σενάριο όπου απαιτούνται αξιόπιστες προβλέψεις, κάποιος θα είχε στη διάθεσή του πολλαπλές μεθόδους. Το κεντρικό ερώτημα λοιπόν, εστιάζει στο αν ή όχι η χρήση βαθμολογιών συναισθήματος βελτιώνει τις προβλέψεις, και αν ναι, σε ποιες περιπτώσεις. Επιπλέον, εγείρεται το ερώτημα του ποια μορφή θα πρέπει να έχει η χρονοσειρά των βαθμολογιών συναισθήματος, ανάλογα με το εκάστοτε σενάριο πρόβλεψης.

Στο πρώτο σκέλος, η απάντηση είναι ξεκάθαρη. Οι διαμορφώσεις που χρησιμοποιούν ανάλυση συναισθήματος βελτιώνουν τις προβλέψεις σε πιο απαιτητικά σενάρια πρόβλεψης, δηλαδή όταν ο χρονικός ορίζοντας μεγαλώνει. Όσον αφορά το δεύτερο σκέλος του ερωτήματος, φαίνεται ότι στις μακροπρόθεσμες προβλέψεις, από τα εξαγόμενα αποτελέσματα ενισχύεται η επιλογή της χρήσης κινητών μέσων όρων των απλών μεθόδων αποτίμησης των συναισθημάτων. Μπορεί αυτή η βελτίωση να μην φαίνεται εκ πρώτης όψεως σημαντική, όμως πρέπει να σημειωθεί ότι όταν το χρονικό παράθυρο πρόβλεψης αυξάνεται, ακόμη και φαινομενικά μικρές βελτιώσεις, όπως αυτές που παρατηρούνται στη μελέτη μας, μπορεί να αποκτήσουν ιδιαίτερη σημασία.

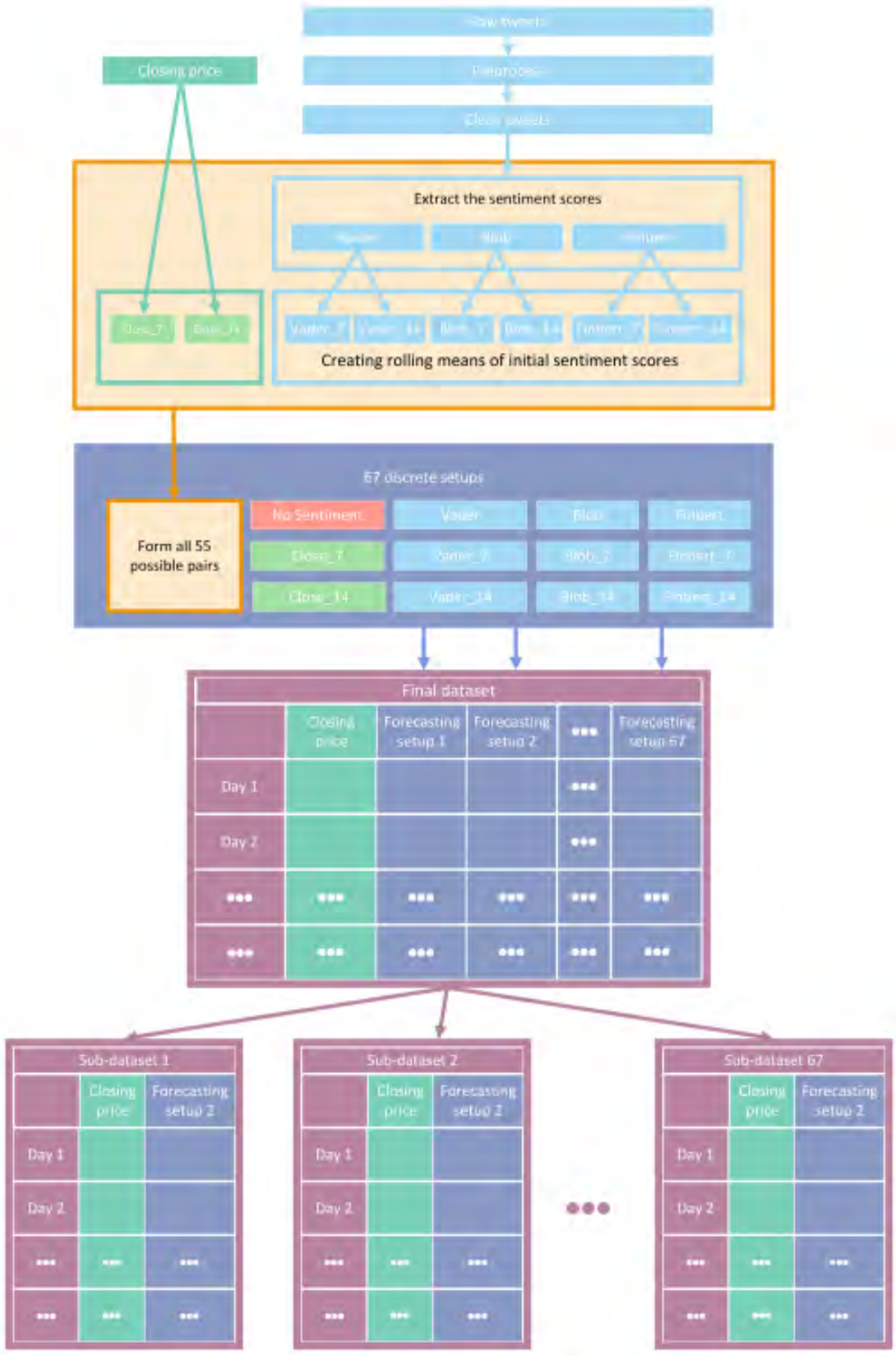
5.5.4 Χρήση Αλγορίθμων Βαθιάς Μάθησης

Σε συνέχεια της προηγούμενης συγκριτικής μελέτης και με έναυσμα την υπεροχή των μεθόδων Βαθιάς Μάθησης, υλοποιήθηκε μια δεύτερη εκτεταμένη πειραματική διάταξη, αυτή τη φορά με χρήση αποκλειστικά τεχνικών Βαθιάς Μάθησης. Χρησιμοποιήθηκαν εκ νέου οι 16 μετοχές που χρησιμοποιήθηκαν και στην προηγούμενη πειραματική διαδικασία, μόνο που αυτή τη φορά – και πάλι με τη χρήση κινητών μέσων, δημιουργήθηκαν περισσότερα χαρακτηριστικά. Συγκεκριμένα, μετά τη δημιουργία των τριών αρχικών χρονοσειρών πολικοτήτων, εφαρμόστηκαν κινητοί μέσοι όροι 7 και 14 ημερών στις παραπάνω εξαγόμενες χρονοσειρές. Αυτό είχε ως αποτέλεσμα την παραγωγή εννέα διαφορετικών χρονοσειρών συναισθηματικών δεδομένων, οι οποίες, μαζί με τις αρχικές τιμές κλεισίματος των μετοχών και τις δικές τους αντίστοιχες κινητές μέσες τιμές, οδήγησαν σε 12 συνολικά χαρακτηριστικά. Οι διαφορετικοί συνδυασμοί των παραπάνω χαρακτηριστικών, καθώς και η εξέταση του σεναρίου όπου δεν λαμβάνεται υπόψιν η ανάλυση συναισθήματος, οδήγησαν στη διαμόρφωση 67 διαφορετικών διαμορφώσεων εισόδου, οι οποίες χρησιμοποιήθηκαν στην πειραματική διαδικασία για κάθε έναν από τους συγκρινόμενους αλγορίθμους. Η παραπάνω διαδικασία εμφανίζεται διαγραμματικά στο Σχήμα ;;.

5.5.4.1 Αλγόριθμοι

Στη συνέχεια, παρατίθενται οι μέθοδοι, τα αλγοριθμικά σχήματα και οι αρχιτεκτονικές που χρησιμοποιήθηκαν στην πειραματική διαδικασία, καθώς και επιπλέον λεπτομέρειες σχετικά με το πλαίσιο υλοποίησης και τα εργαλεία που χρησιμοποιήθηκαν.

Όσον αφορά τους αλγορίθμους που χρησιμοποιήθηκαν, συγκρίθηκαν συνολικά 30 διαφορετικές σύγχρονες μέθοδοι και παραλλαγές τους. Όλοι οι αλγόριθμοι υλοποιήθηκαν μέσω της βιβλιοθήκης της Python tsAI, όπου και υπάρχει πλήρης περιγραφή των μεθόδων. Ωστόσο, η πληθώρα των μεθόδων καθιστά αδύνατη την εκτενή παρουσίασή τους στο παρόν κείμενο. Ο Πίνακας 5.16 περιλαμβάνει τους κύριους αλγορίθμους που



Σχήμα 5.25: Διαγραμματική αναπαράσταση της διαδικασίας εξαγωγής των 67 συναισθηματικών διαμορφώσεων

χρησιμοποιήθηκαν κατά τη διάρκεια της πειραματικής διαδικασίας, μαζί με τις αντίστοιχες παραπομπές τους. Εκεί, μεταξύ άλλων, μπορεί κανείς να παρατηρήσει ότι, εκτός από έναν μεγάλο αριθμό σύγχρονων μεθόδων, χρησιμοποιήθηκαν επίσης και υλοποιήσεις που συνδυάζουν διαφορετικές αρχιτεκτονικές. Σε αντιστοιχία με τα προηγούμενα, θα παρουσιαστεί συνοπτικά μονάχα η κυρίαρχη μέθοδος, όπως αυτή εντοπίζεται στην πειραματική διαδικασία.

Πίνακας 5.16: Αλγόριθμοι Βαθιάς Μάθησης

No.	Abbreviation	Algorithm
1	FCN	Fully Convolutional Network [288]
2	FCNPlus	Fully Convolutional Network Plus [186]
3	IT	Inception Time [81]
4	ITPlus	Inception Time Plus [187]
5	MLP	Multilayer Perceptron [288]
6	RNN	Recurrent Neural Network [191]
7	LSTM	Long Short-Term Memory [119]
8	GRU	Gated Recurrent Unit [49]
9	RNNPlus	Recurrent Neural Network Plus [191]
10	LSTMPlus	Long Short-Term Memory Plus [191]
11	GRUPlus	Gated Recurrent Unit Plus [191]
12	RNN_FCN	Recurrent Neural—Fully Convolutional Network [189]
13	LSTM_FCN	Long Short-Term Memory—Fully Convolutional Network [134]
14	GRU_FCN	Gated Recurrent Unit—Fully Convolutional Network [76]
15	RNN_FCNPlus	Recurrent Neural—Fully Convolutional Network Plus [190]
16	LSTM_FCNPlus	Long Short-Term Memory—Fully Convolutional Network Plus [190]
17	GRU_FCNPlus	Gated Recurrent Unit—Fully Convolutional Network Plus [190]
18	ResCNN	Residual—Convolutional Neural Network [315]
19	ResNet	Residual Network [288]
20	RestNetPlus	Residual Network Plus [188]
21	TCN	Temporal Convolutional Network [21]
22	TST	Time Series Transformer [307]
23	TSTPlus	Time Series Transformer Plus [194]
24	TSiTPlus	Time Series Vision Transformer Plus [193]
25	Transformer	Transformer Model [192]
26	XCM	Explainable Convolutional Neural Network [80]
27	XCMPlus	Explainable Convolutional Neural Network Plus [196]
28	XceptionTime	Xception Time Model [220]
29	XceptionTimePlus	Xception Time Plus [195]
30	OmniScaleCNN	Omni-Scale 1D-Convolutional Neural Network [268]

Όπως αναφέρθηκε, την εκτέλεση των πειραμάτων χρησιμοποιήθηκε η βιβλιοθήκη tsAI. Η tsAI, που είναι ένα πακέτο ανοικτού κώδικα, βασίζεται στις δημοφιλείς βιβλιοθήκες PyTorch και Fastai, εστιάζοντας σε σύγχρονες τεχνικές για εργασίες πρόβλεψης χρονοσειρών, ταξινόμησης και παλινδρόμησης. Παρέχει επίσης εργαλεία προεπεξεργασίας δεδομένων, ενσωματωμένες μεθόδους για εκπαίδευση και αξιολόγηση μοντέλων, καθώς και πληθώρα ενσωματωμένων μετρικών αξιολόγησης.

Στο παρόν πλαίσιο, η διαδικασία πρόβλεψης αντιμετωπίστηκε ως ένα πρόβλημα πα-

λινδρόμησης. Κατά την εκτέλεση των πειραμάτων, οι αρχικές παράμετροι των αντίστοιχων μεθόδων από τη βιβλιοθήκη διατηρήθηκαν, διατηρώντας σταθερό το περιβάλλον υλοποίησης για όλα τα αλγοριθμικά σχήματα. Έτσι, όλοι οι αλγόριθμοι συγκρίθηκαν στη βασική τους διαμόρφωση, ενισχύοντας την εύκολη αναπαραγωγή των πειραμάτων, αλλά και τη σύγκριση με επιπλέον μεθόδους.

Επιπλέον, κατά την πειραματική διαδικασία, τα δεδομένα χωρίστηκαν σε αναλογία 80% - 20% σε σύνολα εκπαίδευσης και ελέγχου αντίστοιχα.

Όσον αφορά τους χρονικούς ορίζοντες πρόβλεψης, εφαρμόστηκαν τρεις διαφορετικές στρατηγικές: πρόβλεψη μιας ημέρας (single-step), 7 και 14 ημερών (multi-step).

Τα αποτελέσματα αξιολογήθηκαν με βάση τις μετρικές απόδοσης MAE, MAPE, MSE, RMSE, RMSLE και R^2.

5.5.4.2 Temporal Convolutional Networks (TCNs)

Τα Temporal Convolutional Networks (TCN) αποτελούν μια σύγχρονη κατηγορία συνελικτικών νευρωνικών δικτύων που έχουν σχεδιαστεί ειδικά για την επεξεργασία ακολουθιακών δεδομένων, όπως χρονοσειρές και γενικά δεδομένα που φέρουν πληροφορία μέσω της χρονικής διάστασης. Η βασική καινοτομία των TCN έγκειται στο ότι επιτυγχάνουν την επεξεργασία ακολουθιών χρησιμοποιώντας συνελίξεις, αντί για τις επαναληπτικές (αναδρομικές) δομές που συναντάμε στα παραδοσιακά RNNs και LSTMs). Έτσι, τα TCNs προσφέρουν τη δυνατότητα παραλληλοποίησης κατά την εκπαίδευση, βελτιωμένη σταθερότητα και συνήθως καλύτερη διαχείριση μακροπρόθεσμων εξαρτήσεων, που αποτελούν διαχρονικό ζητούμενο στη μοντελοποίηση ακολουθιακών δεδομένων.

Στον πυρήνα της αρχιτεκτονικής των TCNs βρίσκεται η έννοια της αιτιακής συνέλιξης (causal convolution), όπου κάθε έξοδος σε χρονικό βήμα t εξαρτάται αποκλειστικά από τα δεδομένα που έχουν προηγηθεί ή συμπίπτουν χρονικά ($t' \leq t$). Αυτή η ιδιότητα εγγυάται ότι δεν χρησιμοποιείται μελλοντική πληροφορία κατά την πρόβλεψη ή την παραγωγή της εξόδου, κάτι που είναι θεμελιώδες σε εφαρμογές πρόβλεψης χρονοσειρών και ελέγχου. Επιπλέον, τα TCNs ενσωματώνουν την έννοια της αραιωμένης συνελίξεως (dilated convolution), όπου το συνελικτικό φίλτρο εφαρμόζεται σε μη διαδοχικά δείγματα της ακολουθίας, με το πλήθος των παραλείψεων (αραίωση) να ορίζεται από έναν παράγοντα d. Έτσι, το δίκτυο μπορεί να «βλέπει» μακρινά χρονικά σημεία χωρίς να αυξάνεται εκθετικά το πλήθος των παραμέτρων ή το υπολογιστικό κόστος.

Η μαθηματική περιγραφή ενός TCN για μια ακολουθία $x = \{x_0, x_1, x_2, ..., x_{T-1}\}$ και για ένα συνελικτικό φίλτρο μεγέθους K με παράγοντα αραίωσης d στο χρονικό βήμα t δίνεται ως εξής:

$$y_t = \sum_{k=1}^{K-1} W_k x_{t-dk} \quad \text{όπου} \quad t - dk \geq 0 \tag{5.40}$$

Ο περιορισμός $t - dk \geq 0$ διασφαλίζει ότι το φίλτρο χρησιμοποιεί μόνο παρελθοντικά ή παρόντα δείγματα της ακολουθίας, και επομένως παραμένει αιτιακό. Ο παράγοντας αραίωσης d αυξάνει το αντιληπτικό πεδίο του φίλτρου, επιτρέποντας στο δίκτυο να λαμβάνει υπόψη του ένα μεγαλύτερο ιστορικό εισόδου χωρίς την ανάγκη αύξησης του αριθμού των επιπέδων ή του μεγέθους του φίλτρου.

Ένα ακόμη σημαντικό χαρακτηριστικό των TCNs είναι η υιοθέτηση των residual ή skip connections, οι οποίες επιτρέπουν την κατασκευή βαθύτερων δικτύων, ενισχύοντας τη ροή πληροφορίας ανάμεσα στα επίπεδα και περιορίζοντας το φαινόμενο της

εξαφάνισης βαθμίδων (vanishing gradients) κατά την διαδικασία της οπισθοδιάδοσης. Με αυτόν τον τρόπο, τα TCNs συνδυάζουν τα πλεονεκτήματα των βαθιών συνελικτικών αρχιτεκτονικών με τις ανάγκες της αιτιακής επεξεργασίας ακολουθιών.

Σε πρακτικό επίπεδο, τα TCNs έχουν αποδειχθεί αποτελεσματικά σε πληθώρα εφαρμογών όπου απαιτείται η ανάλυση χρονοεξαρτώμενων δεδομένων. Η αρχιτεκτονική τους επιτρέπει μεγαλύτερη παραλληλοποίηση, σταθερότερη εκπαίδευση και, χάρη στη χρήση διαδοχικών αραιωμένων συνελίξεων, δίκτυα με πολύ μεγάλο αντιληπτικό πεδίο. =

5.5.5 Αποτελέσματα

Ακολουθώντας την περίπτωση που παρουσιάστηκε προηγουμένως, τα αποτελέσματα της εκτεταμένης πειραματικής διαδικασίας θα παρουσιαστούν με τρόπο τέτοιο που αντικατοπτρίζει το διττό στόχο της: αφενός, να αποτελέσει σύγκριση ενός μεγάλου αριθμού σύγχρονων αλγορίθμων πρόβλεψης χρονοσειρών, και αφετέρου να διερευνήσει το αν και κατά πόσο η γνώση της κοινής γνώμης, όπως ανταναακλάται στα μέσα κοινωνικής δικτύωσης και ποσοτικοποιείται μέσω τριών διαφορετικών μεθόδων ανάλυσης συναισθήματος, μπορεί να βελτιώσει τις παραγόμενες προβλέψεις. Αντίστοιχα, η παρουσίαση των αποτελεσμάτων της πειραματικής διαδικασίας χωρίζεται σε δύο διακριτά μέρη, όπως ακριβώς συνέβη και με την προηγούμενη περίπτωση που αφορούσε τις κλασσικές μεθόδους. Το πρώτο αφορά τη σύγκριση των μεθόδων, ενώ το δεύτερο τη σύγκριση των διαφορετικών σχημάτων συναισθηματικής ανάλυσης, με στόχο να διερευνηθεί η αποτελεσματικότητά τους.

Ωστόσο, πρέπει να σημειωθεί ότι ο μεγάλος αριθμός συγκρίσεων που πραγματοποιήθηκαν οδήγησε και πάλι σε έναν τεράστιο όγκο αποτελεσμάτων. Συγκεκριμένα, σε κάθε περίπτωση, αξιολογήθηκε η απόδοση 30 διαφορετικών αλγοριθμικών σχημάτων και 67 διαφορετικών διαμορφώσεων χαρακτηριστικών για τρεις διαφορετικούς χρονικούς ορίζοντες πρόβλεψης (1, 7 και 14 ημερών). Κάθε ένα από αυτά τα σχήματα αξιολογήθηκε χρησιμοποιώντας έξι διαφορετικές μετρικές απόδοσης, ενώ η διαδικασία επαναλήφθηκε για κάθε ένα από τα σύνολα δεδομένων. Στη συνέχεια, ακολουθεί μία απαραίτητη σύνοψη της διαδικασίας, πριν την παρουσίαση των εξαγόμενων αποτελεσμάτων.

Όπως έχει ήδη αναφερθεί, κατά τη διαδικασία, για κάθε μία από τις μετοχές ακολουθήθηκε η εξής στρατηγική:

- Κάθε ένας από τους 30 αλγορίθμους πρόβλεψης εκτελέστηκε 67 φορές, κάθε φορά λαμβάνοντας ως είσοδο έναν από τους 67 διαφορετικούς συνδυασμούς χαρακτηριστικών.
- Αυτή η διαδικασία επαναλήφθηκε τρεις φορές, μία για κάθε χρονικό παράθυρο πρόβλεψης.
- Για κάθε εκτέλεση, υπολογίστηκαν οι έξι μετρικές απόδοσης.

Η σύγκριση των αλγορίθμων πραγματοποιήθηκε χρησιμοποιώντας στατιστικούς ελέγχους Friedman, οι οποίοι εφαρμόστηκαν ξεχωριστά για κάθε σύνολο χαρακτηριστικών και κάθε χρονικό διάστημα πρόβλεψης. Έτσι, για την ανάλυση των αλγορίθμων, εκτελέστηκαν συνολικά:

$$67 \times 6 \times 3 = 1206 \ \ \text{Friedman tests.} \tag{5.41}$$

Αντίστοιχα, η αξιολόγηση της αποδοτικότητας των συνδυασμών χαρακτηριστικών έγινε με βάση τη μέση απόδοση των 30 αλγορίθμων για κάθε έναν από τους 67 διαφορετικούς συνδυασμούς. Σε αυτή την περίπτωση, εκτελέστηκαν συνολικά:

$$30 \times 6 \times 3 = 540 \ \text{ Friedman tests.} \tag{5.42}$$

Για να συνοψιστούν τα αποτελέσματα, υπολογίστηκε η μέση κατάταξη των μεθόδων και των συνόλων χαρακτηριστικών σε κάθε χρονικό διάστημα. Η κατάταξη βασίστηκε στις θέσεις που παρήγαγε ο έλεγχος Friedman, ενώ στη συνέχεια, χρησιμοποιήθηκε ο post-hoc έλεγχος Nemenyi για να ελεγχθούν κατά ζεύγη οι σημαντικές διαφορές μεταξύ των μοντέλων. Τα αποτελέσματα του ελέγχου Nemenyi απεικονίζονται σε διαγράμματα κρίσιμων διαφορών, στα οποία οι μέθοδοι που δεν διαφέρουν σημαντικά μεταξύ τους συνδέονται με μαύρες οριζόντιες γραμμές. Κατά τα γνωστά, δύο μέθοδοι θεωρούνται μη σημαντικά διαφορετικές, εάν η διαφορά στις μέσες κατατάξεις τους είναι μικρότερη από την τιμή της κρίσιμης διαφοράς.

Στη συνέχεια, παρουσιάζονται τα αποτελέσματα που προέκυψαν από τις δύο μελέτες περίπτωσης: πρώτα η σύγκριση των αλγορίθμων πρόβλεψης (Υποενότητα 5.5.5.1) και στη συνέχεια η σύγκριση των συνδυασμών χαρακτηριστικών και ανάλυσης συναισθήματος (Υποενότητα 5.5.5.2).

5.5.5.1 Σύγκριση μεθόδων

Η παρούσα υποενότητα αφορά τα αποτελέσματα της πειραματικής διαδικασίας από τη σκοπιά της σύγκρισης των μεθόδων, με στόχο να εντοπιστούν τυχόν μέθοδοι που υπερτερούν σε σχέση με τις υπόλοιπες. Τα αποτελέσματα παρουσιάζονται ανά χρονικό παράθυρο πρόβλεψης. Σε κάθε περίπτωση, παρατίθενται τα αποτελέσματα της κατάταξης Friedman για κάθε μία από τις έξι μετρικές. Για εξοικονόμηση χώρου, αναφέρονται μόνο οι μέθοδοι που καταλαμβάνουν τις δέκα πρώτες θέσεις της κατάταξης, ενώ οι πλήρεις πίνακες των κατατάξεων είναι διαθέσιμοι στη διεύθυνση: shorturl.at/FTU06.

Παρατίθενται επίσης τα διαγράμματα κρίσιμων διαφορών του post-hoc στατιστικού ελέγχου Nemenyi, που ακολούθησε τον στατιστικό έλεγχο Friedman. Σε αυτά αποδίδεται οπτικά το ποιες από τις μεθόδους παρουσιάζουν παρόμοια συμπεριφορά και ποιες διαφέρουν σημαντικά.

Τέλος, για τα αποτελέσματα κάθε μετρικής παρέχονται τα αντίστοιχα θηκογράμματα, και πάλι για τις 10 καλύτερες μεθόδους. Τα θηκογράμματα παρέχουν μια γραφική και συνοπτική απεικόνιση της κατανομής των δεδομένων, δηλαδή, στην περίπτωσή μας, των μέσων τιμών των διαμορφώσεων συναισθήματος ανά αλγόριθμο για όλες τις μετοχές. Συγκεκριμένα, από τα διαγράμματα μπορούν να εξαχθούν πληροφορίες σχετικά με τη μέγιστη και ελάχιστη τιμή της απόδοσης των μεθόδων σε κάθε μετρική, τη διάμεση τιμή της, καθώς και το 1ο και 3ο τεταρτημόριο, που διαχωρίζουν το 25% και το 75% των προαναφερθέντων τιμών.

Χρονικός ορίζοντας 1 ημέρας Όσον αφορά τις προβλέψεις μίας ημέρας, τα αποτελέσματα της κατάταξης Friedman για τις 10 κορυφαίες μεθόδους ανά μετρική εμφανίζονται στον Πίνακα 5.17. Είναι εμφανές πως αν και καμία μεμονωμένη μέθοδος δεν κυριαρχεί σε όλες τις μετρικές και διαπιστώνονται σημαντικές ανακατατάξεις στις θέσεις του πίνακα, η μέθοδος TCN επιτυγχάνει την καλύτερη κατάταξη σε τρεις από τις έξι μετρικές (MAPE, R^2 και RMSLE) και βρίσκεται πάντα μέσα στις τέσσερις πρώτες θέσεις. Αυτό διαπιστώνεται εύκολα και από τα διαγράμματα box plot, όπου είναι σαφές

ότι το TCN παρουσιάζει το μικρότερο εύρος τιμών από όλες τις μεθόδους. Επιπλέον, παρατηρείται πως σε όλες τις μετρικές, η μέθοδος GRU_FCN βρίσκεται σταθερά μέσα στις πέντε πρώτες θέσεις, ενώ οι μέθοδοι LSTM_FCN και LSTMPlus παρουσιάζουν επίσης καλή απόδοση. Η LSTM_FCN εμφανίζεται σταθερά στην 5η θέση σε όλες τις μετρικές, ενώ η μέθοδος LSTMPlus κατατάσσεται πρώτη στις μετρικές MAE και RMSE. Στις μετρικές R^2 και RMSLE, καταλαμβάνει τη δεύτερη θέση, ενώ στη μετρική MSE βρίσκεται στην τρίτη θέση. Ωστόσο, στη μετρική MAPE δεν συγκαταλέγεται καν στις 10 καλύτερες μεθόδους.

Με βάση τα παραπάνω, είναι ξεκάθαρη η υπεροχή της TCN.

Τα αποτελέσματα που παρήγαγε ο στατιστικός έλεγχος Friedman για τις έξι μετρικές παρουσιάζονται στον Πίνακα 5.17, ενώ τα αντίστοιχα CD διαγράμματα και θηκογράμματα απεικονίζονται στα Σχήματα 5.27 και 5.26

Πίνακας 5.17: Κατάταξη Friedman (Μέθοδοι) — Χρονικός ορίζοντας μιας ημέρας

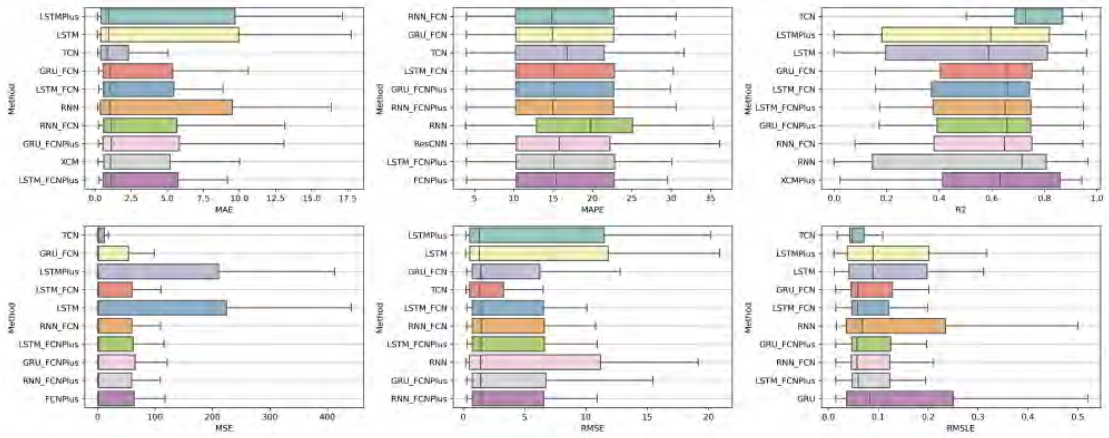
No.	MAE		MAPE		R^2	
	Μέθοδος	**Friedman Score**	**Μέθοδος**	**Friedman Score**	**Μέθοδος**	**Friedman Score**
1	LSTMPlus	8.266667	RNN_FCN	10.4	TCN	22.4
2	LSTM	8.533333	GRU_FCN	10.533333	LSTMPlus	21.13333
3	TCN	9.466667	TCN	11	LSTM	20.6
4	GRU_FCN	10	LSTM_FCN	11.2	GRU_FCN	20.53333
5	LSTM_FCN	10.2	GRU_FCNPlus	11.6	LSTM_FCN	19.73333
6	RNN	10.73333	RNN_FCNPlus	11.73333	LSTM_FCNPlus	19.13333
7	RNN_FCN	11.13333	RNN	11.93333	GRU_FCNPlus	18.93333
8	GRU_FCNPlus	11.33333	ResCNN	12.13333	RNN_FCN	18.93333
9	XCM	11.33333	LSTM_FCNPlus	12.13333	RNN	18.93333
10	LSTM_FCNPlus	11.4	FCNPlus	12.46667	XCMPlus	18.66667

No.	MSE		RMSE		RMSLE	
	Μέθοδος	**Friedman Score**	**Μέθοδος**	**Friedman Score**	**Μέθοδος**	**Friedman Score**
1	TCN	9	LSTMPlus	9.066667	TCN	7.733333
2	GRU_FCN	9.266667	GRU_FCN	9.6	LSTMPlus	9.333333
3	LSTMPlus	9.6	RNN_FCN	9.6	LSTM	9.8
4	LSTM_FCN	9.8	LSTM_FCN	9.733333	GRU_FCN	10
5	LSTM	9.933333	LSTM	10.066667	LSTM_FCN	10.2
6	RNN_FCN	10.33333	RNN	10.93333	RNN	11.13333
7	LSTM_FCNPlus	10.46667	LSTM_FCNPlus	11.06667	GRU_FCNPlus	11.26667
8	GRU_FCNPlus	10.8	RNN	11.13333	RNN_FCN	11.26667
9	RNN_FCNPlus	11.33333	GRU_FCNPlus	11.2	LSTM_FCNPlus	11.26667
10	FCNPlus	11.46667	RNN_FCNPlus	11.86667	GRU	12

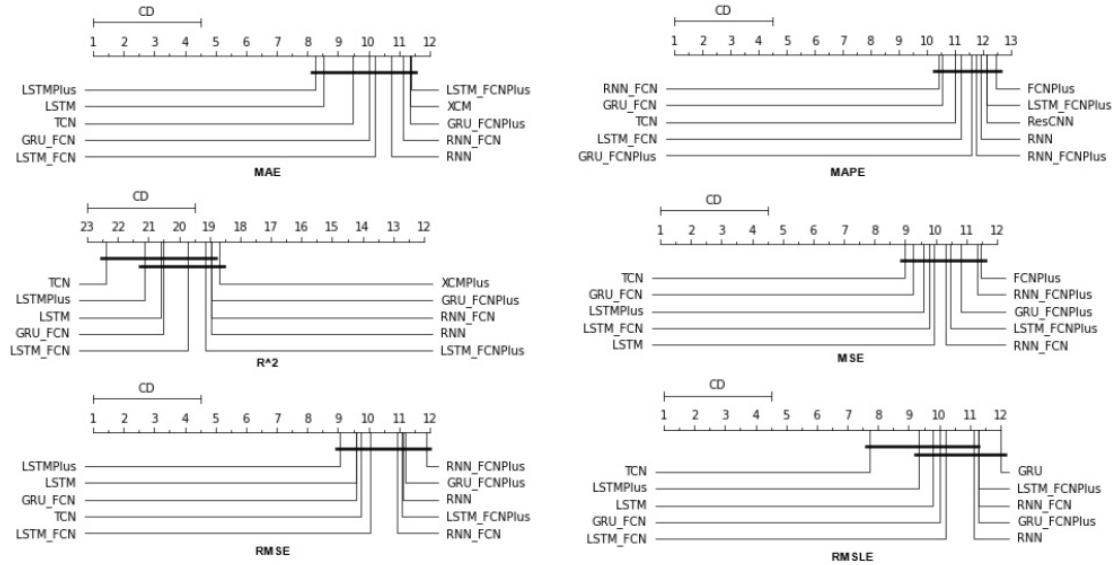
Χρονικός ορίζοντας 7 ημερών Στο χρονικό πλαίσιο πρόβλεψης μίας εβδομάδας, οι αλγόριθμοι που καταλαμβάνουν τις κορυφαίες θέσεις στην κατάταξη φαίνεται να έχουν σταθεροποιηθεί, όπως προκύπτει από τα αποτελέσματα του στατιστικού ελέγχου. Η αντίστοιχη κατάταξη του στατιστικού ελέγχου Friedman, όσον αφορά τις δέκα καλύτερες μεθόδους σε σχέση με τις έξι μετρικές απόδοσης, παρουσιάζεται στον Πίνακα 5.18.

Είναι ξεκάθαρο πως πλέον σε όλες τις μετρικές, η μέθοδος TCN καταλαμβάνει την πρώτη θέση.

Από τα CD διαγράμματα, μπορεί να παρατηρηθεί ότι σε όλες τις μετρικές—εκτός από την R^2—η υπεροχή της TCN επιβεβαιώνεται, καθώς διαφέρει σημαντικά από τις υπόλοιπες μεθόδους. Τα box plots δείχνουν επίσης ότι η TCN παρουσιάζει τη μικρότερη διακύμανση γύρω από τη διάμεσο.



Σχήμα 5.26: Θηκογράμματα για την απόδοση των αλγορίθμων - Χρονικός ορίζοντας μιας ημέρας



Σχήμα 5.27: CD διαγράμματα για την απόδοση των αλγορίθμων - Χρονικός ορίζοντας μιας ημέρας

Άλλες μέθοδοι που εμφανίζουν σαφή κυριαρχία έναντι των υπολοίπων, όσον αφορά τις επιδόσεις τους, είναι η TSTPlus, η οποία καταλαμβάνει τη δεύτερη θέση σε όλες τις μετρικές, εκτός από τη MAPE, και οι XCMPlus και XCM, που βρίσκονται στην πρώτη πεντάδα στην πλειοψηφία των περιπτώσεων. Σε γενικές γραμμές, οι ίδιες μέθοδοι εμφανίζονται σε παρόμοιες θέσεις σε όλες τις μετρικές, με μικρές διαφοροποιήσεις στις κατατάξεις. Τα σχετικά αποτελέσματα απεικονίζονται με τη μορφή θηκογραμμάτων και CD διαγραμμάτων στα Σχήματα 5.29 και 5.28.

Πίνακας 5.18: Κατάταξη Friedman (Μέθοδοι) — Χρονικός ορίζοντας μιας εβδομάδας

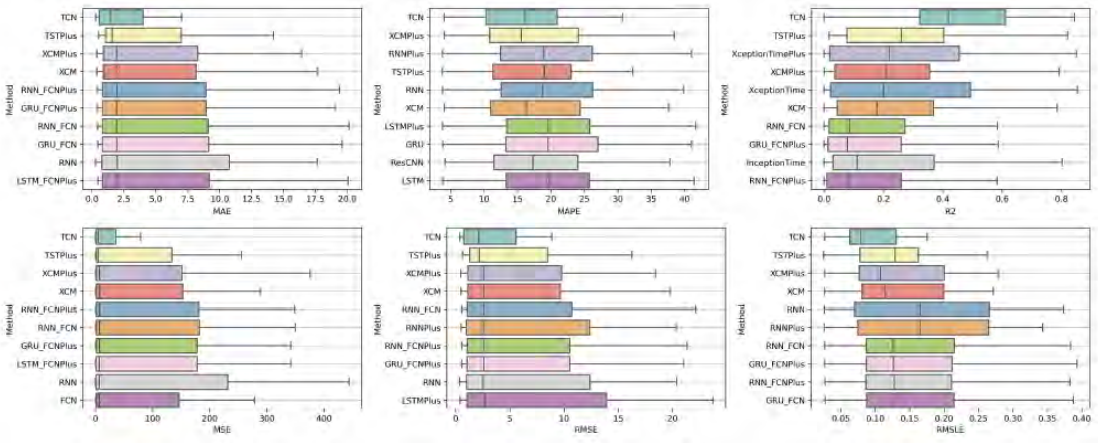
No.	**MAE**		**MAPE**		R^2	
	Μέθοδος	**Friedman Score**	**Μέθοδος**	**Friedman Score**	**Μέθοδος**	**Friedman Score**
1	TCN	3.733333	TCN	6.133333	TCN	25.86667
2	TSTPlus	8.266667	XCMPlus	9.866667	TSTPlus	25.8
3	XCMPlus	8.866667	RNNPlus	10.93333	XceptionTimePlus	19.7
4	XCM	10.53333	TSTPlus	11	XCMPlus	19.66667
5	RNN_FCNPlus	12.06667	RNN	11.06667	XceptionTime	19.53333
6	GRU_FCNPlus	12.13333	XCM	11.26667	XCM	18.53333
7	RNN_FCN	12.26667	LSTMPlus	13	RNN_FCN	16.9
8	GRU_FCN	13.2	GRU	13.66667	GRU_FCNPlus	16.66667
9	RNN	13.53333	ResCNN	13.86667	InceptionTime	16.5
10	LSTM_FCNPlus	13.53333	LSTM	14.06667	RNN_FCNPlus	16.16667

No.	**MSE**		**RMSE**		**RMSLE**	
	Μέθοδος	**Friedman Score**	**Μέθοδος**	**Friedman Score**	**Μέθοδος**	**Friedman Score**
1	TCN	3.666666667	TCN	3.933333333	TCN	3.8
2	TSTPlus	8.733333333	TSTPlus	8.066666667	TSTPlus	8.066666667
3	XCMPlus	8.933333333	XCMPlus	9.066666667	XCMPlus	9.4
4	XCM	11.93333333	XCM	10.8	XCM	10.6
5	RNN_FCNPlus	12.06666667	RNN_FCN	12.26666667	RNN	12.13333333
6	RNN_FCN	12.2	RNNPlus	12.8	RNNPlus	12.46666667
7	GRU_FCNPlus	12.6	RNN_FCNPlus	12.86666667	RNN_FCN	12.46666667
8	LSTM_FCNPlus	13	GRU_FCNPlus	13	GRU_FCNPlus	12.86666667
9	RNN	13.26666667	RNN	13.06666667	RNN_FCNPlus	13.06666667
10	FCN	13.4	LSTMPlus	13.66666667	GRU_FCN	13.06666667

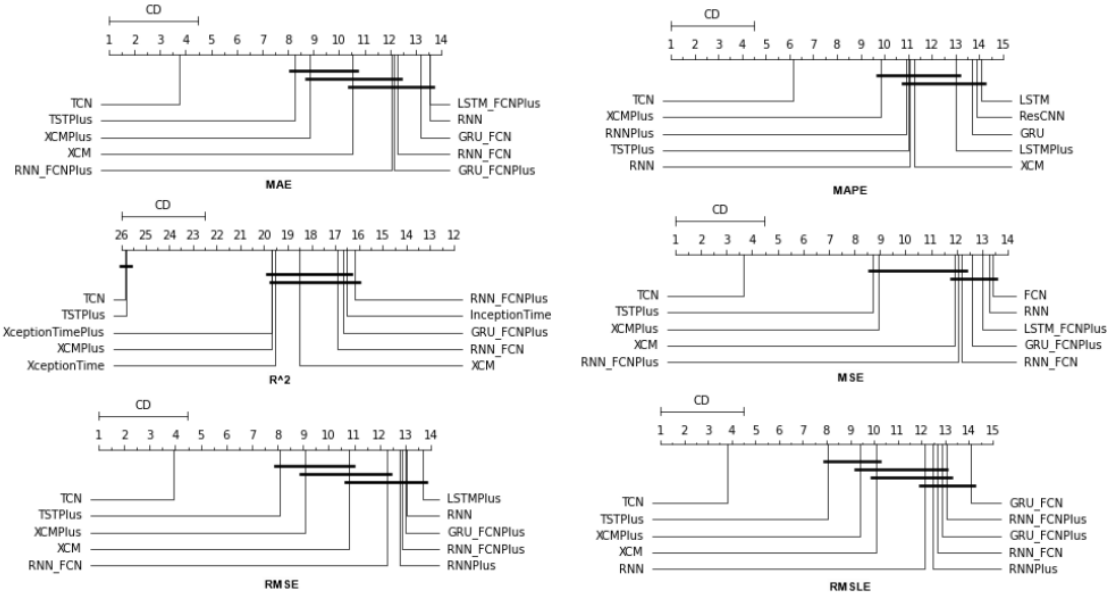
Χρονικός ορίζοντας 14 ημερών Στα αποτελέσματα πρόβλεψης χρονικού ορίζοντα δύο εβδομάδων, παρατηρείται σχετική συμφωνία στους κορυφαίους αλγορίθμους σε σύγκριση με τα αποτελέσματα του χρονικού πλαισίου μίας εβδομάδας. Η κατάταξη που προκύπτει από το στατιστικό έλεγχο Friedman για τις δέκα καλύτερες μεθόδους σε σχέση με τις έξι μετρικές απόδοσης παρουσιάζεται στον Πίνακα 5.19.

Για άλλη μια φορά, η μέθοδος TCN καταλαμβάνει την πρώτη θέση σε όλες τις μετρικές. Η TSTPlus κατατάσσεται ξανά δεύτερη σε όλες τις μετρικές, εκτός από την R^2, όπου καταλαμβάνει την τρίτη θέση. Επιπλέον, στην πλειοψηφία των περιπτώσεων, οι μέθοδοι XCMPlus και RNNPlus εμφανίζονται στην πρώτη πεντάδα. Όπως και στους δύο προηγούμενους χρονικούς ορίζοντες, υπάρχει συμφωνία στις θέσεις των μεθόδων στις διάφορες μετρικές.Σύμφωνα με τα παραπάνω αποτελέσματα που αφορούν και τους τρεις χρονικούς ορίζοντες, η κυριαρχία της μεθόδου TCN καθίσταται εμφανής.

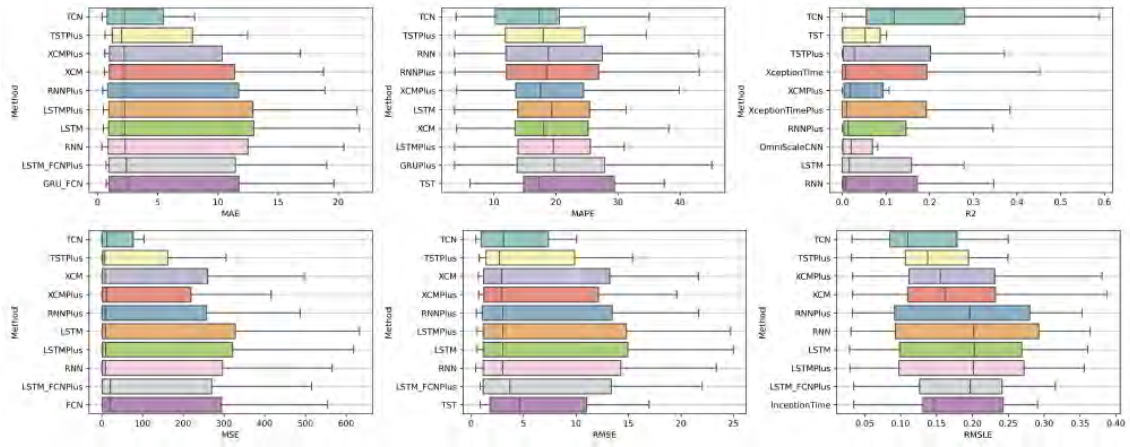
Τα αντίστοιχα CD διαγράμματα και θηκογράμματα για τους δέκα καλύτερους αλγορίθμους και την απόδοσή τους σε χρονικό ορίζοντα δύο εβδομάδων παρουσιάζονται στα Σχήματα 5.31 και 5.30.



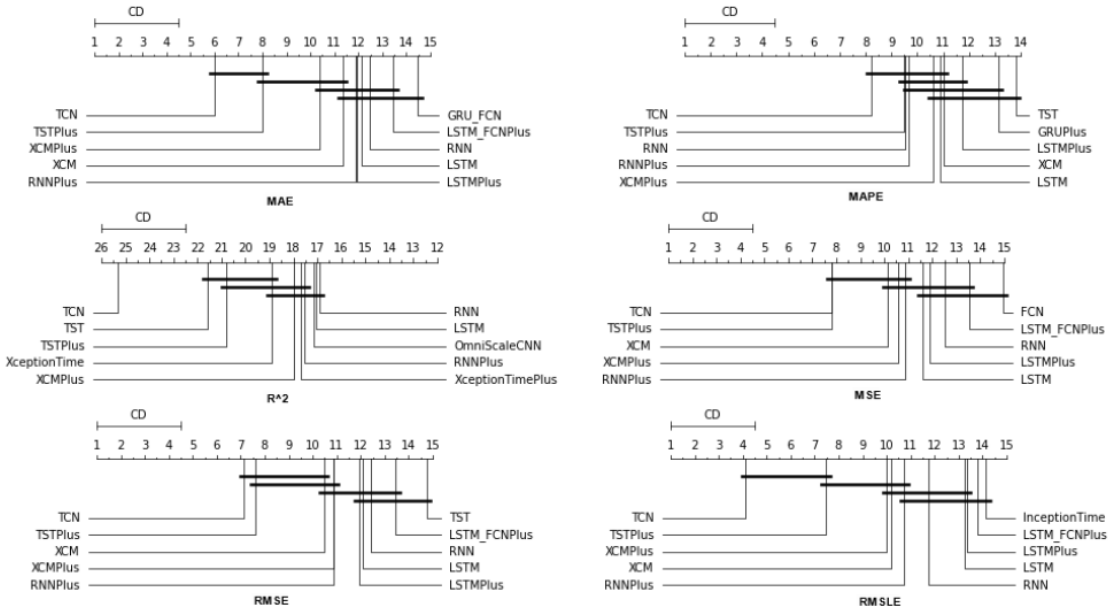
Σχήμα 5.28: Θηκογράμματα για την απόδοση των αλγορίθμων - Χρονικός ορίζοντας μιας εβδομάδας



Σχήμα 5.29: CD διαγράμματα για την απόδοση των αλγορίθμων - Χρονικός ορίζοντας μιας εβδομάδας



Σχήμα 5.30: Θηκογράμματα για την απόδοση των αλγορίθμων - Χρονικός ορίζοντας 2 εβδομάδων



Σχήμα 5.31: CD διαγράμματα για την απόδοση των αλγορίθμων - Χρονικός ορίζοντας δύο εβδομάδων

Πίνακας 5.19: Κατάταξη Friedman (Μέθοδοι) — Χρονικός ορίζοντας δύο εβδομάδων

No.	MAE		MAPE		R^2	
	Μέθοδος	**Friedman Score**	**Μέθοδος**	**Friedman Score**	**Μέθοδος**	**Friedman Score**
1	TCN	6	TCN	8.2	TCN	25.333333
2	TSTPlus	8	TSTPlus	9.466666667	TST	21.6
3	XCMPlus	10.4	RNN	9.533333333	TSTPlus	20.8
4	XCM	11.333333	RNNPlus	9.666666667	XceptionTime	18.9
5	RNNPlus	11.866666667	XCMPlus	10.6	XCMPlus	17.966667
6	LSTMPlus	11.933333333	LSTM	10.866666667	XceptionTimePlus	17.7
7	LSTM	12.133333333	XCM	11	RNNPlus	17.533333
8	RNN	12.466666667	LSTMPlus	11.733333333	OmniScaleCNN	17.133333
9	LSTM_FCNPlus	13.466666667	GRUPlus	13.133333333	LSTM	17.033333
10	GRU_FCN	14.466666667	TST	13.8	RNN	16.933333

No.	MSE		RMSE		RMSLE	
	Μέθοδος	**Friedman Score**	**Μέθοδος**	**Friedman Score**	**Μέθοδος**	**Friedman Score**
1	TCN	7.8	TCN	7.133333333	TCN	4.133333333
2	TSTPlus	7.8	TSTPlus	7.6	TSTPlus	7.466666667
3	XCM	10.133333333	XCM	10.466666667	XCMPlus	9.4
4	XCMPlus	10.6	XCMPlus	10.866666667	XCM	10.2
5	RNNPlus	10.866666667	RNNPlus	10.866666667	RNNPlus	10.733333333
6	LSTM	11.6	LSTM	11.933333333	RNN	11.733333333
7	LSTMPlus	11.866666667	LSTMPlus	12.066666667	LSTM	13.266666667
8	RNN	12.533333333	RNN	12.466666667	LSTM_FCNPlus	13.8
9	LSTM_FCNPlus	13.533333333	LSTM_FCNPlus	13.466666667	InceptionTime	14.133333333
10	FCN	14.933333333	TST	14.733333333	TST	14.733333333

5.5.5.2 Αποτελέσματα συναισθηματικών διαμορφώσεων

Προχωρώντας στα ευρήματα της δεύτερης μελέτης περίπτωσης, η οποία αφορά, αφενός, τη διερεύνηση του κατά πόσο η χρήση της ανάλυσης συναισθήματος συμβάλλει στη βελτίωση των παραγόμενων προβλέψεων, και αφετέρου, την αναγνώριση συγκεκριμένων συνδυασμών χαρακτηριστικών, των οποίων η χρήση βελτιώνει την προβλεπτική ικανότητα του μοντέλου, τα αποτελέσματα, όπως και στα προηγούμενα, θα παρουσιαστούν ανά χρονικό ορίζοντα πρόβλεψης.

Επιπλέον, για ακόμη μια φορά, λόγω του μεγάλου όγκου αποτελεσμάτων, θα αναφερθούν μόνο οι 10 πιο υποσχόμενοι συνδυασμοί χαρακτηριστικών, όπως αυτοί προέκυψαν βάσει της κατάταξης Friedman, η οποία υπολογίστηκε για τον μέσο όρο της απόδοσής τους, λαμβάνοντας υπόψη τις προβλέψεις που προέκυψαν από τη χρήση των 30 αλγορίθμων πρόβλεψης. Η πλήρης κατάταξη και των 67 συνδυασμών χαρακτηριστικών μπορεί να βρεθεί στο shorturl.at/alqwx.

Χρονικός ορίζοντας 1 ημέρας Ξεκινώντας με τα αποτελέσματα που αφορούν την πρόβλεψη μίας ημέρας, από τον Πίνακα 5.20 παρατηρείται ότι σε όλες τις μετρικές υπερτερούν σενάρια όπου γίνεται χρήση ανάλυσης συναισθήματος, εκτός από την περίπτωση της μετρικής R^2, όπου στην πρώτη θέση βρίσκεται το σενάριο μη χρήσης Ανάλυσης Συναισθήματος (univariate version), όπου οι προβλέψεις βασίζονται αποκλειστικά στις τιμές κλεισίματος των προηγούμενων ημερών. Στην πραγματικότητα, σε τρεις από τις έξι μετρικές, η μονοπαραγοντική εκδοχή δεν συγκαταλέγεται καν στις πρώτες είκοσι θέσεις της κατάταξης, όπως διαπιστώνεται από τους πλήρεις πίνακες της κατάταξης και επίσης διαφαίνεται στα Σχήματα 5.33 και 5.32.

Ένα ακόμη ενδιαφέρον εύρημα είναι ότι, αν και παρατηρούνται ανακατατάξεις στις θέσεις των διαφορετικών διαμορφώσεων Ανάλυσης Συναισθήματος όσον αφορά τις έξι

μετρικές, η διάταξη Blob_RM_7_Blob—δηλαδή, ο συνδυασμός Blob, Rolling Mean 7 Blob και των τιμών κλεισίματος—αν και δεν έχει καλή επίδοση στην κατάταξη όσον αφορά την R^2, καταλαμβάνει, αφενός, την κορυφαία θέση σε τέσσερις μετρικές (MAE, MSE, RMSE, RMSLE) και, αφετέρου, τη δεύτερη θέση στη μετρική MAPE.

Επιπλέον, από τα αποτελέσματα καθίσταται σαφές ότι μπορεί να διατυπωθεί ένα ισχυρό επιχείρημα υπέρ της χρήσης της Ανάλυσης Συναισθήματος στην πρόβλεψη χρονοσειρών χρηματοοικονομικού χαρακτήρα, ακόμη και στην περίπτωση που οι προβλέψεις αφορούν βάθος μίας ημέρας.

Παράλληλα, η χρήση εξομαλυμένων εκδοχών τόσο των χρονοσειρών συναισθήματος όσο και των τιμών κλεισίματος των μετοχών φαίνεται να είναι γενικά επωφελής, καθώς βλέπουμε τις αντίστοιχες εκδοχές να κατέχουν υψηλές θέσεις σε όλες τις μετρικές.

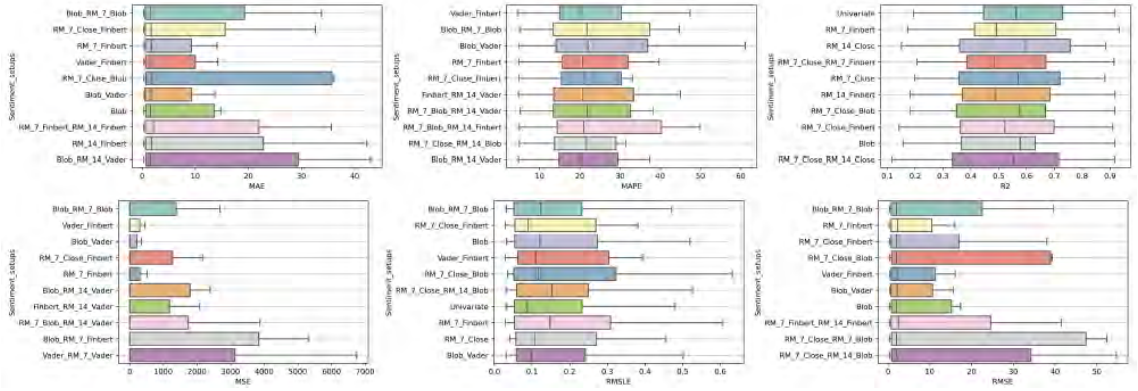
Πίνακας 5.20: Κατάταξη Friedman (Συναισθηματικές Διαμορφώσεις) — Χρονικός ορίζοντας μιας ημέρας

No.	**MAE**		**MAPE**		R^2	
	Feature Setup	**Friedman Score**	**Feature Setup**	**Friedman Score**	**Feature Setup**	**Friedman Score**
1	B_RM7B	19.733333	V_F	19.2	U	54.933333
2	RM7C_F	24.133333	B_RM7B	20.066667	RM7F	50.533333
3	RM7F	24.533333	B_V	21.8	RM14C	47.8
4	V_F	25.333333	RM7F	24.8	RM7C_RM7F	47.733333
5	RM7C_B	26.8	RM7C_F	26.933333	RM7C	47.133333
6	B_V	27.266667	F_RM14V	27.2	RM14F	46.4
7	B	28.4	RM7B_RM14V	28.133333	RM7C_B	45.866667
8	RM7F_RM14F	28.4	RM7B_RM14F	28.2	RM7C_F	43.6
9	RM14F	30	RM7C_RM14B	28.266667	B	43.4
10	B_RM14V	30	B_RM14V	29.2	RM7C_RM14C	43.133333

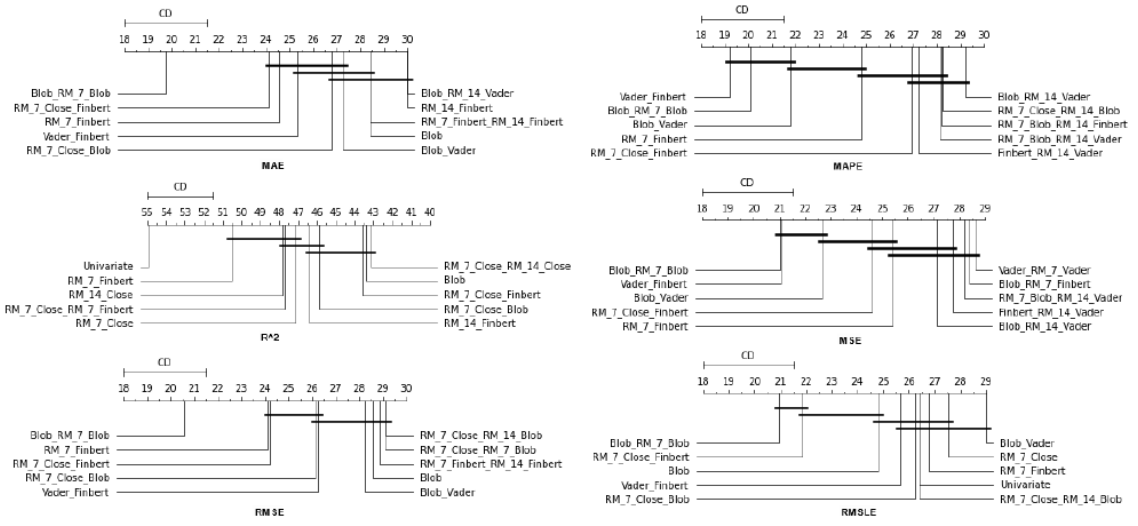
No.	**MSE**		**RMSE**		**RMSLE**	
	Feature Setup	**Friedman Score**	**Feature Setup**	**Friedman Score**	**Feature Setup**	**Friedman Score**
1	B_RM7B	21	B_RM7B	20.6	B_RM7B	20.933333
2	V_F	21.066667	RM7F	24.133333	RM7C_F	21.866667
3	B_V	22.666667	RM7C_F	24.2	B	24.8
4	RM7C_F	24.6	RM7C_B	26.133333	V_F	25.666667
5	RM7F	25.4	V_F	26.266667	RM7C_B	26.266667
6	B_RM14V	27.133333	B_V	28.266667	RM7C_RM14B	26.4
7	F_RM14V	27.733333	RM7F_RM14F	28.866667	U	26.4
8	RM7B_RM14V	28.2	RM7C_RM7B	29.066667	RM7F	27.533333
9	B_RM7F	28.4	RM7C_RM14B	29.133333	RM7C	27.533333
10	V_RM7V	28.6	B_V	29.133333	B_V	29

Χρονικός ορίζοντας 7 ημερών Όσον αφορά το χρονικό πλαίσιο της μίας εβδομάδας, από την κατάταξη Friedman του Πίνακα 5.21 μπορεί κανείς να παρατηρήσει ότι η εκδοχή στην οποία δεν γίνεται χρήση Ανάλυσης Συναισθήματος, καταλαμβάνει οριακά την πρώτη θέση σε τρεις μετρικές, συγκεκριμένα στις R^2, RMSE και RMSLE. Αντίθετα, στις δύο μετρικές, το Vader sentiment setup καταλαμβάνει την πρώτη θέση στην κατάταξη, καταλαμβάνοντας ταυτόχρονα τη δεύτερη θέση στη MAPE και RMSE και την πέμπτη θέση στη RMSLE, όπως είναι εμφανές στα Σχήματα 5.35 και 5.34 .

Αξιοσημείωτο είναι επίσης ότι η Blob_RM_7_Blob, η οποία είχε εξαιρετική απόδοση στη χρονική μετατόπιση της μίας ημέρας, παραμένει στην τριάδα κορυφής σε πέντε από τις έξι μετρικές. Σε γενικές γραμμές, παρατηρούνται και πάλι ανακατατάξεις,



Σχήμα 5.32: Θηκογράμματα για την απόδοση των συναισθηματικών διαμορφώσεων - Χρονικός ορίζοντας μιας ημέρας



Σχήμα 5.33: CD διαγράμματα για την απόδοση των αλγορίθμων - Χρονικός ορίζοντας μιας ημέρας

ειδικά στις κεντρικές θέσεις του πίνακα κατάταξης. Ωστόσο, δεδομένων των μικρών διαφορών απόδοσης μεταξύ των διαφορετικών διαμορφώσεων χαρακτηριστικών, αυτό δεν θα πρέπει να θεωρείται μη αναμενόμενο.

Συνολικά, η εικόνα εξακολουθεί να ευνοεί τη χρήση πολυπαραγοντικών εισόδων που περιέχουν δεδομένα Ανάλυσης Συναισθήματος.

Πίνακας 5.21: Κατάταξη Friedman (Συναισθηματικές Διαμορφώσεις) — Χρονικός ορίζοντας μιας εβδομάδας

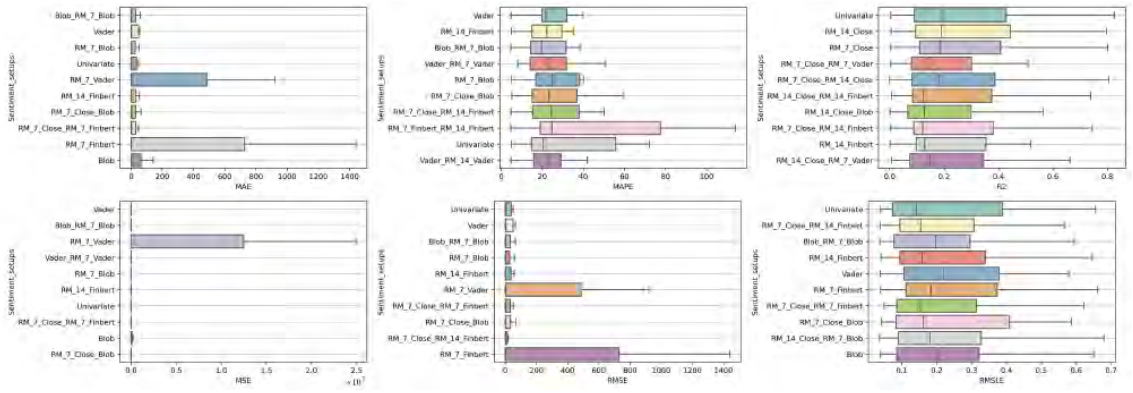
No.	MAE		MAPE		R^2	
	Feature Setup	Friedman Score	Feature Setup	Friedman Score	Feature Setup	Friedman Score
1	B_RM7B	21.533333	V	22.333333	U	55.066667
2	V	22.8	RM14F	24.8	RM14C	54.933333
3	RM7B	24.466667	B_RM7B	25.2	RM7C	54
4	U	24.866667	V_RM7V	25.933333	RM7C_RM7V	49.333333
5	RM7V	25.333333	RM7B	26.333333	RM7C_RM14C	47.233333
6	RM14F	25.533333	RM7C_B	26.666667	RM14C_RM14F	45.466667
7	RM7C_B	25.866667	RM7C_RM14F	26.8	RM14C_B	45.333333
8	RM7C_RM7F	26.666667	RM7F_RM14F	27.533333	RM7C_RM14F	45.266667
9	RM7F	26.866667	U	27.733333	RM14F	45.133333
10	B	27.133333	V_RM14V	27.933333	RM14C_RM7V	44.933333

No.	MSE		RMSE		RMSLE	
	Feature Setup	Friedman Score	Feature Setup	Friedman Score	Feature Setup	Friedman Score
1	V	21.933333	U	22.733333	U	17.6
2	B_RM7B	23.133333	V	23.133333	RM7C_RM14F	21.133333
3	RM7V	24	B_RM7B	23.666667	B_RM7B	22.533333
4	V_RM7V	24.8	RM7B	24.466667	RM14F	22.733333
5	RM7B	25.533333	RM14F	24.933333	V	23.2
6	RM14F	26.733333	RM7V	25.4	RM7F	24.466667
7	U	27.2	RM7C_RM7F	25.8	RM7C_RM14B	25.733333
8	RM7C_RM7F	27.733333	RM7C_B	25.933333	RM7C_B	26.4
9	B	27.866667	RM7C_RM14F	26.6	RM14C_RM7B	26.733333
10	RM7C_B	27.933333	RM7F	27.2	B	26.866667

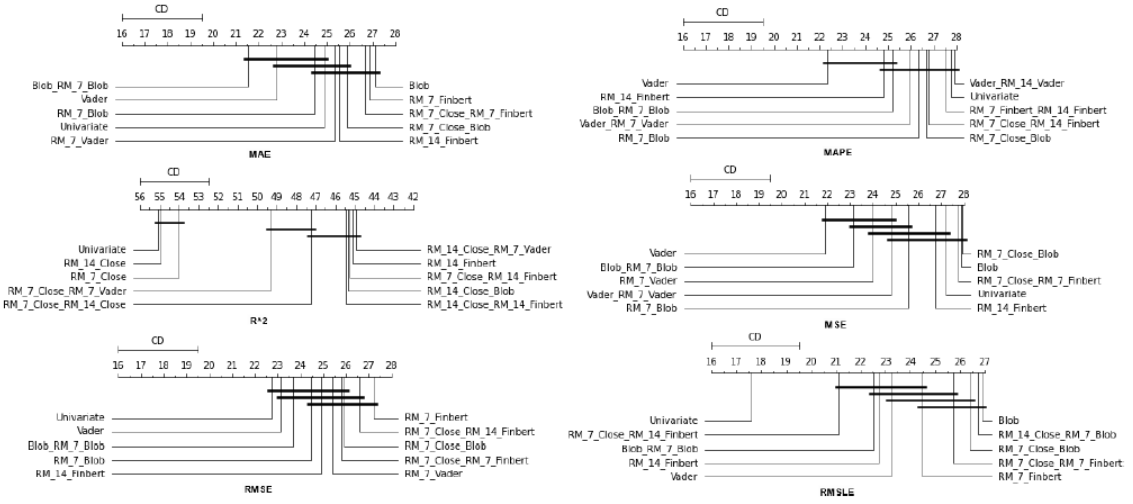
Χρονικός ορίζοντας 14 ημερών Τέλος, όσον αφορά το χρονικό πλαίσιο των δύο εβδομάδων, μια πρώτη παρατήρηση είναι ότι, σε σχέση με τη μετρική R^2, κυριαρχεί η και πάλι, όπως στις προηγούμενες δύο περιπτώσεις, η εκδοχή που δεν περιέχει Ανάλυσης Συναισθήματος (Πίνακας 5.22. Επιπλέον, αν και η συγκεκριμένη εκδοχή εμφανίζεται στις δέκα πρώτες θέσεις σε ορισμένες μετρικές, η διαφορά στην απόδοσή της σε σύγκριση με τις πρώτες θέσεις είναι αρκετά σημαντική. Αυτό φαίνεται ξεκάθαρα από τα CD διαγράμματα όπως φαίνεται στο Σχήμα 5.37, καθώς δεν υπάρχει καμία σύνδεση με τις κορυφαίες διαμορφώσεις χαρακτηριστικών. Επιπροσθέτως, τα αντίστοιχα θηκογράμματα παρουσιάζονται στο Σχήμα 5.36.

Στην πρόβλεψη με ορίζοντα 7 ημερών, παρατηρήθηκε ότι η εκδοχή όπου η ανάλυση συναισθήματος δεν χρησιμοποιείται, κατέλαβε την πρώτη θέση σε τρεις μετρικές. Ωστόσο, στη χρονική μετατόπιση των 14 ημερών, γίνεται σαφές ότι η υπεροχή των μεθόδων που χρησιμοποιούν δεδομένα Ανάλυσης Συναισθήματος ενισχύεται.

Παράλληλα, οι συνδυασμοί που περιλαμβάνουν τις τιμές κλεισίματος εμφανίζονται πιο συχνά στις πρώτες θέσεις της κατάταξης, σε σύγκριση με τις δύο προηγούμενες διαμορφώσεις. Επιπλέον, ο συνδυασμός RM_7_Close_Blob κυριαρχεί σε τέσσερις από τις έξι μετρικές (MAE, MAPE, MSE και RMSE), ενώ επιπλέον καλή συμπεριφορά παρουσιάζουν και οι εξομαλυμένες εκδοχές του. Συνεπώς, η χρήση κινητών



Σχήμα 5.34: Θηκογράμματα για την απόδοση των συναισθηματικών διαμορφώσεων - Χρονικός ορίζοντας μιας εβδομάδας



Σχήμα 5.35: CD διαγράμματα για την απόδοση των αλγορίθμων - Χρονικός ορίζοντας μιας εβδομάδας

μέσων όρων στις αρχικές χρονοσειρές, σε συνδυασμό με την ενσωμάτωση δεδομένων Ανάλυσης Συναισθήματος, αποδεικνύεται ως η βέλτιστη προσέγγιση, ανεξάρτητα από τη συγκεκριμένη διάταξη που χρησιμοποιείται. Μεθοδολογικά, η αξιοποίηση και των δύο αυτών στοιχείων έχει σαφή βελτιωτική επίδραση.

Πίνακας 5.22: Κατάταξη Friedman (Συναισθηματικές Διαμορφώσεις) — Χρονικός ορίζοντας δύο εβδομάδων

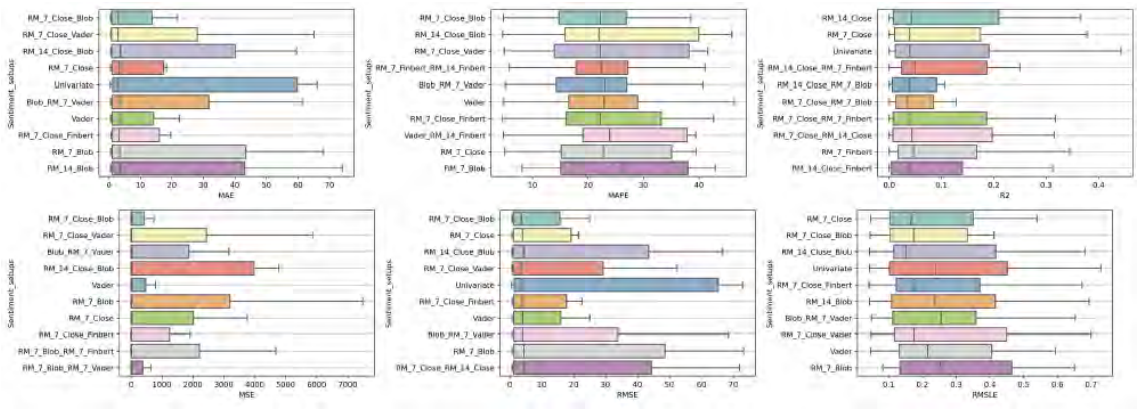
No.	**MAE**		**MAPE**		R^2	
	Feature Setup	**Feature Setup**	**Feature Setup**	**Friedman Score**	**Φεατυρε Σετυπ**	**Friedman Score**
1	RM7C_B	17.46667	RM7C_B	18.53333	RM14C	50.5
2	RM7C_V	21.53333	RM14C_B	22.2	RM7C	48.9
3	RM14C_B	22.26667	RM7C_V	22.93333	U	48.76667
4	RM7C	22.26667	RM7F_RM14F	23.73333	RM14C_RM7F	48.16667
5	U	23.73333	B_RM7V	24.2	RM14C_RM7B	46.83333
6	B_RM7V	23.8	V	25.6	RM7C_RM7B	46.06667
7	V	24.46667	RM7C_F	26.33333	RM7C_RM7F	45.76667
8	RM7C_F	24.86667	V_RM14F	27.4	RM7C_RM14C	45.56667
9	RM7B	25.86667	RM7C	27.66667	RM7F	43.83333
10	RM14B	27.33333	RM7B	28.13333	RM14C_F	43.36667

No.	**MSE**		**RMSE**		**RMSLE**	
	Feature Setup	**Friedman Score**	**Feature Setup**	**Friedman Score**	**Feature Setup**	**Friedman Score**
1	RM7C_B	18.26667	RM7C_B	15.86667	RM7C	13.8
2	RM7C_V	21.26667	RM7C	20.33333	RM7C_B	16
3	B_RM7V	21.6	RM14C_B	21.26667	RM14C_B	21.06667
4	RM14C_B	23.86667	RM7C_V	21.4	U	22.33333
5	V	25.06667	U	22	RM7C_F	23.13333
6	RM7B	25.86667	RM7C_F	24	RM14B	24.33333
7	RM7C	26.2	V	24.06667	B_RM7V	24.46667
8	RM7C_F	26.26667	B_RM7V	24.46667	RM7C_V	26.26667
9	RM7B_RM7F	26.33333	RM7B	26	RM7B	26.46667
10	RM7B_RM7V	26.66667	RM7C_RM14C	27	RM7B	26.46667

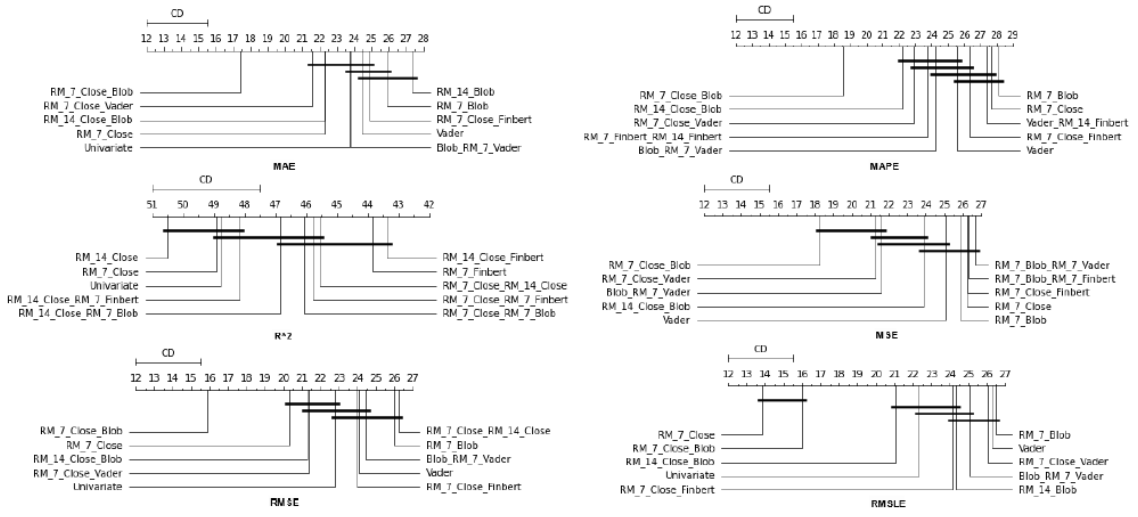
5.5.5.3 Συμπεράσματα

Σχετικά με τις μεθόδους πρόβλεψης Συνοψίζοντας τα αποτελέσματα της πειραματικής διαδικασίας σχετικά με τον πρώτο προς διερεύνηση στόχο, δηλαδή αν υπάρχει κάποια μέθοδος που να παρουσιάζει συγκριτικά καλύτερη προβλεπτική συμπεριφορά από τις υπόλοιπες, παρατηρούμε πως τα εξαγόμενα αποτελέσματα επιτρέπουν με ασφάλεια το συμπέρασμα ότι η μέθοδος TCN υπερέχει έναντι των υπολοίπων. Αυτό ισχύει, καθώς στη συντριπτική πλειονότητα των συγκρίσεων η μέθοδος TCN υπερτερεί των άλλων, καταλαμβάνοντας κυρίως την κορυφή της κατάταξης Friedman. Συγκεκριμένα, οι μόνες περιπτώσεις στις οποίες δεν ξεπερνά όλες τις υπόλοιπες εντοπίζονται στις προβλέψεις χρονικού ορίζοντα μίας ημέρας. Μάλιστα, από τα CD διαγράμματα, μπορεί να εξαχθεί το επιπλέον συμπέρασμα ότι, σε πολλές περιπτώσεις, η υπεροχή της TCN συνοδεύεται από στατιστικά σημαντική διαφορά έναντι των υπόλοιπων μεθόδων.

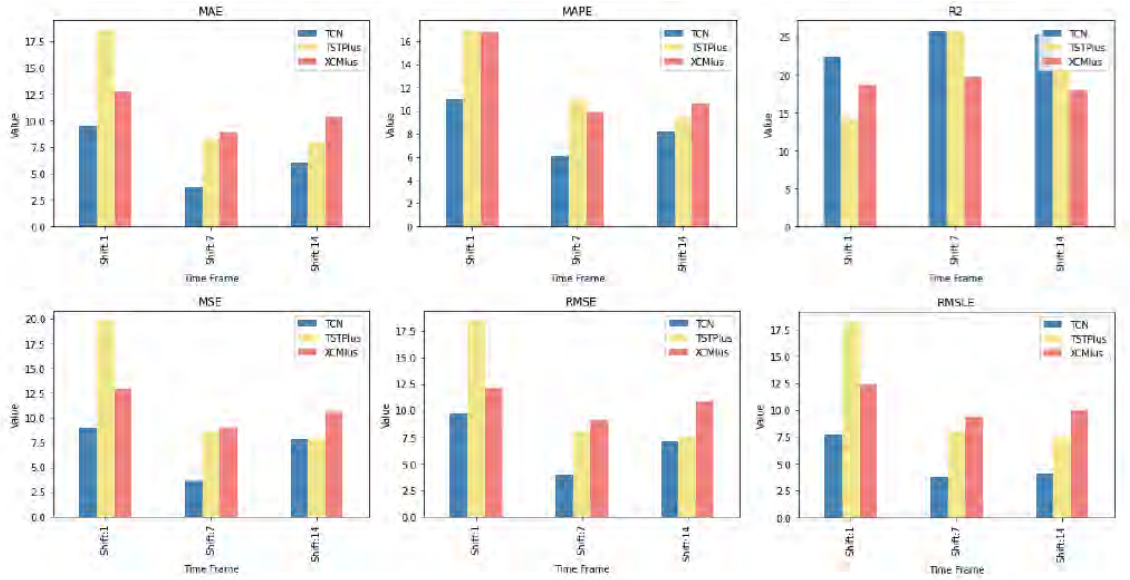
Επιπλέον, εκτός από τη μέθοδο TCN, εντοπίστηκαν και άλλες μέθοδοι με αξιόλογες προβλεπτικές ικανότητες. Μία από αυτές είναι η TSTPlus, καθώς παράγει αξιοσημείωτα αποτελέσματα, ιδιαίτερα για μεγαλύτερους χρονικούς ορίζοντες. Μια άλλη μέθοδος είναι η XCMPlus, η οποία στην πλειοψηφία των περιπτώσεων έχει ανάλογα καλή συμπεριφορά. Στο Σχήμα 5.38, μπορεί κανείς να δει τις σχετικές κατατάξεις αυτών των τριών μεθόδων ανά χρονικό ορίζοντα πρόβλεψης, αντικατοπτρίζοντας συνοπτικά τα αποτελέσματα όπως αυτά αποτυπώνονται στους Πίνακες 5.17, 5.18 και 5.19.



Σχήμα 5.36: Θηκογράμματα για την απόδοση των συναισθηματικών διαμορφώσεων - Χρονικός ορίζοντας δύο εβδομάδων



Σχήμα 5.37: CD διαγράμματα για την απόδοση των αλγορίθμων - Χρονικός ορίζοντας δύο εβδομάδων



Σχήμα 5.38: Σχετική κατάταξη των TCN, TSTPlus και XCMPlus

Όσον αφορά το χρονικό παράθυρο πρόβλεψης μίας ημέρας, η LSTMPlus αποτελεί μέθοδο που παρουσιάζει αξιοσημείωτη απόδοση, ενώ έπεται και ο συνδυασμός GRU και FCN. Ωστόσο, σημειωθεί πως διαφορές μεταξύ των μεμονωμένων μεθόδων δεν είναι πάντα ξεκάθαρα σημαντικές, όπως γίνεται αντιληπτό από τα διαγράμματα κρίσιμων διαφορών 5.27, 5.29 και 5.31.

Ανάλογα, μπορούν επίσης να εξαχθούν συμπεράσματα σχετικά με μεθόδους των οποίων η απόδοση αξιολογήθηκε, κατά μέσο όρο, ως μη ικανοποιητική. Συγκεκριμένα, εντοπίστηκαν συγκεκριμένες μέθοδοι που κατατάσσονται σταθερά στις τελευταίες θέσεις σε όλα τα σενάρια. Μια από αυτές και συγκεκριμένα το πιο αντιπροσωπευτικό παράδειγμα είναι η TSiTPlus, η οποία κατατάσσεται στην τελευταία θέση και στα τρία σενάρια, ανεξαρτήτως μετρικής. Επιπλέον, υπάρχουν και άλλες μέθοδοι, όπως το Transformer Model, XceptionTime και XceptionTimePlus, οι οποίες βρίσκονται σχεδόν πάντα στο κάτω μέρος του πίνακα κατάταξης.

Σχετικά με τις διαμορφώσεις συναισθηματικών χαρακτηριστικών Σε σχέση με τη δεύτερη μελέτη περίπτωσης που αφορά, αφενός, το αν η χρήση πληροφορίας που προέρχεται από την Ανάλυση Συναισθήματος δεδομένων από κοινωνικά δίκτυα συμβάλλει στη βελτίωση των προβλέψεων και, αφετέρου, τη διερεύνηση του αν υπάρχει κάποια συγκεκριμένη διαμόρφωση συναισθηματικών χαρακτηριστικών που να υπερτερεί των υπολοίπων, τα αποτελέσματα της πειραματικής διαδικασίας οδηγούν στα συμπεράσματα που διατυπώνονται στη συνέχεια.

Το βασικό συμπέρασμα που προκύπτει είναι πως η χρήση πληροφοριών που προέρχονται τόσο από εξομαλυμένες εκδοχές των αρχικών χρονοσειρών όσο και από την Ανάλυση Συναισθήματος έχει, στις περισσότερες περιπτώσεις, ευεργετική επίδραση στις παραγόμενες προβλέψεις. Η μη χρήση συναισθηματικών δεδομένων στη διαμόρφωση των χαρακτηριστικών υπερέχει έναντι των υπόλοιπων μόνο σε έναν μικρό αριθμό περιπτώσεων, και όπως επιβεβαιώνεται από τα CD διαγράμματα 5.33, 5.35 και 5.37, μόνο σε δύο από αυτές η διαφορά είναι στατιστικά σημαντική.

Επιπλέον, η απάντηση στο ερώτημα αν η χρήση συναισθηματικών διαμορφώσεων οδηγεί συγκεκριμένα στην εξαγωγή πιο ακριβών προβλέψεων, όπως αποδεικνύεται α-

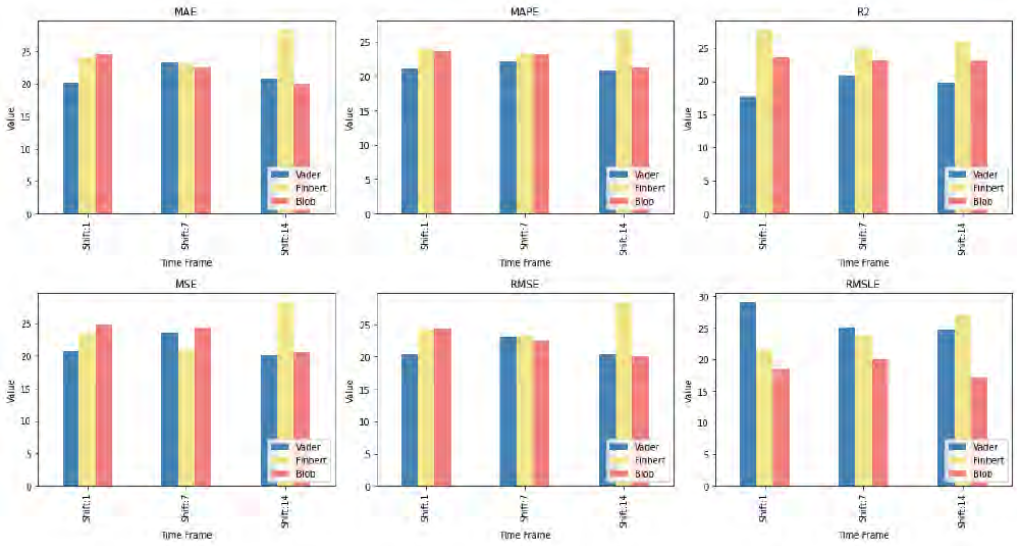
πό τα επιμέρους αποτελέσματα των σταθμισμένων μέσων όρων, φαίνεται να είναι ότι, γενικά, η Ανάλυση Συναισθήματος βελτιώνει τις προβλέψεις. Παρόλα αυτά, δεν εντοπίζεται κάποια συγκεκριμένη μέθοδος που να υπερτερεί γενικά στα διάφορα σενάρια, συνεπώς το συγκεκριμένο ερώτημα χρήζει περαιτέρω διερεύνησης. Ακόμη και αν προχωρήσουμε σε περισσότερες αναλύσεις στο πλαίσιο του πειραματικού σχεδιασμού που παρουσιάστηκε εδώ, δεν είναι βέβαιο ότι μπορούν να εξαχθούν σαφή συμπεράσματα. Εδώ, αν και μπορούν να βρεθούν συγκεκριμένες βέλτιστες διαμορφώσεις για κάθε χρονικό ορίζοντα, δεν υπάρχει μία που να κυριαρχεί σε όλους τους χρονικούς ορίζοντες.

Για να απεικονιστεί η σχετική κατάταξη των τριών μεθοδολογιών ανάλυσης συναισθήματος, ανεξάρτητα από την εκάστοτε παραλλαγή τους, δημιουργήθηκε μια ακόμη οπτική αναπαράσταση. Για να δοθεί μια σαφέστερη εικόνα της απόδοσης των τριών τεχνικών ανάλυσης συναισθήματος, δημιουργήθηκαν τρεις κατηγορίες συναισθήματος, καθεμία από τις οποίες αντιστοιχεί σε μία από τις τρεις χρησιμοποιούμενες μεθόδους Ανάλυσης Συναισθήματος. Ο αριθμητικός μέσος όρος όλων των διαμορφώσεων χαρακτηριστικών που περιέχουν αποκλειστικά διαφορετικές παραλλαγές μίας μόνο από τις τρεις τεχνικές Ανάλυσης Συναισθήματος χρησιμοποιήθηκε για να αντιπροσωπεύσει κάθε κατηγορία ανά μετρική.

Με άλλα λόγια, κάθε κατηγορία αντιπροσωπεύει μία μέθοδο Ανάλυσης Συναισθήματος, και κάθε κατηγορία περιλαμβάνει έξι διαμορφώσεις χαρακτηριστικών που περιέχουν αποκλειστικά παραλλαγές της συγκεκριμένης τεχνικής. Συγκεκριμένα, μια αντιπροσωπευτική τιμή μιας κατηγορίας, όπως αυτή αφορά μία συγκεκριμένη μέθοδο, προκύπτει από το ακόλουθο σύνολο {Μέθοδος, RM7_Μέθοδος, RM14_Μέθοδος, Μέθοδος + RM7_Μέθοδος, Μέθοδος + RM14_Μέθοδος, RM7_Μέθοδος + RM14_Μέθοδος }. Το άθροισμα αυτών διαιρείται δια έξι (δηλαδή το πλήθος των διαμορφώσεων), και αυτό το αποτέλεσμα είναι η τιμή εξόδου που απεικονίζεται. Με αυτόν τον τρόπο, εξαιρούνται οι διαμορφώσεις που παράγονται είτε από συνδυασμούς διαφορετικών μεθόδων ανάλυσης συναισθήματος είτε από χρήση της εξαρτημένης μεταβλητής με κινητούς μέσους όρους, ώστε να συγκριθούν αποκλειστικά οι σχετικές αποδόσεις των τριών τεχνικών Ανάλυσης Συναισθήματος και των παραλλαγών τους.

Το Σχήμα 5.39 απεικονίζει αυτές τις σχετικές κατατάξεις των τριών μεθόδων Ανάλυσης Συναισθήματος ανά χρονικό ορίζοντα, όπως αυτές διαμορφώθηκαν μέσω της προαναφερθείσας διαδικασίας.

Μπορεί κανείς να παρατηρήσει ότι οι κλάσεις των Blob και Vader καταλαμβάνουν την κορυφή της κατάταξης επτά φορές η καθεμία, ενώ η κλάση των διαμορφώσεων του FinBERT κερδίζει μόνο τέσσερις φορές. Και πάλι, δεν φαίνεται να προκύπτει ένα ξεκάθαρο γενικό συμπέρασμα υπέρ μιας συγκεκριμένης μεθόδου. Παρ' όλα αυτά, η ταυτοποίηση ομάδων διαμορφώσεων χαρακτηριστικών, έστω και σε επίπεδο συγκεκριμένου χρονικού πλαισίου, μπορεί να είναι ιδιαίτερα χρήσιμη.



Σχήμα 5.39: Σχετική κατάταξη των κλάσεων συναισθηματικών διαμορφώσεων TextBlob, VADER και FinBERT

6. Η ανάλυση συναισθήματος σε χρηματοοικονομικές προβλέψεις από τη σκοπιά της ταξινόμησης – Κλασικές vs. μέθοδοι Βαθιάς Μάθησης

Ένα από τα χαρακτηριστικά φαινόμενα της σύγχρονης εποχής είναι η μεταφορά μεγάλου μέρους των δραστηριοτήτων που παραδοσιακά πραγματοποιούνταν διά ζώσης στον ψηφιακό κόσμο, όπως οι τραπεζικές συναλλαγές, οι αγορές και η επικοινωνία. Αυτή η διαδικασία, γνωστή και ως ψηφιακός μετασχηματισμός (digital transformation), εκτός από την ευκολία που παρέχει σε πλήθος δραστηριοτήτων, οι οποίες μπορούν πλέον να εκτελούνται από την άνεση του σπιτιού, έχει οδηγήσει στη δημιουργία τεράστιου όγκου δεδομένων. Τα δεδομένα αυτά εμπεριέχουν πολύτιμες πληροφορίες, οι οποίες μπορούν να αποδειχθούν εξαιρετικά χρήσιμες, εφόσον υποβληθούν σε κατάλληλη επεξεργασία και αξιοποιηθούν. Η πρόσβαση σε αυτές τις πληροφορίες επιτρέπει την αποτύπωση της κοινής γνώμης μέσα από την ανάλυση των δεδομένων, μια γνώση που συχνά είναι κρίσιμη για τη λήψη ορθών ή βέλτιστων αποφάσεων σε διάφορους τομείς.

Σε συγκεκριμένους τομείς, όπως η βιομηχανία, οι αγορές και η ψηφιακή ψυχαγωγία, οι απόψεις των πελατών ή των χρηστών υπηρεσιών διαδραματίζουν καθοριστικό ρόλο. Η σημασία τους δεν περιορίζεται μόνο στη συλλογή ανατροφοδότησης, αλλά επεκτείνεται στη βελτίωση της ποιότητας των υπηρεσιών και στην ενίσχυση της ανάπτυξης των προϊόντων. Σε αυτό το πλαίσιο αναδύθηκε η ανάγκη για συστηματική εξαγωγή ουσιαστικών πληροφοριών από αξιολογήσεις χρηστών για προϊόντα, ταινίες και κινηματογραφικές παραγωγές. Η συμμετοχή των χρηστών στα κοινωνικά δίκτυα και οι δραστηριότητές τους σε αυτές τις πλατφόρμες αποτελούν μια τεράστια πηγή πληροφορίας. Η πηγή αυτή μπορεί να αξιοποιηθεί αποτελεσματικά για την εξαγωγή πληροφοριών σχετικά με τις απόψεις και τις προθέσεις των ατόμων, μια διαδικασία που πριν την εξάπλωση του διαδικτύου απαιτούσε σημαντική προσπάθεια, χρόνο και οικονομικούς πόρους. Η ανάγκη για εξαγωγή πληροφοριών από τα σχόλια χρηστών σε κοινωνικές πλατφόρμες, όπως το Facebook [200], το YouTube [271] και το Twitter [129], [311], [142] οδήγησε επίσης στην ανάπτυξη νέων μεθοδολογιών και εργαλείων στον τομέα της Ανάλυσης Συναισθήματος. Πέρα από τα μέσα κοινωνικής δικτύωσης – αν και σε ορισμένες περιπτώσεις μέσω δημοσιεύσεων σε αυτές τις πλατφόρμες – η Ανάλυση Συναισθήματος βρήκε εφαρμογές σε τομείς όπως η υγεία, η πολιτική, η ψυχολογία, το μάρκετινγκ και η οικονομία.

Στον τομέα της οικονομίας συγκεκριμένα, η Ανάλυση Συναισθήματος έχει αναδειχθεί ως εργαλείο με ιδιαίτερη σημασία. Εφαρμόζεται σε διάφορα οικονομικά πεδία, όπως η πρόβλεψη της πορείας των χρηματιστηριακών αγορών [202], η αξιολόγηση της αστάθειας των αγορών [63] καθώς και η ανάλυση της αγοράς κρυπτονομισμάτων [52]. Η αξιοποίηση τεχνικών Ανάλυσης Συναισθήματος στα οικονομικά δεδομένα προσφέρει στους επενδυτές, τις επιχειρήσεις και τους αναλυτές τη δυνατότητα λήψης πιο ενημερωμένων και στρατηγικών αποφάσεων, λαμβάνοντας υπόψη τις απόψεις και τις τάσεις της αγοράς. Ωστόσο, οι εξειδικευμένες απαιτήσεις του οικονομικού τομέα, όπως η

χρήση τεχνικής ορολογίας και η ανάγκη ακριβούς επεξεργασίας δεδομένων, καθιστούν απαραίτητη την ανάπτυξη προσαρμοσμένων μεθόδων και εργαλείων που μπορούν να ανταποκριθούν στις ιδιαίτερες προκλήσεις αυτού του πεδίου. Για τους παραπάνω λόγους αναπτύχθηκε ένα υποπεδίο της Ανάλυσης Συναισθήματος που επικεντρώνεται στην εφαρμογή μεθόδων και τεχνικών της Ανάλυσης Συναισθήματος σε δεδομένα που αντλούνται από τον κόσμο της οικονομίας, δηλαδή σε έγγραφα χρηματοοικονομικής φύσης, η Οικονομική Ανάλυση Συναισθήματος (Financial Sentiment Analysis - FSA). Η FSA αποτελεί πολύτιμο εργαλείο για επενδυτές, χρηματοπιστωτικά ιδρύματα και αναλυτές, καθώς προσφέρει τη δυνατότητα εκτίμησης του συναισθήματος της αγοράς, αξιολόγησης των κινδύνων και λήψης πιο ενημερωμένων οικονομικών αποφάσεων. Σε αντίθεση με την πρόβλεψη χρονοσειρών με τη χρήση Ανάλυσης Συναισθήματος που συζητήθηκε εκτενώς στο προηγούμενο κεφάλαιο, η Οικονομική Ανάλυση Συναισθήματος εστιάζει σε προβλήματα ταξινόμησης, και όχι παλινδρόμησης.

Η φιλοσοφία που στηρίζει την άνοδο της FSA στον χρηματοοικονομικό τομέα σχετίζεται με την υποχώρηση της Υπόθεσης Αποτελεσματικής Αγοράς (Efficient Market Hypothesis [298], EMH) υπέρ της Υπόθεσης Προσαρμοστικής Αγοράς (Adaptive Market Hypothesis, AMH), ως απόρροια της κριτικής από τον τομέα της Συμπεριφορικής Οικονομίας. Σύμφωνα με τη θεωρία της EMH, οι αγορές είναι πλήρως αποδοτικές και οι τιμές των περιουσιακών στοιχείων αντανακλούν πάντα όλες τις διαθέσιμες πληροφορίες. Η AMH, ωστόσο, αμφισβητεί τις αυστηρές υποθέσεις της EMH, αναγνωρίζοντας ότι οι συμμετέχοντες στις αγορές δεν είναι πάντα απολύτως ορθολογικοί και ότι η δυναμική των αγορών μπορεί να μεταβάλλεται με την πάροδο του χρόνου. Η AMH υποστηρίζει ότι οι συμμετέχοντες προσαρμόζονται στις νέες πληροφορίες και στις συνθήκες της αγοράς, οδηγώντας έτσι σε μεταβολές στις τιμές των περιουσιακών στοιχείων.

Σε ένα δυναμικό και διαρκώς εξελισσόμενο περιβάλλον, η ενεργή επεξεργασία και αξιολόγηση κάθε νέας πληροφορίας θεωρείται ζωτικής σημασίας για τη διαμόρφωση μελλοντικών αποφάσεων. Συνεπώς, η χρήση τεχνικών Ανάλυσης Συναισθήματος σε χρηματοοικονομικά δεδομένα μπορεί να αποδειχθεί εξαιρετικά ωφέλιμη και προσοδοφόρα, καθοδηγώντας τους ενδιαφερόμενους φορείς προς πιο τεκμηριωμένες και στρατηγικές επιλογές. Η FSA βρίσκει εφαρμογή σε ποικίλους τομείς του χρηματοοικονομικού κόσμου. Ενδεικτικά, χρησιμοποιείται για την πρόβλεψη χρηματιστηριακών τάσεων [224], την πρόβλεψη συναλλαγματικών ισοτιμιών (FOREX) [207], την εκτίμηση της μεταβλητότητας της αγοράς [63], την κατανομή περιουσιακών στοιχείων [160], [300], την αξιολόγηση πιστοληπτικής ικανότητας [309], την εκτίμηση αξίας αρχικών δημόσιων εγγραφών [22] και την ανάλυση κρυπτονομισμάτων [139], [228], [18].

Η Ανάλυση Συναισθήματος στον χρηματοοικονομικό τομέα συνοδεύεται από ιδιαίτερες προκλήσεις που την καθιστούν πιο απαιτητική σε σύγκριση με άλλες εφαρμογές. Η περιορισμένη διαθεσιμότητα μεγάλων συνόλων δεδομένων, σε συνδυασμό με την ανάγκη εξειδικευμένης γνώσης για την ετικετικοποίηση χρηματοοικονομικών κειμένων, την ξεχωρίζει από πιο γενικές χρήσεις της Ανάλυσης Συναισθήματος [299]. Τα χρηματοοικονομικά έγγραφα, όπως αναφορές, δημοσιεύσεις σε κοινωνικά δίκτυα και ειδησεογραφικά άρθρα, περιέχουν συχνά τεχνική ορολογία και εξειδικευμένο οικονομικό λεξιλόγιο, καθιστώντας την ανάλυσή τους μια σύνθετη διαδικασία.

Λαμβάνοντας υπόψιν τα παραπάνω, η ανάπτυξη αποτελεσματικών μεθόδων προσαρμοσμένων στις ανάγκες του χρηματοοικονομικού τομέα είναι κρίσιμη. Η αξιολόγηση των υφιστάμενων τεχνικών συμβάλλει στον εντοπισμό των πιο κατάλληλων εργαλείων, ενώ η τροποποίηση και προσαρμογή αυτών των εργαλείων ανοίγει τον δρόμο για

πιο εξειδικευμένες και αποδοτικές προσεγγίσεις. Σε αυτό το πλαίσιο, η ακρίβεια στην ανάλυση συναισθήματος αποκτά καθοριστική σημασία, καθώς λανθασμένα αποτελέσματα μπορούν να οδηγήσουν σε σημαντικές οικονομικές απώλειες. Η παραπάνω διαπίστωση αποτέλεσε έναυσμα της εργασίας που θα παρουσιαστεί σε αυτό το κεφάλαιο, που στοχεύει στην σύγκριση ενός μεγάλου πλήθους μεθόδων ταξινόμησης σε οικονομικά κείμενα, τόσο κλασσικών όσο και σύγχρονων από το πεδίο των transformers της οικογένειας BERT.

6.1 Περιγραφή της πειραματικής διαδικασίας

Στην υποενότητα αυτή παρέχονται πληροφορίες σχετικά με τα δεδομένα και τις μεθόδους που χρησιμοποιήθηκαν στο πλαίσιο της πειραματικής διαδικασίας, ενώ περιγράφονται επίσης τα σενάρια που εξετάστηκαν.

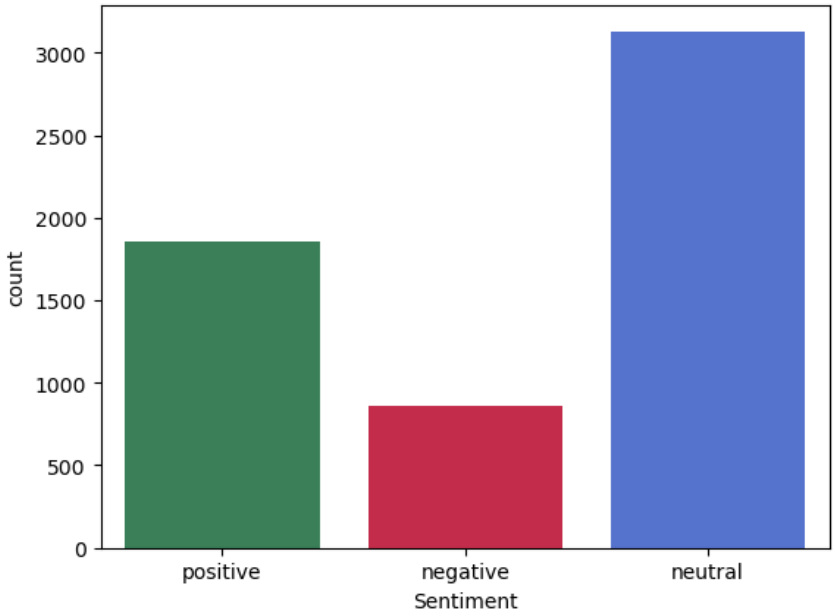
6.1.1 Σύνολο Δεδομένων

Για τις ανάγκες της εργασίας, χρησιμοποιήθηκε ένα δημόσιο σύνολο δεδομένων [84] που αποτελεί συνένωση δύο ξεχωριστών συνόλων δεδομένων που η χρήση τους είναι συχνή σε προβλήματα ανάλυσης συναισθήματος που αφορούν χρηματοοικονομικά δεδομένα. Τα σύνολα δεδομένων είναι:

- FiQA (Financial Question Answering and Opinion Mining) [157]: Το σύνολο δεδομένων FiQA δημιουργήθηκε στο πλαίσιο του διαγωνισμού Financial Opinion Mining and Question Answering που διοργανώθηκε στο συνέδριο WWW 2018, είναι δημόσια διαθέσιμο και χρησιμοποιείται ευρέως για την εκπαίδευση και την αξιολόγηση μοντέλων Επεξεργασίας Φυσικής Γλώσσας σε χρηματοοικονομικά περιβάλλοντα, ενώ είναι κατάλληλο για ένα φάσμα εργασιών όπως η κατηγοριοποίηση συναισθήματος, η ανάλυση απόψεων και η απάντηση ερωτήσεων. Αποτελείται από ερωτήσεις και απαντήσεις που σχετίζονται με χρηματοοικονομικές εκθέσεις, νέα, άρθρα, και συζητήσεις, συνοδευόμενες από τις αντίστοιχες συναισθηματικές πολικότητες, οι οποίες αποτυπώνονται σε ένα συνεχές φάσμα από το -1 έως το 1, όπου η τιμή -1 εκφράζει αρνητική πολικότητα, ενώ η τιμή 1 θετική.

- Financial PhraseBank [161]: Το σύνολο δεδομένων Financial PhraseBank περιέχει χρηματοοικονομικές φράσεις και είναι σχεδιασμένο ειδικά για την εκπαίδευση και αξιολόγηση μοντέλων Ανάλυσης Συναισθήματος στον χρηματοοικονομικό τομέα. Αποτελείται από 4.845 χρηματοοικονομικές προτάσεις που έχουν επιλεγεί από άρθρα και αναφορές από τη βάση δεδομένων LexisNexis [145], κάθε μια από τις οποίες συνοδεύεται από την αντίστοιχη πολικότητα. Η πολικότητα σε αυτό το σύνολο δεδομένων μπορεί να πάρει μια από τρεις διακριτές τιμές: 1, 0 ή -1 που δηλώνουν αντίστοιχα τη θετική, την ουδέτερη και την αρνητική πολικότητα. Ο χαρακτηρισμός των προτάσεων έχει από 16 ειδικούς στον τομέα της οικονομίας, όπου η πλειοψηφία των απόψεων χαρακτήριζε την πρόταση, ενώ σε περιπτώσεις ισοψηφίας για το χαρακτηρισμό, ακολουθούσε συζήτηση μεταξύ των ειδικών.

Το τελικό σύνολο δεδομένων αποτελείται από 5.842 εγγραφές, συνοδευόμενες με την αντίστοιχη πολικότητα. Καθώς παρατηρήθηκε το φαινόμενο να υπάρχουν διπλοεγγραφές, με διαφορετική όμως πολικότητα, κατά τη φάση της προεργασίας αφαιρέθηκαν



Σχήμα 6.1: Κατανομή των δειγμάτων σε κλάσεις

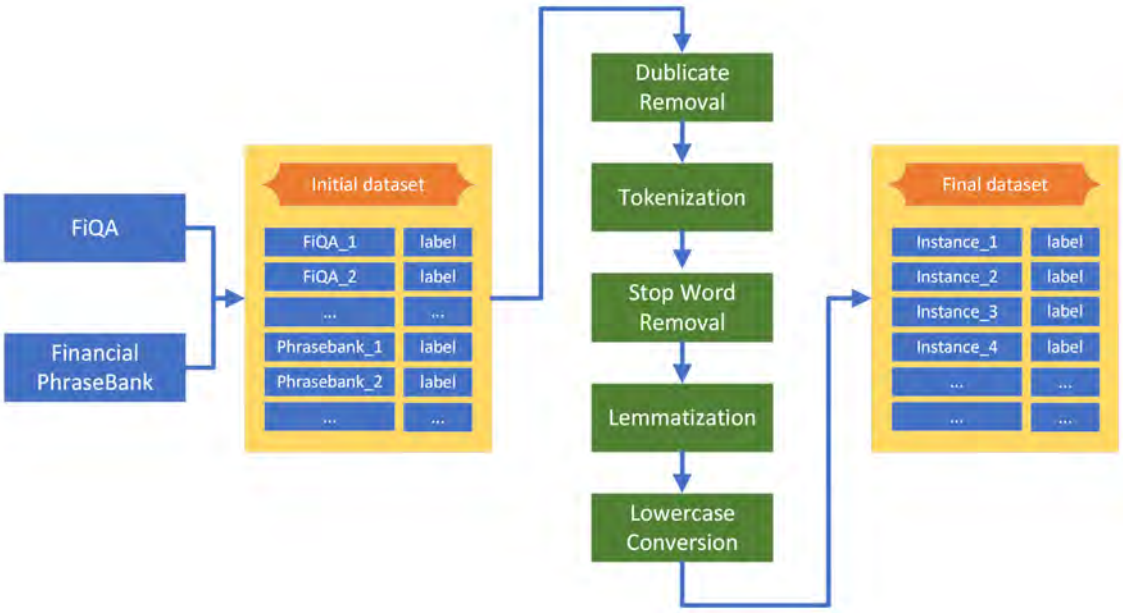
από το σύνολο δεδομένων. Η ύπαρξη διπλοεγγραφών πιθανότατα προέκυψε από τη μετατροπή των συνεχών τιμών του FiQA σε διακριτές. Για παράδειγμα, η πρόταση «However, the company saw its net profit for the third quarter decline to EUR 1.4 million from EUR 1.5 million for the corresponding period of 2009» εμφανίζεται δύο φορές, όπου την πρώτη χαρακτηρίζεται ως ουδέτερη και τη δεύτερη ως αρνητική. Επιπλέον, ακόμη ένα ζήτημα που θα πρέπει να αναφερθεί, είναι η παρουσία παρόμοιων προτάσεων, στις οποίες όμως έχει ανατεθεί διαφορετική πολικότητα, όπως για παράδειγμα οι προτάσεις «Earnings per share (EPS) amounted to a loss of EUR0.05» και «Earnings per share (EPS) amounted to a loss of EUR0.06» οι οποίες έχουν παρόμοια σημασιολογία και θα έπρεπε να φέρουν την ίδια πολικότητα. Το γεγονός αυτό, αυξάνει τη δυσκολία του προβλήματος, καθώς δύναται να μειώσει την αποδοτικότητα των μεθόδων. Τέλος, πρέπει να αναφερθεί πως οι εγγραφές του συνόλου δεδομένων δεν μοιράζονται ισάξια στις τρεις κλάσεις – πολικότητες, καθιστώντας το μη ισορροπημένο. Όπως φαίνεται και στο Σχήμα 6.1, στην πλειοψηφία των εγγραφών έχει αποδοθεί ουδέτερη πολικότητα, ενώ από τις υπόλοιπες εγγραφές, το πλήθος των θετικών είναι διπλάσιο από αυτό των αρνητικών.

Στη συνέχεια θα περιγραφεί η διαδικασία προεπεξεργασίας επί του σθνόλου δεδομένων, η οποία φαίνεται στο διάγραμμα ροής στην εικόνα 6.2.

Κατά το στάδιο της προεξεργασίας ακολουθήθηκαν τα παρακάτω βήματα:

1. **Αφαίρεση Διπλότυπων Προτάσεων:** Εντοπίστηκαν περίπου 520 διπλότυπες προτάσεις που είχαν πιθανώς διαφορετικά συναισθήματα. Επιλέχθηκε η αφαίρεση αυτών των διπλότυπων προτάσεων, ώστε να αποφευχθεί η σύγχυση στα δεδομένα και να διασφαλιστεί η ομοιομορφία.

2. **Διαχωρισμός σε Λέξεις (Tokenization):** Ο διαχωρισμός κειμένου σε λέξεις (tokens) είναι μια βασική τεχνική στην Επεξεργασία Φυσικής Γλώσσας. Για τον σκοπό αυτό, χρησιμοποιήσαμε τη βιβλιοθήκη NLTK της Python, η οποία διευκολύνει τον κατακερματισμό των κειμένων σε μεμονωμένες λέξεις ή φράσεις.



Σχήμα 6.2: Διαγραμματική απεικόνιση της προεπεξεργασίας του συνόλου δεδομένων

3. **Αφαίρεση των Stop Words (Stop Word Removal):** Η αφαίρεση των stop words είναι μια συνήθης πρακτική στην προεπεξεργασία δεδομένων. Αυτές οι λέξεις, όπως άρθρα, προθέσεις και σύνδεσμοι, συχνά δεν προσθέτουν ουσιαστική σημασία στο κείμενο. Η αφαίρεσή τους μειώνει τον συνολικό αριθμό των λέξεων και δίνει έμφαση στις πιο σημαντικές λέξεις.

4. **Λημματοποίηση (Lemmatization):** Επιλέχθηκε η λημματοποίηση αντί της διαδικασίας αποκοπής (stemming). Ενώ και οι δύο τεχνικές στοχεύουν στην κανονικοποίηση του κειμένου, η λημματοποίηση παρέχει καλύτερα αποτελέσματα με λιγότερες υπολογιστικές απαιτήσεις. Χρησιμοποιήσαμε το WordNet Lemmatizer της βιβλιοθήκης NLTK, το οποίο αποδίδει τις λέξεις στη βασική τους μορφή (λήμμα) με ακρίβεια.

5. **Μετατροπή σε Πεζά (Lowercase Conversion):** Η μετατροπή όλων των χαρακτήρων σε πεζά είναι μια τυπική τεχνική κανονικοποίησης του κειμένου, καθώς η σημασία των λέξεων παραμένει ίδια ανεξαρτήτως κεφαλαίων ή πεζών γραμμάτων. Μελέτες έχουν δείξει ότι η συγκεκριμένη τεχνική μπορεί να βελτιώσει την απόδοση των μοντέλων Ανάλυσης Συναισθήματος.

6.1.2 Γλωσσικές αναπαραστάσεις

Τα δεδομένα κειμένου είναι εγγενώς μη δομημένα και δύσκολα επεξεργάσιμα από μαθηματικά μοντέλα. Συνεπώς, βασικό βήμα της προεπεξεργασίας αποτελεί η μετατροπή τους σε μορφές κατάλληλες να αποτελέσουν είσοδο σε υπολογιστικά συστήματα εν γένει και στην περίπτωσή μας συγκεκριμένα σε αλγορίθμους κατηγοριοποίησης. Υπάρχουν διάφορα ήδη γλωσσικών αναπαραστάσεων, καθένα από τα οποία παρουσιάζει πλεονεκτήματα και μειονεκτήματα. Στο πλαίσιο της έρευνας, χρησιμοποιήθηκαν δύο γλωσσικές αναπαραστάσεις οι Bag of Words (BoW) [216] και Term Frequency-Inverse Document Frequency (TF-IDF) [221].

	BoW	TF-IDF
Βασική Ιδέα	Καταμέτρηση της συχνότητας εμφάνισης κάθε λέξης	Βάρη στις λέξεις βάσει της συχνότητας και της σπανιότητάς τους
Υπολογιστική Πολυπλοκότητα	Χαμηλή, κατάλληλη για μεγάλα σύνολα δεδομένων	Υψηλότερη, λόγω της υπολογιστικής απαίτησης του
Αντίληψη Σημασιολογίας	Αγνοεί τη σημασιολογία και τη σειρά των λέξεων	Ανιχνεύει τη σημασία λέξεων με βάση τη σχετική συχνότητα
Καταλληλότητα	Κατάλληλο για βασικές εφαρμογές (π.χ. ανάλυση συναισθήματος	Κατάλληλο για εξειδικευμένες εφαρμογές (π.χ. ανίχνευση ρητορικής μίσους)
Πρόβλημα Θορύβου	Ευάλωτο σε κοινές λέξεις (π.χ. "και", "ή")	Αντιμετωπίζει το πρόβλημα θορύβου αποδίδοντας χαμηλά βάρη στις κοινές λέξεις
Διάσταση Δεδομένων	Υψηλής διαστασιμότητας διανύσματα	Παρόμοια διανύσματα, αλλά με πληροφοριακά βάρη
Ευκολία Υλοποίησης	Εύκολη στην κατανόηση και εφαρμογή	Πιο πολύπλοκη, λόγω της ενσωμάτωσης δύο στατιστικών μετρήσεων

Πίνακας 6.1: Διαφορές των Bow και TF-IDF

Η πρώτη τεχνική, Bag of Words (BoW), δημιουργεί αναπαραστάσεις των κειμένων υπολογίζοντας την εμφάνιση κάθε μοναδικής λέξης σε αυτά. Αρχικά, δημιουργείται ένα λεξιλόγιο από όλες τις μοναδικές λέξεις που εμφανίζονται στη συλλογή εγγράφων. Στη συνέχεια, κάθε κείμενο αναπαρίσταται ως ένα διάνυσμα, όπου κάθε διάσταση αντιστοιχεί σε μία λέξη του λεξιλογίου, και η τιμή της διάστασης είναι ο αριθμός εμφανίσεων της λέξης στο συγκεκριμένο κείμενο. Αυτή η τεχνική είναι απλή, υπολογιστικά αποδοτική, και ιδιαίτερα κατάλληλη για βασικές εφαρμογές. Ωστόσο, η κύρια αδυναμία της είναι ότι δεν λαμβάνει υπόψη τη σημασιολογία ή τις σχέσεις μεταξύ των λέξεων, γεγονός που περιορίζει την αποτελεσματικότητά της σε πιο πολύπλοκες εφαρμογές.

Αντίθετα, η τεχνική Term Frequency-Inverse Document Frequency (TF-IDF) εισάγει μια πιο εξελιγμένη προσέγγιση, προσδίδοντας βάρη στις λέξεις ανάλογα με τη συχνότητα και τη σπανιότητά τους. Το TF υπολογίζει πόσο συχνά εμφανίζεται μια λέξη σε ένα κείμενο, ενώ το IDF εκτιμά τη σπανιότητα της λέξης σε όλο το σύνολο των κειμένων. Ο συνδυασμός αυτών των μετρήσεων παράγει τη βαθμολογία TF-IDF για κάθε λέξη, που αντανακλά τη σχετική σημασία της στο κείμενο. Αυτή η τεχνική είναι ιδιαίτερα χρήσιμη σε εφαρμογές που απαιτούν μεγαλύτερη κατανόηση του περιεχομένου και της σχετικότητας των λέξεων.

Παρά τα πλεονεκτήματά τους, και οι δύο τεχνικές έχουν τον περιορισμό της απώλειας του σημασιολογικού πλαισίου, καθώς αγνοούν τη σειρά των λέξεων μέσα στο κείμενο. Επιπλέον, η υψηλή διαστατικότητα που χαρακτηρίζει τη BoW και οι αυξημένες υπολογιστικές απαιτήσεις της TF-IDF μπορεί να αποτελέσουν πρόκληση σε πολύ μεγάλα σύνολα δεδομένων. Συνοπτικά, οι διαφορές των δύο μεθόδων, παρουσιάζονται στον Πίνακα 6.1.

6.1.3 Μοντέλα Ταξινόμησης

Στην υποενότητα αυτή θα περιγραφούν οι μέθοδοι Μηχανικής Μάθησης και συγκεκριμένα οι μέθοδοι ταξινόμησης που χρησιμοποιήθηκαν στο πλαίσιο της εργασίας, οι οποίες εμπίπτουν σε δύο κατηγορίες. Η πρώτη αφορά κλασικές μεθόδους ταξινόμησης, αλλά και ομάδες ταξινομητών αποτελούμενες από αυτούς. Η δεύτερη περιλαμβάνει μια σειρά από προεκπαιδευμένα μοντέλα Βαθιάς Μάθησης, τα οποία αποτελούν μια σύγχρονη και υποσχόμενη προσέγγιση σε προβλήματα Επεξεργασίας Φυσικής Γλώσσας.

	Μέθοδος	Συντομογραφία
1	AdaBoost Classifier	AdaBoost
2	CatBoost Classifier	CatBoost
3	Decision Tree Classifier	DT
4	Dummy Classifier	DC
5	Extra Trees Classifier	ET
6	Extreme Gradient Boosting	XGBoost
7	Gradient Boosting Classifier	GBC
8	K-Nearest Neighbors	KNN
9	Light Gradient Boosting Machine	LightGBM
10	Linear Discriminant Analysis	LDA
11	Logistic Regression	LR
12	Naive Bayes	NB
13	Quadratic Discriminant Analysi	QDA
14	Random Forest Classifier	RF
15	Ridge Classifier	Ridge
16	Support Vector Machine	SVM

Πίνακας 6.2: Αλγόριθμοι Μηχανικής Μάθησης

6.1.3.1 Κλασικές μέθοδοι ταξινόμησης

Ξεκόντας με την παρουσίαση των κλασικών μεθόδων ταξινόμησης, πρέπει να γίνει αναφοράς την βιβλιοθήκη μέσω της οποίας αυτά υλοποιήθηκαν, η οποία είναι η Pycaret [8]. Η PyCaret είναι μια ισχυρή βιβλιοθήκη ανοιχτού κώδικα σε πψτηον, σχεδιασμένη να απλοποιεί τη δημιουργία, εκπαίδευση και αξιολόγηση μοντέλων μέσω μιας λοω-ςοδε προσέγγισης. Υποστηρίζει επιβλεπόμενη μάθηση (ταξινόμηση, παλινδρόμηση), μη επιβλεπόμενη μάθηση (ομαδοποίηση, εντοπισμός ανωμαλιών), παρέχοντας ενσωματωμένη υποστήριξη για περισσότερους από 25 αλγορίθμους, όπως Decision Trees, Random Forests, Support Vector Machines, Gradient Boosting (XGBoost, LightGBM, CatBoost). Η βιβλιοθήκη αυτή επιτρέπει στους χρήστες να συγκρίνουν μοντέλα, να εκτελούν διασταυρούμενη επικύρωση και να βελτιστοποιούν υπερπαραμέτρους μέσω αυτόματων διαδικασιών. Επιπλέον, διαθέτει ισχυρές δυνατότητες προεπεξεργασίας δεδομένων, όπως η διαχείριση ελλιπών τιμών, η κλιμάκωση χαρακτηριστικών, η αφαίρεση μεταβλητών και η μετατροπή κατηγορηματικών δεδομένων σε αριθμητικά δεδομένα, αλλά και εργαλεία οπτικοποίησης για τη σύγκριση της απόδοσης των μοντέλων. Ένα επιπλέον πλεονέκτημά της είναι η δυνατότητα ενσωμάτωσης με άλλες βιβλιοθήκες, όπως οι scikit-learn, pandas και matplotlib. Από τη βιβλιοθήκη, αξιοποιήθηκαν οι παρακάτω μέθοδοι κατηγοριοποίησης. Όπως φαίνονται στον Πίνακα 6.2, μαζί με τις συντομογραφίες τους, με τις οποίες από εδώ κι έπειτα θα αναφέρονται.

Στο σημείο αυτό, πρέπει να υπενθυμίσουμε πως το σύνολο δεδομένων που χρησιμοποιήθηκε παρουσιάζει ανομοιομορφία ως προς τον αριθμό των δειγμάτων που ανήκουν στις τρεις κλάσεις. Η ύπαρξη μη ισορροπημένων κλάσεων, μπορεί να επηρεάσει σε μεγάλο βαθμό την αποτελεσματικότητα των αλγορίθμων, γι' αυτό το λόγο, εκτός από την χρήση stratified k-fold cross-validation που χρησιμοποιήθηκε κατά την πειραματική διαδικασία, αποφασίστηκε να διερευνηθεί και η χρήση της τεχνικής Synthetic Mino-

rity Over-sampling Technique (SMOTE) [86], [244]. Η τεχνική αυτή έχει σχεδιαστεί ειδικά για την εξισορρόπηση των κλάσεων δημιουργώντας συνθετικά δείγματα για τη μειονοτική κλάση. Σε αντίθεση με απλές τεχνικές oversampling, όπως η αντιγραφή υπαρχόντων δεδομένων, η τεχνική SMOTE δημιουργεί νέα συνθετικά δείγματα μέσω παρεμβολής. Συγκεκριμένα, για κάθε δείγμα της μειονοτικής κλάσης, εντοπίζει τους k πλησιέστερους γείτονές του στον χώρο χαρακτηριστικών και δημιουργεί νέα σημεία δεδομένων σε τυχαίες θέσεις κατά μήκος των γραμμών που συνδέουν το αρχικό σημείο με τους γείτονες. Αυτή η διαδικασία όχι μόνο ενισχύει τη κλάση με τα λιγότερα παραδείγματα, αλλά επίσης αποτρέπει το υπερταίριασμα που μπορεί να προκύψει από την απλή αντιγραφή δεδομένων. Τα συνθετικά δείγματα εισάγουν περισσότερη ποικιλία στα δεδομένα της μειονοτικής κλάσης, βοηθώντας τα μοντέλα Μηχανικής Μάθησης να μάθουν καλύτερα τα χαρακτηριστικά αυτής της κλάσης και να βελτιώσουν την ικανότητά τους να γενικεύουν. Η τεχνική SMOTE έχει αποδειχθεί εξαιρετικά αποτελεσματική σε διάφορες εφαρμογές ανάλυσης συναισθήματος, επιτυγχάνοντας σημαντική βελτίωση στην απόδοση των ταξινομητών, όπως καταγράφεται και στη βιβλιογραφία. Επιπλέον, στο πλαίσιο της μελέτης μας, η απόδοση της SMOTE συγκρίνεται με μια δεύτερη προσέγγιση που βασίζεται αποκλειστικά στη stratified k-fold cross-validation χωρίς την προσθήκη συνθετικών δεδομένων. Αυτή η σύγκριση επιτρέπει την αξιολόγηση της επίδρασης της SMOTE στη βελτίωση των αποτελεσμάτων ταξινόμησης.

Καθώς η βιβλιογραφία παρέχει ισχυρές ενδείξεις πως η δημιουργία ομαδοποιητών ταξινομητών (ensembles) δύναται να παρέχει καλύτερα αποτελέσματα από τη χρήση ενός μόνο ταξινομητή, θεωρήθηκε εύλογο να δημιουργηθούν τέτοια συνδυαστικά μοντέλα, χρησιμοποιώντας ως βάση τους ταξινομητές του Πίνακα 6.2. Συγκεκριμένα, για κάθε περίπτωση χρησιμοποιήθηκαν οι καλύτεροι ταξινομητές σύμφωνα με την μετρική MCC, και διαμορφώθηκαν από αυτούς όλες οι δυνατές τριάδες. Για τη συγχώνευση των αποτελεσμάτων των ταξινομητών, χρησιμοποιούνται κυρίως δύο βασικές τεχνικές: το σκληρό voting (hard voting) και το ήπιο voting (soft voting).

- Hard voting: Γνωστό και ως ¨πλειοψηφική ψήφος,' σε αυτήν την προσέγγιση κάθε μοντέλο-ταξινομητής ρίχνει μία ψήφο για μία κλάση, και η κλάση που λαμβάνει τις περισσότερες ψήφους αποτελεί την τελική πρόβλεψη.

- Soft voting: Σε αντίθεση, αυτή η μέθοδος υπολογίζει τον μέσο όρο των πιθανοτήτων για κάθε κλάση από τους επιμέρους ταξινομητές. Η κλάση με την υψηλότερη μέση πιθανότητα επιλέγεται ως η τελική πρόβλεψη.

Και οι δύο αυτές τεχνικές έχουν εφαρμοστεί ευρέως στον τομέα της ανάλυσης συναισθημάτων. Στο [19], η τεχνική του ήπιου voting χρησιμοποιήθηκε για τη δημιουργία ενός μετα-ενσεμβλε ταξινομητή για ανάλυση συναισθημάτων σε κριτικές ταινιών, ενώ στο [284], η τεχνική του σκληρού voting εφαρμόστηκε σε ανάλυση συναισθημάτων δεδομένων Twitter σχετικά με υπηρεσίες αεροπορικών εταιρειών. Στην παρούσα έρευνα, επιλέγουμε τη μέθοδο του ήπιουvoting, καθώς λαμβάνει υπόψη τα επίπεδα εμπιστοσύνης κάθε ταξινομητή, καθιστώντας την πιο εξελιγμένη και κατάλληλη επιλογή για τη συγκεκριμένη ανάλυση.

Λαμβάνοντας υπόψιν όλα τα παραπάνω, στο πλαίσιο της διεύρυνσης της απόδοσης των κλασσικών μεθόδων, εξετάστηκαν τα παρακάτω σενάρια:

1. Κλασικοί ταξινομητές ML + Bag of Words (BoW) + SMOTE.

2. Κλασικοί ταξινομητές ML + TF-IDF + SMOTE.

3. Κλασικοί ταξινομητές ML + Bag of Words (BoW).
4. Κλασικοί ταξινομητές ML + TF-IDF.
5. Soft Voting (συνδυασμός τριών μοντέλων) + Bag of Words (BoW) + SMOTE.
6. Soft Voting (συνδυασμός τριών μοντέλων) + TF-IDF + SMOTE.
7. Soft Voting (συνδυασμός τριών μοντέλων) + Bag of Words (BoW).
8. Soft Voting (συνδυασμός τριών μοντέλων) + TF-IDF.

6.1.3.2 Προεκπαιδευμένα γλωσσικά μοντέλα

Τα προεκπαιδευμένα μοντέλα είναι μια κεντρική έννοια στον τομέα της επεξεργασίας φυσικής γλώσσας και όχι μόνο γιατί χρησιμοποιούνται ευρέως και για άλλες εργασίες όπως για παράδειγμα η υπολογιστική όραση ή η αναγνώριση ομιλίας. Τα μοντέλα αυτά, αντί να εκπαιδεύονται από το μηδέν για κάθε νέα εφαρμογή, βασίζονται στη φιλοσοφία μεταφοράς γνώσης (transfer learning). Η κεντρική ιδέα της μεταφοράς γνώσης, είναι ότι ένα μοντέλο που έχει εκπαιδευτεί για ένα συγκεκριμένο καθήκον χρησιμοποιείται ως βάση για να λυθούν άλλα συναφή με το αρχικό, προβλήματα. Με τον τρόπο αυτό, τα μοντέλα δεν χρειάζεται να εκπαιδευτούν από την αρχή σε κάθε εργασία, καθώς μεταφέρουν τις γνώσεις τους από ένα τομέα σε έναν άλλο, μειώνοντας έτσι τον χρόνο εκπαίδευσης και τις απαιτήσεις δεδομένων. Συνεπώς, η ουσία αυτής της καινοτομίας είναι ότι τα μοντέλα μπορούν να μάθουν γενικές παραστάσεις που είναι χρήσιμες σε πολλές εφαρμογές. Για παράδειγμα, σε μια εργασία που αφορά εικόνες, το μοντέλο μπορεί να έχει μάθει να αναγνωρίζει γενικά χαρακτηριστικά, όπως άκρα ή σχήματα, τα οποία μπορούν να είναι χρήσιμα για την αναγνώριση διαφορετικών αντικειμένων σε νέες εικόνες.

Αντίστοιχα, στο πεδίο της Επεξεργασίας Φυσικής Γλώσσας, τα μοντέλα μαθαίνουν γενικές γλωσσικές δομές και παραστάσεις από μεγάλες ποσότητες κειμένων και στη συνέχεια μπορούν να προσαρμοστούν για να εκτελούν εξειδικευμένες εργασίες χωρίς να απαιτείται πλήρης εκπαίδευση από το μηδέν. Προεκπαιδευμένα μοντέλα, όπως το BERT (Bidirectional Encoder Representations from Transformers), έχουν εκπαιδευτεί σε τεράστιες βάσεις δεδομένων κειμένων και είναι σε θέση να κατανοούν τις γλωσσικές σχέσεις και τις έννοιες σε βάθος. Αυτά τα μοντέλα χρησιμοποιούνται για ένα ευρύ φάσμα εργασιών, από την κατηγοριοποίηση κειμένων [5], μέχρι πιο σύνθετες εφαρμογές όπως η απάντηση σε ερωτήματα (question answering) [168] και η μετάφραση κειμένων [295].

Αφού το μοντέλο αρχικά εκπαιδευτεί, ακολουθεί σε δεύτερη φάση η διαδικασία του fine-tuning, όπου το μοντέλο αναπροσαρμόζεται χρησιμοποιώντας δεδομένα που σχετίζονται με την νέα ή εξειδικευμένη εφαρμογή και προσαρμόζεται με αυτόν τον τρόπο στις ανάγκες του νέου προβλήματος. Πριν όμως περάσουμε στην παρουσίαση των προεκπαιδευμένων μοντέλων που χρησιμοποιήθηκαν στο πλαίσιο της έρευνας, ας δούμε πως λειτουργούν οι transformers, στη λειτουργία των απόιων βασίζονται αυτά τα μοντέλα.

6.1.3.3 Αρχιτεκτονική των transformers

Όπως ήδη συζητήθηκε στο Κεφάλαιο 4, οι transformers αποτελούν μια επαναστατική προσέγγιση στον τομέα της φυσικής επεξεργασίας γλώσσας. Η αρχιτεκτονική ενός

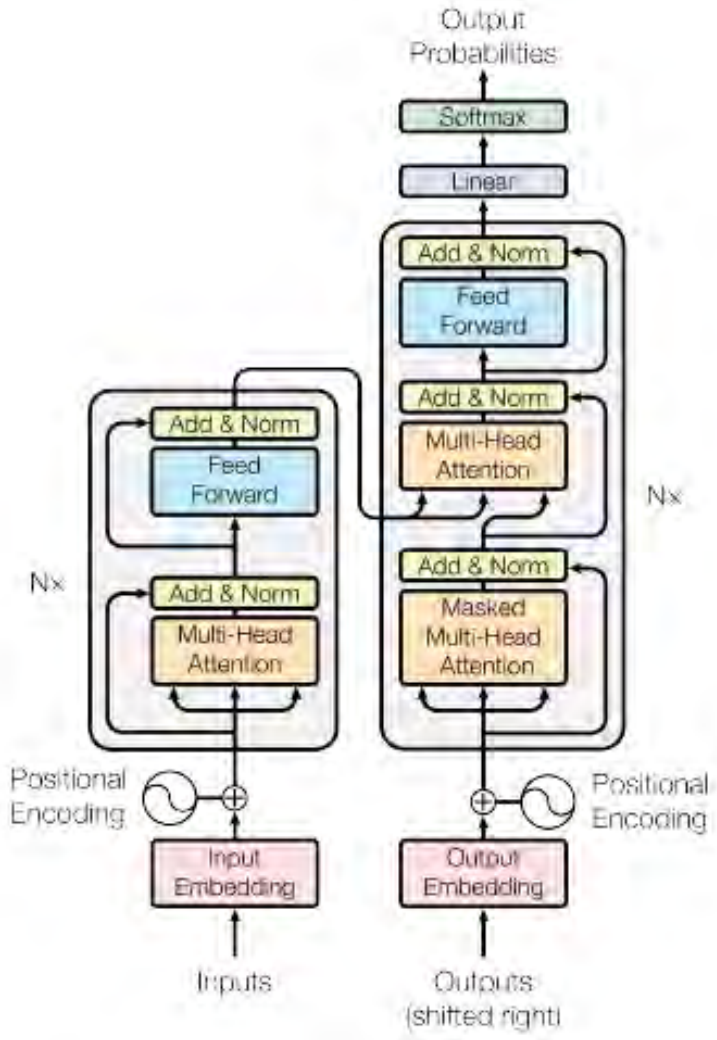
transformer αποτελείται από δύο βασικά μέρη, τον κωδικοποιητή (encoder) και τον αποκωδικοποιητή (decoder).

1. *Ο κωδικοποιητής* λαμβάνει την είσοδο – για παράδειγμα μια πρόταση – και τη μετατρέπει σε μια σειρά από αναπαραστάσεις υψηλού επιπέδου. Η αρχιτεκτονική του κάθε κωδικοποιητή περιλαμβάνει πολλαπλά επίπεδα, το καθένα από τα οποία συνίσταται από τα εξής στοιχεία:

    - Μηχανισμός Αυτο-Προσοχής (Self-Attention Mechanism): Ο μηχανισμός αυτο-προσοχής επιτρέπει στο μοντέλο να επικεντρώνεται σε κρίσιμα μέρη της εισόδου, αναλύοντας τη σχέση κάθε λέξης με όλες τις άλλες στην ακολουθία. Με αυτόν τον τρόπο, ενισχύεται η κατανόηση των εξαρτήσεων εντός της πρότασης.
    - Πλήρως Συνδεδεμένο Δίκτυο (Feed-Forward Neural Network): Μετά τη φάση της προσοχής, το δίκτυο εφαρμόζει μη γραμμικούς μετασχηματισμούς στα δεδομένα για να δημιουργήσει πιο εκφραστικές αναπαραστάσεις.
    - Κανονικοποίηση και Υπόλοιπο (Add & Norm): Κάθε υπομονάδα (π.χ., προσοχή ή πλήρως συνδεδεμένο δίκτυο) συνοδεύεται από ένα υπόλοιπο (residual connection) που διευκολύνει τη ροή πληροφορίας και σταθεροποιεί την εκπαίδευση. Η κανονικοποίηση βοηθά στη διατήρηση της ομοιόμορφης κατανομής των τιμών κατά την εκπαίδευση.

2. *Ο αποκωδικοποιητής* χρησιμοποιεί την έξοδο του κωδικοποιητή για να παράγει την τελική έξοδο (π.χ. μια πρόβλεψη λέξης). Η κύρια διαφορά του αποκωδικοποιητή από τον κωδικοποιητή είναι ότι περιλαμβάνει έναν επιπλέον μηχανισμό προσοχής, το μηχανισμό προσοχής με μάσκα (Masked Multi-Head Attention), ο οποίος βοηθά στη διαχείριση της ακολουθιακής δημιουργίας κειμένου.

    - Μηχανισμός Αυτο-Προσοχής με Μάσκα (Masked Multi-Head Self-Attention): Εφαρμόζεται για να διασφαλιστεί ότι, κατά την παραγωγή κάθε λέξης, το μοντέλο δεν "βλέπει" τις μελλοντικές λέξεις, αφού επιβάλλεται μια μάσκα που αποκλείει την πρόσβαση σε αυτές.
    - Μηχανισμός Προσοχής Κωδικοποιητή-Αποκωδικοποιητή (Encoder-Decoder Attention): Σε αυτό το στάδιο, ο αποκωδικοποιητής "αναζητά" τις σχετικές πληροφορίες στην έξοδο του κωδικοποιητή, επιτρέποντάς του να εστιάζει στα κρίσιμα τμήματα της εισόδου για την παραγωγή της σωστής πρόβλεψης.
    - Πλήρως Συνδεδεμένο Δίκτυο (Feed-Forward Neural Network): Μετά τους μηχανισμούς προσοχής, το αποτέλεσμα περνάει από ένα πλήρως συνδεδεμένο δίκτυο που εφαρμόζει μη γραμμικούς μετασχηματισμούς, βελτιώνοντας την ικανότητα του μοντέλου να συλλαμβάνει σύνθετες σχέσεις.
    - Κανονικοποίηση και Υπόλοιπο (Add & Norm): Κάθε υπομονάδα του αποκωδικοποιητή συνοδεύεται από μια σύνδεση υπόλοιπου και εφαρμογή κανονικοποίησης, βοηθώντας στη σταθεροποίηση της εκπαίδευσης και στη συνεχή ροή της πληροφορίας.
    - Τελικό Γραμμικό Επίπεδο και Softmax: Στο τελευταίο βήμα, οι τελικές αναπαραστάσεις μετασχηματίζονται μέσω ενός γραμμικού επιπέδου και της συνάρτησης softmax σε πιθανότητες για κάθε πιθανή έξοδο (π.χ. επόμενη λέξη), ολοκληρώνοντας έτσι τη διαδικασία παραγωγής.



Σχήμα 6.3: Αρχιτεκτονική ενός transformer

Η αρχιτεκτονική ενός transformer όπως αρχικά παρουσιάστηκε στην εργασία που τους εισήγαγε [280] παρουσιάζεται στο Σχήμα 6.3.

BERT Το BERT (Bidirectional Encoder Representations from Transformers) [65] αποτελεί ένα από τα πιο προηγμένα μοντέλα αναπαράστασης γλώσσας, το οποίο αξιοποιεί την αρχιτεκτονική των Transformers [280]. Ως προεκπαιδευμένο μοντέλο, το BERT είναι σε θέση να συλλαμβάνει τις συμφραζόμενες σχέσεις και τα νοήματα των λέξεων μέσα σε προτάσεις και μπορεί να προσαρμοστεί (fine-tuning) σε εξειδικευμένες εργασίες NLP, όπως η ταξινόμηση κειμένου, η απάντηση σε ερωτήσεις κ.ά. Επιπλέον, ο όρος αμφίδρομος υποδηλώνει την ικανότητα του μοντέλου να λαμβάνει υπόψη τόσο το αριστερό όσο και το δεξί συμφραζόμενο μιας λέξης κατά τον καθορισμό του νοήματός της, οδηγώντας έτσι σε πιο ακριβείς συμφραζόμενες αναπαραστάσεις. Υπάρχουν δύο κύριες παραλλαγές του μοντέλου BERT: το Base, το οποίο χαρακτηρίζεται από 12 στρώματα, 768 κρυφές καταστάσεις, 12 μηχανισμούς προσοχής και συνολικά 110 εκατομμύρια παραμέτρους, και το Large, το οποίο διαθέτει 24 στρώματα, 1024 κρυφές καταστάσεις, 16 μηχανισμούς προσοχής και συνολικά 340 εκατομμύρια παραμέτρους. Το μοντέλο BERT Large παρουσιάζει βελτιωμένες δυνατότητες ανίχνευσης σύνθετων γλωσσικών σχέσεων, αν και απαιτεί περισσότερους υπολογιστικούς πόρους.

Το BERT υφίσταται μια διαδικασία εκπαίδευσης σε δύο στάδια. Στο πρώτο στάδιο, υποβάλλεται σε προεκπαίδευση μέσω ενός πλαισίου διπλής εργασίας που περιλαμβάνει το Masked Language Modeling και την πρόβλεψη της επόμενης πρότασης (Next Sentence Prediction). Αυτή η φάση αξιοποιεί τον πλούτο των γραπτών πόρων, όπως το BooksCorpus και την Αγγλική Wikipedia, με στόχο την απόκτηση μιας στιβαρής κατανόησης των γενικών γλωσσικών προτύπων και δομών. Στη συνέχεια, το μοντέλο προσαρμόζεται σε συγκεκριμένες εργασίες NLP μέσω fine-tuning, όπου προστίθενται ειδικά στρώματα και ενημερώνονται τα προεκπαιδευμένα βάρη ώστε να ανταποκρίνονται στις ιδιαιτερότητες της εργασίας.

Επιπρόσθετα, το BERT έχει επιδείξει εξαιρετικές επιδόσεις σε διάφορα πρότυπα δοκιμών NLP (π.χ. GLUE, SQuAD v1.1, SQuAD v2.0, SWAG) [65]. Παρόλο που δεν έχει σχεδιαστεί αποκλειστικά για οικονομικά κείμενα, έχει χρησιμοποιηθεί με επιτυχία σε έρευνες χρηματοοικονομικής ανάλυσης συναισθήματος [143].

RoBERTa Το RoBERTa (A Robustly Optimized BERT Pretraining Approach) [151] αποτελεί μια εξέλιξη του αρχικού μοντέλου BERT, η οποία βασίζεται σε σειρά λεπτομερών τροποποιήσεων όπως η επιμήκυνση της διάρκειας εκπαίδευσης, η χρήση ενός διευρυμένου συνόλου δεδομένων, η αύξηση του μεγέθους των batches και η ενσωμάτωση μεγαλύτερων ακολουθιών, καθώς και η απόρριψη της διεργασίας Next Sentence Prediction υπέρ της δυναμικής μάσκας. Οι συγκεκριμένες βελτιώσεις δεν επηρεάζουν τις ταξινομικές ικανότητες του μοντέλου· αντιθέτως, το RoBERTa επιδεικνύει ανώτερες επιδόσεις σε σημαντικά πρότυπα δοκιμών NLP, όπως τα SQuAD v1.1, SQuAD v2.0, GLUE, RACE, MNLI-m και SST-2. Ένα αξιοσημείωτο χαρακτηριστικό του RoBERTa είναι η πρόσβαση σε ένα εξαιρετικά διευρυμένο σύνολο κειμένων (π.χ. CC-News, Open WebText, STORIES) που συγκεντρώνει συνολικά 160 ΓΒ δεδομένων [151]. Όπως και ο BERT, το RoBERTa δεν έχει σχεδιαστεί εξειδικευμένα για οικονομικά κείμενα, ωστόσο έχει χρησιμοποιηθεί με εξαιρετικά θετικά αποτελέσματα σε εφαρμογές χρηματοοικονομικής ανάλυσης συναισθήματος [34, 255, 106].

Μοντέλα FinBERT Τα μοντέλα FinBERT είναι εξειδικευμένα μοντέλα αναπαράστασης γλώσσας που έχουν σχεδιαστεί για εφαρμογές στον χρηματοοικονομικό τομέα και βασίζονται στην αρχιτεκτονική του BERT. Στόχος τους είναι η ανάλυση και κατανόηση κειμένων που αφορούν χρηματοοικονομικές αγορές, μετοχές, επενδύσεις και οικονομικές ειδήσεις, με σκοπό την αξιολόγηση του συναισθήματος και την εξαγωγή πολύτιμων πληροφοριών. Η τυπική διαδικασία εκπαίδευσής τους περιλαμβάνει μια αρχική φάση προεκπαίδευσης, κατά την οποία το μοντέλο αποκτά γενικές γλωσσικές αναπαραστάσεις μέσω της επεξεργασίας ενός ευρέος συνόλου χρηματοοικονομικών κειμένων. Στη συνέχεια, το μοντέλο υποβάλλεται σε fine-tuning σε συγκεκριμένες εργασίες ανάλυσης συναισθήματος χρησιμοποιώντας ετικετοποιημένα σύνολα δεδομένων.

Σημειώνεται ότι ο όρος FinBERT περιλαμβάνει διάφορες διακριτές παραλλαγές, τρεις από τις οποίες χρησιμοποιούνται στην παρούσα εργασία. Συνολικά, στο πειραματικό πλαίσιο χρησιμοπιούνται πέντε μέθοδοι - παραλλαγές του BERT:

1. BERT
2. RoBERTa [151]
3. FinBERT Araci Model [15]
4. FinBERT (Huang et al. Model) [123]

5. FinancialBERT [113]

Η υλοποίηση των παραπάνω μοντέλων είναι διαθέσιμη μέσω των αποθετηρίων του Hugging Face AI [215, 114].

Για την εκπαίδευση των παραπάνω προεκπαιδευμένων μοντέλων χρησιμοποιήθηκε ο βελτιστοποιητής ADAM [136].

6.2 Αποτελέσματα

Στην παρούσα ενότητα, παρουσιάζονται τα αποτελέσματα της πειραματικής διαδικασίας που πραγματοποιήθηκε στο πλαίσιο της μελέτης. Ο κύριος στόχος είναι να αξιολογηθεί η αποτελεσματικότητα τόσο των κλασικών μεθόδων Μηχανικής Μάθησης όσο και ομαδοποιητών αυτών, αλλά και σύγχρονων προεκπαιδευμένων μοντέλων Βαθιάς Μάθησης στον τομέα της οικονομικής ανάλυσης συναισθήματος. Επιπλέον, εξετάζεται η απόδοση διαφορετικών γλωσσικών αναπαραστάσεων, όπως οι τεχνικές BoW και TF-IDF, αλλά και η χρήση της τεχνικής SMOTE για την αντιμετώπιση του προβλήματος της ανισοκατανομής δεδομένων.

Η ενότητα αυτή αποσκοπεί στο να αναδείξει τα δυνατά και αδύνατα σημεία κάθε προσέγγισης, να εξετάσει τη σημασία της επιλογής της κατάλληλης μεθοδολογίας και να παρέχει εμπεριστατωμένα συμπεράσματα για τη χρήση τους στον χρηματοοικονομικό τομέα. Τα αποτελέσματα παρουσιάζονται από τη σκοπιά έξι διαφορετικών μετρικών, κλασσικές για την αξιολόγηση προβλημάτων ταξινόμησης, ενώ για την ερμηνεία τους λαμβάνονται υπόψη οι ιδιαιτερότητες του εξεταζόμενου συνόλου δεδομένων.

6.2.1 Κλασσικές μέθοδοι ταξινόμησης + BoW + SMOTE

Στην αρχική πειραματική διάταξη, ο CatBoost υπερέχει έναντι όλων των άλλων μεθόδων, τόσο στα μεμονωμένα μοντέλα όσο και στους ομαδοποιητές ταξινομητών. Συγκεκριμένα, ο CatBoost σημειώνει ακρίβεια 63.71% και AUC 79.08%, τα οποία είναι τα υψηλότερα μεταξύ των μεμονωμένων αλγορίθμων. Η μόνη εξαίρεση αφορά τη μετρική AUC, όπου οι συνδυασμοί CatBoost+LR+XGBoost (AUC 79.51%) και CatBoost+XGBoost+SVM (AUC 79.33%) ξεπερνούν ελαφρώς την απόδοση του CatBoost, αν και με μικρή διαφορά. Οι μέθοδοι LR και XGBoost ακολουθούν τον CatBoost με ακρίβεια 62.15% και 61.37% αντίστοιχα, καταλαμβάνοντας τη δεύτερη και τρίτη θέση μεταξύ των μεμονωμένων αλγορίθμων.

Σχετικά με τους ομαδοποιητές ταξινομητών, ο συνδυασμός CatBoost+LR+XGBoost ξεπερνά άλλες συνδυαστικές μεθόδους, σημειώνοντας τις υψηλότερες επιδόσεις στις περισσότερες μετρικές, με εξαίρεση την precision και το F1-score. Πολύ κοντά ακολουθεί ο συνδυασμός CatBoost+XGBoost+SVM, ο οποίος καταλαμβάνει τη δεύτερη καλύτερη θέση στην κατηγορία των ομαδοποιητών, καθώς έπεται του συνδυασμού CatBoost+LR+XGBoost σε όλες τις μετρικές, εκτός των precision και AUC όπου και υπερτερεί όλων των ομαδοποιητών. Οι υπόλοιποι ομαδοποιητές καταγράφουν επίσης γενικά καλύτερες επιδόσεις σε σύγκριση με το σύνολο των μεμονωμένων μοντέλων, ειδικά όταν συνδυάζονται αλγόριθμοι όπως οι CatBoost, XGBoost, και SVM. Αναλυτικά, τα αποτελέσματα για το σύνολο των μεμονωμένων μεθόδων, καθώς και των ομαδοποιητών τους, στο σενάριο αυτό παρουσιάζονται στον Πίνακα 6.3

Μέθοδος	Accuracy	AUC	Recall	Precision	F1	Καππα	MCC
Κλασσικές Μέθοδοι Μηχανικής Μάθησης							
CatBoost	0.6371	0.7908	0.6371	0.6684	0.6481	0.4074	0.411
LR	0.6215	0.7816	0.6215	0.6422	0.6296	0.3785	0.3804
XGBoost	0.6137	0.7847	0.6137	0.6473	0.6263	0.3723	0.3758
LightGBM	0.6101	0.7721	0.6101	0.6336	0.6193	0.36	0.362
SVM	0.6045	0.7584	0.6045	0.642	0.6187	0.3549	0.3585
GBC	0.5705	0.763	0.5705	0.6391	0.5905	0.3302	0.3415
ET	0.6022	0.6807	0.6022	0.6142	0.6073	0.333	0.3337
Ridge	0.581	0.7433	0.581	0.6285	0.5985	0.3273	0.3323
RF	0.5754	0.7149	0.5754	0.5961	0.5845	0.2971	0.2981
AdaBoost	0.5125	0.6881	0.5125	0.664	0.5424	0.2678	0.2958
NB	0.5072	0.6286	0.5072	0.5687	0.5249	0.2348	0.2424
DT	0.5115	0.6199	0.5115	0.5489	0.5266	0.2128	0.2151
KNN	0.2833	0.5537	0.2833	0.6057	0.2519	0.0884	0.1317
LDA	0.268	0.357	0.268	0.2973	0.279	0.0754	0.077
QDA	0.4937	0.5191	0.4937	0.4987	0.4354	0.0487	0.0573
DC	0.1472	0.5	0.1472	0.0217	0.0378	0	0
Ομαδοποιητές κλασσικών μεθόδων							
CatBoost+LR+XGBoost	0.6326	0.7951	0.6326	0.6581	0.6422	0.3983	0.4008
CatBoost+XGBoost+SVM	0.6317	0.7933	0.6317	0.6629	0.6433	0.3968	0.4
CatBoost+XGBoost+Ridge	0.6257	0.7889	0.6257	0.66	0.6386	0.3905	0.3942
CatBoost+LR+SVM	0.6244	0.7896	0.6244	0.6514	0.6349	0.3826	0.385
LR+XGBoost+SVM	0.6221	0.7911	0.6221	0.6507	0.6332	0.38	0.3826
LR+XGBoost+Ridge	0.6142	0.7862	0.6142	0.6442	0.626	0.3698	0.3724
CatBoost+LR+Ridge	0.6142	0.7845	0.6142	0.6432	0.6256	0.3694	0.372
CatBoost+SVM+Ridge	0.6105	0.7811	0.6105	0.6434	0.6233	0.364	0.367
XGBoost+SVM+Ridge	0.6088	0.7832	0.6088	0.6425	0.6218	0.3616	0.3646
LR+SVM+Rigde	0.6054	0.7772	0.6054	0.6369	0.6178	0.3544	0.357

Πίνακας 6.3: Κλασσικές μέθοδοι + BoW + SMOTE

6.2.2 Κλασσικές μέθοδοι ταξινόμησης + TF-IDF + SMOTE

Σχετικά με το δεύτερο σενάριο, στο οποίο χρησιμοποιούνται κλασικές μέθοδοι Μηχανικής Μάθησης αλλά πλέον συνδυασμό με τις τεχνικές TF-IDF και SMOTE, ο συνδυασμός CatBoost+LR+SVM καταλαμβάνει την πρώτη θέση στις μετρικές AUC (83.65%), precision (70.14%), και F1-score (69.20%). Ο ίδιος συνδυασμός έρχεται δεύτερος σε όλες τις υπόλοιπες μετρικές, με μικρή διαφορά από τα κορυφαία μεμονωμένα μοντέλα. Εν τω μεταξύ, το μεμονωμένο μοντέλο CatBoost σημειώνει την υψηλότερη ακρίβεια (68.82%) και recall (68.82%) μεταξύ όλων των μεθόδων.

Αξίζει να σημειωθεί ότι σε μη εξισορροπημένα σύνολα δεδομένων, όπως αυτό που εξετάζεται εδώ, οι μετρικές recall, precision και F1-score παρέχουν καλύτερη εικόνα για τα αποτελέσματα από οτι η ακρίβεια. Η precision εκφράζει το ποσοστό των θετικών προβλέψεων που είναι πράγματι σωστές, ενώ η recall καταδεικνύει πόσα από τα πραγματικά θετικά παραδείγματα ανιχνεύθηκαν σωστά. Η F1-score, ως αρμονικός μέσος των precision και recall, προσφέρει ισορροπημένη αξιολόγηση μεταξύ των δύο.

Μεταξύ των μεμονωμένων μεθόδων, τα CatBoost και LR επιδεικνύουν καλύτερες επιδόσεις σε σύγκριση με τα υπόλοιπα. Στους ομαδοποιητές, ο συνδυασμός CatBoost+LR+SVM μαζί με τον συνδυασμό CatBoost+LR+Ridge παρουσιάζουν τις υψηλότερες επιδόσεις, με precision 70.14% και F1-score 69.02% για τον πρώτο και precision 69.14% για τον δεύτερο.

Συνολικά, οι ομαδοποιητές αποδεικνύονται αποδοτικότεροι από τα μεμονωμένα μοντέλα, με μικρές βελτιώσεις σε όλες τις μετρικές. Αυτή η τάση επιβεβαιώνει τη σημασία των συνδυασμών ταξινομητών, ειδικά σε σενάρια με ανισόρροπα δεδομένα. Τα αποτελέσματα που αφορούν όλες τις μετρικές παρουσιάζονται στον Πίνακα 6.4.

6.2.3 Κλασσικές μέθοδοι ταξινόμησης + BoW

Στο συγκεκριμένο σενάριο, ο μεμονωμένος ταξινομητής CatBoost καταγράφει ανώτερες επιδόσεις σε σχέση με άλλες μεμονωμένες και συνδυαστικές μεθόδους στις μετρικές ακρίβειας (69.31%), precision (68.90%), recall (69.31%) και MCC (45.42%). Ωστόσο, σε ό,τι αφορά τις μετρικές AUC (83.39%), F1-score (66.57%) και Kappa (42.71%), το CatBoost καταλαμβάνει τη δεύτερη θέση, ενώ αντίστοιχα στην κορυφή της κατάταξης για την μετρική AUC συναντάμε τον ομαδοποιητή CatBoost+XGBoost+SVM (0.8363%) και για τις μετρικές F1-score και Kappa τον CatBoost+XGBoost+Ridge (με ποσοστά 0.666% και 0.4275% αντίστοιχα).

Μεταξύ των ομαδοποιητών μεθόδων, ο συνδυασμός CatBoost+
XGBoost+SVM εμφανίζεται να υπερέχει, όχι μόνο σε σχέση με άλλες ομαδοποιήσεις αλλά και έναντι της πλειονότητας των μεμονωμένων μοντέλων. Ένας εναλλακτικός συνδυασμός, ο CatBoost+XGBoost+Ridge, παρουσιάζει επίσης υψηλές επιδόσεις. Όπως ήδη σχολιάστηκε, οι δύο αυτοί ομαδοποιητές είναι οι μόνοι που ξεπερνούν τον CatBoost σε κάποιες από τις μετρικές.

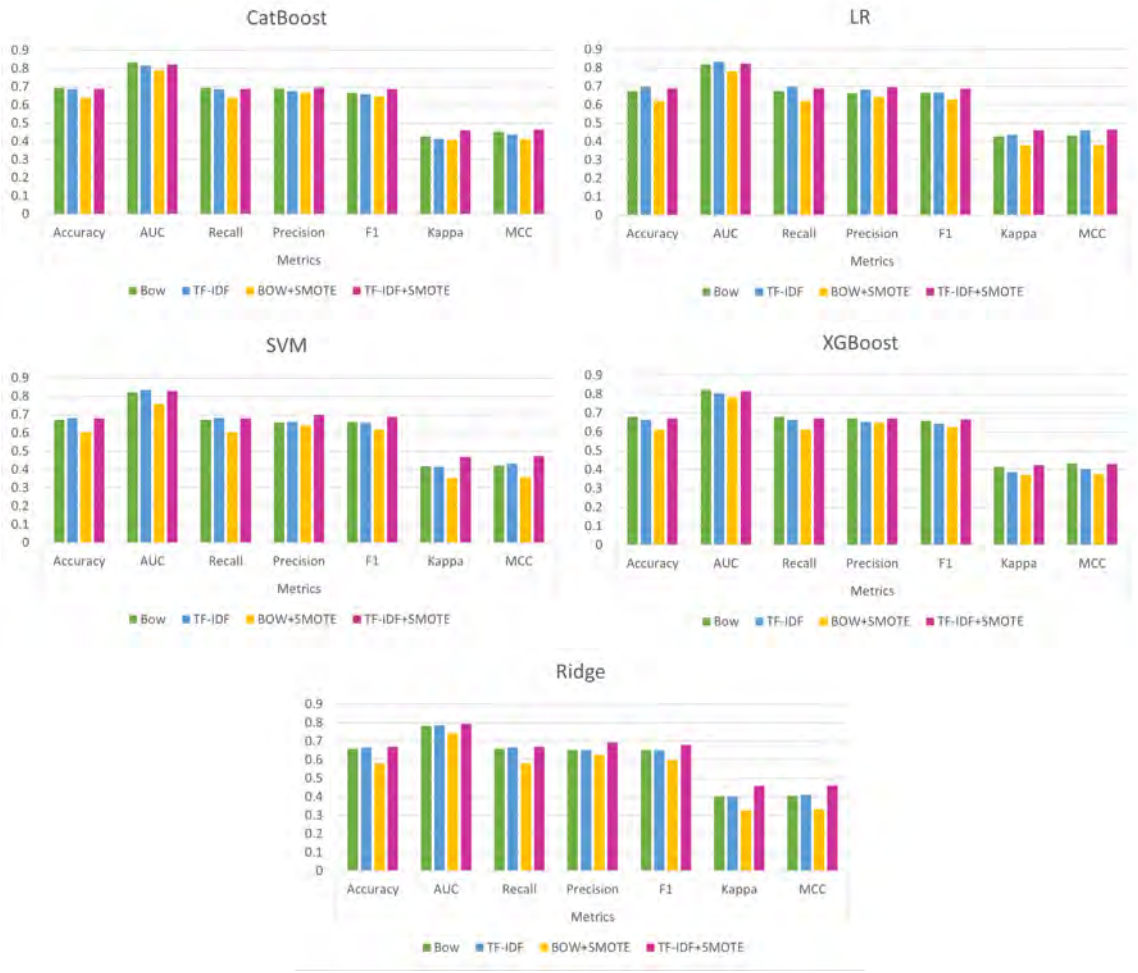
Αν και οι ομαδοποιητές εν γένει έχουν καλύτερη απόδοση, είναι σημαντικό να σημειωθεί ότι ο μεμονωμένος ταξινομητής CatBoost επιτυγχάνει σταθερά τις υψηλότερες τιμές στις περισσότερες μετρικές, γεγονός που υποδεικνύει την ισχύ του. Αναλυτικά, τα αποτελέσματα για όλες τις μετρικές παρουσιάζονται στον Πίνακα 6.5.

Μοντέλο	**Accuracy**	**AUC**	**Recall**	**Precision**	**F1**	**Kappa**	**MCC**
Κλασσικές Μέθοδοι Μηχανικής Μάθησης							
LR	0.6826	0.8285	0.6826	0.6986	0.6887	0.4727	0.4741
SVM	0.6794	0.8294	0.6794	0.6987	0.6862	0.4704	0.4726
CatBoost	0.6882	0.8223	0.6882	0.6941	0.6865	0.4619	0.4658
Ridge	0.6713	0.7947	0.6713	0.6938	0.6794	0.4594	0.462
XGBoost	0.6715	0.8155	0.6715	0.6723	0.666	0.4242	0.4294
LightGDM	0.6638	0.8121	0.6638	0.6623	0.6595	0.4155	0.4186
GBC	0.6629	0.791	0.6629	0.6733	0.6549	0.4048	0.4151
ET	0.6492	0.7123	0.6492	0.6472	0.6435	0.3861	0.3903
RF	0.645	0.7727	0.645	0.6433	0.6369	0.3732	0.3794
AdaBoost	0.6306	0.6746	0.6306	0.6384	0.6153	0.3435	0.3568
DT	0.5927	0.6642	0.5927	0.6038	0.5974	0.3162	0.3167
KNN	0.3775	0.6447	0.3775	0.662	0.3636	0.1744	0.2276
NB	0.4946	0.6164	0.4946	0.5533	0.5129	0.2118	0.218
LDA	0.4256	0.5844	0.4256	0.4749	0.4376	0.1074	0.1115
DC	0.1472	0.5	0.1472	0.0217	0.0378	0	0
QDA	0.4868	0.4924	0.4868	0.5209	0.3978	-0.0082	-0.0104
Ομαδοποιητές Κλασσικών Μεθόδων Μηχανικής Μάθησης							
CatBoost+LR+Ridge	0.6839	0.8346	0.6839	0.7014	0.6902	0.4753	0.4773
CatBoost+LR+SVM	0.688	0.8365	0.688	0.7014	0.692	0.4753	0.4772
CatBoost+XGBoost+Ridge	0.6871	0.8335	0.6871	0.701	0.6918	0.4754	0.4771
CatBoost+SVM+Ridge	0.6831	0.8335	0.6831	0.7014	0.6896	0.4748	0.4769
LR+XGBoost+SVM	0.6865	0.8357	0.6865	0.6979	0.6894	0.47	0.4719
LR+XGBoost+Ridge	0.6814	0.8346	0.6814	0.698	0.6873	0.47	0.4719
LR+SVM+Ridge	0.6775	0.8298	0.6775	0.6976	0.6848	0.4683	0.4706
XGBoost+SVM+Ridge	0.6786	0.8339	0.6786	0.696	0.6848	0.4661	0.4681
CatBoost+XGBoost+SVM	0.6854	0.8345	0.6854	0.6937	0.6862	0.4621	0.4646
CatBoost+LR+XGBoost	0.6854	0.8354	0.6854	0.6926	0.6855	0.4603	0.463

Πίνακας 6.4: Κλασσικές μέθοδοι + TF-IDF + SMOTE

Μέθοδοι		**AUC**	**Recall**	**Precision**	**F1**	**Kappa**	**MCC**
Κλασσικές Μέθοδοι Μηχανικής Μάθησης							
CatBoost	0.6931	0.8339	0.6931	0.689	0.6657	0.4271	0.4542
XGBoost	0.6818	0.8232	0.6818	0.6722	0.6594	0.4148	0.4337
LR	0.6741	0.8194	0.6741	0.6634	0.6648	0.4273	0.4314
SVM	0.6711	0.8204	0.6711	0.6558	0.6588	0.4177	0.4224
GBC	0.6702	0.8028	0.6702	0.6837	0.6321	0.3671	0.413
Ridge	0.6591	0.7832	0.6591	0.6519	0.6516	0.4014	0.4052
AdaBoost	0.6606	0.6788	0.6606	0.6699	0.6295	0.3604	0.3953
ET	0.6535	0.717	0.6535	0.6356	0.6406	0.3879	0.3921
LightGBM	0.6542	0.7933	0.6542	0.6376	0.6382	0.3805	0.388
RF	0.6448	0.785	0.6448	0.6233	0.6262	0.3607	0.3692
DT	0.6137	0.6807	0.6137	0.6165	0.6143	0.3444	0.345
NB	0.5191	0.6365	0.5191	0.5769	0.5363	0.2498	0.2569
KNN	0.5716	0.6558	0.5716	0.5681	0.5063	0.1605	0.2001
LDA	0.4089	0.5747	0.4089	0.4522	0.4212	0.073	0.0751
DC	0.5358	0.5	0.5358	0.2871	0.3739	0	0
QDA	0.233	0.4855	0.233	0.3782	0.2423	-0.0237	-0.0313
Ομαδοποιητές Κλασικών Μεθόδων Μηχανικής Μάθησης							
CatBoost+XGBoost+SVM	0.6865	0.8363	0.6865	0.673	0.6644	0.425	0.4419
CatBoost+XGBoost+Ridge	0.685	0.8324	0.685	0.6714	0.666	0.4275	0.4409
CatBoost+LR+SVM	0.6816	0.8303	0.6816	0.6678	0.6664	0.4284	0.4373
CatBoost+LR+XGBoost	0.6831	0.8347	0.6831	0.6695	0.6627	0.4215	0.4366
CatBoost+SVM+Ridge	0.6799	0.8282	0.6799	0.6654	0.6647	0.4258	0.4342
LR+XGBoost+SVM	0.6792	0.8282	0.6792	0.6651	0.6647	0.426	0.4339
XGBoost+SVM+Ridge	0.679	0.8271	0.679	0.6656	0.665	0.426	0.4336
CatBoost+LR+Ridge	0.6773	0.828	0.6773	0.664	0.6637	0.4236	0.4309
LR+XGBoost+Ridge	0.6737	0.8266	0.6737	0.6617	0.6616	0.4195	0.4259
LR+SVM+Ridge	0.6696	0.8187	0.6696	0.6564	0.6588	0.4166	0.421

Πίνακας 6.5: Κλασσικές μέθοδοι + BoW



Σχήμα 6.4: Απόδοση προεκπαιδευμένων μοντέλων

6.2.4 Κλασσικές μέθοδοι ταξινόμησης + TF-IDF

Στο συγκεκριμένο σενάριο, ο μεμονωμένος ταξινομητής LR καταλαμβάνει την κορυφαία θέση στις περισσότερες μετρικές, σημειώνοντας ακρίβεια 69.74%, precision 68.20%, recall 69.74%, και MCC 46.08%. Ωστόσο, στις μετρικές AUC (83.38%) και F1-score (66.71%), ο LR έρχεται στη δεύτερη θέση. Μεταξύ των μεμονωμένων μεθόδων, οι CatBoost και Ridge ακολουθούν το LR, με τον CatBoost να σημειώνει ακρίβεια 68.46% και το Ridge 66.66%.

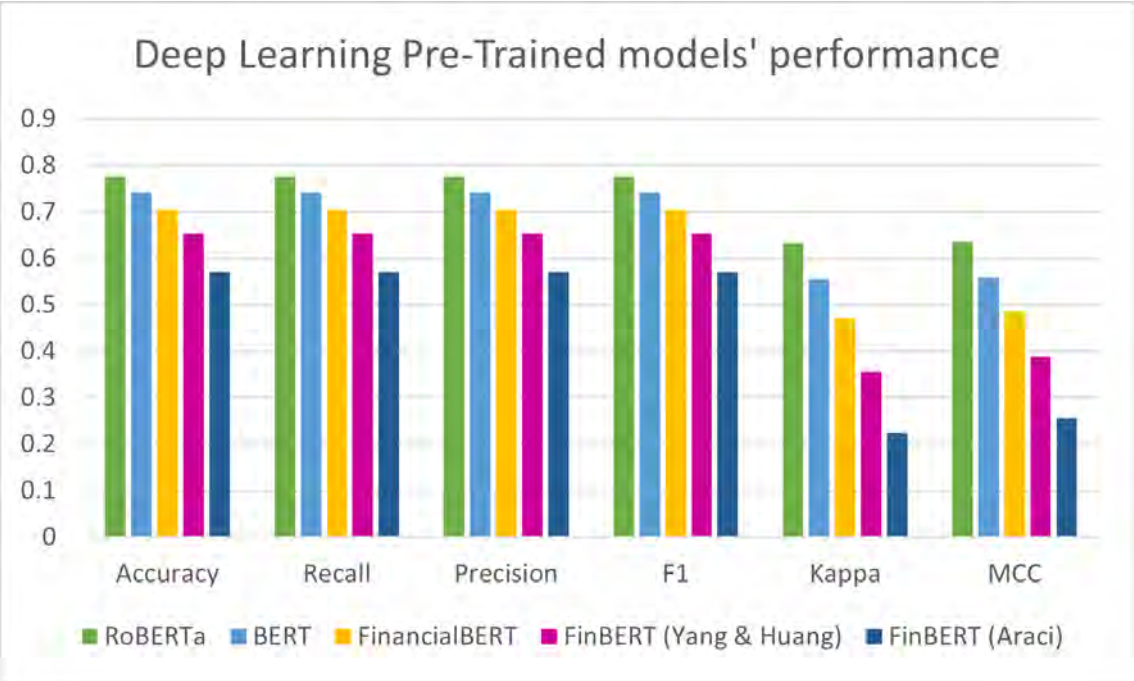
Στους ομαδοποιητές, ο συνδυασμός CatBoost+LR+SVM ξεχωρίζει, καταγράφοντας τις υψηλότερες επιδόσεις στις περισσότερες μετρικές, όπως AUC 84.08%, precision 67.00%, και F1-score 65.88%. Ο συγκεκριμένος συνδυασμός αποδεικνύεται εξαιρετικά αποδοτικός, ξεπερνώντας όλους τους άλλους ομαδοποιητές.

Παρόλο που ο LR αναδεικνύεται σε σημαντικό ανταγωνιστή στο συγκεκριμένο σενάριο, οι ομαδοποιητές συνεχίζουν να παρουσιάζουν συνολικά καλύτερες επιδόσεις σε σχέση με τα μεμονωμένα μοντέλα. Αξιοσημείωτο είναι ότι μόνο λίγες μεμονωμένες μέθοδοι καταφέρνουν να ξεπεράσουν ακόμη και τους ασθενέστερους ομαδοποιητές, γεγονός που υπογραμμίζει τη σημασία της χρήσης συνδυαστικών ταξινομητών. Αναλυτικά, τα αποτελέσματα για όλες τις μετρικές παρουσιάζονται στον Πίνακα 6.6.

Για καλύτερη κατανόηση των αποτελεσμάτων, στο Σχήμα 6.4 παρουσιάζονται γραφικά τα αποτελέσματα των πέντε κορυφαίων μεθόδων (σύμφωνα με το MCC).

Μέθοδοι	**Accuracy**	**AUC**	**Recall**	**Precision**	**F1**	**Kappa**	**MCC**
Κλασσικές Μέθοδοι Μηχανικής Μάθησης							
LR	0.6974	0.8338	0.6974	0.682	0.6671	0.4375	0.4608
CatBoost	0.6846	0.8168	0.6846	0.6764	0.6582	0.4138	0.4374
SVM	0.682	0.8347	0.682	0.6598	0.6546	0.415	0.4314
Ridge	0.6666	0.7863	0.6666	0.6524	0.651	0.402	0.4108
XGBoost	0.6647	0.806	0.6647	0.6537	0.6443	0.3874	0.4031
ET	0.6578	0.7096	0.6578	0.6387	0.6424	0.3899	0.3963
GBC	0.6604	0.7937	0.6604	0.6653	0.6202	0.3484	0.3923
AdaBoost	0.6514	0.6774	0.6514	0.6609	0.6212	0.3449	0.3778
RF	0.6463	0.7702	0.6463	0.6281	0.6281	0.3626	0.372
KNN	0.6306	0.7701	0.6306	0.6278	0.6233	0.3541	0.3587
LightGBM	0.6364	0.7772	0.6364	0.6211	0.6213	0.35	0.3572
DT	0.5878	0.6645	0.5878	0.6016	0.5939	0.3103	0.3109
NB	0.5076	0.6252	0.5076	0.5639	0.5256	0.2288	0.2349
LDA	0.2429	0.3461	0.2429	0.2736	0.2521	0.0447	0.046
DC	0.5358	0.5	0.5358	0.2871	0.3739	0	0
QDA	0.2093	0.4684	0.2093	0.3587	0.2267	-0.0534	-0.0716
Ομαδοποιητές Κλασικών Μεθόδων Μηχανικής Μάθησης							
CatBoost+LR+SVM	0.6882	0.8408	0.6882	0.67	0.6588	0.4216	0.4432
CatBoost+LR+Ridge	0.6837	0.837	0.6837	0.6638	0.6595	0.4216	0.4369
LR+XGBoost+SVM	0.6837	0.8372	0.6837	0.6647	0.6575	0.4177	0.436
CatBoost+SVM+Ridge	0.682	0.8362	0.682	0.6611	0.6581	0.4192	0.4338
CatBoost+XGBoost+SVM	0.682	0.8349	0.682	0.6655	0.6562	0.4134	0.4328
LR+XGBoost+Ridge	0.6801	0.8341	0.6801	0.6591	0.6578	0.4181	0.4312
CatBoost+LR+XGBoost	0.6809	0.8351	0.6809	0.6635	0.6543	0.4105	0.4306
CatBoost+XGBoost+Ridge	0.679	0.831	0.679	0.6605	0.6566	0.4146	0.429
LR+SVM+Ridge	0.6784	0.8338	0.6784	0.6561	0.6557	0.4155	0.4277
XGBoost+SVM+Ridge	0.6771	0.8338	0.6771	0.6562	0.6552	0.4132	0.4258

Πίνακας 6.6: Κλασσικές μέθοδοι + TF-IDF



Σχήμα 6.5: Απόδοση προεκπαιδευμένων μοντέλων

6.2.5 Deep learning pre-trained models

Σχετικά με τα προεκπαιδευμένα μοντέλα βαθιάς μάθησης (Deep Learning Pre-Trained Models), το RoBERTa εμφανίζει εξαιρετική απόδοση, ξεπερνώντας όλα τα υπόλοιπα μοντέλα σε όλες τις μετρικές. Συγκεκριμένα, σημειώνει ακρίβεια 77.59%, precision 77.59%, recall 77.59%, F1-score 77.59%, και MCC 63.49%, καθιστώντας το τον απόλυτο νικητή στο συγκεκριμένο σενάριο. Το BERT ακολουθεί στη δεύτερη θέση, με ακρίβεια 74.16% και F1-score 74.16%, ενώ τα FinancialBERT και FinBERT (Yang & Hung) καταλαμβάνουν την τρίτη και τέταρτη θέση, αντίστοιχα, με χαμηλότερες επιδόσεις. Το FinBERT (Araci) κατατάσσεται τελευταίο σε όλες τις μετρικές, με ακρίβεια 57.05% και MCC 25.60%.

Η αδιαμφισβήτητη υπεροχή του RoBERTa είναι προφανής, καθώς καταφέρνει να ξεπεράσει τους ανταγωνιστές του με διαφορές που κυμαίνονται από 2% έως 7% στις διάφορες μετρικές, όπως φαίνεται στον Πίνακα 6.7.

Αξιοσημείωτο είναι ότι οι τιμές ROC-AUC για τα μοντέλα FinBERT δεν μπόρεσαν να υπολογιστούν, λόγω της απουσίας των απαραίτητων πιθανοτήτων κλάσης, οι οποίες είναι κρίσιμες για τον υπολογισμό αυτής της μετρικής. Το γεγονός αυτό περιορίζει τη σύγκριση με τα υπόλοιπα μοντέλα.

Η συνολική σύγκριση απεικονίζεται επίσης γραφικά στο Σχήμα 6.5, όπου η απόδοση του RoBERTa ξεχωρίζει οπτικά, με σημαντικά υψηλότερες τιμές στις περισσότερες μετρικές. Στο γράφημα, είναι εμφανές ότι το BERT διατηρεί μια σταθερή δεύτερη θέση, ενώ τα FinancialBERT και FinBERT καταγράφουν χαμηλότερες αλλά παραπλήσιες επιδόσεις. Παρόλο που το FinBERT (Araci) κατατάσσεται τελευταίο, τα αποτελέσματα υποδεικνύουν περιθώρια βελτίωσης, ειδικά με την προσαρμογή των μοντέλων στα δεδομένα του χρηματοοικονομικού τομέα.

Μέθοδοι	Accuracy	AUC	Recall	Precision	F1	Kappa	MCC
RoBERTa	0.775877	0.920793	0.775877	0.775877	0.775877	0.632183	0.634966
BERT	0.74166	0.89992	0.74166	0.74166	0.74166	0.554974	0.558077
FinancialBERT	0.704876	-	0.704876	0.704876	0.704876	0.47108	0.486317
FinBERT (Yang & Huang)	0.65355	-	0.65355	0.65355	0.65355	0.355566	0.389141
FinBERT (Araci)	0.570573	-	0.570573	0.570573	0.570573	0.224854	0.256037

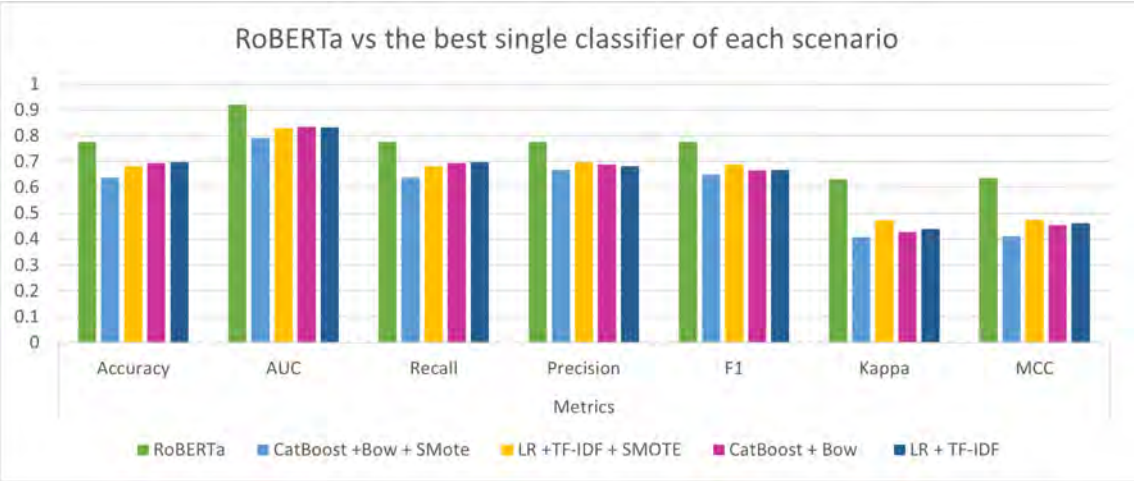
Πίνακας 6.7: Απόδοση προεκπαιδευμένων μοντέλων

6.3 Συμπεράσματα

Σε αυτό το κεφάλαιο πραγματοποιήθηκε μια εκτεταμένη σύγκριση μεταξύ παραδοσιακών αλγορίθμων Μηχανικής Μάθησης και σύγχρονων προεκπαιδευμένων μοντέλων, στο πρόβλημα της Οικονομικής Συναισθηματικής Ανάλυσης. Για το σκοπό αυτό, χρησιμοποιήθηκαν 15 δημοφιλείς μέθοδοι Μηχανικής Μάθησης μέσω της βιβλιοθήκης Pycaret) και 5 προεκπαιδευμένα μοντέλα Βαθιάς Μάθησης - όλα βαριάντες του BERT σε τέσσερα σενάρια, τα οποία συνδυάζουν δύο τεχνικές αναπαράστασης κειμένου - τις (BoW και TF-IDF). Επιπλέον, καθώς το αρχικό σύνολο δεδομένων του προβλήματος παρουσιάζει ανισορροπία στην κατανομή των παραδειγμάτων ανάμεσα στις τρεις κλάσεις, κρίθηκε ενδιαφέρουσα η χρήση και η αξιολόγηση της απόδοσης της τεχνικής SMOTE, η οποία χρησιμοποιείται σε προβλήματα με μη ισοσσοπημένες κλάσεις.

Σχετικά με τη χρήση του BoW ή του TF-IDF (όταν δεν εφαρμόζεται το SMOTE), φαίνεται ότι σε ορισμένες μεθόδους, όπως οι CatBoost και XGBoost, το BoW παρουσιάζει ανώτερη απόδοση, ενώ σε άλλες το TF-IDF αποδεικνύεται πιο αποτελεσματικό. Ωστόσο, όταν εφαρμόζεται το SMOTE, το TF-IDF υπερέχει του BoW, βελτιώνοντας σημαντικά την απόδοση σε όλες τις μετρικές. Επιπλέον, αναφορικά με τα πιθανά οφέλη από την εφαρμογή του SMOTE για την αύξηση των δεδομένων ώστε να αντιμετωπιστεί το πρόβλημα του αρχικά μη ισορροπημένου συνόλου δεδομένων, η απάντηση που δίνεται είναι αρκετά σαφής. Όταν χρησιμοποιείται η τεχνική BoW, η εφαρμογή του SMOTE τείνει να οδηγεί σε μείωση της απόδοσης του μοντέλου σε όλες τις μετρικές. Σε ορισμένες περιπτώσεις, όπως όταν χρησιμοποιείται ο Ridge Classifier, αυτή η μείωση είναι αρκετά σημαντική. Όταν χρησιμοποιείται το TF-IDF όσον αφορά την ακρίβεια, την F1 Score, την Kappa και το MCC, τα αποτελέσματα υποδηλώνουν αδιαμφισβήτητα ότι η εφαρμογή του SMOTE ενισχύει την απόδοση του μοντέλου. Ωστόσο, όταν εξετάζονται μετρικές όπως η ακρίβεια, η AUC και η ανάκληση, η εικόνα γίνεται λιγότερο ξεκάθαρη, καθώς παρά το γεγονός ότι στις περισσότερες περιπτώσεις η χρήση του SMOTE αποδεικνύεται επωφελής, υπάρχουν περιπτώσεις όπου φαίνεται να έχει μια μικρή αρνητική επίδραση στην απόδοση των μοντέλων ως προς αυτές τις τρεις μετρικές.

Όσον αφορά τη χρήση ομαδοποιητών, πρέπει να σημειωθούν δύο βασικές παρατηρήσεις. Πρώτον, οι ομαδοποιητές τείνουν γενικά να αποδίδουν καλύτερα από το σύνολο των μεμονωμένων μοντέλων Μηχανικής Μάθησης. Ωστόσο, δεν υπερβαίνουν πάντα την καλύτερη απόδοση κάθε μεμονωμένης μεθόδου σε κάθε συγκεκριμένο σενάριο, εκτός από τη μετρική AUC, όπου οι ομαδοποιητές αποδίδουν σταθερά καλύτερα από τις μεμονωμένες μεθόδους σε όλες τις περιπτώσεις. Τα ευρήματά μας επιβεβαιώνουν τη γενική ισχυρή απόδοση των ομαδοποιητών, αλλά στο πλαίσιο της παρούσας μελέτης η συμβολή τους δεν είναι ιδιαίτερα σημαντική και ενδέχεται να μην αποτελούν την προτιμώμενη επιλογή. Αξίζει επίσης να αναφερθεί ότι οι ομαδοποιητές που χρησιμοποιήθηκαν βασίστηκαν σε τεχνικές soft voting και ότι ενδεχομένως πιο εξελιγμένες



Σχήμα 6.6: Σύγκριση της μεθόδου RoBERTa με την επικρατούσα κλασική μέθοδο σε κάθε σενάριο

τεχνικές ενσωμάτωσης θα μπορούσαν να αποδώσουν καλύτερα αποτελέσματα.

Αναφορικά με την σύγκριση των παραδοσιακών αλγορίθμων ταξινόμησης Μηχανικής Μάθησης με τα εξελιγμένα προεκπαιδευμένα μοντέλα Βαθιάς Μάθησης, το αποτέλεσμα φαίνεται ιδιαίτερα ενθαρρυντικό. Τα προεκπαιδευμένα μοντέλα αποδίδουν σταθερά καλύτερα από τους παραδοσιακούς τους αντιπάλους σε όλες τις μετρικές. Συγκεκριμένα, το RoBERTa είναι το μοντέλο με την καλύτερη απόδοση συγκεντρωτικά αφού υπερέχει της κορυφαίας μεθόδου σε κάθε ένα από τα τέσσερα που εξετάστηκαν. Η υπεροχή είναι καθολική καθώς το RoBERTa παρουσιάζει ουσιαστική βελτίωση σε κάθε μετρική. Για παράδειγμα, επιτυγχάνει περίπου 7% υψηλότερη ακρίβεια σε σύγκριση με τη μέθοδο που συνδυάζει την Logistic Regression με τον TF-IDF. Όσον αφορά τη μετρική MCC, τα αποτελέσματα είναι ακόμη πιο εντυπωσιακά, με το RoBERTa να ξεπερνά τον καλύτερο ανταγωνιστή του κατά 17%. Στο Σχήμα 6.6 απεικονίζονται τα αντίστοιχα αποτελέσματα όλων των μετρικών.

Ωστόσο, αξίζει να επισημανθεί ένα απροσδόκητο αλλά πολύτιμο αποτέλεσμα: τόσο το RoBERTa όσο και το BERT αποδίδουν καλύτερα από τις τρεις παραλλαγές του FinBERT. Αυτή η ανακάλυψη θέτει ενδιαφέροντα ερωτήματα και απαιτεί περαιτέρω διερεύνηση, ιδιαίτερα δεδομένου ότι τα μοντέλα FinBERT έχουν προεκπαιδευθεί και προσαρμοστεί ειδικά για χρηματοοικονομικές εργασίες.

7. Ελαχιστοποιώντας την Ανάγκη Ετικετοποιημένων Δεδομένων στην Κατηγοριοποίηση Κειμένων, με χρήση Ενεργής Μάθησης σε BERT-based transformers

7.1 Εισαγωγή

Ένα από τα θεμελιώδη προβλήματα στον τομέα της Επεξεργασίας Φυσικής Γλώσσας, είναι αυτό της ταξινόμησης, με προβλήματα που ανήκουν σε ένα μεγάλο φάσμα εφαρμογών που εκτείνονται από την Ανάλυση Συναισθήματος και την ανίχνευση ανεπιθύμητων μηνυμάτων, έως πιο εξειδικευμένες εφαρμογές όπως την κατηγοριοποίηση ιατρικών εγγράφων και την ανάλυση νομικών κειμένων. Καθώς οι απαιτήσεις αυξάνονται και τα μοντέλα Μηχανικής Μάθησης γίνονται συνθετότερα, η ανάγκη για δεδομένα υψηλής ποιότητας καθίσταται ολοένα και επιτακτικότερη. Αφού γίνεται λόγος για μοντέλα ταξινόμησης, είναι προφανές ότι η ανάγκη αυτή αφορά επισημασμένα δεδομένα. Ωστόσο, αν και πλέον τα δεδομένα είναι παντού γύρω μας, αυτό δε σημαίνει πως είναι πάντοτε επισημασμένα. Συνήθως, η απόδοση των ετικετών είναι διαδικασία χρονοβόρα και δαπανηρή – σε κόστος και χρόνο, ιδίως σε εξειδικευμένους τομείς όπου απαιτείται ανθρώπινη εποπτεία από ειδικούς. Αυτή η πρόκληση έχει οδηγήσει στην αναζήτηση πιο αποδοτικών μεθόδων που επιτυγχάνουν υψηλές επιδόσεις με περιορισμένο αριθμό επισημασμένων δειγμάτων.

Στο Κεφάλαιο 3, έγινε αναφορά στο αντικείμενο της Ενεργής Μάθησης και τη φιλοσοφία του. Όπως αναφέρθηκε εκεί, η Ενεργή Μάθηση είναι ευεργετική σε προβλήματα όπου αν και υπάρχει πληθώρα διαθέσιμων δεδομένων, η συντριπτική πλειοψηφία αυτών δεν είναι επισημασμένα. Επιπροσθέτως, το κόστος απόκτησης των εν λόγω ετικετών είναι είτε πολύ μεγάλο, είτε η απόκτησή τους πρακτικά αδύνατη. Με έναυσμα αυτή την προβληματική, αναπτύχθηκε το πεδίο της Ενεργής Μάθησης που αφορά το σύνολο των στρατηγικών με τις οποίες μπορούμε να αποκτήσουμε αποδοτικά μοντέλα, με λιγότερα διαθέσιμα επικετοποιημένα δεδομένα, επιλέγοντας με «έξυπνο» τρόπο τα δεδομένα των οποίων η ετικέτα θα αναζητηθεί.

Είναι εύλογο λοιπόν να αναρωτηθεί κανείς, αν χρησιμοποιώντας τεχνικές Ενεργής Μάθησης μπορούμε να μειώσουμε αισθητά την ανάγκη για επικετοποιημένα δεδομένα σε προβλήματα Επεξεργασίας Φυσικής Γλώσσας. Επιπλέον, είναι επίσης εύλογη η επιθυμία στην παραπάνω προβληματική να ενσωματωθούν όσο το δυνατόν πιο σύγχρονα μοντέλα από το προαναφερθέν πεδίο, ώστε να διερευνηθεί το αν ο συνδυασμός των δύο αυτών πεδίων μπορεί να οδηγήσει σε ενδιαφέροντα ευρήματα.

Με δεδομένο ότι μια από τις πιο επαναστατικές εξελίξεις στην Επεξεργασία Φυσικής Γλώσσας τα τελευταία χρόνια ήταν η εισαγωγή των αρχιτεκτονικών μετασχηματιστών (Transformers), όπως τα μοντέλα BERT (Bidirectional Encoder Representations from Transformers), τα οποία παρά την εντυπωσιακή τους ακρίβεια και ανθεκτικότητα, εξακολουθούν να απαιτούν μεγάλα σύνολα επισημασμένων δεδομένων για τη διαδικασία της εκπαίδευσής τους. Η μείωση αυτής της εξάρτησης, χωρίς υποβάθ-

μιση της απόδοσης, αποτελεί ένα κρίσιμο ερευνητικό ζήτημα, ιδιαίτερα σε πραγματικές εφαρμογές όπου η διαθεσιμότητα επισημασμένων δεδομένων είναι περιορισμένη ή η απόκτησή τους είναι ιδιαίτερα δαπανηρή.

Το κεφάλαιο αυτό εστιάζει στην ενσωμάτωση της Ενεργής Μάθησης στη διαδικασία της ταξινόμησης κειμένου μέσω μοντέλων που βασίζονται σε μετασχηματιστές BERT, με στόχο τη μείωση των απαιτήσεων για επισημασμένα δεδομένα. Πρέπει να τονιστεί πως δεν εξετάζεται η βελτίωση των μοντέλων από τη σκοπιά των επιδόσεών τους στις διάφορες μετρικές αξιολόγησης, αλλά το κατά πόσο με τη χρήση της Ενεργής Μάθησης μπορούμε να επιτύχουμε εξίσου καλά αποτελέσματα, με χρήση λιγότερων επισημασμένων δεδομένων. Για τον σκοπό αυτό σχεδιάστηκε και εκτελέστηκε μια εκτεταμένη πειραματική διάταξη, η οποία ενσωματώνει Ενεργή Μάθηση με δειγματοληψία αβεβαιότητας και εμπλέκει έναν μεγάλο αριθμό συνόλων δεδομένων και παραλλαγές του μοντέλου BERT, με στόχο τη διερεύνηση του παραπάνω ζητήματος. Επιπλέον, προτείνεται και ένας νέος μέτα-δείκτης, ο Δείκτης Απόδοσης Προσαρμοσμένης Αποδοτικότητας (Efficiency-Adjusted Performance Index - EAPI), με στόχο την καλύτερη αξιολόγηση της πειραματικής διαδικασίας, καθώς δεν υπάρχει κάποιος ήδη υπάρχων τρόπος να αξιολογηθούν τέτοια σενάρια, λαμβάνοντας υπόψη όλες τις επιμέρους παραμέτρους – δηλαδή τόσο το ποσοστό μείωσης των δεδομένων εκπαίδευσης, όσο και την απόδοση των μεθόδων στις εκάστοτε μετρικές.

Το κεφάλαιο διαμορφώνεται ως εξής: αρχικά, παρουσιάζονται τα δεδομένα και οι παραλλαγές της μεθόδου BERT που χρησιμοποιήθηκαν στο πλαίσιο της πειραματικής διαδικασίας. Στη συνέχεια, περιγράφεται ο τρόπος υλοποίησης των πειραμάτων, καθώς και ο προτεινόμενος μέτα-δείκτης EAPI, ενώ έπονται τα αποτελέσματα και τα συμπεράσματα που εξάγονται από την πειραματική διαδικασία.

7.2 Πειραματική Διαδικασία

Σε αυτήν την ενότητα περιγράφεται η πειραματική διαδικασία, ώστε να δοθεί μια σαφής εικόνα για το πώς σχεδιάστηκαν και υλοποιήθηκαν τα πειράματα. Είναι σημαντικό να τονιστεί ότι η διαδικασία διαμορφώθηκε ειδικά για να απαντήσει στο κεντρικό ερώτημα αυτής της μελέτης: αν η Ενεργή Μάθηση μπορεί να επιτύχει συγκρίσιμα αποτελέσματα με το παραδοσιακό σχήμα εκπαίδευσης που χρησιμοποιεί το πλήρες σύνολο εκπαίδευσης, όσον αφορά τις μετρικές αξιολόγησης.

Το ιδανικό σενάριο θα ήταν να επιτευχθούν παρόμοιες τιμές στις μετρικές αξιολόγησης με εκείνες που λαμβάνονται όταν χρησιμοποιείται το πλήρες σύνολο εκπαίδευσης, αλλά με σημαντική μείωση των δεδομένων. Φυσικά, η μελέτη εξετάζει και δευτερεύοντα ερωτήματα, αλλά πρέπει να τονιστεί πως ο κύριος σχεδιασμός των πειραμάτων επικεντρώνεται αποκλειστικά σε αυτό το κεντρικό ερώτημα.

7.2.1 Υλοποίηση της Πειραματικής Διάταξης

Κατά την εκτέλεση των πειραμάτων, κάθε σύνολο δεδομένων χωρίστηκε σε σύνολο εκπαίδευσης και σύνολο ελέγχου σε αναλογία 80% / 20%. Αρχικά, από το σύνολο εκπαίδευσης, χρησιμοποιήθηκε μόλις 1% των δεδομένων για τη διαδικασία fine-tuning του εκάστοτε BERT μοντέλου. Στα υπόλοιπα δεδομένα, η ετικέτα κρύβεται τεχνιέντως, ώστε να είναι δυνατή η προσομοίωση ενός σεναρίου Ενεργής Μάθησης, όπου οι ετικέτες του συνόλου δεδομένων είναι άγνωστες και γίνονται διαθέσιμες μέσω κάποιου ειδικού, όταν αυτές ζητηθούν. Στη συνέχεια, η απόδοση του μοντέλου αξιολογείται

με τρεις μετρικές απόδοσης: την ακρίβεια (accuracy), το F1-score και τον συντελεστή συσχέτισης Matthews (MCC - Matthews Correlation Coefficient).

Εν συνεχεία, τέθηκε σε εφαρμογή ένα επαναληπτικό σχήμα μέσω του οποίου, σε κάθε κύκλο επανεκπαίδευσης, το μοντέλο εκπαιδευόταν ξανά καθώς το σύνολο των δεδομένων αυξανόταν κατά 1% επί του αρχικού συνόλου εκπαίδευσης, καθώς σε αυτό προστίθενται δεδομένα επιλεγμένα μέσω uncertainty sampling. Τα δεδομένα που επιλέγονται αναχτούν την ετικέτα τους – σε ένα πραγματικό σενάριο, η ανάχτηση αυτή θα επέφερε κάποιο σεβαστό κόστος – και το μοντέλο αναπροσαρμόζεται.

Η παραπάνω διαδικασία επαναλαμβάνεται, μέχρις ότου το συνεχώς αυξανόμενο σύνολο εκπαίδευσης να φτάσει το 50% του αρχικού συνόλου εκπαίδευσης. Ο στόχος αυτής της στρατηγικής ήταν να αξιολογηθεί η απόδοση του μοντέλου χρησιμοποιώντας το πολύ το μισό των διαθέσιμων δεδομένων εκπαίδευσης, καθώς η Ενεργή Μάθηση αποσκοπεί σε σημαντική μείωση των αναγκών για δεδομένα με γνωστή ετικέτα, αναγνωρίζοντας το σημαντικό κόστος της απόκτησής τους. Συνεπώς, περαιτέρω αύξηση του συνόλου δεδομένων, σε ποσοστά που αγγίζουν το αρχικό, δεν θεωρήθηκε επικερδής.

Όσον αφορά το σενάριο Ενεργής Μάθησης, η στρατηγική που επιλέχθηκε ήταν η δειγματοληψία με βάση την αβεβαιότητα (uncertainty sampling). Στη δειγματοληψία με αβεβαιότητα, ο αλγόριθμος επιλέγει τα πιο κρίσιμα δεδομένα προς επισήμανση, εστιάζοντας στις περιπτώσεις για τις οποίες έχει τη μεγαλύτερη αβεβαιότητα ως προς την ετικέτα τους. Αυτή η μέθοδος αξιοποιεί τη γνώση του μοντέλου για να εντοπίσει και να δώσει προτεραιότητα σε περιπτώσεις που είναι ασαφείς, βελτιστοποιώντας έτσι τη μαθησιακή διαδικασία, καθώς εστιάζει στις περιοχές όπου το μοντέλο είναι πιο αδύναμο. Με αυτόν τον τρόπο, μειώνεται το κόστος ετικετοποίησης και βελτιώνεται παράλληλα η απόδοση.

Αυτή η προσέγγιση είναι ιδιαίτερα χρήσιμη σε σενάρια με περιορισμένα ετικετοποιημένα δεδομένα, καθώς μεγιστοποιεί το πληροφοριακό κέρδος από κάθε ετικετοποιημένο δείγμα.

Πριν προχωρήσουμε στις λεπτομέρειες που αφορούν τα σύνολα δεδομένων και τις παραλλαγές της μεθόδου BERT που χρησιμοποιήθηκαν στα πειράματα, πρέπει να τονιστεί ξανά πως αυτή η πειραματική διαδικασία έχει σχεδιαστεί αποκλειστικά για ερευνητικούς σκοπούς. Σε πραγματικές εφαρμογές, θα ήταν μη ρεαλιστικό να χρησιμοποιηθεί αυτή η προσέγγιση, καθώς η πλήρης γνώση των βαθμολογιών μετρικών από το συνολικό σύνολο εκπαίδευσης δεν θα ήταν διαθέσιμη ή θα είχε απαγορευτικό κόστος. Ωστόσο, αυτή η στρατηγική επιτρέπει τη διερεύνηση του τρόπου με τον οποίο η Ενεργή Μάθηση συγκρίνεται με τη μέγιστη δυνατή απόδοση των μοντέλων, παρέχοντας πολύτιμες γνώσεις για τις δυνατότητες αυτής της τεχνικής.

7.2.1.1 Σύνολα δεδομένων

Κατά τη διάρκεια της πειραματικής διαδικασίας, χρησιμοποιήθηκαν δέκα σύνολα δεδομένων, διαθέσιμα μέσω του αποθετηρίου Hugging Face. Η επιλογή αυτή έγινε για να διασφαλιστεί ότι τα σύνολα δεδομένων είναι δημόσια προσβάσιμα, διευκολύνοντας έτσι την αναπαραγωγή των αποτελεσμάτων της πειραματικής διαδικασίας. Τα σύνολα δεδομένων επιλέχθηκαν με γνώμονα τη διαφορετικότητά τους όσον αφορά το περιεχόμενο, τον αριθμό των δειγμάτων και τον αριθμό των κλάσεων που περιλαμβάνουν. Συγκεκριμένα, τα μισά από αυτά χρησιμοποιούνται για δυαδική ταξινόμηση, ενώ τα υπόλοιπα περιλαμβάνουν δεδομένα που ανήκουν σε τρεις έως έξι κλάσεις. Το μικρότερο σύνολο

δεδομένων περιέχει λιγότερα από 1.000 δείγματα, ενώ το μεγαλύτερο περιλαμβάνει περίπου 67.000 δείγματα. Αυτή η ποικιλομορφία είναι κρίσιμη για να διασφαλιστεί ότι τα αποτελέσματα της πειραματικής διαδικασίας δεν επηρεάζονται από τα χαρακτηριστικά ενός συγκεκριμένου συνόλου δεδομένων.

Αναλυτικά, τα σύνολα δεδομένων που χρησιμοποιήθηκαν είναι τα εξής:

- poem sentiment [250]: Το σύνολο δεδομένων poem sentiment, το οποίο προέρχεται από το Project Gutenberg, αποτελείται από στίχους ποιημάτων, αντιστοιχισμένους με τις κατάλληλες ετικέτες συναισθηματικής ανάλυσης. Είναι κατάλληλο για εργασίες συναισθηματικής ταξινόμησης, επιτρέποντας την ανάλυση και κατηγοριοποίηση του συναισθηματικού τόνου σε ποιητικά κείμενα. Οι κατηγορίες του περιλαμβάνουν θετικό, αρνητικό, ουδέτερο συναίσθημα, αλλά και μικτά συναισθήματα.

- tweet eval emotion [174]: Το υποσύνολο tweet eval emotion της συλλογής TweetEval χρησιμοποιείται για ανάλυση συναισθήματος, ταξινομώντας τα tweets σε τέσσερις κατηγορίες: θυμός (anger), χαρά (joy), αισιοδοξία (optimism) και λύπη (sadness).

- rotten tomatoes [204]: Το σύνολο δεδομένων rotten tomatoes, γνωστό και ως Movie Review (MR) dataset, περιέχει κριτικές ταινιών που χρησιμοποιούνται σε εργασίες ανάλυσης συναισθήματος. Προέρχεται από το Πανεπιστήμιο Cornell και οι κριτικές του ταξινομούνται σε θετικές και αρνητικές, αποτελώντας έτσι ένα δυαδικό πρόβλημα.

- tweet eval hate [25]: Ένα άλλο υποσύνολο του TweetEval, το tweet eval hate, αφορά την ανίχνευση ρητορικής μίσους. Ταξινομεί τα tweets σε δύο κατηγορίες: ρητορική μίσους (hate speech) και μη ρητορική μίσους (non-hate speech). Ο κύριος στόχος του είναι να εντοπίσει ρητορική μίσους στο διαδίκτυο, μια ιδιαίτερα σημαντική εφαρμογή για την καταπολέμηση της παρενόχλησης και των διακρίσεων στο διαδίκτυο.

- tweet sentiment airlines [278]: Το σύνολο δεδομένων tweet sentiment airlines αποτελείται από tweets σχετικά με αεροπορικές εταιρείες, εφοδιασμένο με τις ανάλογες ετικέτες ανάλυσης συναισθήματος. Οι ετικέτες συναισθήματος περιλαμβάνουν τρεις κατηγορίες: θετικό, αρνητικό και ουδέτερο. Είναι ιδιαίτερα χρήσιμο για την ανάλυση σχολίων και αξιολογήσεων πελατών στον τομέα των αερομεταφορών.

- emotion(dair-ai) [243]: Το emotion(dair-ai) αποτελείται από αγγλικά μηνύματα στο Twitter, τα οποία έχουν ταξινομηθεί σε έξι βασικά συναισθήματα: θυμός (anger), φόβος (fear), χαρά (joy), αγάπη (love), λύπη (sadness) και έκπληξη (surprise). Περιλαμβάνει διαφορετικές αναλογίες εμφανίσεων σε κάθε κατηγορία και χρησιμοποιείται για πολυκατηγορική ταξινόμηση συναισθημάτων.

- ade corpus v2 [107]: Το ade corpus v2, γνωστό και ως Adverse Drug Reaction Data, έχει σχεδιαστεί για τη διευκόλυνση της έρευνας σε θέματα εξόρυξης κειμένου και σχέσεων στον τομέα της βιοϊατρικής. Εστιάζει στον εντοπισμό προτάσεων που σχετίζονται με ανεπιθύμητες ενέργειες φαρμάκων (ADEs) και στην εξαγωγή σχέσεων μεταξύ φαρμάκων και παρενεργειών ή μεταξύ φαρμάκων και δόσεων.

Πίνακας 7.1: Σύνολα Δεδομένων

Dataset	Instances	Classes	Labels
poem_sentiment	892	4	Positive, Negative, No Impact, Mixed (both Positive and Negative)
tweet_eval__emotion	3257	4	Anger, Joy, Optimism, Sadness
rotten_tomatoes	8530	2	Positive, Negative
tweet_eval__hate	9000	2	Hate, Non hate
tweet_sentiment_airlines	11712	3	Positive, Neutral, Negative
Emotion(dair-ai)	16000	6	Anger, Joy, Sadness, Love, Fear, Surprise
Ade_corpus_v2	23516	2	Related, Non related
Imdb	25000	2	Positive, Negative
nyu_mll_glue_sst2	67349	2	Positive, Negative
tweet_eval__sentiment	45615	3	Positive, Neutral, Negative

- Imdb [155]: Το σύνολο δεδομένων Imdb έχει σχεδιαστεί για δυαδική ταξινόμηση συναισθήματος και περιέχει κριτικές ταινιών με ετικέτες θετικό ή αρνητικό συναίσθημα. Χρησιμοποιείται ευρέως για εκπαίδευση και αξιολόγηση αλγορίθμων ανάλυσης συναισθήματος.

- nyu mll glue sst2 [287]: Το nyu mll glue sst2 είναι ένα υποσύνολο του Stanford Sentiment Treebank (SST) και περιλαμβάνεται στη σουίτα GLUE (General Language Understanding Evaluation). Εστιάζει στη δυαδική ταξινόμηση συναισθήματος, κατηγοριοποιώντας προτάσεις από κριτικές ταινιών ως θετικές ή αρνητικές.

- tweet eval sentiment [231]: Ένα ακόμη υποσύνολο της συλλογής TweetEval, το tweet eval sentiment, έχει σχεδιαστεί για ανάλυση συναισθήματος σε tweets, ταξινομώντας τα σε τρεις κατηγορίες: θετικό, αρνητικό και ουδέτερο.

Οι βασικές πληροφορίες για τα χαρακτηριστικά των συνόλων δεδομένων, όπως ο όγκος των δεδομένων και ο αριθμός των κλάσεων που περιλαμβάνει κάθε ένα, συνοψίζονται στον Πίνακα 7.1.

7.2.1.2 Μοντέλα BERT

Στην παρούσα μελέτη, χρησιμοποιήθηκαν επτά παραλλαγές του BERT, οι οποίες είναι διαθέσιμες στο αποθετήριο Hugging Face. Ακολουθεί μια σύντομη περιγραφή των διαφορετικών παραλλαγών του BERT που χρησιμοποιήθηκαν στην έρευνα:

- **bert-base-model-cased** [65]: Η κλασική έκδοση του BERT, η οποία διατηρεί τη διάκριση μεταξύ κεφαλαίων και πεζών χαρακτήρων (case-sensitive). Χρήσιμο σε περιπτώσεις όπου η διάκριση αυτή έχει σημασία, όπως κύρια ονόματα και ακρωνύμια.

- **bert-base-model-uncased** [65]: Παραλλαγή του BERT όπου δεν διατηρείται η διάκριση μεταξύ κεφαλαίων και πεζών χαρακτήρων (case-insensitive). Όλες οι λέξεις μετατρέπονται σε πεζά γράμματα πριν από την επεξεργασία.

- **distilbert-base-cased** [241]: Μια ελαφριά και πιο αποδοτική έκδοση του BERT (DistilBERT). Είναι 40% μικρότερο από το κανονικό BERT, 60% ταχύτερο, αλλά διατηρεί 97% της απόδοσής του.

- **distilbert-base-uncased** [241]: Η uncased έκδοση του προηγούμενου μοντέλου, κατάλληλη για περιπτώσεις όπου η διάκριση κεφαλαίων-πεζών δεν είναι απαραίτητη.

- **bert-base-multilingual-uncased-sentiment** [184]: Ένα πολυγλωσσικό μοντέλο BERT, ειδικά εκπαιδευμένο για Ανάλυση Συναισθήματος σε προϊόντα σε έξι γλώσσες (Αγγλικά, Ολλανδικά, Γερμανικά, Γαλλικά, Ισπανικά και Ιταλικά).

- **distilbert-base-multilingual-cased-sentiments-student** [184]: Μια αποσταγμένη (distilled) πολυγλωσσική έκδοση του BERT, σχεδιασμένη για αποδοτική και ακριβή Ανάλυση Συναισθήματος σε πολλές γλώσσες.

- **bertweet-base-sentiment-analysis** [210]: Ένα μοντέλο BERT βελτιστοποιημένο για την Ανάλυση Συναισθήματος σε tweets. Είναι εκπαιδευμένο σε 40.000 tweets και κατηγοριοποιεί τα συναισθήματα ως θετικά, αρνητικά και ουδέτερα.

7.2.1.3 Αξιολόγηση

Κατά τη φάση της αξιολόγησης, χρησιμοποιήθηκαν τρεις παραδοσιακές μετρικές που χρησιμοποιούνται ευρέως σε εργασίες ταξινόμησης: η Accuracy, η F1-score και ο Συντελεστής Matthews (MCC). Αυτές οι μετρικές παρέχουν μια ολοκληρωμένη εκτίμηση της απόδοσης του μοντέλου. Η Accuracy αντικατοπτρίζει το ποσοστό σωστών προβλέψεων, αλλά μπορεί να είναι παραπλανητική σε μη ισορροπημένα δεδομένα. Το F1-score σκοπεύει να καλύψει αυτό το κενό, καθώς συνδυάζει την ακρίβεια και την ανάκληση, καθιστώντας το πιο αξιόπιστο όταν υπάρχουν σημαντικές αποκλίσεις μεταξύ των κλάσεων. Τέλος, ο MCC λαμβάνει υπόψη όλες τις κατηγορίες του πίνακα σύγχυσης και προσφέρει μια πιο αντικειμενική εκτίμηση της συσχέτισης μεταξύ πραγματικών και προβλεπόμενων τιμών, καθιστώντας τον ιδανικό για περιπτώσεις με ανισομερή κατανομή των δεδομένων στις κλάσεις. Η συνδυαστική χρήση αυτών των μετρικών διασφαλίζει μια δίκαιη και ακριβή αξιολόγηση της απόδοσης του μοντέλου.

Ωστόσο, δεδομένου ότι ο βασικός στόχος αυτής της μελέτης είναι να εξετάσει εάν η χρήση Ενεργής Μάθησης μπορεί να μειώσει το μέγεθος του συνόλου εκπαίδευσης σε σενάρια ταξινόμησης κειμένου, χωρίς να μειώσει την απόδοση του μοντέλου, κρίθηκε αναγκαία η χρήση ενός πιο εξειδικευμένου δείκτη για βαθύτερη ανάλυση των αποτελεσμάτων. Για τον σκοπό αυτό, χρησιμοποιήθηκε ο Δείκτης Απόδοσης Προσαρμοσμένος στην Αποδοτικότητα (Efficiency-Adjusted Performance Index), ο οποίος έχει σχεδιαστεί ειδικά για την αξιολόγηση του συμβιβασμού μεταξύ μεγέθους συνόλου δεδομένων και απόδοσης του μοντέλου σε περιβάλλοντα Ενεργής Μάθησης. Αυτός ο δείκτης προσφέρει μια πιο ολοκληρωμένη εικόνα των αποδοτικών κερδών που επιτυγχάνονται μέσω της Ενεργής Μάθησης κατά τη διάρκεια της εκτεταμένης πειραματικής διαδικασίας.

Η αποδοτικότητα της Ενεργής Μάθησης εξαρτάται από την ικανότητά της να μειώνει την ανάγκη για επισημειωμένα δεδομένα, διατηρώντας παράλληλα υψηλά επίπεδα απόδοσης του μοντέλου. Ωστόσο, οι παραδοσιακές μετρικές αξιολόγησης, δεν λαμβάνουν υπόψη το μέγεθος του επισημασμένου συνόλου δεδομένων που απαιτείται για

την επίτευξη συγκεκριμένης απόδοσης. Για την πληρέστερη αξιολόγηση της αποδοτικότητας της χρήσης ενεργής Μάθσης προτείνουμε τον Δείκτη Απόδοσης Προσαρμοσμένης Αποδοτικότητας (Efficiency-Adjusted Performance Index - EAPI).

Ο EAPI διατυπώνεται ως εξής:

$$EAPI = \left(1 + \frac{PostAL_{metric} - PreAL_{metric}}{PreAL_{metric}}\right) * (1 - Proportion\ of\ AL_{trainset}) \quad (7.1)$$

που απλοποιείται ως εξής:

$$EAPI = \left(\frac{PostAL_{metric}}{PreAL_{metric}}\right) * (1 - Proportion\ of\ AL_{trainset}) \quad (7.2)$$

όπου:

- $PreAL_{metric}$ η απόδοση του μοντέλου, εκφρασμένη μέσω μιας μετρικής πριν τη χρήση Ενεργής Μάθησης, όταν το μοντέλο εκπαιδεύεται σε ολόκληρο το σύνολο εκπαίδευσης
- $PostAL_{metric}$ η απόδοση του μοντέλου, εκφρασμένη μέσω της ίδιας μετρικής στο σενάριο όπου χρησιμοποιείται Ενεργή Μάθηση
- $Proportion\ of\ AL_{trainset}$ το ποσοστό το συνόλου εκπαίδευσης που χρησιμοποιήθηκε στο σενάριο Ενεργής Μάθησης

Ο EAPI επιτρέπει τη συγκριτική αξιολόγηση διαφορετικών στρατηγικών Ενεργής Μάθησης λαμβάνοντας υπόψη τόσο την απόδοση όσο και την αποδοτικότητα της χρήσης δεδομένων. Είναι ένας μέτα-δείκτης, καθώς μπορεί να χρησιμοποιηθεί στα αποτελέσματα οποιασδήποτε μετρικής και να δώσσει χρήσιμη πληροφορία για το πως η ενεργή μάθηση επιδρά στις επιδώσεις στην συγκεκριμένη μετρική. Υψηλότερες τιμές του EAPI υποδηλώνουν καλύτερη ισορροπία μεταξύ ακρίβειας και απαιτήσεων επισήμανσης. Έτσι, ο προτεινόμενος δείκτης επιτρέπει την αξιολόγηση διαφορετικών στρατηγικών Ενεργής Μάθησης σε συνάρτηση με την απαιτούμενη ποσότητα επισημασμένων δεδομένων. Όταν δύο μοντέλα επιτυγχάνουν παρόμοια απόδοση, το μοντέλο με την υψηλότερη βαθμολογία EAPI θεωρείται πιο αποδοτικό, καθώς απαιτεί λιγότερα επισημασμένα δεδομένα. Επιπλέον, η χρήση του δείκτη καθιστά δυνατή την αξιολόγηση της γενικότητας διαφόρων μεθόδων ενεργής μάθησης σε διαφορετικά μεγέθη δεδομένων και συνθήκες ανισορροπίας κατηγοριών.

Θεωρητικά, ο δείκτης EAPI μπορεί να κυμαίνεται από το αρνητικό έως το θετικό άπειρο. Ωστόσο, στο πλαίσιο της μελέτης μας, όπου το σενάριο Ενεργής Μάθησης χρησιμοποιεί μεταξύ 1% και 50% του συνόλου εκπαίδευσης, παρατηρείται ότι το πρώτο συνθετικό του EAPI είναι συνήθως κοντά στο 1, ενώ το δεύτερο κυμαίνεται μεταξύ 0.5 και 0.99. Συνεπώς, αναμένουμε οι τιμές του EAPI στη μελέτη μας να βρίσκονται μεταξύ 0 και 1.

Πρέπει να τονιστεί πως ο EAPI προορίζεται σαφώς για σενάρια σαν αυτό που παρουσιάζεται εδώ - καθαρά εκπαιδευτικά. Δεν είναι δυνατή η χρήση του σε προβλήματα καθημερινής ζωής, καθώς ο υπολογισμός του απαιτεί την γνώση της τιμής της εκάστοτε μετρικής στο πλήρες σύνολο εκπαίδευσης. Αν κάτι τέτοιο είναι δυνατό, η χρήση της Ενεργής Μάθησης περιττεύει.

7.3 Αποτελέσματα

Στη συνέχεια παρουσιάζονται τα αποτελέσματα της πειραματικής διαδικασίας. Δεδομένου του εκτεταμένου όγκου δεδομένων, τα ευρήματα παρουσιάζονται μέσω ενός συνδυασμού αριθμητικών και γραφικών απεικονίσεων. Αυτή η προσέγγιση υιοθετείται ώστε να διευκολυνθεί μια πιο ολοκληρωμένη κατανόηση των αποτελεσμάτων, καθώς οι παραδοσιακές μέθοδοι παρουσίασης εκτενών πινάκων με αριθμητικές τιμές μπορεί να είναι αποπροσανατολιστικές και δύσκολες στην ερμηνεία.

Ο βασικός στόχος αυτής της μελέτης είναι να διερευνηθεί εάν η Ενεργή Μάθηση μπορεί να επιτύχει μετρικές απόδοσης συγκρίσιμες με εκείνες που λαμβάνονται σε σενάρια όπου αυτή δεν εφαρμόζεται, δηλαδή, συγκεκριμένα όταν χρησιμοποιείται το πλήρες σύνολο εκπαίδευσης. Σημειώνεται ότι κατά τη διερεύνηση αυτού του βασικού ερωτήματος προέκυψαν και επιπρόσθετα ευρήματα, τα οποία επίσης παρουσιάζονται σε αυτήν την ενότητα. Επιπλέον, πριν από τη συζήτηση των κύριων ερευνητικών ερωτημάτων της μελέτης μας, κρίθηκε χρήσιμο να αξιολογηθεί η συνολική απόδοση των χρησιμοποιούμενων μεθόδων σε διαφορετικά σύνολα δεδομένων χωρίς την εφαρμογή της Ενεργής Μάθησης, όπου η διαδικασία fine-tuning των μοντέλων γίνεται στο πλήρες σύνολο εκπαίδευσης. Η προκαταρκτική αυτή ανάλυση παρέχει στον αναγνώστη μια σαφέστερη εικόνα της βασικής απόδοσης των μοντέλων, διευκολύνοντας έτσι την κατανόηση των μεταγενέστερων ευρημάτων που σχετίζονται με την χρήση Ενεργής Μάθησης. Η σύγκριση αυτή, θα παρουσιαστεί πρώτη στη συνέχεια. Πρέπει να σημειωθεί ότι η σύγκριση αυτή αν και εκ πρώτης όψεως δεν ευθυγραμμίζεται με το σκοπό της εργασίας, εκτός από μια ευκαιρία εντοπισμού αποδοτικών μεθόδων στην ταξινόμηση κειμένου, επιπλέον αποτελεί απαραίτητο βήμα στον υπολογισμό του προτεινόμενου μέτα-δείκτη.

Στη συνέχεια, έπονται τα αποτελέσματα από τη σκοπιά της χρήσης της Ενεργής Μάθησης. Συγκεκριμένα, εξετάζεται αν οι μετρικές αξιολόγησης παραμένουν εντός αποδεκτών ορίων απόκλισης σε σχέση με τα σενάρια όπου δε χρησιμοποιήθηκε Ενεργή Μάθηση. Επιπλέον, αναλύεται η αποδοτικότητα της χρήσης την Ενεργής Μάθησης στη χρήση μικρότερου ποσοστού δεδομένων για την επίτευξη παρόμοιων αποτελεσμάτων και τέλος, διερευνάται η επίδραση χαρακτηριστικών των συνόλων δεδομένων, όπως το μέγεθος, η πολυπλοκότητα και ο αριθμός κατηγοριών, στη χρήση Ενεργητικής Μάθησης.

7.3.1 Σύγκριση μοντέλων χωρίς τη χρήση Ενεργής Μάθησης

Όπως αναφέρθηκε, η παρουσίαση αποτελεσμάτων θα ξεκινήσει με σύγκριση της απόδοσης διαφορετικών εκδόσεων του BERT σε σενάρια όπου δεν χρησιμοποιείται Ενεργή Μάθηση και συνεπώς το εκάστοτε μοντέλο εκπαιδεύεται στο πλήρες σύνολο εκπαίδευσης. Στο Σχήμα 7.1, απεικονίζονται τα ραβδογράμματα για κάθε μια από τις τρεις μετρικές με τις αντίστοιχες βαθμολογίες τους σε αυτές για κάθε μέθοδο και σύνολο δεδομένων. Είναι δύσκολο να εντοπιστεί μία μέθοδος που υπερέχει σταθερά έναντι των υπολοίπων, καθώς οι περισσότερες παρουσιάζουν μόνο μικρές διακυμάνσεις στην απόδοσή τους. Μια πιο ξεκάθαρη εικόνα για την κατάταξη των μεθόδων ανά μετρική, αποκτάται με τα αποτελέσματα του στατιστικού ελέγχου Friedman παρουσιάζεται στον Πίνακα 7.2. Σύμφωνα με αυτήν, το μοντέλο bert-base-uncased αναδείχθηκε ως το καλύτερο σε όλες τις μετρικές, ακολουθούμενο από το distilbert-base-uncased, ενώ

Πίνακας 7.2: Κατάταξη Friedman με fine-tuning στο πλήρες σύνολο δεδομένων

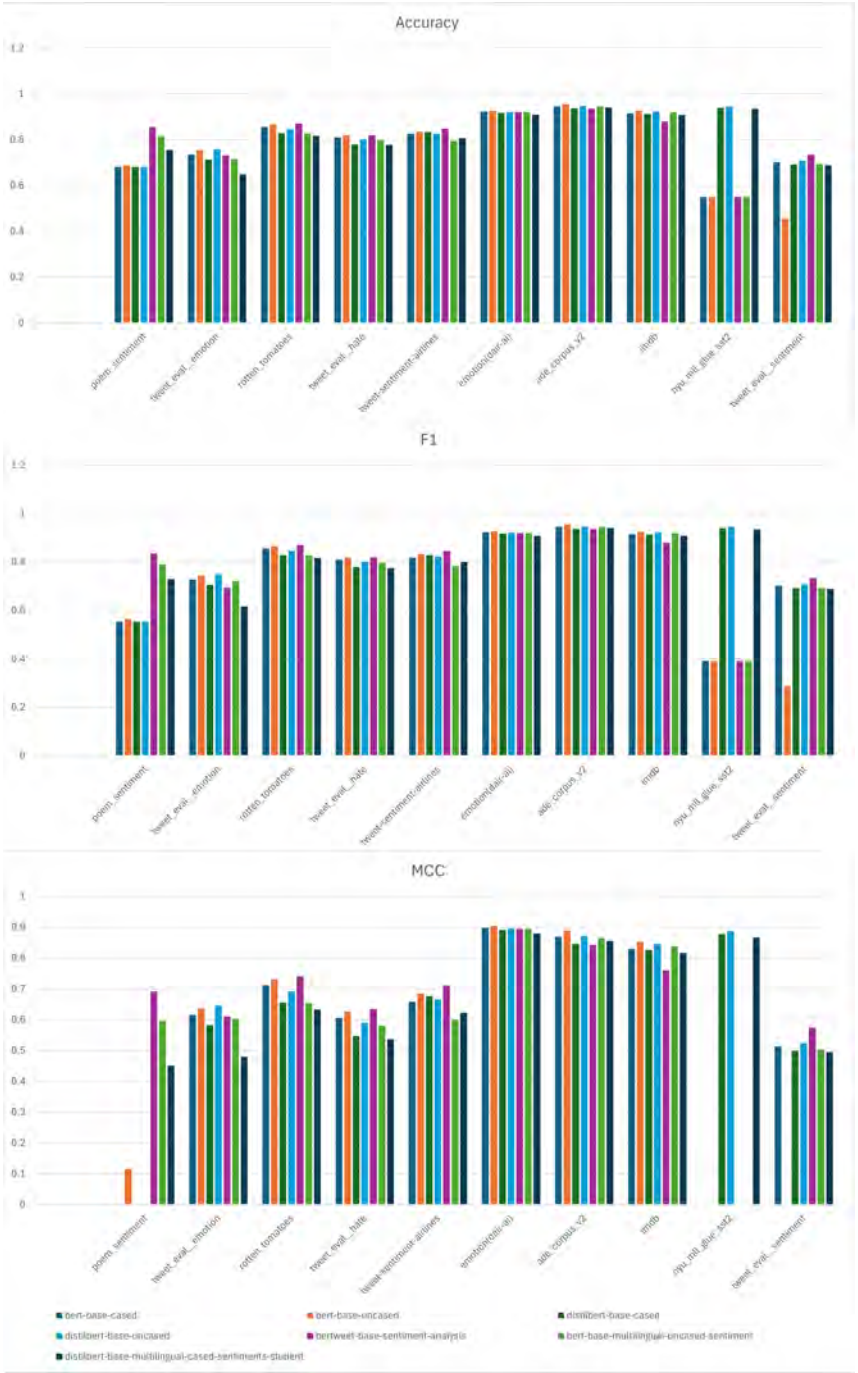
	Accuracy	
Rank	**Method**	**Friedman score**
1στ	bert-base-uncased	5.35
2νδ	distilbert-base-uncased	5.10
3ρδ	bertweet-base-sentiment-analysis	4.60
4τη	bert-base-cased	4.25
5τη	bert-base-multilingual-uncased-sentiment	3.40
6τη	distilbert-base-cased	3.00
7τη	distilbert-base-multilingual-cased-sentiments-student	2.30
	F1	
Rank	**Method**	**Friedman score**
1στ	bert-base-uncased	5.25
2νδ	distilbert-base-uncased	5.10
3ρδ	bertweet-base-sentiment-analysis	4.45
4τη	bert-base-cased	4.25
5τη	bert-base-multilingual-uncased-sentiment	3.55
6τη	distilbert-base-cased	3.10
7τη	distilbert-base-multilingual-cased-sentiments-student	2.30
	MCC	
Rank	**Method**	**Friedman score**
1στ	bert-base-uncased	5.25
2νδ	distilbert-base-uncased	5.10
3ρδ	bertweet-base-sentiment-analysis	4.65
4τη	bert-base-cased	4.25
5τη	bert-base-multilingual-uncased-sentiment	3.45
6τη	distilbert-base-cased	3.00
7τη	distilbert-base-multilingual-cased-sentiments-student	2.30

το bertweet-base-sentiment-analysis κατέλαβε την τρίτη θέση.

Επιπλέον, για να διαπιστωθεί αν οι μέθοδοι ανά ζεύγη διαφέρουν στην απόδοσή τους και η διαφορά αυτή είναι στατιστικά σημαντική, πραγματοποιήθηκε ο post-hoc έλεγχος Bonferroni-Dunn. Τα αποτελέσματα αυτού του ελέγχου παρουσιάζονται στο Σχήμα 7.2. Στην απεικόνιση των αποτελεσμάτων, οι μέθοδοι με παρόμοια απόδοση συνδέονται με έντονες οριζόντιες γραμμές. Ο έλεγχος έδειξε ότι όλα τα ζεύγη μεθόδων είναι στατιστικά όμοια στην απόδοσή τους, με μοναδική εξαίρεση το ζεύγος που αποτελείται από τα μοντέλα bert-base-uncased και distilbert-base-multilingual-cased-sentiments-student που κατατάσσονται πρώτο και τελευταίο αντίστοιχα. Επομένως, μόνο αυτά τα δύο ζεύγη εμφανίζουν σημαντικές στατιστικές διαφορές στην απόδοσή τους, ενώ η υπάρχουσα στατιστική ομοιότητα μεταξύ όλων των άλλων ζευγών δείχνει ότι οι μέθοδοι έχουν σχεδόν παρόμοια συμπεριφορά όσον αφορά τη γενική αξιολόγηση των

προβλεπτικών τους δυνατοτήτων.

Αξιολογώντας τις πραγματικές τιμές των κατατάξεων, μπορεί να παρατηρηθεί ένα ακόμα ενδιαφέρον εύρημα: οι uncased εκδόσεις των bert-base και distilBERT-base υπερέχουν των cased εκδόσεών τους, αν και αυτή η διαφορά δεν είναι στατιστικά σημαντική. Επίσης, τα μοντέλα bert-base (cased και uncased) ξεπερνούν συνολικά τα distilbert αντίστοιχα. Αυτό το αποτέλεσμα ήταν εν μέρει αναμενόμενο, καθώς τα μοντέλα bert-base είναι προεκπαιδευμένα σε μεγαλύτερα σύνολα δεδομένων. Ωστόσο, υπάρχουν περιπτώσεις όπου ελαφρύτερα μοντέλα έχουν εμφανίσει παρόμοια αποδοτι-



Σχήμα 7.1: Απόσοδη αλγορίθμων στο σενάριο όπου δε χρησιμοποιείται μηχανική μάθηση

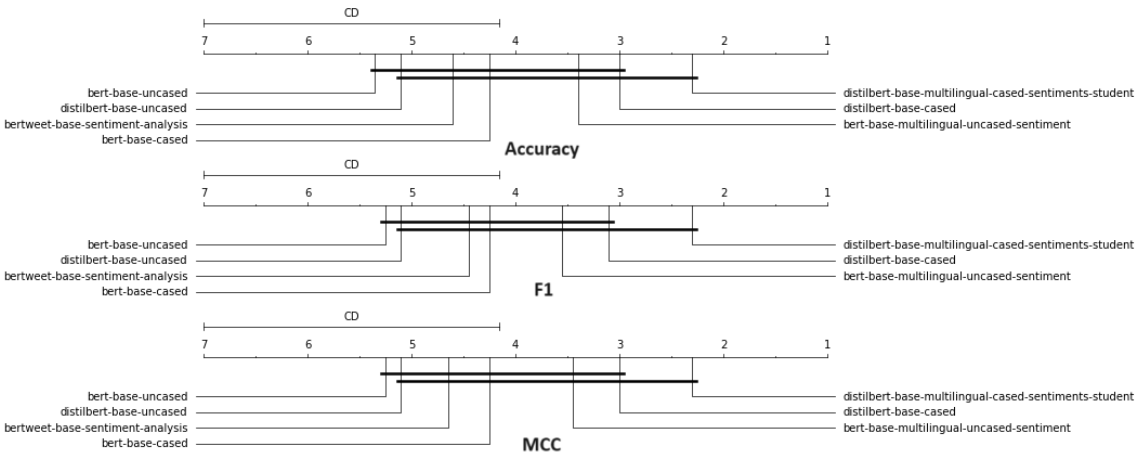
κότητα, ανάλογα με το πλαίσιο εφαρμογής. Σε κάθε περίπτωση, η έλλειψη σημαντικής διαφοράς στην απόδοση υπογραμμίζει ότι τα μοντέλα distilBERT — αν και ελαφρύτερα — αποτελούν μια ιδιαίτερα αποδοτική εναλλακτική λύση.

7.3.2 Συγκρίσεις όπου εμπλέκεται Ενεργή Μάθηση

Προχωρώντας στα αποτελέσματα που σχετίζονται με το βασικό σκέλος της έρευνας, η παράθεση των ευρημάτων, θα ξεκινήσει με το κεντρικό ερώτημα του αν η χρήση της Ενεργής Μάθησης μπορεί τελικά να επιφέρει το ίδιο ικανοποιητικά αποτελέσματα σε σχέση με τις μετρικές αξιολόγησης, με τη χρήση όμως σημαντικά λιγότερων ετικετοποιημένων δεδομένων. Για να μπορέσουμε όμως να απαντήσουμε σε αυτό το ερώτημα, θα πρέπει να ορίζουμε με σαφήνεια το τι σημαίνει «σχετικά παρόμοια αποτελέσματα» στο πλαίσιο της έρευνάς μας. Έτσι, στο πλαίσιο της έρευνας θεωρούμε ότι έχουμε σχετικά παρόμοια συμπεριφορά όταν τα αποτελέσματα των αντίστοιχων μετρικών διαφέρουν λιγότερο από 1%.

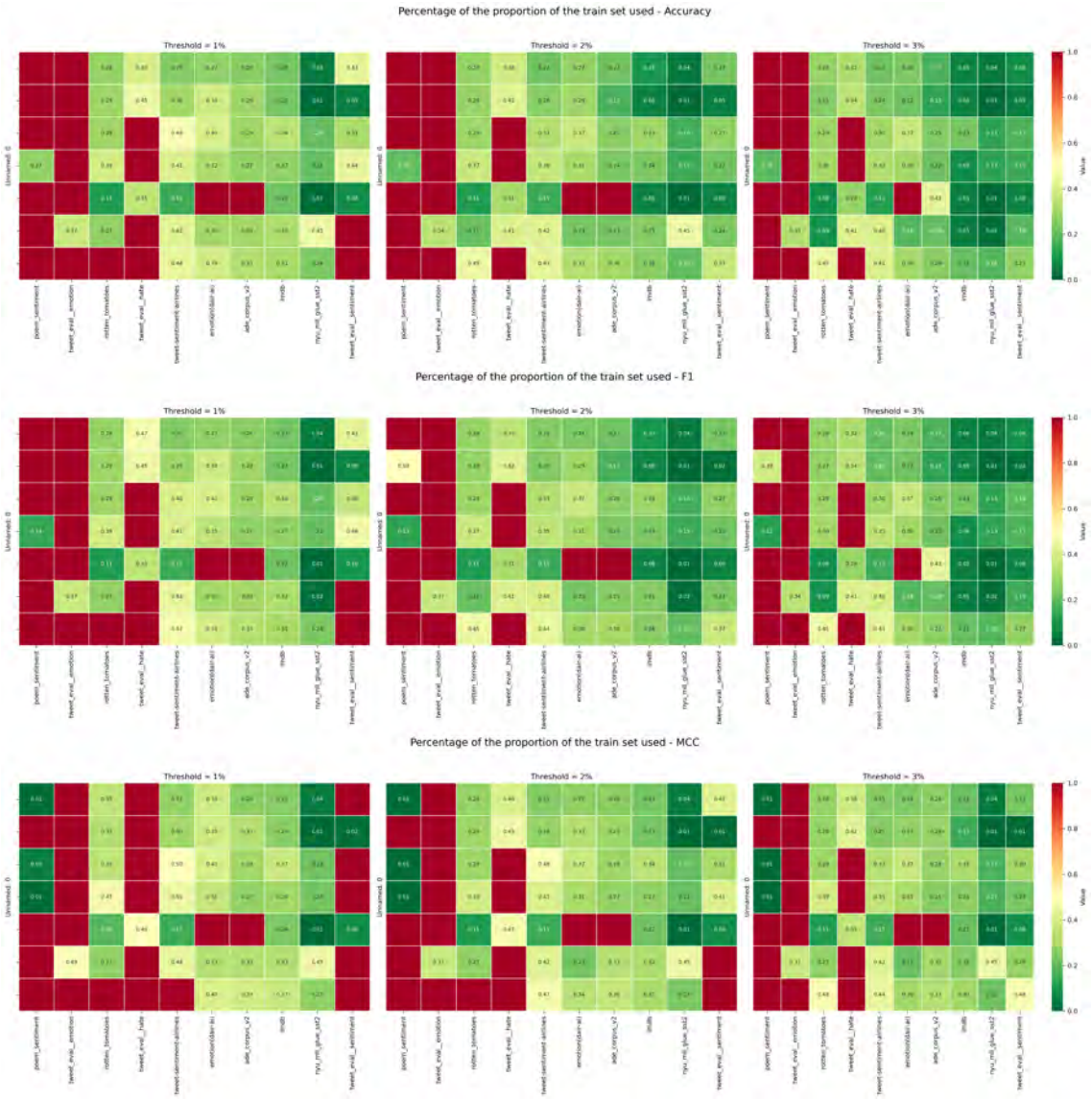
Για να έχουμε μια ξεκάθαρη εικόνα για το πότε το προαναφερθέν κριτήριο ικανοποιείται, δημιουργήθηκαν τα heatmaps του Σχήματος 7.3. Κάθε τριάδα heatmap αφορά μια μετρική, όταν το παραπάνω κατώφλι ανεκτικότητας για να θεωρηθεί μια περίπτωση επιτυχημένη έχει τεθεί όχι μόνο στο 1%, αλλά και στο 2% και 3% αντίστοιχα, ώστε να υπάρχει καλύτερη εικόνα του αν η αυθαίρετη αυτή επιλογή επηρεάζει σε μεγάλο βαθμό τα αποτελέσματα. Κάθε τετράγωνο αντιστοιχεί σε ένα ζεύγος αλγορίθμου και συνόλου δεδομένων. Ο κόκκινος χρωματισμός υποδεικνύει ότι ακόμη και όταν χρησιμοποιήθηκε το 50% των δεδομένων, η χρήση της Ενεργής Μάθησης δεν υπήρξε ευεργετική και δεν καταφέραμε να έχουμε παρόμοια απόδοση με λιγότερα δεδομένα. Αντίθετα, οι αποχρώσεις του πράσινου δείχνουν ότι έχουμε αριθμητικά παρόμοια αποτελέσματα, ενώ ο αριθμός στο κέντρου του τετραγώνου δηλώνει την απόλυτη διαφορά της τιμής της μετρικής στα δύο σενάρια – με και χωρίς χρήση Ενεργής Μάθησης.

Από τα προαναφερθέντα heatmaps παρατηρούμε ότι όσον αφορά την ακρίβεια (accuracy), η Ενεργή Μάθηση επέτρεψε την επίτευξη του ίδιου σκορ στο 70% των περιπτώσεων (49 από 70) με κατώφλι 1%, ενώ το ποσοστό αυτό αυξήθηκε στο 76% (53 από 70) με κατώφλι 2% και στο 77% (54 από 70) με κατώφλι 3%. Παρόμοια τάση παρατηρήθηκε για το F1-score, όπου το 70% (49 από 70) των περιπτώσεων έφτασε το επιθυμητό σκορ



Σχήμα 7.2: Διάγραμμα κρίσιμων διαφορών της σύγκρισης των μεθόδων στο σενάριο όπου δε χρησιμοποιείται Ενεργή Μάθηση

με κατώφλι 1%, ενώ τα ποσοστά αυξήθηκαν στο 77% (54 από 70) με κατώφλι 2% και στο 78,5% (55 από 70) με κατώφλι 3%. Για τον MCC, ακολουθήθηκε το ίδιο μοτίβο, με 64% (45 από 70) των περιπτώσεων να φτάνουν το στόχο με κατώφλι 1%, ενώ τα ποσοστά αυξήθηκαν στο 73% (51 από 70) με κατώφλι 2% και στο 77% (54 από 70) με κατώφλι 3%. Ωστόσο, σε περιπτώσεις όπου οι επιθυμητές μετρικές δεν επιτεύχθηκαν με κατώφλι 1%, αναλύθηκε η ελάχιστη διαφορά μεταξύ των σκορ που επιτεύχθηκαν χωρίς Ενεργή Μάθηση (με το πλήρες σύνολο εκπαίδευσης) και των σκορ που επιτεύχθηκαν με Ενεργή Μάθηση, ώστε να καθοριστεί το ελάχιστο ποσοστό δεδομένων που απαιτείται για την επίτευξη του βέλτιστου επιπέδου απόδοσης. Συγκεκριμένα, για την ακρίβεια (accuracy) και το F1-score, η επιθυμητή απόδοση επιτεύχθηκε στο 50% των περιπτώσεων (35 από 70) χρησιμοποιώντας λιγότερο από 35% των αρχικών δεδομένων εκπαίδευσης, ενώ για τον MCC, το όριο αυτό ήταν επαρκές στο 43% των περιπτώσεων (30 από 70).



Σχήμα 7.3: Ποσοστό χρήσης του αρχικού συνόλου δεδομένων στο σενάριο της Ενεργής Μάθησης στο οποίο επιτυγχάνεται παρόμοια ακρίβεια με απόκλιση 1%, 2% και 3% αντίστοιχα

Συνεχίζοντας με τις οπτικοποιήσεις στο Σχήμα 7.4, παρουσιάζονται ραβδογράμματα που απεικονίζουν την κατανομή των συνόλων δεδομένων που πέτυχαν το επιθυμητό σκορ με κατώφλι 1% (πράσινο χρώμα). Αντίθετα, οι περιπτώσεις όπου αυτό δεν επιτεύχθηκε απεικονίζονται με ροζ, υποδεικνύοντας ότι απαιτήθηκε μεγαλύτερο ποσοστό δεδομένων εκπαίδευσης για να φτάσουν στο κοντινότερο δυνατό σκορ. Είναι εμφανές ότι στις περισσότερες περιπτώσεις, η επίτευξη των επιθυμητών αποτελεσμάτων απαιτεί τη χρήση περισσότερων δεδομένων εκπαίδευσης, αν και η μείωση είναι σημαντική. Μια εναλλακτική οπτικοποίηση αυτών των πληροφοριών, που εστιάζει σε συγκεκριμένα ζεύγη μεθόδων και συνόλων δεδομένων, φαίνεται στο Σχήμα 7.5, όπου κάθε κύκλος αντιπροσωπεύει έναν συνδυασμό μεθόδου και συνόλου δεδομένων. Πράσινοι κύκλοι υποδηλώνουν περιπτώσεις όπου το ίδιο σκορ επιτεύχθηκε με λιγότερο από 1% δεδομένων, σε σύγκριση με το σενάριο χωρίς Ενεργή Μάθηση, ενώ οι ροζ κύκλοι δείχνουν περιπτώσεις όπου το κατώφλι δεν ικανοποιήθηκε. Το μέγεθος κάθε κύκλου αντιπροσωπεύει το ποσοστό του πλήρους συνόλου εκπαίδευσης που χρησιμοποιήθηκε στο σενάριο χρήσης Ενεργής Μάθησης. Για τους πράσινους κύκλους, το μέγεθος του κύκλου δείχνει το ποσοστό του συνόλου εκπαίδευσης που απαιτήθηκε για την επίτευξη του επιθυμητού σκορ, ενώ για τους ροζ κύκλους, το μέγεθος δείχνει το ποσοστό δεδομένων που απαιτήθηκε για την επίτευξη του πλησιέστερου δυνατού σκορ. Συνεπώς, όσο μεγαλύτερος είναι ο κύκλος, τόσο μεγαλύτερο ποσοστό δεδομένων χρησιμοποιήθηκε στην περίπτωση χρήσης Ενεργής Μάθησης.

Αξιοσημείωτο είναι ότι οι περισσότεροι ροζ κύκλοι σχετίζονται με τα σύνολα δεδομένων poem sentiment και tweet eval emotion, τα οποία έχουν τον μικρότερο αριθμό δειγμάτων. Αντίθετα, η χρήση Ενεργής Μάθησης παρείχε σημαντικά οφέλη σε μεγαλύτερα σύνολα δεδομένων, επιτυγχάνοντας παρόμοια σκορ με εκείνα που προέκυψαν από την εκπαίδευση σε ολόκληρο το σύνολο δεδομένων. Αυτό υποδηλώνει ότι η χρήση Ενεργής Μάθησης είναι ιδιαίτερα αποτελεσματική σε περιβάλλοντα με μεγαλύτερα σύνολα δεδομένων, καθώς επιτρέπει τη μείωση του απαιτούμενου όγκου δεδομένων εκπαίδευσης, διατηρώντας παράλληλα την απόδοση.

Αν και οι παραπάνω οπτικοποιήσεις δίνουν μια ικανοποιητική εικόνα των αποτελεσμάτων και του πως η Ενεργή Μάθηση βοηθά στην μείωση της ανάγκης για ετικετοποιημένα δεδομένα. Παρ΄ όλ΄ αυτά είναι ζωτικής σημασίας να υπάρχει μια μέθοδος παρακολούθησης της αποδοτικότητας της Ενεργής Μάθησης, ιδιαίτερα σε περιπτώσεις όπως η παρούσα όπου χρησιμοποιείται για τη μείωση του όγκου των δεδομένων εκπαίδευσης. Μια τέτοια μέθοδος πρέπει να λαμβάνει υπόψη όχι μόνο την απόδοση του μοντέλου βάσει συγκεκριμένων μετρικών, αλλά και το ποσοστό δεδομένων που απαιτήθηκε για την επίτευξη αυτής της απόδοσης. Για τον σκοπό αυτό, όπως αναφέρθηκε νωρίτερα, χρησιμοποιήσαμε τον Δείκτη Απόδοσης Προσαρμοσμένο στην Αποδοτικότητα (Efficiency-Adjusted Performance Index - EAPI), έναν νέο μέτα-δείκτη που αξιολογεί την αποτελεσματικότητα της Ενεργής Μάθησης στη μείωση των απαιτούμενων δεδομένων εκπαίδευσης, διατηρώντας ή βελτιώνοντας τα επίπεδα απόδοσης.

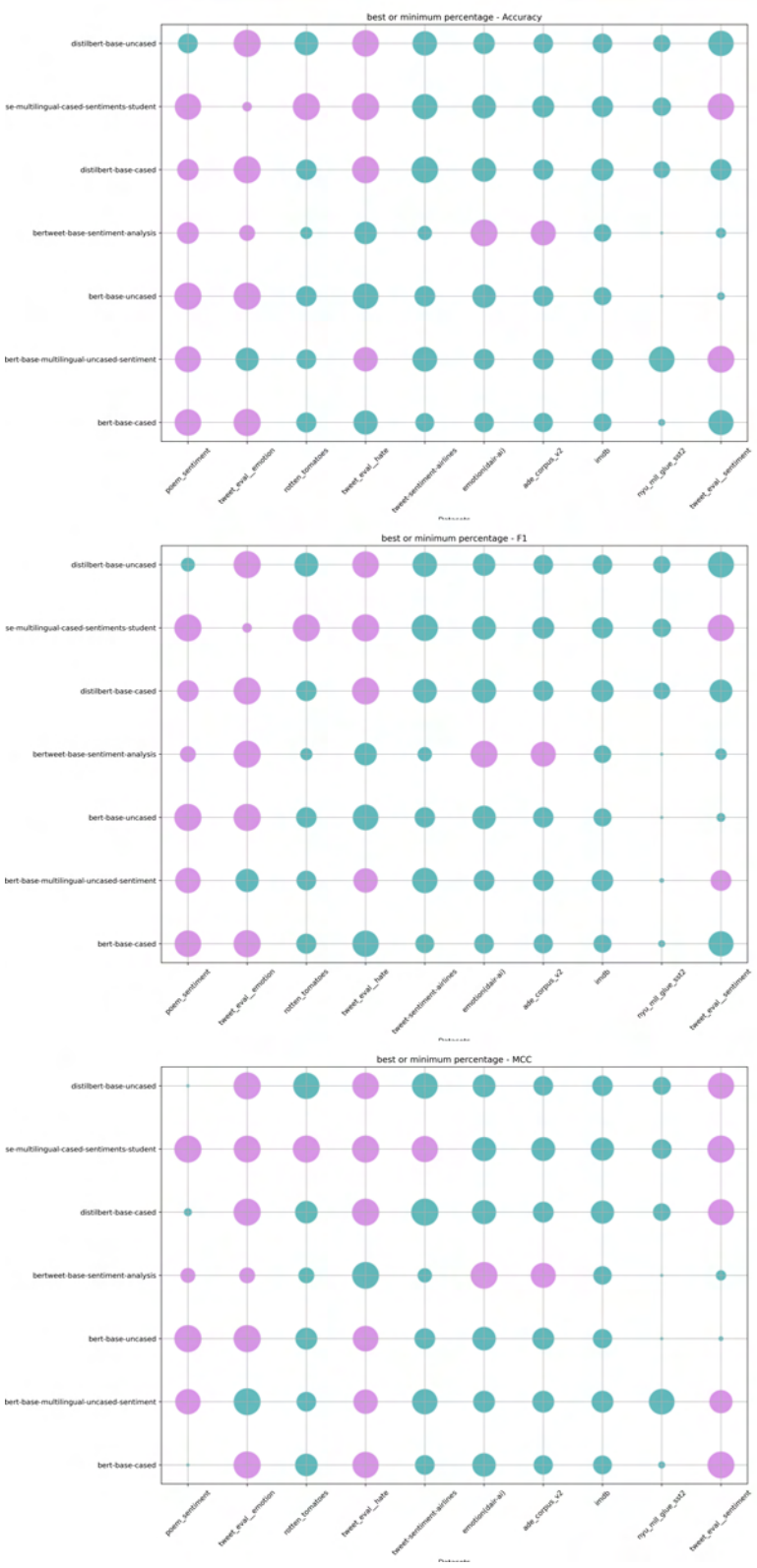
Σε αντίθεση με τις παραδοσιακές μετρικές, οι οποίες εστιάζουν αποκλειστικά στην ακρίβεια των προβλέψεων, ο EAPI ενσωματώνει και την ποσότητα των δεδομένων που χρησιμοποιήθηκαν, επιτρέποντας μια πιο ολοκληρωμένη εκτίμηση της απόδοσης και της αποδοτικότητας της διαχείρισης των δεδομένων. Ο υπολογισμός του βασίζεται στη σύγκριση της απόδοσης του μοντέλου με και χωρίς χρήση Ενεργής Μάθησης, ενώ παράλληλα περιλαμβάνει έναν παράγοντα ποινής που εξαρτάται από το ποσοστό του αρχικού συνόλου δεδομένων που χρησιμοποιήθηκε. Έτσι, υψηλότερη τιμή EAPI υποδηλώνει ότι η Ενεργή Μάθηση επιτυγχάνει παρόμοια ή καλύτερη απόδοση με

τη χρήση μικρότερου ποσοστού δεδομένων, υποδεικνύοντας βελτιστοποιημένη διαχείριση των δεδομένων. Αντίθετα, χαμηλότερη τιμή EAPI δείχνει περιορισμένα κέρδη απόδοσης ή αυξημένη χρήση δεδομένων, γεγονός που υποδεικνύει λιγότερο αποδοτική αξιοποίηση του συνόλου εκπαίδευσης.

Τα heatmaps στο Σχήμα 7.6 απεικονίζουν τη βαθμολογία EAPI για κάθε συνδυασμό μεθόδου και συνόλου δεδομένων, βάσει των τριών βασικών μετρικών αξιολόγησης. Ο δείκτης αυτός αντιπροσωπεύει το ελάχιστο ποσοστό δεδομένων που χρησιμοποιήθηκε, διατηρώντας συγκρίσιμη απόδοση με το σενάριο χωρίς χρήση Ενεργής Μάθησης. Σε περιπτώσεις όπου δεν ήταν δυνατή η επίτευξη συγκρίσιμης απόδοσης, ο EAPI υπολογίστηκε με βάση το ελάχιστο ποσοστό στο οποίο καταγράφηκε η μικρότερη διαφορά μεταξύ των σκορ με και χωρίς Ενεργή Μάθηση. Αυτοί οι θερμικοί χάρτες παρέχουν μια συνολική εικόνα του συμβιβασμού μεταξύ αποδοτικότητας δεδομένων και απόδοσης



Σχήμα 7.4: Ραβδογράμματα κατανομής των περιπτώσεων σε κλάσεις ποσοστού χρήσης του συνολικού συνόλου δεδομένου στις οποίες το σενάριο Μηχανικής Μάθησης επιτυγχάνει με απόκλιση μικρότερη του 1% (πράσινο) ή μεγαλύτερη του 1% (ροζ) παραπλήσιο σκορ στην κάθε μετρική σε σχέση με το σενάριο όπου αυτή δε χρησιμοποιείται



Σχήμα 7.5: Οπτική απεικόνηση του ποσοστού του πλήρους συνόλου δεδομένων για κάθε ζεύγος μεθόδου - συνόλου δεδομένων, στο οποίο η χρήση Ενεργής Μάθησης έχει παρόμοιο αποτέλεσμα με όταν αυτή δε χρησιμοποιείται με απόκλιση μικρότερη του 1% (πράσινο) και μεγαλύτερη του 1% (ροζ)

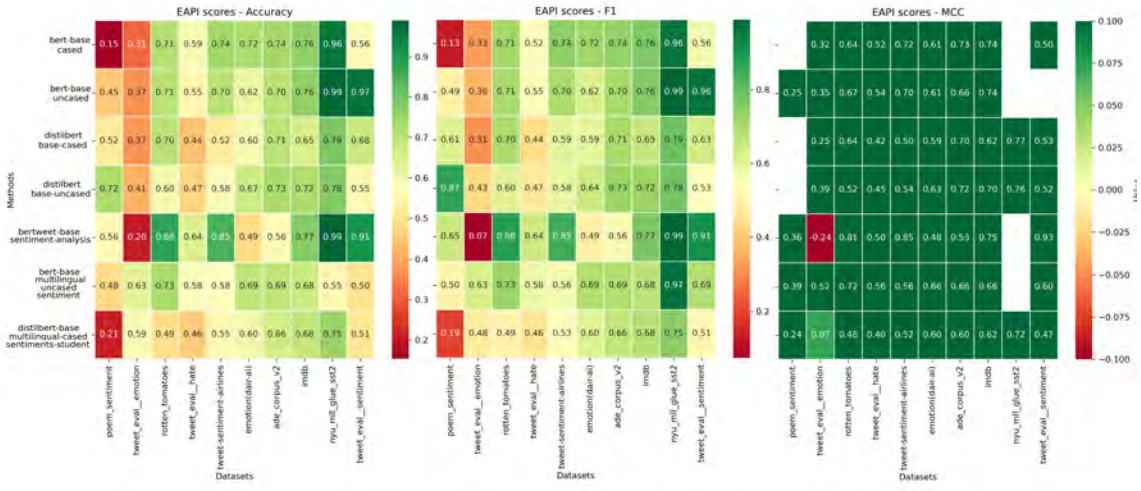
του μοντέλου, αναδεικνύοντας τις περιπτώσεις όπου η χρήση Ενεργής Μάθησης ήταν περισσότερο ή λιγότερο ωφέλιμη.

Στις περισσότερες περιπτώσεις, ο EAPI δείχνει ότι η Ενεργή Μάθηση ήταν επωφελής, καθώς οι σχετικά υψηλές τιμές του καταδεικνύουν αποτελεσματική μείωση δεδομένων χωρίς σημαντική απώλεια απόδοσης. Ωστόσο, σε συγκεκριμένα σενάρια, χαμηλές τιμές EAPI υποδηλώνουν ότι η χρήση Ενεργής Μάθησης δεν λειτούργησε αποτελεσματικά, καθώς τα μοντέλα δεν πέτυχαν τις ίδιες επιδόσεις με το σενάριο χωρίς χρήση Ενεργής Μάθησης. Αυτή η διαφορά μεταξύ υψηλών και χαμηλών τιμών EAPI επιτρέπει την αναγνώριση των συνθηκών υπό τις οποίες η Ενεργή Μάθηση είναι πραγματικά χρήσιμη ή, αντίθετα, αποδεικνύεται λιγότερο αποδοτική.

Στο πλαίσιο της ανάλυσης του MCC, παρατηρήθηκαν περιπτώσεις όπου τα heatmaps δεν εμφάνιζαν χρωματισμούς. Αυτό συνέβη όταν είτε μία είτε και οι δύο τιμές MCC (για τα σενάρια με και χωρίς χρήση Ενεργής Μάθησης) ήταν μηδενικές, καθιστώντας αδύνατο τον υπολογισμό μιας ουσιαστικής τιμής MCC. Αυτές οι περιπτώσεις αντικατοπτρίζουν κακή απόδοση του μοντέλου. Παρόλα αυτά, ο κύριος στόχος της παρούσας ανάλυσης δεν είναι η απόλυτη απόδοση των μοντέλων, αλλά η σύγκριση της απόδοσής τους μεταξύ των σεναρίων με και χωρίς χρήση Ενεργής Μάθησης, ώστε να αναδειχθούν οι περιπτώσεις όπου η Ενεργής Μάθησης συμβάλλει ουσιαστικά στη βελτιστοποίηση της διαδικασίας εκπαίδευσης.

Ένα επιπλέον εύρημα που διαφαίνεται στα αποτελέσματα είναι το πως το μέγεθος του συνόλου δεδομένων φαίνεται να επηρεάζει σημαντικά την αποδοτικότητα της Ενεργής Μάθησης στη μείωση του μεγέθους εκπαίδευσης. Στα Σχήματα 7.3 και 7.5 είναι εμφανές ότι σε μικρότερα σύνολα δεδομένων, σχεδόν όλες οι μέθοδοι αποτυγχάνουν να επιτύχουν τις ίδιες βαθμολογίες μετρικών σε σύγκριση με τα σενάρια χωρίς AL. Αυτό δεν προκαλεί έκπληξη, καθώς η χρήση του ίδιου ποσοστού του συνόλου εκπαίδευσης σε σύνολα δεδομένων διαφορετικού μεγέθους οδηγεί σε μεταβλητό αριθμό παραδειγμάτων εκπαίδευσης, επηρεάζοντας την ποιότητα της μάθησης.

Επιπλέον, τα σύνολα δεδομένων αυτά είναι έντονα ανισοβαρή, γεγονός που καθιστά το πρόβλημα ταξινόμησης πιο απαιτητικό, ειδικά όσον αφορά την πρόβλεψη της μειονοτικής κλάσης. Στο σύνολο δεδομένων poem sentiment, η μειονοτική κλάση αποτελεί μόλις το 5,5% των δεδομένων, ενώ στο tweet eval sentiment, το ποσοστό αυτό



Σχήμα 7.6: Βαθμολογία EAPI για κάθε ζεύγος μεθόδου - συνόλου δεδομένων ανά μετρική

είναι περίπου 9%. Αυτή η ανισορροπία αυξάνει την πολυπλοκότητα της πρόβλεψης, καθώς το μοντέλο ενδέχεται να μην έχει επαρκή εκπαίδευση σε δείγματα της μειονοτικής κλάσης. Αντίθετα, η διαφορά μεταξύ δυαδικής ταξινόμησης και ταξινόμησης πολλαπλών κλάσεων δεν φαίνεται να επηρεάζει σημαντικά την ικανότητα της Ενεργής Μάθησης να μειώνει τον όγκο των απαιτούμενων δεδομένων εκπαίδευσης. Αν και τα δύο σύνολα δεδομένων με τις χαμηλότερες επιδόσεις είναι πολλαπλών κλάσεων (4 κλάσεις το καθένα), το σύνολο δεδομένων «emotion(dair-ai)», το οποίο διαθέτει 6 κλάσεις, παρουσιάζει υποσχόμενα αποτελέσματα. Αξίζει να σημειωθεί ότι και αυτό το σύνολο είναι ιδιαίτερα ανισοβαρές, με τη μειονοτική κλάση να αντιπροσωπεύει μόλις το 3,6% των δεδομένων.

Τα παραπάνω ευρήματα υποδηλώνουν ότι παράγοντες πέρα από τον αριθμό των κλάσεων—όπως τα χαρακτηριστικά του ίδιου του συνόλου δεδομένων και η ανισορροπία μεταξύ των κλάσεων—ενδέχεται να διαδραματίζουν πιο κρίσιμο ρόλο στην επιτυχία της Ενεργής Μάθησης σε αυτά τα σενάρια.

7.4 Συμπεράσματα

Η παρούσα μελέτη μέσω της εκτεταμένης πειραματικής διαδικασίας επιβεβαίωσε ότι η ενσωμάτωση της Ενεργής Μάθησης στη διαδικασία ταξινόμησης κειμένων με μοντέλα βασισμένα σε BERT μπορεί να μειώσει σημαντικά την ανάγκη για επισημασμένα δεδομένα, διατηρώντας παράλληλα υψηλή απόδοση. Συγκεκριμένα, τα πειραματικά αποτελέσματα έδειξαν ότι σε ποσοστό 70% των περιπτώσεων, ήταν δυνατή η μείωση του μεγέθους του συνόλου εκπαίδευσης κατά τουλάχιστον 50% χωρίς αξιοσημείωτη επίδραση στις μετρικές αξιολόγησης. Επιπλέον, η αποδοτικότητα της χρήσης Ενεργής Μάθησης παρατηρήθηκε κυρίως σε μεγαλύτερα σύνολα δεδομένων, ενώ η απόδοσή της επηρεάστηκε αρνητικά σε μικρότερα και ανισοκατανεμημένα σύνολα, λόγω της δυσκολίας στη σωστή κατηγοριοποίηση των λιγότερο εκπροσωπούμενων κατηγοριών. Προτάθηκε επίσης ένας νέος μέτα-δείκτης EAPI που μπορεί να χρησιμοποιηθεί στην αξιολόγηση παρόμοιων σεναρίων όπως εξετάζεται η αποδοτικότητα του σεναρίου στο ποιο αξιολογείται η χρήση της ενεργητικής μάθησης έναντι του απλού σεναρίου, σε σχέση με το ποσοστό χρήσης των δεδομένων που χρησιμοποιήθηκαν. Στην παρούσα μελέτη, ο δείκτης αυτός έδειξε ότι η Ενεργή Μάθηση μπορεί να είναι ιδιαίτερα αποδοτική, επιτυγχάνοντας υψηλές επιδόσεις με λιγότερα δεδομένα. Ωστόσο, υπήρχαν περιπτώσεις όπου το όφελος ήταν περιορισμένο, υποδηλώνοντας ότι η επιλογή των δεδομένων που προστίθενται κατά τη διαδικασία χρήσης Ενεργής Μάθησης είναι κρίσιμη για τη συνολική αποδοτικότητα της μεθόδου. Επιπλέον, παρατηρήθηκε σχέση του μεγέθους του αρχικού συνόλου δεδομένων, καθώς στα μικρότερα σύνολα δεδομένων η χρήση Ενεργής Μάθησης δεν ήταν ευεργετική. Ως προς την εγκυρότητα των αποτελεσμάτων, η χρήση δημόσιων συνόλων δεδομένων και ανοιχτού κώδικα μοντέλων διασφαλίζει την αναπαραγωγιμότητα, αν και η ανάλυση δεν κάλυψε εξαιρετικά μεγάλα σύνολα δεδομένων λόγω υπολογιστικών περιορισμών. Μελλοντικές επεκτάσεις θα μπορούσαν να περιλαμβάνουν τη δοκιμή πιο εξελιγμένων στρατηγικών ενεργής μάθησης και την εφαρμογή της σε διαφορετικά είδη δεδομένων, όπως εικόνες ή δεδομένα βιοϊατρικής.

8. Πρόβλεψη Σφαλμάτων Λογισμικού: Βελτιστοποίηση Προβλέψεων με Ενεργή Μάθηση και Ομαδοποιητές Ταξινομητών

8.1 Εισαγωγή

Στο παρόν κεφάλαιο διερευνάται η εφαρμογή της Ενεργητικής Μάθησης σε συνδυασμό με τη χρήση ομαδοποιητών ταξινομητών για το πρόβλημα της πρόβλεψης σφαλμάτων λογισμικού. Ο σκοπός της μελέτης είναι διττός: αφενός, υποστηρίζεται ότι η αξιοποίηση τεχνικών Μηχανικής Μάθησης μπορεί να οδηγήσει σε εξίσου αξιόπιστα αποτελέσματα πρόβλεψης με μειωμένες απαιτήσεις εκπαίδευσης· αφετέρου, κρίνεται σκόπιμη η ενσωμάτωση ομαδοποιητών ταξινομητών στην πειραματική διαδικασία, προκειμένου να διερευνηθεί κατά πόσο η συγκεκριμένη προσέγγιση μπορεί να παρέχει καλύτερα αποτελέσματα από τη χρήση απλών ταξινομητών.

Ως σφάλμα λογισμικού ορίζεται μια αστοχία ή μια βλάβη σε ένα λογισμικό, η οποία οδηγεί σε απρόσμενα ή εσφαλμένα αποτελέσματα [141]. Η ύπαρξη σφαλμάτων, μπορεί να οδηγήσει σε σοβαρά προβλήματα που μπορεί να επηρεάζουν τόσο τους τελικούς χρήστες, όσο και τις εταιρίες που αναπτύσσουν λογισμικό. Η ύπαρξη σφαλμάτων εκτός από δυσκολίες στην χρήση αλλά και μειωμένη ικανοποίηση των χρηστών μπορεί να οδηγήσει και σε ποικίλες άλλες μη επιθυμητές καταστάσεις όπως απρόσμενες διακοπές στη λειτουργία του λογισμικού, απώλεια δεδομένων αλλά και δημιουργία τρωτών σημείων που μειώνουν την ασφάλεια και την αξιοπιστία του λογισμικού. Η ύπαρξη σφαλμάτων σε εξειδικευμένα λογισμικά που χρησιμοποιούνται από εταιρίες ή άλλους οργανισμούς μπορεί να επιφέρει οικονομική ζημία αλλά και να μειώσει την εμπιστοσύνη των συμβαλλομένων. Οι επιπτώσεις μπορεί να είναι ακόμη πιο κρίσιμες, σε τομείς όπως η υγεία όπου αστοχίες λογισμικού μπορεί να οδηγήσουν σε ατυχήματα ή και απώλειες ζωών ή στη βιομηχανία, όπου σφάλματα μπορεί να δημιουργήσουν διαρροές ή περιβαλλοντικές καταστροφές.

Σε κάθε περίπτωση, η ύπαρξη σφαλμάτων οδηγεί στην ανάγκη για αντιμετώπιση και διόρθωσή τους, διαδικασία που συνοδεύεται με κόστος σε χρόνο και χρήμα. Έρευνα του 2012, αναφέρει πως το κόστος εντοπισμού και διόρθωσης σφαλμάτων είναι από τις πιο ακριβές δραστηριότητες στην ανάπτυξη λογισμικού. Επιπλέον, το κόστος διόρθωσης ενός σφάλματος λογισμικού αυξάνεται ανάλογα με το στάδιο ανάπτυξης του λογισμικού. Η διόρθωση ενός σφάλματος έχει χαμηλότερο κόστος στη φάση συγγραφής κώδικα, ενώ το κόστος πολλαπλασιάζεται στη φάση των δοκιμών του λογισμικού και γίνεται ακόμη μεγαλύτερο κατά τη διαδικασία συντήρησης [283]. Ενδεικτικά, η διαδικασία εντοπισμού και διόρθωσης ενός σφάλματος κοστίζει 977$ στη φάση συγγραφής του κώδικα, 7,136$ κατά τη φάση δοκιμών, ενώ το ποσό σχεδόν διπλασιάζεται και αγκίζει τις 7,136$ στη φάση συντήρησης [30]. Επιπλέον, είναι επίσης προφανές πως ο εντοπισμός και η διόρθωση σφαλμάτων κοστίζει πολύ λιγότερο κατά την ανάπτυξη του λογισμικού και το κόστος εκτοξεύεται μετά την κυκλοφορία του. Συνεπώς, ο

έγκαιρος εντοπισμός των σφαλμάτων, αποτελεί κομβικό σημείο κατά την δημιουργία λογισμικού.

8.2 Πειραματική διαδικασία

Στην υποενότητα αυτή θα δοθούν οι λεπτομέρειες της πειραματικής διαδικασίας που ακολουθήθηκε για την διερεύνηση των ζητημάτων που αποτέλεσαν έναυσμα για την ενασχόλησή μας με το συγκεκριμένο θέμα. Θα ακολουθήσουν πληροφορίες για τις μεθόδους, τα σύνολα δεδομένων τα σχήματα ενεργητικής μάθησης, αλλά και πληροφορίες για τον νέο δείκτη που προτείναμε για την αξιολόγηση των σχημάτων ενεργητικής μάθησης που υλοποιήθηκαν.

8.2.1 Δεδομένα

Όσον αφορά τα σύνολα δεδομένων που χρησιμοποιήθηκαν, είναι σημαντικό να επισημανθεί ότι όλα επιλέχθηκαν με γνώμονα να είναι δημόσια διαθέσιμα. Η διαθεσιμότητά τους εξασφαλίζει την αναπαραγωγιμότητα των ερευνητικών αποτελεσμάτων και επιτρέπει τη σύγκριση με παρόμοιες μελέτες, είτε ήδη δημοσιευμένες είτε μελλοντικές, καθώς η πρακτική του διαμοιρασμού δεδομένων προάγει τη διαφάνεια και διευκολύνει την περαιτέρω ερευνητική πρόοδο.

Στην παρούσα μελέτη, χρησιμοποιήθηκαν συνολικά δέκα σύνολα δεδομένων, τα οποία τροποποιήθηκαν με τρόπο τέτοιο ώστε να είναι κατάλληλα για δυαδική ταξινόμηση. Τα πρώτα πέντε σύνολα δεδομένων προήλθαν από το PROMISE Software Engineering Repository [251], το οποίο εμπεριείχε δεδομένα από το NASA Metrics Data Program. Τα υπόλοιπα πέντε σύνολα δεδομένων προέρχονται από το bug prediction dataset [55] και αφορούν πέντε ξεχωριστά συστήματα λογισμικού. Η επιλογή αυτών των συνόλων δεδομένων διασφαλίζει την ποικιλομορφία και τη συνάφεια των πειραματικών δεδομένων, συμβάλλοντας σε μια ισχυρή και ολοκληρωμένη ερευνητική μεθοδολογία.

Τα πρώτα πέντε σύνολα δεδομένων αντλήθηκαν από το NASA Metrics Data Program και καθένα από αυτά περιλαμβάνει πληροφορίες από ξεχωριστά όργανα ή συστήματα. Συγκεκριμένα:

- **CM1**: Το CM1 προέρχεται από ένα έργο λογισμικού της NASA για συστήματα ελέγχου. Είναι υλοποιημένο σε γλώσσα C.
- **JM1**: Το JM1 είναι ένα σύστημα πραγματικού χρόνου πρόβλεψης εδάφους που χρησιμοποιεί προσομοιώσεις για τη δημιουργία προβλέψεων. Όπως και το CM1, είναι υλοποιημένο σε C.
- **KC1**: Το KC1 περιέχει δεδομένα από ένα σύστημα C++ που αφορά τη διαχείριση αποθήκευσης για τη λήψη και επεξεργασία δεδομένων εδάφους.
- **KC2**: Το KC2 αποτελεί τμήμα του ίδιου έργου με το KC1, το οποίο, αν και μοιράζεται με αυτό κάποιες βιβλιοθήκες τρίτων, δεν έχει κάποια άλλη αλληλοεπικάλυψη.
- **PC1**: Το PC1 περιέχει δεδομένα που εξήχθησαν από συναρτήσεις C στο λογισμικό πτήσης ενός δορυφόρου σε τροχιά γύρω από τη Γη.

Σύνολο δεδομένων	Χαρακτηριστικά	Μέγεθος	Defected	Non-defected
CM1	21	498	49	449
JM1	21	10885	2106	8779
KC1	21	2019	326	1783
KC2	21	522	107	415
PC1	21	1109	77	1032
Eclipse JDT Core	17	997	206	791
Eclipse PDE UI	17	1497	209	1288
Equinox framework	17	324	129	195
Mylyn	17	1862	245	1617
Apache Lucene	17	691	64	627

Πίνακας 8.1: Χαρακτηριστικά των συνόλων δεδομένων

Και τα πέντε σύνολα δεδομένων περιέχουν πληροφορίες όπως ο αριθμός γραμμών του κώδικα, μετρικές πολυπλοκότητας (π.χ., McCabe, Halstead), αλλά και άλλες μετρικές λογισμικού, ενώ η κλάση στόχου υποδεικνύει αν μια μονάδα είναι επιρρεπής σε σφάλματα ή όχι.

Τα υπόλοιπα πέντε σύνολα δεδομένων που προέρχονται από το bug prediction dataset. Περιλαμβάνουν πληροφορίες σχετικά με το ιστορικό ανάπτυξης λογισμικού, κώδικα, και αναφορές σφαλμάτων. Τα πέντε συστήματα που καλύπτονται στο πλαίσιο της έρευνάς μας είναι τα Eclipse JDT Core, Eclipse PDE UI, Equinox, Mylyn, και Apache Lucene και είναι όλα υλοποιημένα σε Java. Συγκεκριμένα:

• **Eclipse JDT Core**: Το Eclipse JDT Core είναι το βασικό μέρος του Java Development Tools (JDT) του Eclipse IDE. Περιλαμβάνει βασικές λειτουργίες για την ανάλυση και μεταγλώττιση Java κώδικα. Είναι από τα πιο δημοφιλή σύνολα δεδομένων για έρευνα στην πρόβλεψη σφαλμάτων και χρησιμοποιείται εκτεταμένα για την ανάλυση της ποιότητας λογισμικού και την πρόβλεψη σφαλμάτων.

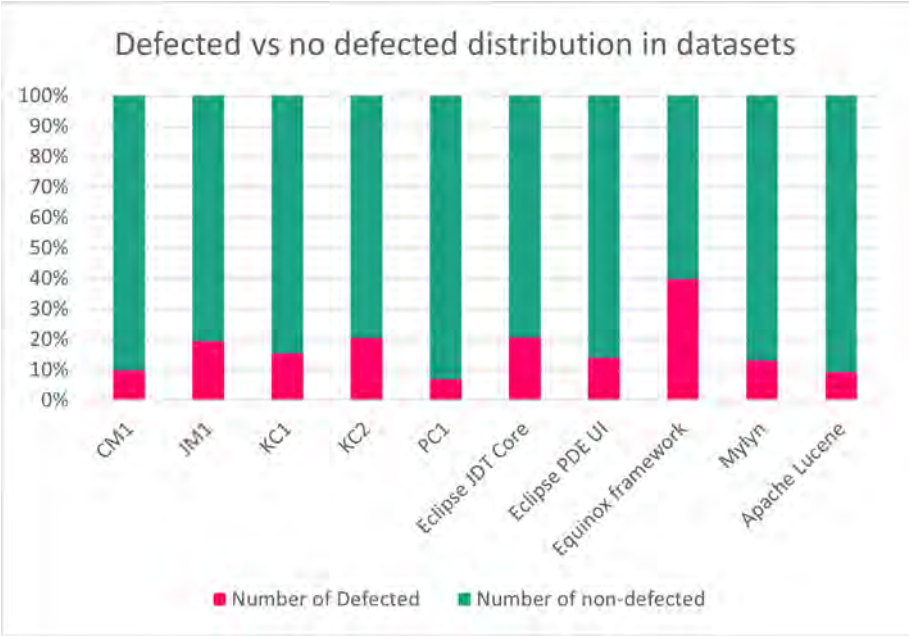
• **Eclipse PDE UI**: Το Eclipse PDE UI είναι ένα υποσύστημα του Eclipse που υποστηρίζει την ανάπτυξη και διαχείριση plugin λογισμικού. Σχεδιάστηκε για την ανάλυση προβλημάτων και σφαλμάτων σε λογισμικό που βασίζεται σε plugins.

• **Equinox**: Το Equinox είναι το OSGi-based runtime που αποτελεί τη βάση της Eclipse platform. Παρέχει μια modular προσέγγιση για την ανάπτυξη λογισμικού. Εστιάζει στην πρόβλεψη σφαλμάτων σε ένα modular περιβάλλον, ενώ συχνά χρησιμοποιείται για την ανάλυση εξαρτήσεων.

• **Mylyn**: Το Mylyn είναι ένα πλαίσιο διαχείρισης εργασιών (task management) για το Eclipse. Συνεργάζεται με σύστημα παρακολούθησης σφαλμάτων και παρέχει εργαλεία για προγραμματιστές. Εστιάζει κυρίως στις δυναμικές πτυχές της ανάπτυξης λογισμικού.

• **Apache Lucene**: Το Apache Lucene είναι μια high-performance βιβλιοθήκη για αναζήτηση κειμένου, που χρησιμοποιείται ευρέως σε μεγάλες πλατφόρμες και εφαρμογές. Συχνά χρησιμοποιείται για την πρόβλεψη σφαλμάτων σε λογισμικό μεγάλης κλίμακας και για τη μελέτη της σχέσης μεταξύ μετρικών πολυπλοκότητας και σφαλμάτων.

Και τα πέντε παραπάνω σύνολα περιέχουν μετρικές που αφορούν τον κώδικα, όπως ο αριθμός γραμμών του κώδικα, τον δείκτη πολυπλοκότητας McCabe, τον αριθμό μεθόδων (NOM) αλλά και τον αριθμό κλάσεων (NOC) που περιλαμβάνονται στο αρχείο,



Σχήμα 8.1: Κατανομή των δειγμάτων σε κλάσεις

καθώς και μετρικές Halstead, οι οποίες αφορούν την πολυπλοκότητα των τελεστών και των τελεστέων. Εκτός από μετρικές που αφορούν τον κώδικα, περιλαμβάνουν και μετρικές αλλαγών, όπως για παράδειγμα τον αριθμό των φορών που έχει τροποποιηθεί ένα αρχείο, τον αριθμό των αλλαγών που έχει υποστεί ο κώδικας (π.χ. προσθήκη ή αφαίρεση γραμμών), τον αριθμό των πρόσφατων επιδιορθώσεων σφαλμάτων, αλλά και τον αριθμό προγραμματιστών που έχουν εργαστεί στο αρχείο. Περιέχονται επίσης μετρικές που αφορούν το ιστορικό σφαλμάτων, όπως ο αριθμός σφαλμάτων ανά γραμμή ή το μέγεθος αρχείου και ο αριθμός σφαλμάτων που έχουν εντοπιστεί σε κάθε κλάση, καθώς και δομικές μετρικές που αποτυπώνουν σχέσεις εξάρτησης και σύζευξης με άλλες κλάσεις ή αρχεία.

Σημειώνεται ότι ορισμένα χαρακτηριστικά, όπως οι πληροφορίες για τη σοβαρότητα των υπαρχόντων σφαλμάτων, τροποποιήθηκαν σε ένα ενιαίο δυαδικό χαρακτηριστικό για να ευθυγραμμιστούν με το συγκεκριμένο πρόβλημα ταξινόμησης υπό εξέταση. Αυτή η προσαρμογή έγινε επειδή τα σύνολα δεδομένων είχαν αρχικά διαμορφωθεί για χρήση σε εργασίες παλινδρόμησης.

Χαρακτηριστικά όπως ο αριθμός των παραδειγμάτων, των χαρακτηριστικών αλλά και των αριθμό παραδειγμάτων σε κάθε μια από τις κλάσεις παρουσιάζονται στον πίνακα 8.1. Επιπλέον, στο σχήμα 8.1 γίνεται αντιληπτή η πληροφορία που αποτυπώνει το βαθμό δυσκολίας του εκάστοτε προβλήματος ταξινόμησης, δηλαδή του ποσοστό της κλάσης που εμφανίζεται σπανιότερα επί του συνόλου των παραδειγμάτων.

8.2.2 Ταξινομητές

Στο πλαίσιο της πειραματικής διαδικασίας χρησιμοποιήθηκαν δεκαπέντε κλασσικοί ταξινομητές οι οποίοι αναφέρονται συγκεντρωτικά στον Πίνακα 8.2 , μαζί με τις συντομογραφίες τους, με τις οποίες θα αναφέρονται από εδώ και στο εξής. Οι ίδιοι ταξινομητές χρησιμοποιήθηκαν και ως μέθοδοι βάσης στους ομαδοποιητές που δημιουργήθηκαν. Για την υλοποίηση των ταξινομητών χρησιμοποιήθηκε η βιβλιοθήκη PyCaret, πληροφορίες για την οποία έχουν δοθεί σε προηγούμενο κεφάλαιο.

№	Μέθοδος	Συντομογραφία
1	AdaBoost Classifier	ADA
2	Decision Tree Classifier	DT
3	Extra Trees Classifier	ET
4	Gradient Boosting Classifier	GBC
5	K-Nearest Neighbors	KNN
6	Light Gradient Boosting Machine	LightGBM
7	Linear Discriminant Analysis	LDA
8	Logistic Regression	LR
9	Multilayer Perceptron Classifier	MLP
10	Naive Bayes	NB
11	Quadratic Discriminant Analysis	QDA
12	Random Forest Classifier	RF
13	Ridge Classifier	RIDGE
14	Support Vector Machine - Linear Kernel	SVM
15	Support Vector Machine - Radial Kernel	RBFSVM

Πίνακας 8.2: Λίστα των μεθόδων ταξινόμησης που χρησιμοποιήθηκαν

8.2.3 Πρώτο στάδιο πειραματικής διαδικασίας: Σύγκριση απλών μεθόδων και δημιουργία ομαδοποιητών

Το πρώτο μέρος της πειραματικής διαδικασίας αφορά τη σύγκριση των κλασικών μεθόδων ταξινόμησης που αναφέρονται στον πίνακα 8.2, καθώς και τη δημιουργία και την επακόλουθη σύγκριση ομαδοποιητών ταξινομητών που βασίζονται στους παραπάνω ταξινομητές.

Για την αρχική σύγκριση των μεθόδων, κάθε σύνολο δεδομένων χωρίστηκε σε δύο υποσύνολα: το σύνολο εκπαίδευσης, το οποίο περιλάμβανε το 80% των παραδειγμάτων, και το σύνολο ελέγχου, το οποίο αποτελούσε το υπόλοιπο 20%. Αρχικά, κάθε ταξινομητής εκπαιδεύτηκε στο σύνολο εκπαίδευσης και στη συνέχεια αξιολογήθηκε ως προς την απόδοσή του στο σύνολο ελέγχου. Η διαδικασία εκπαίδευσης πραγματοποιήθηκε με τη χρήση στρωματοποιημένης k-fold διασταυρούμενης επικύρωσης (stratified K-fold cross-validation) με k = 10. Αυτή η τεχνική αποτελεί παραλλαγή της κλασικής k-fold διασταυρούμενης επικύρωσης και δημιουργεί στρωματοποιημένα σύνολα επικύρωσης, διασφαλίζοντας τη διατήρηση των ποσοστών των δειγμάτων κάθε κλάσης όπως υπάρχουν στα αρχικά δεδομένα. Επιπλέον, η διαδικασία διατηρεί τις σειριακές εξαρτήσεις στα δεδομένα, ενώ υποστηρίζει επίσης στρατηγικές τυχαίας κατάτμησης για μεγαλύτερη ευελιξία.

Για τη δημιουργία ομαδοποιητών, υλοποιήθηκαν ομαδοποιητές αποτελούμενοι από τρεις ταξινομητές βάσης, χρησιμοποιώντας τη διαδικασία της «ήπιας ψηφοφορίας» (soft voting). Με τη μέθοδο αυτή για το συνδυασμό των αποφάσεων των επί μέρους μεθόδων, λαμβάνονται υπόψη οι πιθανότητες που προβλέπουν τα μοντέλα για κάθε κλάση. Η τελική πρόβλεψη καθορίζεται από την κλάση με τη μεγαλύτερη αθροιστική πιθανότητα. Έπειτα ακολουθήθηκε η ίδια τακτική χωρισμού των αρχικών συνόλων σε σύνολα εκπαίδευσης και ελέγχου με της προαναφερθείσα αναλογία, και με χρήση στρωματοποιημένης 10 – διασταυρωμένης επικύρωσης.

8.2.4 Δεύτερο στάδιο πειραματικής διαδικασίας: Σχήματα Ενεργητικής Μάθησης

Το δεύτερο μέρος της πειραματικής διαδικασίας, αφορά τη χρήση τεχνικών Ενεργής Μάθησης για την διερεύνηση του αν η χρήση της μπορεί να οδηγήσει στη μείωση του πλήθους ετικετοποιημένων παραδειγμάτων κατά τη φάση της εκπαίδευσης, χωρίς όμως να υπάρξει σημαντική απώλεια πληροφορίας.

Για την ανάπτυξη των σχημάτων Ενεργής Μάθησης, χρησιμοποιήθηκε η τεχνική της δειγματοληψίας από «δεξαμενή» (pool-based sampling), η οποία επικεντρώνεται στη στρατηγική επιλογή δεδομένων προς επισήμανση, χρησιμοποιώντας ένα μη επισημασμένο σύνολο δεδομένων, το οποίο ονομάζεται «δεξαμενή» (pool). Τα παραδείγματα της δεξαμενής δεν συνοδεύονται από τις ετικέτες τους. Στην υλοποίησή μας, για κάθε σύνολο δεδομένων χρησιμοποιήθηκε το 80%, ενώ το υπόλοιπο 20% κρατήθηκε για την αξιολόγηση της μεθόδου. Από το 80% της «δεξαμενής» τραβήχτηκε τυχαία το 10% και αποτέλεσε το αρχικό σύνολο εκπαίδευσης. Το 10% των παραδειγμάτων αυτού ανακτά τις ετικέτες του, ώστε να αναπαρασταθεί τεχνηέντως η εμπλοκή ενός γνώστη του αντικειμένου, ο οποίος μπορεί να αποδώσει ετικέτες σε μη γνωστά παραδείγματα.

Έπειτα σε κάθε επανάληψη επιλεγόταν το 1% που αποτελούσε τα δείγματα με την περισσότερη πληροφορία και αφού τους ανατεθεί η ετικέτα, ενσωματώνονταν στο σύνολο εκπαίδευσης και το μοντέλο εκ νέου εκπαιδευόταν. Για την επιλογή των παραδειγμάτων που θα ετικετοποιηθούν μέσω της επαναληπτικής διαδικασίας, χρησιμοποιήθηκαν τρεις στρατηγικές – οπότε η όλη διαδικασία εκτελέστηκε τρεις φορές, μία για την κάθε στρατηγική επιλογής. Οι στρατηγικές επιλογής ήταν: δειγματοληψία με αβεβαιότητα (uncertainty sampling), δειγματοληψία περιθωρίου (margin sampling) και δειγματοληψία σύμφωνα με την εντροπία (entropy sampling).

Η επαναληπτική διαδικασία τερματιζόταν είτε όταν το εκπαιδευόμενο μοντέλο σημείωνε τα ίδια αποτελέσματα από τη σκοπιά της μετρικής AUC – με ανεκτή διαφορά 0.01, είτε όταν είχε χρησιμοποιηθεί το 50% των δειγμάτων για την εκπαίδευση του μοντέλου, καθώς στόχος μας ήταν να μελετήσουμε την ουσιαστική μείωση των παραδειγμάτων που θα πρέπει να ετικετοποιηθούν. Στο κριτήριο τερματισμού προτιμήθηκε η AUC έναντι των άλλων μετρικών, λόγω της ακαταλληλότητάς της σε προβλήματα με μεγάλη διαφορά των δειγμάτων στις κατανομές των κλάσεων, όπου για παράδειγμα η χρήση της ακρίβειας (accuracy) θα ήταν πιθανόν παραπλανητική. Παρέχει επίσης μια συνολική εικόνα της ικανότητας του μοντέλου να ξεχωρίζει τις κλάσεις. Έτσι, αντί να εξετάζεται η απόδοση σε μεμονωμένα σημεία (όπως π.χ. η F1-score που εξαρτάται από το κατώφλι), αξιολογεί τη συμπεριφορά του μοντέλου συνολικά.

8.2.5 Αξιολόγηση

Για την αξιολόγηση των μεθόδων, χρησιμοποιήθηκαν τρεις κλασικές μετρικές οι Area Under the Curve (AUC), τηε Matthews correlation coefficient (MCC), και Cohen's Kappa coefficient. Η επιλογή κατάλληλων μετρικών αξιολόγησης είναι κρίσιμη για τη σωστή αποτίμηση της απόδοσης ενός μοντέλου μηχανικής μάθησης. Οι κλασικές μετρικές, όπως η ακρίβεια, η ανάκληση , και η F1-score, είναι συχνά οι πρώτες επιλογές κυρίως λόγω της απλότητάς τους. Ωστόσο, αυτές οι μετρικές είναι συχνά παραπλανητικές σε σενάρια με ανισοκατανεμημένες κλάσεις όπως αυτά που εξετάζουμε στην παρούσα προβληματική.

Εν αντιθέσει, οι μετρικές που χρησιμοποιήθηκαν για την αξιολόγηση των μεθόδων

στα πειραματικά πλαίσια που παρουσιάζονται, υπερτερούν σημαντικά σε σενάρια όπου η κατανομή των παραδειγμάτων στις κλάσεις είναι ανομοιόμορφη. Η AUC προσφέρει μια γενική εκτίμηση της ικανότητας του μοντέλου να διαχωρίζει τις θετικές και αρνητικές περιπτώσεις στα δυαδικά προβλήματα ταξινόμησης ανεξαρτήτως κατωφλίων, γεγονός που την καθιστά κατάλληλη για σενάρια με υψηλές απαιτήσεις διακριτικής ικανότητας. Ο MCC λαμβάνει υπόψη όλες τις κατηγορίες του πίνακα σύγχυσης, προσφέροντας μια συμμετρική και ισορροπημένη μέτρηση που είναι ανθεκτική σε μη ισορροπημένες κατανομές των παραδειγμάτων στις κλάσεις. Παράλληλα, ο συντελεστής Cohen's Kappa διορθώνει τη συμφωνία για τυχαίες προβλέψεις, παρέχοντας μια δίκαιη αποτίμηση της συνολικής απόδοσης. Για το λόγο αυτό, για τα σενάρια όπου αξιολογούνται τα απλά σενάρια ταξινόμησης επιλέχθηκαν αυτές οι τρεις μετρικές, καθώς ο συνδυασμός τους μπορεί να δώσει μια ολοκληρωμένη εικόνα των αποτελεσμάτων της πειραματικής διαδικασίας.

Για να αξιολογηθεί το κύριο ζητούμενο που αφορά το αν με λιγότερα ετικετοποιημένα παραδείγματα μπορούμε να πετύχουμε το ίδιο καλά αποτελέσματα, αυτές οι μετρικές δεν είναι αρκετές. Οι μετρικές αυτές αποτυπώνουν την απόδοση ενός μοντέλου, δηλαδή το πόσο καλά μπορεί να φέρει εις πέρας την εργασία για την οποία προορίζεται. Το ζητούμενο εδώ όμως είναι όχι να εκτιμηθεί το πόσο καλά ή όχι μια μέθοδος μπορεί να διακρίνει ανάμεσα στις δύο κλάσεις, αλλά το αν η εφαρμογή της Ενεργής Μάθησης μπορεί να βελτιώσει την όλη διαδικασία, όχι από τη σκοπιά των μετρικών αξιολόγησης, αλλά από την σκοπιά του όγκου των ετικετοποιημένων δεδομένων που χρειάζονται για να πετύχουμε το ίδιο αποτέλεσμα.

Το σενάριο που εξετάζουμε έχει καθαρά θεωρητική βάση – σε πραγματικές εφαρμογές δε θα μπορούσε να γίνει αυτή η σύγκριση καθώς θα απαιτούσε την ετικετοποίηση του συνόλου των δεδομένων – κατάσταση την οποία προφανώς θέλουμε να αποφύγουμε. Στο πειραματικό όμως πλαίσιο, η σύγκριση αυτή είναι εφικτή. Χρειαζόμαστε όμως έναν δείκτη που να λαμβάνει υπόψιν, όχι μόνο πόσο καλά ο ταξινομητής διεκπεραιώνει την εργασία της ταξινόμησης, αλλά και ποιο ποσοστό του αρχικού συνόλου των δεδομένων αξιοποιεί για να το πετύχει αυτό.

Για το σκοπό αυτό προτείνουμε ένα νέο μετα-δείκτη, τον Δείκτη Απόδοσης Προσαρμοσμένο στην Αποδοτικότητα (Efficiency-Adjusted Performance Index (EAPI)), πληροφορίες για τον οποίο παρουσιάστηκαν στο προηγούμενο κεφάλαιο. Υπενθυμίζεται ότι ο EAPI είναι ένας μετα-δείκτης που αξιοποιεί:

1. Την απόδοση του μοντέλου πριν και μετά την εφαρμογή ενεργής μάθησης.

2. Την ποσότητα των δεδομένων εκπαίδευσης που χρησιμοποιήθηκαν κατά τη διαδικασία εφαρμογής της ενεργητικής μάθησης.

Ουσιαστικά, ο EAPI είναι μια μετα-μετρική που ποσοτικοποιεί τη βελτίωση που επιτυγχάνεται μέσω της ενεργής μάθησης με βάση μια δεδομένη μετρική απόδοσης. Υψηλότερες τιμές του δείκτη EAPI υποδηλώνουν καλύτερη απόδοση, κάτι που σημαίνει ότι τα σχήματα Ενεργής Μάθησης όχι μόνο πλησιάζουν την απόδοση των μεθόδων χωρίς Ενεργής Μάθησης, αλλά το επιτυγχάνουν χρησιμοποιώντας μικρότερο σύνολο δεδομένων, αντανακλώντας την αποδοτικότητα σε σχέση με τις μειωμένες απαιτήσεις σε πλήθος ετικετοποιημένων δεδομένων. Αντίθετα, χαμηλότερες τιμές δείκτη υποδηλώνουν περιορισμένη βελτίωση ή αυξημένη χρήση δεδομένων, γεγονός που απαιτεί λεπτομερή εξέταση της διαδικασίας Ενεργής Μάθησης στο πλαίσιο συγκεκριμένων στόχων ταξινόμησης.

Πρέπει να αναγνωρίσουμε ότι μια πιθανή απειλή για την εγκυρότητα του EAPI είναι ότι σε πραγματικά σενάρια όπου εφαρμόζονται σχήματα ενεργής μάθησης, δεν υπάρχει πλήρες ετικετοποιημένο σύνολο εκπαίδευσης διαθέσιμο για την εκπαίδευση και κατόπιν αξιολόγηση μέσω του δείκτη, ή η απόκτησή του είναι ιδιαίτερα κοστοβόρα. Στην πραγματικότητα, αυτή είναι και η βασική ιδέα στην οποία στηρίζεται η ενεργητική μάθηση. Ωστόσο, η πλήρης συμβολή του EAPI έγκειται στην αξιολόγηση της απόδοσης των μεθόδων στο πλαίσιο διαδοχικών εκπαίδευσης σε ολοένα και μεγαλύτερα σύνολα εκπαίδευσης που παράγονται μέσω ενεργής μάθησης. Με άλλα λόγια, ο EAPI μπορεί να χρησιμοποιηθεί για την αξιολόγηση διαφορετικών συνόλων εκπαίδευσης που παράγονται από ενεργή μάθηση.

Στην παρούσα εργασία, εφαρμόστηκαν δύο κριτήρια τερματισμού όσον αφορά το σενάριο της ενεργητικής μάθησης, τα οποία επηρεάζουν την τιμή του δείκτη EAPI:

- Η πρώτη συνθήκη ορίζει ότι η απόλυτη διαφορά στο AUC μεταξύ των σεναρίων με και χωρίς AL πρέπει να είναι μικρότερη από 0.01.
- Η δεύτερη συνθήκη εφαρμόζεται αν η πρώτη δεν ικανοποιηθεί ακόμα και όταν χρησιμοποιηθεί το 50% των αρχικών δεδομένων εκπαίδευσης, οδηγώντας στον τερματισμό των επαναλήψεων.

Ως απόρροια του πρώτου, ο πρώτος συντελεστής του EAPI τείνει να βρίσκεται κοντά στο 1, υποδεικνύοντας ελάχιστη απόκλιση στο AUC μεταξύ των δύο σεναρίων. Εν τω μεταξύ, ο δεύτερος συντελεστής κυμαίνεται μεταξύ 0.5 και 0.9, αντανακλώντας το γεγονός ότι, σε σενάρια AL, χρησιμοποιήθηκε μόνο ένα ποσοστό μεταξύ 10% και 50% του αρχικού συνόλου δεδομένων. Ως αποτέλεσμα, στη μελέτη μας, αναμένουμε ο EAPI να λαμβάνει κυρίως τιμές εντός του εύρους 0 και 1. Το γεγονός αυτό έρχεται σε συμφωνία με τη φύση του πειράματός μας, όπου ο πρωταρχικός στόχος είναι να μειωθεί αποδοτικά η ποσότητα των δεδομένων εκπαίδευσης διατηρώντας παράλληλα την απόδοση.

Συμπερασματικά, είναι σημαντικό να τονίσουμε ότι η ευελιξία του EAPI του επιτρέπει την εφαρμογή του σε διάφορες μελέτες περιπτώσεων ενεργής μάθησης, περιλαμβάνοντας σενάρια όπου η βελτίωση στη δεδομένη μετρική μπορεί να είναι πιο σημαντική ή να χρησιμοποιούνται διαφορετικά ποσοστά του αρχικού συνόλου εκπαίδευσης. Συνολικά, η προσαρμοστικότητα του EAPI τον καθιστά ένα πολύτιμο εργαλείο για την αξιολόγηση μεθόδων ενεργής μάθησης σε ένα ευρύ φάσμα σεναρίων, συμβάλλοντας σε βαθύτερη κατκατανόηση της αποτελεσματικότητάς τους και στη λήψη τεκμηριωμένων αποφάσεων σε εργασίες μηχανικής μάθησης.

8.3 Αποτελέσματα

Η παρουσίαση των αποτελεσμάτων χωρίζεται σε τρία στάδια και ακολουθεί τη δομή των ερευνητικών ερωτημάτων που εξετάζει η μελέτη και κατ' επέκταση τη σειρά με την οποία υλοποιήθηκαν τα πειράματα κατά την εκτέλεση του εκτεταμένου πειραματικού πλαισίου. Αρχικά, παρουσιάζονται οι επιδόσεις των μεμονωμένων σχημάτων ταξινόμησης χωρίς τη χρήση τεχνικών Ενεργής Μάθησης. Αυτά τα αποτελέσματα αποτελούν τη βάση για την επιλογή ενός υποσυνόλου μεθόδων που συγκροτούν τους βασικούς ταξινομητές για τα ομαδοποιημένα σχήματα που δοκιμάστηκαν. Οι μέθοδοι που εμφάνισαν ανώτερη απόδοση σε όλες τις μετρικές χρησιμοποιήθηκαν για τη δημιουργία των ομαδοποιημένων σχημάτων, η απόδοση των οποίων παρουσιάζεται στη συνέχεια. Σημειώνεται ότι αυτά

τα αποτελέσματα περιλαμβάνουν όλα τα δοκιμασμένα σχήματα, τόσο τα βασικά όσο και τα συνδυασμένα. Ακολούθως, παρουσιάζονται τα αποτελέσματα που σχετίζονται με την εφαρμογή τεχνικών Ενεργής Μάθησης, λαμβάνοντας υπόψη και πάλι το σύνολο των μεθόδων (τόσο μεμονωμένων όσο και ομαδοποιημένων).

8.3.1 Σύγκριση μεμονωμένων ταξινομητών

Για τη σύγκριση των μεμονωμένων ταξινομητών, χρησιμοποιήσαμε την κατάταξη που μας παρέχει ο στατιστικός έλεγχος Friedman. Με αυτόν τον τρόπο διασφαλίζεται ότι οι αξιολογήσεις βασίζονται σε μία συνδυασμένη εκτίμηση της σχετικής απόδοσης κάθε μεθόδου σε όλα τα σύνολα δεδομένων που εξετάστηκαν. Τα αποτελέσματα της εν λόγω κατάταξης, βρίσκονται στον Πίνακα 8.3.

Παρατηρούμε ότι ο LightGBM εμφανίζεται ως η κορυφαία μέθοδος και στις τρεις μετρικές (AUC, Kappa, MCC), αποδεικνύοντας τη σταθερή απόδοσή του σε όλα τα επίπεδα. Η μέθοδος NB κατατάσσεται δεύτερη στο AUC και πέμπτη στις Kappa και MCC, ενώ οι μέθοδοι ET και MLP παρουσιάζουν επίσης ισχυρές αποδόσεις, με υψηλές θέσεις στο Friedman Ranking. Το ET κατατάσσεται τρίτο στο AUC και δεύτερο στο MCC, ενώ το MLP βρίσκεται στην τέταρτη θέση στο AUC και τρίτη στο MCC.

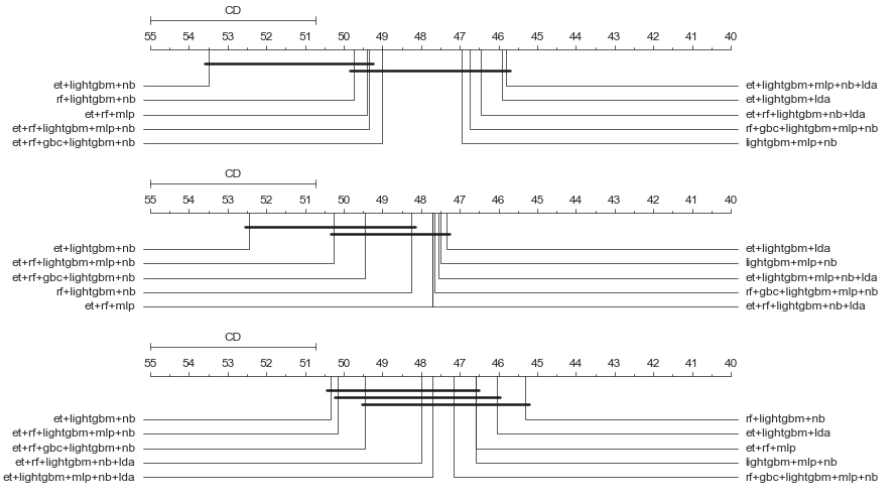
Στον αντίποδα, χαμηλότερες κατατάξεις παρατηρούνται για τις μεθόδους SVM και Ridge που καταλαμβάνουν τις τελευταίες θέσεις στις τρεις μετρικές, υποδεικνύοντας ότι είναι λιγότερο αποδοτικές συγκριτικά με άλλες μεθόδους για τα συγκεκριμένα δεδομένα.

8.3.2 Σύγκριση μεμονωμένων ταξινομητών και ομαδοποιητών

Πίνακας 8.3: Κατάταξη Friedman - Μεμονωμένες μέθοδοι

	AUC		**Kappa**		**MCC**	
	Method	Friedman Ranking	Method	Friedman Ranking	Method	Friedman Ranking
1στ	LightGBM	10.55	LightGBM	10.85	LightGBM	10.35
2νδ	NB	10.5	ET	10.7	ET	10.3
3ρδ	ET	10.4	MLP	10.3	RF	9.9
4τη	MLP	10.15	RF	10	MLP	9.8
5τη	RF	9.8	NB	9.6	NB	9
6τη	DT	9.4	GBC	8.55	LDA	8.85
7τη	GBC	8.4	LDA	8.15	GBC	8.3
8τη	QDA	8.35	DT	7.9	QDA	7.65
9τη	LDA	7.85	QDA	7.85	LR	7.45
10τη	KNN	7.25	KNN	6.85	RBFSVM	7.45
11τη	LR	6.35	LR	6.65	Ridge	7.3
12τη	ADA	5.9	RBFSVM	6.6	KNN	6.35
13τη	RBFSVM	5.55	ADA	5.8	ADA	6.1
14τη	Ridge	5.15	Ridge	5.5	DT	6
15τη	SVM	4.4	SVM	4.7	SVM	5.2

Όπως ήδη αναφέρθηκε, σε συνέχεια της σύγκρισης των μεμονωμένων ταξινομητών, ακολούθησε η δημιουργία ομαδοποιητών ταξινομητών, σε σχήματα «ήπιας ψηφοφορίας», χρησιμοποιώντας κάθε φορά τρεις ή πέντε ταξινομητές ως ταξινομητές βάσης. Για να περιοριστεί ο αριθμός των δημιουργούμενων ταξινομητών, κατά τη διαδικασία αυτοί



Σχήμα 8.2: Διαγράμματα κρίσιμων διαφορών της καλύτερης δεκάδας μεθόδων

χρησιμοποιήθηκαν ως μέθοδοι βάσης, μόνο οι καλύτεροι στην κατάταξη, όπως αυτή αποτυπώθηκε στον Πίνακα 8.3.

Για να εντοπιστούν ομαδοποιητές ταξινομητών που η απόδοσή τους υπερέχει έναντι των υπολοίπων, αλλά και για να απαντηθεί ένα από τα βασικά ερωτήματα της έρευνας, που είναι το αν οι ομαδοποιητές αποδίδουν καλύτερα στο συγκεκριμένο πρόβλημα ταξινόμησης, διεξήχθη ένας στατιστικός έλεγχος Friedman σε ολόκληρο το σύνολο των μεθόδων, μεμονωμένων και ομαδοποιητών. Ενώ ο έλεγχος Friedman αξιολογεί αν οι εξεταζόμενες μέθοδοι έχουν ή όχι παραπλήσια συμπεριφορά, ως παράγωγό του η κατάταξη που προσφέρει είναι ενδεικτική της μέσης απόδοσης των μεθόδων στο σύνολο των συνόλων δεδομένων που εξετάστηκαν. Έτσι, τα Friedman Rankings που παρουσιάζονται στον Πίνακα 8.4 αποτυπώνουν τις σχετικές κατατάξεις μεταξύ των μεθόδων που εξετάστηκαν, αντί να αναπαριστούν τις αντίστοιχες αποδόσεις τους με όρους πραγματικών τιμών μετρικών. Στον πίνακα εμφανίζονται μόνο οι δέκα πρώτες θέσεις της κατάταξης.

Πίνακας 8.4: Κατάταξη Friedman - Μεμονωμένες μέθοδοι και ομαδοποιητές

	ΑΥ͂		**Καππα**		**Μ͂**	
	Method	Friedman Ranking	Method	Friedman Ranking	Method	Friedman Ranking
1στ	et+lightgbm+nb	53.5	et+lightgbm+nb	52.45	et+lightgbm+nb	50.35
2νδ	rf+lightgbm+nb	49.75	et+rf+lightgbm+mlp+nb	50.25	et+rf+lightgbm+mlp+nb	50.15
3ρδ	et+rf+mlp	49.4	et+rf+gbc+lightgbm+nb	49.45	et+rf+gbc+lightgbm+nb	49.45
4τη	et+rf+lightgbm+mlp+nb	49.35	rf+lightgbm+nb	48.25	et+rf+lightgbm+nb+lda	48
5τη	et+rf+gbc+lightgbm+nb	49	et+rf+lightgbm+nb+lda	47.7	et+lightgbm+mlp+nb+lda	47.7
6τη	lightgbm+mlp+nb	46.95	et+rf+mlp	47.7	rf+gbc+lightgbm+mlp+nb	47.15
7τη	rf+gbc+lightgbm+mlp+nb	46.75	rf+gbc+lightgbm+mlp+nb	47.65	et+rf+mlp	46.6
8τη	et+rf+lightgbm+nb+lda	46.45	et+lightgbm+mlp+nb+lda	47.55	lightgbm+mlp+nb	46.6
9τη	et+lightgbm+lda	45.9	lightgbm+mlp+nb	47.5	et+lightgbm+lda	46.05
10τη	et+lightgbm+mlp+nb+lda	45.8	et+lightgbm+lda	47.35	rf+lightgbm+nb	45.3

Επιπροσθέτως, χρησιμοποιήθηκε το Bonferroni-Dunn post-hoc test, ώστε να διαπιστωθεί ανά ζευγάρι μεθόδων, αν έχουν παραπλήσια συμπεριφορά, ή ακριβέστερα αν η αποτελεσματικότητα των μεθόδων, όπως αυτή αποτυπώνεται από την σκοπιά των μετρικών, διαφέρει με στατιστικά σημαντική διαφορά. Τα αποτελέσματα του Bonferroni-Dunn post-hoc test διακρίνονται στην εικόνα 8.2, όπου οι οριζόντιες μαύρες γραμμές που συνδέουν ζευγάρια μεθόδων υποδεικνύουν πως οι εν λόγω μέθοδοι έχουν παρόμοια συμπεριφορά.

Παρατηρείται πως ο συνδυασμός των μεθόδων ET, LightGBM, και NB καταλαμβάνει σταθερά την πρώτη θέση στις τρεις μετρικές, επιβεβαιώνοντας την αποτελεσματικότητά του. Αυτή η απόδοση αντικατοπτρίζει τη δυνατότητα των ομαδοποιητών που χρησιμοποιούν μεθόδους βάσεις από διαφορετικές οικογένειες ταξινομητών να συνδυάζουν τα πλεονεκτήματα διαφορετικών ταξινομητών για να επιτύχουν βέλτιστα αποτελέσματα. Επίσης, στις κορυφαίες θέσεις, παρατηρούνται διάφοροι συνδυασμοί που περιλαμβάνουν ταξινομητές όπως MLP, LDA, και RF, οι οποίοι προσφέρουν σταθερά υψηλές επιδόσεις όταν ενσωματώνονται σε ensembles. Μάλιστα, τα διαγράμματα κρίσιμών διαφορών δείχνουν πως η απόδοση των παραπάνω μεθόδων είναι παραπλήσια.

Ερχόμαστε τώρα στο πρώτο από τα βασικά ερωτήματα αυτής της έρευνας, που αφορά το αν είναι επωφελής η χρήση ομαδοποιητών σε σενάρια πρόβλεψης σφαλμάτων λογισμικού. Όπως μπορεί να διαπιστωθεί από τον Πίνακα 8.4, όπου παρουσιάζονται οι δέκα καλύτερες μέθοδοι ταξινόμησης στο σύνολο των μεθόδων που χρησιμοποιήθηκαν, αλλά και από τα αντίστοιχα διαγράμματα κρίσιμων διαφορών, στο Σχήμα 8.2, καμία μεμονωμένη μέθοδος δεν εμφανίζεται στην πρώτη δεκάδα. Ειδικότερα, στη μετρική AUC, η πρώτη μεμονωμένη μέθοδος που εμφανίζεται είναι η LightGBM, καταλαμβάνοντας την 30ή θέση. Αντίστοιχα, για τις μετρικές Kappa και MCC, οι πρώτες μεμονωμένες μέθοδοι εμφανίζονται στις θέσεις 40 και 41, αντίστοιχα.

Αυτά τα ευρήματα καταδεικνύουν ξεκάθαρα την υπεροχή των σχημάτων ομαδοποιητών, καθώς η πλειονότητά τους υπερτερεί ακόμη και των καλύτερων μεμονωμένων ταξινομητών. Αξιοσημείωτο είναι ότι ο συνδυασμός των ET, LightGBM, και NB καταλαμβάνει σταθερά την πρώτη θέση σε όλες τις μετρικές. Αυτό είναι ιδιαίτερα ενδιαφέρον, δεδομένου ότι, παρά την ύπαρξη άλλων ομαδοποιητών ταξινομητών με συγκρίσιμες επιδόσεις, όπως προτείνεται από τα διαγράμματα κρίσιμων διαφορών στο Σχήμα 8.2, αυτός ο συνδυασμός υπερέχει σταθερά όλων.

8.3.3 Αποτελέσματα σχετικά με τα σχήματα Ενεργής Μάθησης

Για να οργανωθεί καλύτερα η παρουσίαση των αποτελεσμάτων που αφορούν το δεύτερο σκέλος της πειραματικής διαδικασίας, δηλαδή την εφαρμογή της Ενεργής Μάθησης σε σενάρια πρόβλεψης σφαλμάτων λογισμικού, συνοψίζουμε εδώ τα βασικά ερωτήματα τα οποία τέθηκαν κατά το σχεδιασμό της πειραματικής διαδικασίας. Το κυριότερο από αυτά, που αποτελεί και τον έναυσμα για την ενασχόλησή μας με το παρόν θέμα, είναι το παρακάτω:

> «*Συμβάλει η χρήση ενεργητικής μάθησης στη μείωση του αριθμού ετικετοποιημένων δεδομένων, χωρίς να υπάρχει μείωση της απόδοσης των μεθόδων*;»

Αναδιατυπώνοντας, μπορούμε να συμπεράνουμε αν οι ίδιες μέθοδοι όταν χρησιμοποιούνται σε σχήματα Ενεργής Μάθησης πετυχαίνουν το ίδιο καλά αποτελέσματα με τις αρχικές εκδοχές τους όπου δε χρησιμοποιήθηκε Ενεργή Μάθηση, αξιοποιώντας στην περίπτωσή μας λιγότερα από τα μισά δεδομένα του αρχικού συνόλου δεδομένων. Επίσης, θα πρέπει να διαπιστωθεί το κατά πόσο οι δύο εκδοχές – χωρίς χρήση Ενεργής Μάθησης και με χρήση Ενεργής Μάθησης και μικρότερο σύνολο εκπαίδευσης – δε διαφέρουν στατιστικά, ώστε τα αποτελέσματα να είναι ισχυρά και μη αμφιλεγόμενα.

Αν η απάντηση σε αυτό το ερώτημα είναι θετική, εγείρονται με τη σειρά τους νέα υποερωτήματα:



Σχήμα 8.3: Σύγκλιση των μεθόδων

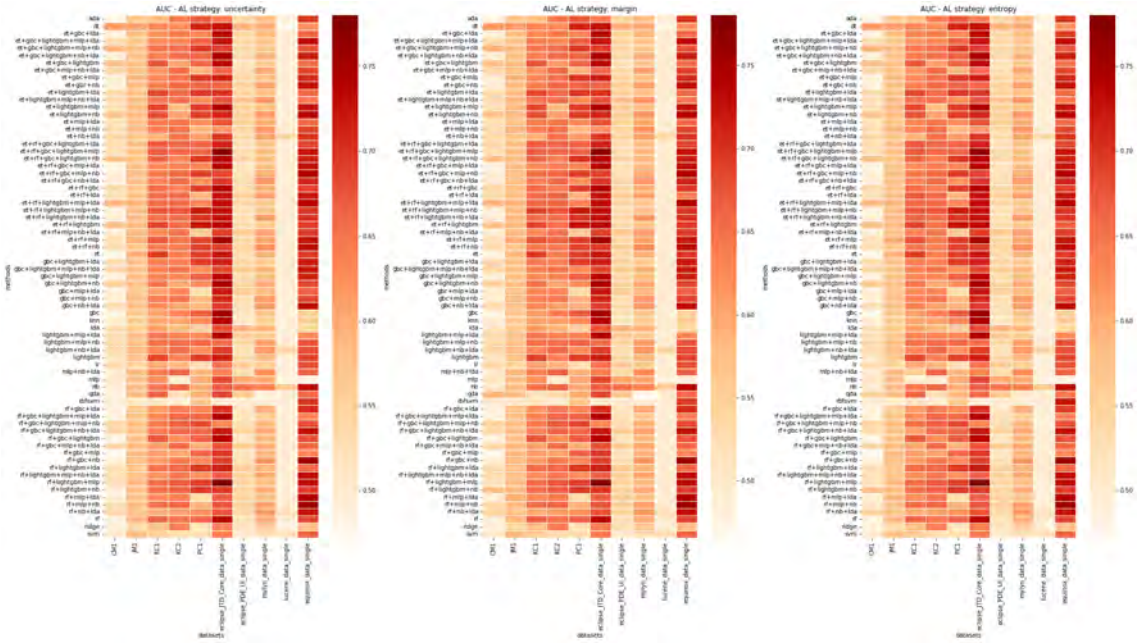
- Υπάρχει κάποια συγκεκριμένη μέθοδος που φαίνεται να υπερτερεί των υπολοίπων σε σενάρια Ενεργής Μάθησης όσον αφορά τον όγκο των δεδομένων που χρησιμοποιούνται για την εκπαίδευση·
- Υπάρχει κάποια συγκεκριμένη στρατηγική ερωτήσεων (query strategy) που φαίνεται να είναι πιο επωφελής όσον αφορά τον όγκο των δεδομένων που χρησιμοποιούνται για την εκπαίδευση·

Εστιάζοντας στην απάντηση του πρώτου ερωτήματος, ξεκινάμε με τη διερεύνηση του κατά πόσο μπορέσαμε να πετύχουμε την ίδια απόδοση των μετρικών, πριν διακοπεί η εκτέλεση των πειραμάτων. Καταχρηστικά, θα αναφέρουμε την κατάσταση αυτή ως σύγκλιση, δηλαδή θα θεωρούμε πως επιτύχαμε το στόχο όταν ένα πείραμα τερματίστηκε γιατί η απόλυτη τιμή της διαφοράς μιας δοσμένης μετρικής ήταν μικρότερη του 0.01 ανάμεσα στα σενάρια χωρίς και με χρήση Ενεργής Μάθησης, και όχι γιατί ενεργοποιήθηκε το δεύτερο κριτήριο τερματισμού, δηλαδή χρησιμοποιήθηκε το ανώτατο όριο του 50% των διαθέσιμων παραδειγμάτων, χωρίς όμως να επιτευχθεί το επιθυμητό αποτέλεσμα.

Αυτή η πληροφορία αποτυπώνεται στο heatmap του Σχήματος 8.3, όπου με πράσινο εμφανίζονται οι περιπτώσεις που η μέθοδος «συνέκλινε», ενώ με κόκκινο εκείνες που ενεργοποιήθηκε το δεύτερο κριτήριο τερματισμού. Οι στήλες του heatmap αντιπροσωπεύουν τα 10 σύνολα δεδομένων, ενώ οι γραμμές αντιστοιχούν στις 71 χρησιμοποιούμενες μεθόδους. Κάθε heatmap αναπαριστά μία από τις τρεις τεχνικές ερωτήσεων.

Ας εστιάσουμε τώρα στον όγκο των δεδομένων που απαιτείται για τη «σύγκλιση» κάθε σχήματος. Στο Σχήμα 8.4, παρουσιάζονται τα heatmaps που απεικονίζουν το ποσοστό των δεδομένων που χρησιμοποιήθηκαν σε κάθε πείραμα, σε σχέση με καθεμία από τις τρεις στρατηγικές ερωτήσεων. Σημειώνεται ότι ως κριτήριο τερματισμού της διαδικασίας χρησιμοποιήθηκε η τιμή της μετρικής AUC. Τα Σχήματα 8.3 και 8.4 πρέπει να μελετώνται σε αντιπαραβολή, ώστε να εντοπίζονται τα πειράματα στα οποία εξαντλήθηκε το 50% του αρχικού συνόλου δεδομένων και έτσι ενεργοποιήθηκε το δεύτερο κριτήριο τερματισμού.

Αν και τα heatmaps παρέχουν ιδιαίτερα χρήσιμη πληροφορία, ο προσδιορισμός της



Σχήμα 8.4: Απόδοση των μεθόδων

πραγματικής συμβολής της Ενεργής Μάθησης αποτελεί πολυπλοκότερη διαδικασία, και αυτό γιατί ορισμένες μέθοδοι παρουσιάζουν πολύ καλή συμπεριφορά σε κάποια σύνολα δεδομένων, χρησιμοποιώντας πολύ μικρό ποσοστό από τα διαθέσιμα δεδομένα, ενώ σε άλλα εξαντλούν το 50% χωρίς ουσιαστικά αποτελέσματα. Για το σκοπό αυτό, και για να εμβαθύνουμε περισσότερο στη φύση των αποτελεσμάτων, χρησιμοποιήθηκαν boxplots που αποτυπώνουν την απόδοση των 10 καλύτερων μεθόδων.

Τα θηκογράμματα που παρουσιάζονται στο Σχήμα 8.5 παρέχουν μια συνοπτική και κατανοητή αναπαράσταση της απόδοσης των μεθόδων για κάθε μια από τις τρεις μετρικές αξιολόγησης για το σενάριο βασισμένο στην εντροπία (entropy sampling). Κάθε παραλληλόγραμμο αντιπροσωπεύει την κατανομή των βαθμολογιών της αντίστοιχης μεθόδου, με τη διάμεσο να αναπαρίσταται από την κάθετη γραμμή και τη μέση τιμή από μια μαύρη κουκκίδα μέσα στο ορθογώνιο. Οι ακραίες τιμές (outliers) παρέχουν πολύτιμες πληροφορίες σχετικά με σύνολα δεδομένων που απαιτούν μεγαλύτερο ποσοστό διαθέσιμων δεδομένων εκπαίδευσης για να επιτύχουν ύ.

Από τα θηκογράμματα επίσης επιβεβαιώνεται ότι συγκεκριμένες μέθοδοι, όπως η rf+gbc+mlp, υπερέχουν, καταλαμβάνοντας σταθερά τις κορυφαίες θέσεις, γεγονός που αποδεικνύεται από τις χαμηλότερες διαμέσους τιμές τους. Αυτή η σταθερή κυριαρχία υποδηλώνει ότι οι μέθοδοι αυτές μπορούν να επιτύχουν αποτελέσματα συγκρίσιμα με εκείνα των μη- ενεργών μεθόδων, χρησιμοποιώντας μόνο ένα μικρό ποσοστό των διαθέσιμων δεδομένων εκπαίδευσης. Ειδικότερα, η rf+gbc+mlp καταφέρνει να επιτύχει τις χαμηλότερες διαμέσους στις περισσότερες περιπτώσεις, χρησιμοποιώντας μόλις 12% των αρχικών δεδομένων εκπαίδευσης.

Ενδιαφέρον παρουσιάζουν επίσης μέθοδοι όπως οι et+rf+gbc+nb+lda και rf+gbc+nb, οι οποίες αν και εμφανίζουν ελαφρώς υψηλότερες διαμέσους, διαθέτουν σημαντικά χαμηλότερους μέσους όρους και μικρότερα τεταρτημόρια, υποδεικνύοντας μεγαλύτερη συνοχή και αποδοτικότητα στη χρήση δεδομένων.

Τα boxplots αναδεικνύουν επίσης την ποικιλομορφία στις απαιτήσεις δεδομένων εκπαίδευσης μεταξύ διαφορετικών μεθόδων και συνόλων δεδομένων. Ορισμένες μέθοδοι

απαιτούν χαμηλότερο ποσοστό δεδομένων για να επιτύχουν υψηλές βαθμολογίες AUC, ενώ άλλες εμφανίζουν μεγαλύτερη εξάρτηση από τα διαθέσιμα δεδομένα, γεγονός που επισημαίνεται από τα outliers. Συνολικά, τα αποτελέσματα επιβεβαιώνουν την αποτελεσματικότητα συγκεκριμένων σχημάτων ομαδοποιητών, ειδικά όταν χρησιμοποιείται περιορισμένος όγκος δεδομένων εκπαίδευσης.

Για να γίνει εμφανέστερη η πληροφορία που αφορά το ποσοστό των δειγμάτων που ήταν αρκετό για να επιτευχθούν παρόμοια αποτελέσματα με το σενάριο στο οποίο δε χρησιμοποιείται Ενεργή Μάθηση, δηλαδή για να διαπιστώσουμε σε τι βαθμό μπορέσαμε να περιορίσουμε το πλήθος των ετικετοποιημένων παραδειγμάτων χωρίς θυσία της απόδοσης, χωρίσαμε σε 8 κλάσεις ισοδυναμίας ίσου πλάτους το διάστημα [10%-50%], που αντιπροσωπεύει το ποσοστό του συνόλου δεδομένων που χρησιμοποιήθηκε, και εξετάσαμε πόσα από τα πειράματα εμπίπτουν σε κάθε κλάση.

Τα αποτελέσματα είναι ενθαρρυντικά, καθώς και για τις τρεις στρατηγικές ερωτημάτων, δηλαδή τις uncertainty, margin, και entropy sampling, σημειώνονται θετικά αποτελέσματα. Συγκεκριμένα, στα σενάρια uncertainty, margin, και entropy, το 56.9%, 57.04%, και 57.32%, αντίστοιχα, των πειραμάτων χρησιμοποίησαν λιγότερο από το 20% των αρχικών δεδομένων εκπαίδευσης για να επιτύχουν συγκρίσιμες βαθμολογίες με τη μετρική AUC.

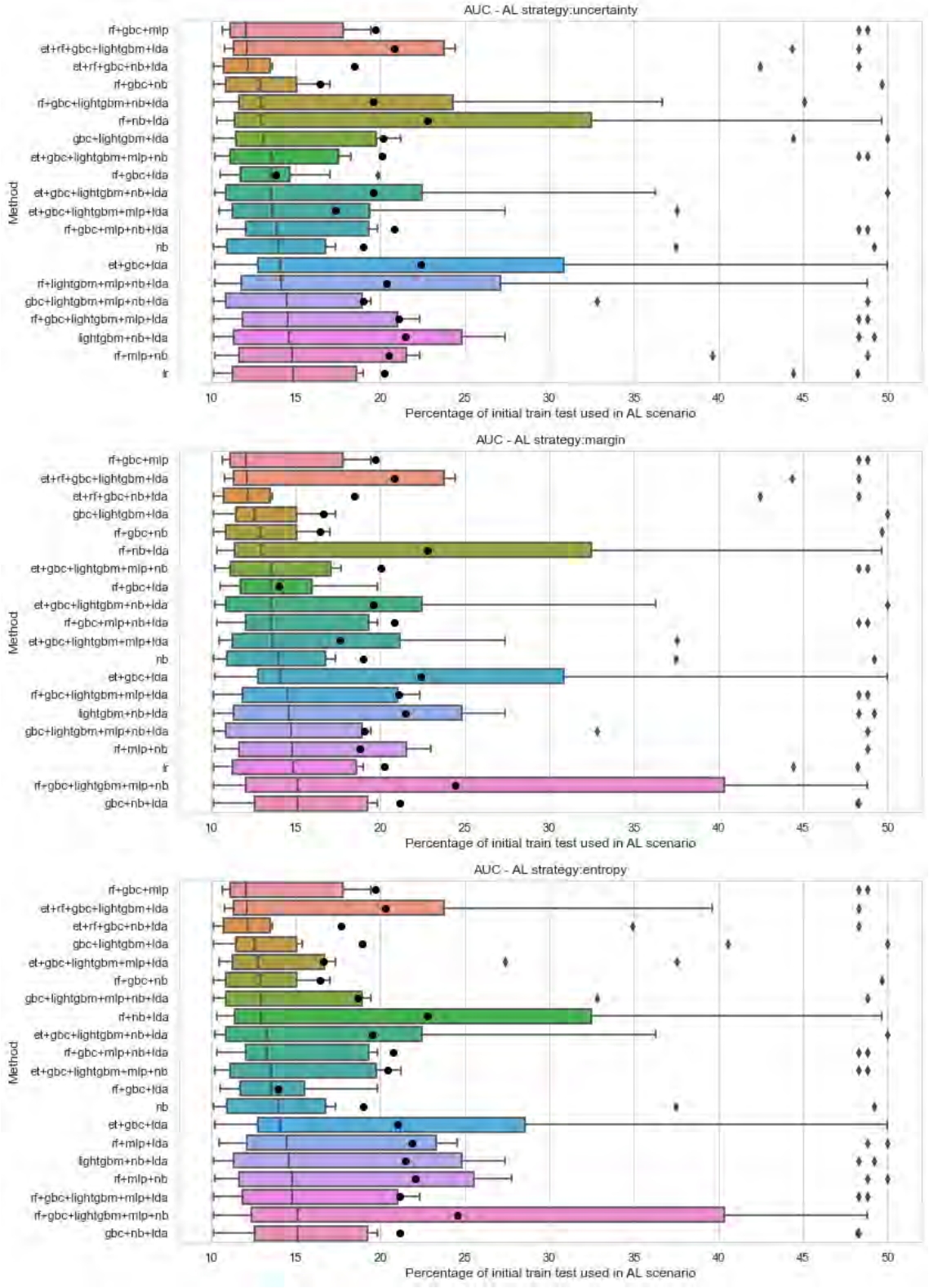
Οι αντίστοιχες γραφικές παραστάσεις παρουσιάζονται στο Σχήμα 8.6. Στην ίδια εικόνα, η δεύτερη στήλη παρέχει πληροφορίες για τις αθροιστικές συχνότητες των προαναφερθέντων αποτελεσμάτων, όπου κάθε ράβδος αντιπροσωπεύει το άθροισμα των πειραμάτων που χρησιμοποιούν ποσοστό του αρχικού συνόλου εκπαίδευσης μικρότερο από το ανώτατο όριο της κατηγορίας.

Ένας ακόμη τρόπος για τη συνολική επισκόπηση της απόδοσης των μεθόδων, παρουσιάζεται στο Σχήμα 8.7. Τα διαγράμματα αυτά αναπαριστούν την απόδοση των μεθόδων τόσο στο σενάριο μη χρήση Ενεργής Μάθησης σε σύγκριση με το σενάριο όπου αυτή χρησιμοποιήθηκε. Κάθε σημείο στα διαγράμματα αντιπροσωπεύει ένα πείραμα, δηλαδή έναν ταξινομητή που εφαρμόστηκε σε συγκεκριμένο σύνολο δεδομένων. Ο πρώτος συντεταγμένος άξονας απεικονίζει τη βαθμολογία της μετρικής χωρίς τη χρήση Ενεργής Μάθησης, ενώ ο δεύτερος αντιστοιχεί στη βαθμολογία της ίδιας μετρικής όταν εφαρμόζεται ενεργητική μάθηση. Οι αποκλίσεις από την ευθεία y=x υποδεικνύουν σημαντικές διαφορές μεταξύ των αποτελεσμάτων στα δύο σενάρια, λειτουργώντας ως ένδειξη μη «σύγκλισης».

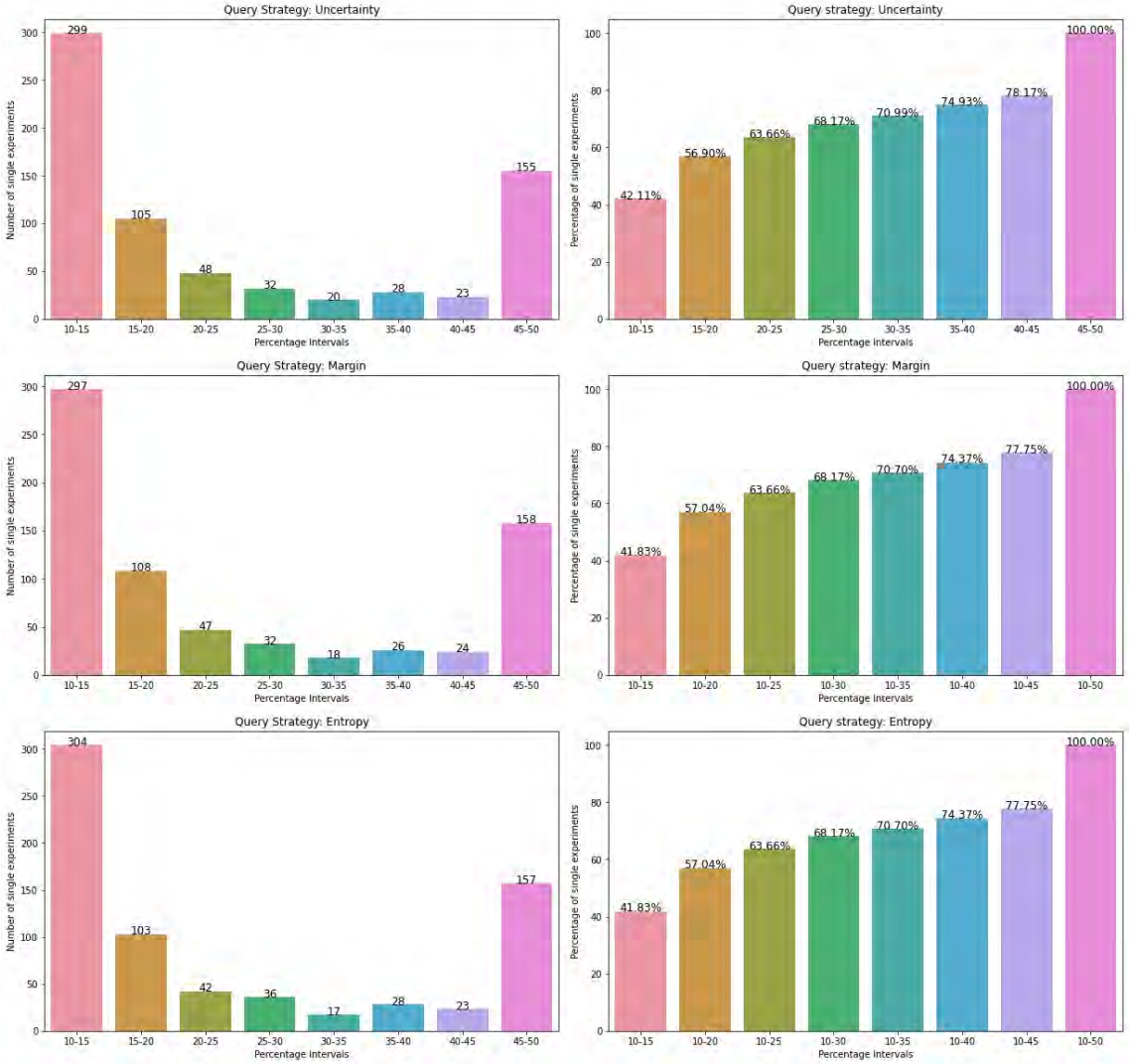
Όπως ήταν αναμενόμενο, με βάση τα προηγούμενα αποτελέσματα, η πλειονότητα των σημείων συγκεντρώνεται κοντά στην ευθεία y=x, γεγονός που επιβεβαιώνει την προηγούμενη παρατήρηση ότι οι περισσότερες μέθοδοι στα σενάρια χρήσης Ενεργής Μάθησης συγκλίνουν στις επιδόσεις των ομόλογών τους, όπου δεν χρησιμοποιήθηκε Ενεργή Μάθηση.

Για να μπορούμε με ασφάλεια να ισχυριστούμε πως η χρήση της Ενεργής Μάθησης μπορεί να ελαττώσει σημαντικά τον αριθμό των δειγμάτων που χρειάζονται για εκπαίδευση χωρίς θυσία στην απόδοσή τους, όπως αυτή αποτυπώνεται στις μετρικές, πρέπει να ελεγχθεί αν σε κάθε πείραμα οι δύο εκδοχές των μεθόδων – αυτή που δεν χρησιμοποιεί Ενεργή Μάθηση και η ομόλογή της που χρησιμοποιεί – είναι στατιστικά όμοιες.

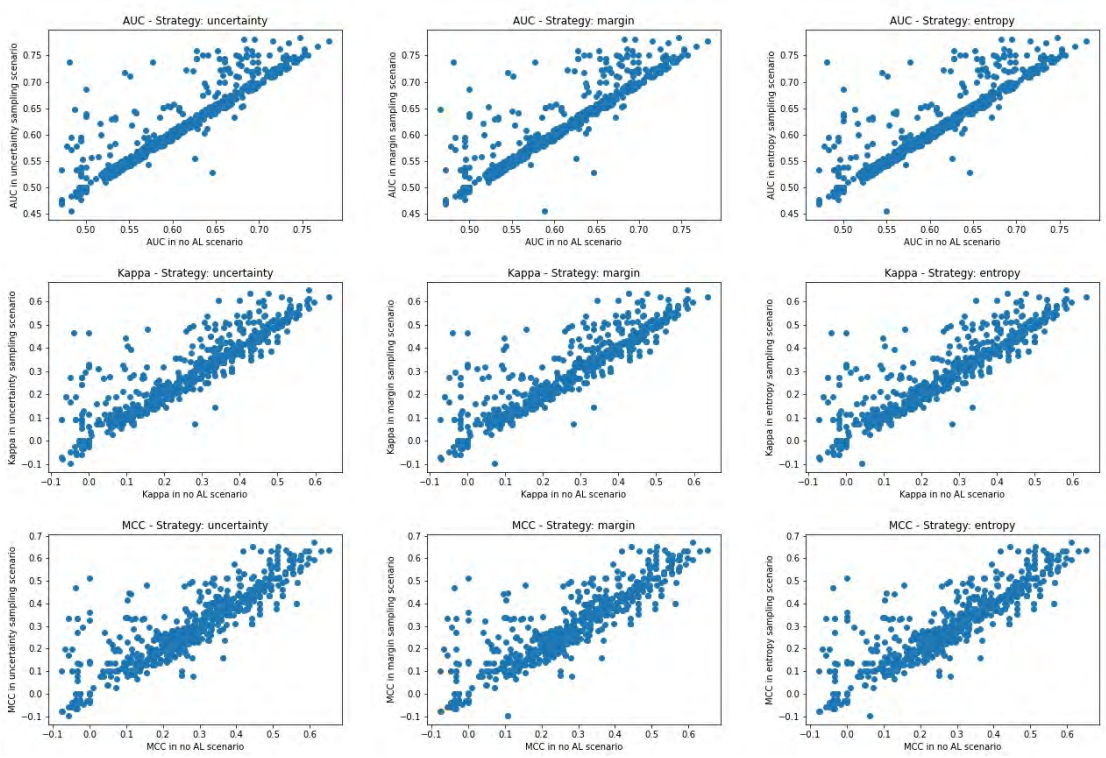
Για το σκοπό αυτό χρησιμοποιήθηκε το Friedman Test, σε συνδυασμό με το Bonferroni-Dunn post hoc test για συγκρίσεις εντός κάθε μεθόδου, είτε επρόκειτο για μεμονωμένες είτε για συναθροιστικές μεθόδους, και για τις τρεις στρατηγικές ερωτημάτων στο σενάριο της Ενεργής Μάθησης. Η μηδενική υπόθεση του Friedman Test



Σχήμα 8.5: Θηκογράμματα που απεικονίζουν το ποσοστό του αρχικού συνόλου που χρησιμοποιήθηκε στο σενάριο Ενεργής Μάθησης



Σχήμα 8.6: Ποσοστό χρήσης δεδομένων χωρισμένο σε κλάσεις ίσου πλάτους



Σχήμα 8.7: Οπτική απεικόνηση πειραμάτων που πετυχαίνουν την μείωση των επικετοποιημένων δεδομένων εκπαίδευσης.

προτείνει ότι οι μέσοι όροι των βαθμολογιών είναι ομοιόμορφοι μεταξύ των ομάδων. Παράλληλα, το Bonferroni-Dunn post hoc test ελέγχει αν οι παρατηρούμενες διαφορές μεταξύ ζευγών μέσων όρων ομάδων είναι στατιστικά σημαντικές.

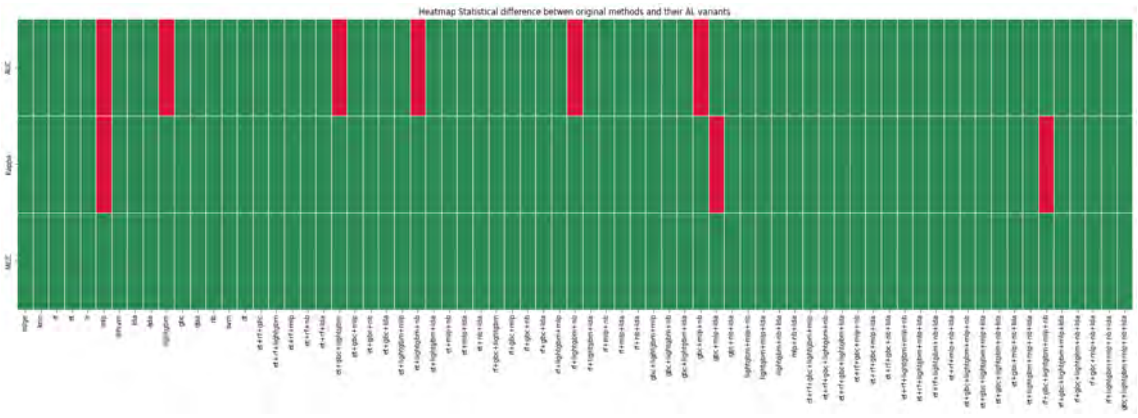
Ως εκ τούτου, η προσδοκία είναι ότι οι μέθοδοι και οι παραλλαγές Ενεργής Μάθησης τους πρέπει να συνδέονται με ευθείες οριζόντιες γραμμές στα παραγόμενα διαγράμματα κρίσιμων διαφορών , εκτός εάν υπάρχουν στατιστικά σημαντικές διαφορές.

Στο Σχήμα 8.8 εμφανίζονται, για το σύνολο των πειραμάτων που πραγματοποιήθηκαν, οι περιπτώσεις όπου οι τρεις παραλλαγές του σεναρίου Ενεργής Μάθησης και η κλασική παραλλαγή της μεθόδου (όπου δε χρησιμοποιήθηκε Ενεργή Μάθηση) παρουσιάζουν ή όχι στατιστικές διαφορές. Με πράσινο χρώμα σημειώνονται οι περιπτώσεις όπου το ζητούμενο επιτυγχάνεται και οι τέσσερις παραλλαγές δε διαφέρουν στατιστικώς. Αντιθέτως, με κόκκινο σημειώνονται οι περιπτώσεις όπου εμφανίζεται στατιστική διαφορά.

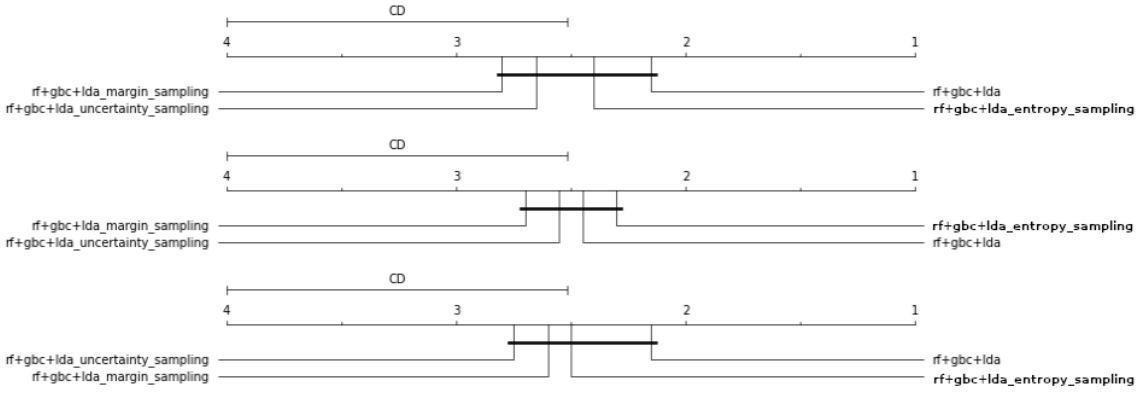
Αξιοσημείωτο είναι ότι, από τις 213 περιπτώσεις που εξετάστηκαν, μόνο 9 εμφάνισαν στατιστικά σημαντικές διαφορές μεταξύ των μεθόδων. Έτσι, στις περισσότερες από τις δοκιμασμένες μεθόδους ομαδοποιητών, δεν παρατηρούνται στατιστικές διαφορές.

Όπως είναι εμφανές στο Σχήμα 8.9, όπου παρουσιάζονται τα CD διαγράμματα της μεθόδου rf+gbc+lda, μπορούμε να παρατηρήσουμε την απόδοση της μεθόδου σε τρία σενάρια Ενεργής Μάθησης, καθένα από τα οποία περιλαμβάνει ελάχιστη χρήση του αρχικού συνόλου εκπαίδευσης. Αυτή η γραφική αναπαράσταση επιβεβαιώνει ότι οι παραλλαγές της Ενεργής Μάθησης δεν διαφέρουν σημαντικά από την αρχική μέθοδο, όπου αυτές δεν εφαρμόστηκαν.

Αφού καταλήξαμε στο θετικό συμπέρασμα ότι η Ενεργή Μάθηση προσφέρει σημαντική ωφέλεια μειώνοντας τις απαιτήσεις σε ετικετοποιημένα παραδείγματα για την



Σχήμα 8.8: Απεικόνηση των περιπτώσεων στατιστικής και μη ανεξαρτησίας ανάμεσα στο σενάριο μη χρήσης ενεργητικής μάθησης και στις παραλλαγές όπου αυτή χρησημοποιείται



Σχήμα 8.9: Στατιστική ανεξαρτησία των παραλλαγών στην περίπτωση του rf+gbc+lda

πρόβλεψη σφαλμάτων λογισμικού, προκύπτει η ανάγκη διερεύνησης δύο σημαντικών υποερωτημάτων. Πρώτον, εάν υπάρχει κάποιο συγκεκριμένο σχήμα μεθόδων ομαδοποιητών που υπερέχει των υπολοίπων, καθιστώντας το ιδανική επιλογή για την επίλυση παρόμοιων προβλημάτων. Δεύτερον, αν μπορεί να εντοπιστεί μία στρατηγική ερωτήσεων προς τη δεξαμενή δεδομένων που εμφανίζει σαφή υπεροχή έναντι των υπολοίπων, προσφέροντας βέλτιστη απόδοση στα σχετικά σενάρια.

Για να απαντήσουμε στο πρώτο υποερώτημα, χρησιμοποιούμε τον δείκτη EAPI. Αρχικά, ο δείκτης αυτός υπολογίστηκε για κάθε μεμονωμένο πείραμα στο οποίο χρησιμοποιήθηκε ενεργητική μάθηση. Στη συνέχεια, για την εξαγωγή συμπερασμάτων σχετικά με την απόδοση των μεθόδων, υπολογίστηκαν οι μέσες τιμές των EAPI για καθεμία από τις τρεις μετρικές που σχετίζονται με τις μεθόδους ταξινόμησης και τις στρατηγικές ερωτήσεων. Αυτές οι μέσες τιμές χρησιμοποιήθηκαν για την κατάταξη των μεθόδων.

Ο Πίνακας 8.3.3 παρουσιάζει τις δέκα κορυφαίες μεθόδους βάσει αυτής της κατάταξης. Αξιοσημείωτη είναι η κυριαρχία της μεθόδου rf+gbc+lda, η οποία καταλαμβάνει σταθερά την πρώτη θέση σε όλες τις περιπτώσεις, με εξαίρεση τη μετρική kappa, όπου κατατάσσεται τρίτη, πίσω από τις μεθόδους et+gbc+lightgbm+mlp+lda και nb.

Η επαναλαμβανόμενη παρουσία συναθροιστικών σχημάτων, όπως τα et+rf+gbc+nb+lda, rf+gbc+nb, και et+gbc+lightgbm+mlp+lda, στις κορυφαίες θέσεις είναι ιδιαίτερα ενδιαφέρουσα. Αυτοί οι ομαδοποιητές εμφανίζουν γενικά χαμηλό ποσοστό χρήσης δεδομένων εκπαίδευσης, γεγονός που τους κάνει ιδιαίτερα ελκυστικούς για ανάλογες προβληματικές.

Σχετικά με το δεύτερο υποερώτημα που αφορά τις διαφορετικές στρατηγικές ερωτήσεων με βάσει της οποίες επιλέγεται το επόμενο παράδειγμα για του οποίου η ετικέτα πρόκειται να ανακτηθεί, τα ευρήματα της μελέτης υποδεικνύουν τη στατιστική ομοιότητά τους. Αυτό επιβεβαιώνεται τόσο από το Friedman Test όσο και από την ανάλυση Bonferroni-Dunn post hoc, που διεξήχθησαν για την αξιολόγηση της απόδοσης. Η post hoc ανάλυση εξετάζει συγκεκριμένα αν ζεύγη μεθόδων —στην περίπτωσή μας, στρατηγικές ερωτήσεων— παρουσιάζουν ομοιότητες ή διαφορές. Σχεδόν σε όλες τις περιπτώσεις, τα αποτελέσματα υποστηρίζουν τη μηδενική υπόθεση υποδεικνύοντας ότι οι στρατηγικές είναι στατιστικά όμοιες. Μόνο ελάχιστες εξαιρέσεις παρατηρούνται, που αφορούν ένα μικρό ποσοστό περιπτώσεων, όπως απεικονίζεται στο σχήμα 8.8 . Συνεπώς, στο εξεταζόμενο πειραματικό πλαίσιο, οι τρεις στρατηγικές ερωτήσεων (uncertainty, margin, και entropy) εμφανίζουν συνολικά συγκρίσιμη απόδοση, επιβεβαιώνοντας ότι καμία από αυτές δεν υπερέχει σημαντικά έναντι των άλλων.

8.4 Συμπεράσματα

Τα αποτελέσματα της παρούσας εργασίας αναδεικνύουν τη δυναμική ενσωμάτωσης τεχνικών Ενεργής Μάθησης στον τομέα της πρόβλεψης σφαλμάτων, καταδεικνύοντας τη δυνατότητα εκπαίδευσης μοντέλων ταξινόμησης με σημαντικά μειωμένα δεδομένα, χωρίς απώλεια απόδοσης. Συγκεκριμένα, επιτεύχθηκαν μειώσεις στο αρχικό σύνολο εκπαίδευσης κατά τουλάχιστον 75% στο 64% των περιπτώσεων, γεγονός που υπογραμμίζει την αποτελεσματικότητα της προσέγγισης. Επιπλέον, η χρήση σχημάτων ομαδοποιητών ταξινόμησης αποδείχθηκε ιδιαίτερα αποδοτική, ξεπερνώντας σε επιδόσεις τα 15 μεμονωμένα μοντέλα που αξιολογήθηκαν, ενώ εντοπίστηκε ένα υποσύνολο αυτών που παρουσίασε εξαιρετική απόδοση υπό την επιρροή των τεχνικών Ενεργής Μάθησης.

	AUC					
	Uncertainty		**Margin**		**Entropy**	
	Method	**Index Score**	**Method**	**Index Score**	**Method**	**Index Score**
1st	rf+gbc+lda	0.863134719	rf+gbc+lda	0.862038403	rf+gbc+lda	0.860376476
2nd	rf+gbc+nb	0.835231113	rf+gbc+nb	0.835231113	rf+gbc+nb	0.835231113
3rd	et+gbc+lightgbm+mlp+lda	0.826148093	et+gbc+lightgbm+mlp+lda	0.824989332	et+gbc+lightgbm+mlp+lda	0.831999764
4th	nb	0.824302255	nb	0.824302255	nb	0.824302255
5th	et+rf+gbc+mlp+lda	0.811464959	gbc+lightgbm+lda	0.82369667	et+rf+gbc+nb+lda	0.818962941
6th	et+rf+gbc+nb+lda	0.811337433	et+rf+gbc+mlp+lda	0.811464959	et+rf+gbc+mlp+lda	0.810575626
7th	gbc+lightgbm+mlp+nb+lda	0.807382069	rf+mlp+nb	0.811361599	gbc+lightgbm+mlp+nb+lda	0.809786535
8th	et+rf+gbc+lightgbm+mlp	0.803953633	et+rf+gbc+nb+lda	0.811337433	et+rf+gbc+lightgbm+mlp	0.804100155
9th	rf+gbc+lightgbm+nb+lda	0.802860858	gbc+lightgbm+mlp+nb+lda	0.806776026	gbc+lightgbm+lda	0.799780964
10th	rf+mlp+nb	0.794501658	et+rf+gbc+lightgbm+mlp	0.803953633	et+rf+gbc+lightgbm+lda	0.792803737
	Kappa					
	Uncertainty		**Margin**		**Entropy**	
	Method	**Index Score**	**Method**	**Index Score**	**Method**	**Index Score**
1st	rf+gbc+lda	0.847511069	rf+gbc+lda	0.847169628	et+gbc+lightgbm+mlp+lda	0.848241987
2nd	et+gbc+lightgbm+mlp+lda	0.847304105	et+gbc+lightgbm+mlp+lda	0.846900592	nb	0.846031914
3rd	nb	0.846031914	nb	0.846031914	rf+gbc+lda	0.84033313
4th	ada	0.808839992	ada	0.808839992	ada	0.808839992
5th	et+rf+lightgbm+mlp+lda	0.801275604	rf+mlp+nb	0.805555937	et+rf+lightgbm+mlp+lda	0.800577236
6th	rf+gbc+mlp+nb+lda	0.794182718	et+rf+lightgbm+mlp+lda	0.801275604	et+rf+gbc+mlp+lda	0.790110985
7th	rf+gbc+nb	0.789758769	rf+gbc+nb	0.789758769	rf+gbc+mlp+nb+lda	0.789806832
8th	et+rf+gbc+mlp+lda	0.788590006	et+rf+gbc+mlp+lda	0.788590006	rf+gbc+nb	0.789758769
9th	rf+mlp+nb	0.786975262	rf+gbc+mlp+nb+lda	0.787317347	gbc+lightgbm+mlp+nb+lda	0.780856161
10th	rf+lightgbm+mlp+nb+lda	0.783655684	gbc+lightgbm+lda	0.785656682	lightgbm+nb+lda	0.78033235
	MCC					
	Uncertainty		**Margin**		**Entropy**	
	Method	**Index Score**	**Method**	**Index Score**	**Method**	**Index Score**
1st	rf+gbc+lda	0.89951329	rf+gbc+lda	0.896390608	rf+gbc+lda	0.89252349
2nd	et+gbc+lightgbm+mlp+lda	0.849184722	et+gbc+lightgbm+mlp+lda	0.844729798	et+gbc+lightgbm+mlp+lda	0.850168054
3rd	rf+gbc+mlp+nb+lda	0.848041873	rf+gbc+mlp+nb+lda	0.8407418	rf+gbc+mlp+nb+lda	0.844462315
4th	nb	0.836449051	nb	0.836449051	nb	0.836449051
5th	lightgbm+nb+lda	0.824515904	lightgbm+nb+lda	0.824515904	rf+mlp+lda	0.832798034
6th	et+rf+lightgbm+mlp+lda	0.817568969	rf+mlp+lda	0.823352278	lightgbm+nb+lda	0.824515904
7th	rf+mlp+lda	0.815697261	rf+mlp+nb	0.820144716	et+rf+lightgbm+mlp+lda	0.815147449
8th	et+rf+gbc+mlp+lda	0.808748503	et+rf+lightgbm+mlp+lda	0.817568969	et+rf+gbc+mlp+lda	0.813475923
9th	gbc+nb+lda	0.804776492	et+rf+gbc+mlp+lda	0.808748503	gbc+nb+lda	0.811221409
10th	rf+mlp+nb	0.801185979	gbc+nb+lda	0.804776492	rf+gbc+nb	0.797554092

Πίνακας 8.5: Η απόδοση των μεθόδων σύμφωνα με τη μετρική EAPI

Ένα από τα σημαντικά επιτεύγματα της μελέτης ήταν η εισαγωγή του δείκτη απόδοσης προσαρμοσμένης αποτελεσματικότητας (EAPI), ο οποίος προσφέρει μια διττή προσέγγιση για την αξιολόγηση των μοντέλων. Ο δείκτης αυτός όχι μόνο επιτρέπει την αναγνώριση των καλύτερων στρατηγικών σε περιβάλλοντα πειραματισμού που βασίζονται στη χρήση ενεργητικής μάθησης, αλλά και παρέχει μια ποσοτική βάση για τη σύγκριση στρατηγικών με διαφορετικά χαρακτηριστικά. Το γεγονός ότι δεν αναδείχθηκε μια σαφής στρατηγική ως ¨καλύτερη' σε όλα τα πειράματα υποδηλώνει ότι η επιλογή στρατηγικής πρέπει να λαμβάνει υπόψη συγκεκριμένους παράγοντες, όπως τα χαρακτηριστικά του εκάστοτε σύνολο δεδομένων, η φύση της ταξινόμησης και οι υπολογιστικοί περιορισμοί. Ο EAPI, με την ικανότητά του να προσαρμόζεται σε διαφορετικές συνθήκες, επιτρέπει μια πιο σφαιρική και αξιόπιστη αξιολόγηση των επιδόσεων, καθιστώντας τον ένα εργαλείο με ευρεία εφαρμογή και πέρα από τον τομέα ττης πρόβλεψης σφαλμάτων λογισμικού.

Συνολικά, η μελέτη συμβάλλει σημαντικά στην κατανόηση των τεχνικών Ενεργής Μάθησης στο πεδίο της πρόβλεψης σφαλμάτων λογισμικού και παρέχει χρήσιμες κατευθύνσεις για τη βέλτιστη αξιοποίησή τους σε ταξινομητικά προβλήματα. Τα ευρήματα μπορούν να αξιοποιηθούν για τον σχεδιασμό μελλοντικών ερευνών, αποσκοπώντας στη βελτιστοποίηση της εφαρμογής των τεχνικών Ενεργής Μάθησης στον συγκεκριμένο

τομέα, αλλά και σε συναφείς επιστημονικές περιοχές, όπου η οικονομία δεδομένων και η υψηλή ακρίβεια είναι ζωτικής σημασίας.

9. Επίλογος

9.1 Κύρια συμπεράσματα της διδακτορικής διατριβής

Στην υποενότητα αυτή, παρουσιάζονται συγκεντρωτικά και επιγραμματικά εν είδει επιλόγου τα κύρια συμπεράσματα της διδακτορικής διατριβής, όπως αυτά εντοπίζονται από τη μελέτη των αποτελεσμάτων των εκτεταμένων πειραματικών διαδικασιών και παρουσιάζονται ανά περίπτωση στο τέλος κάθε κεφαλαίου.

Οι δύο βασικοί άξονες της εργασίας, βασίζονται στα ερωτήματα:

1. Μπορούν οι μέθοδοι Ανάλυσης Συναισθήματος όταν χρησιμοποιούνται στο πρόβλημα πρόβλεψης χρονοσειρών οικονομικών δεδομένων να βελτιώσουν τις εξαγόμενες προβλέψεις;

2. Μπορεί η χρήση τεχνικών Ενεργής Μάθησης να οδηγήσει στη μείωση των απαιτούμενων ετικετοποιημένων παραδειγμάτων διατηρώντας παρόμοια επίπεδα αποτελεσματικότητας όσον αφορά τις διάφορες μετρικές αξιολόγησης;

Σχετικά με το πρώτο ερώτημα, και στις δύο περιπτώσεις που διερευνήθηκαν, η απάντηση είναι καταφατική. Τόσο σε σενάρια χρήσης κλασικών αλγορίθμων Μηχανικής Μάθησης, όσο και σε σενάρια που χρησιμοποιούν σύγχρονες μεθόδους Βαθιάς Μάθησης, η ενσωμάτωση της πληροφορίας της άποψης της κοινής γνώμης, όπως αυτή αντλήθηκε από τα μέσα κοινωνικής δικτύωσης και αποτιμήθηκε μέσω διαφορετικών τεχνικών (γενικής χρήσης, αλλά και εξειδικευμένων στο οικονομικό πεδίο) υπήρξε ευεργετική. Το σενάριο χρήσης Ανάλυσης Συναισθήματος υπερέχει του κλασικού σεναρίου όπου μόνο η τιμή της χρονοσειράς χρησιμοποιείται για την πρόβλεψη, στη συντριπτική πλειοψηφία των περιπτώσεων. Υπήρξαν επίσης, όπως συζητήθηκε στο Κεφάλαιο 5, συγκεκριμένες συναισθηματικές αποτιμήσεις που δείχνουν να υπερέχουν των υπολοίπων, χωρίς ωστόσο να υπάρχει μια καθολική πρωτιά όσον αφορά το σύνολο των χρονικών οριζόντων και μετρικών αποτίμησης. Επιπλέον, η χρήση Ανάλυσης Συναισθήματος υπήρξε περισσότερο ωφέλιμη στα σενάρια όπου ο χρονικός ορίζοντας πρόβλεψης είναι μεγαλύτερος, κάτι ιδιαίτερα χρήσιμο καθώς η δυσκολία τέτοιων προβλεπτικών σεναρίων είναι μεγαλύτερη.

Σχετικά με το δεύτερο ερευνητικό ερώτημα, η απάντηση είναι και πάλι καταφατική. Στα δύο σενάρια που εξετάστηκαν, αυτό της Πρόβλεψης Σφαλμάτων Λογισμικού και εκείνο της χρήσης BERT-based transformers στο πρόβλημα της συναισθηματικής ταξινόμησης κειμένου, στην πλειοψηφία των περιπτώσεων μπορέσαμε να έχουμε το ίδιο καλά αποτελέσματα χρησιμοποιώντας σημαντικά μικρότερο ποσοστό του αρχικού συνόλου δεδομένων. Στο σενάριο μάλιστα της Πρόβλεψης Σφαλμάτων Λογισμικού, μπορέσαμε να αναπαράγουμε τα ίδια αποτελέσματα μετρικών αξιολόγησης χρησιμοποιώντας ποσοστό μικρότερο του 15% του αρχικού συνόλου περίπου στο 41% με 42%

των περιπτώσεων, αναλόγως τη μετρική. Τέλος, προτάθηκε ένας νέος μεταδείκτης, ο Δείκτης Απόδοσης Προσαρμοσμένος στην Αποδοτικότητα (Efficiency-Adjusted Performance Index - EAPI), ο οποίος, αν και μπορεί να χρησιμοποιηθεί μόνο σε ερευνητικά σενάρια όπου υπάρχει διαθέσιμο και επισημασμένο το πλήρες σύνολο δεδομένων, μπορεί να αξιολογήσει την απόδοση της Ενεργής Μάθησης χρησιμοποιώντας πληροφορία που σχετίζεται όχι μόνο με την αποτελεσματικότητα της μετρικής στην οποία χρησιμοποιείται, αλλά και με το ποσοστό μείωσης των επισημασμένων δεδομένων.

9.2 Μελλοντικές επεκτάσεις της διδακτορικής διατριβής

Φτάνοντας στο τέλος της διατριβής, θα ήταν παράληψη να μην αναφερθούν μελλοντικές προεκτάσεις της παρούσας. Σε αυτή την παράγραφο σημειώνονται κάποιες από τις κατευθύνσεις στις οποίες θα μπορούσε να συνεχιστεί η παρούσα έρευνα.

Εφαρμογές σε Διαφορετικά Πεδία Φυσική συνέχεια της παρούσας εργασίας θα ήταν η εφαρμογή των μεθόδων που υλοποιήθηκαν σε διαφορετικά ερευνητικά πεδία που επίσης χαρακτηρίζονται από την ύπαρξη δυναμικών φαινομένων που επηρεάζουν και επηρεάζονται αμφίδρομα από την κοινή γνώμη ή συνδέονται με πληροφορία η οποία μπορεί να αποτυπωθεί στις εκφράσεις των ατόμων μέσω των κοινωνικών δικτύων ή τον τύπο. Κάποια από αυτά, ενδεικτικά, αναφέρονται:

- Υγειονομική Πληροφορική [51]: Η ανάλυση συναισθημάτων σε δεδομένα ιατρικής φύσης (π.χ. ιατρικά φόρουμ, κριτικές υπηρεσιών, δεδομένα από εφαρμογές υγείας) μπορεί να συμβάλει στη διάγνωση ψυχολογικών καταστάσεων ή και σωματικών νοσημάτων που συσχετίζονται με το στρες ή άλλους δείκτες. Η Ενεργή Μάθηση βοηθά στην επιτάχυνση της διαδικασίας ανάλυσης, επιτρέποντας στο σύστημα να εστιάζει στα πλέον ασαφή ή κρίσιμα δείγματα που χρειάζονται ανθρώπινη επιβεβαίωση.
- Νομική Τεχνολογία (LegalTech) [262]: Η αυτοματοποιημένη κατηγοριοποίηση και επεξεργασία νομικών εγγράφων (π.χ. αποφάσεις, συμβόλαια, δικόγραφα) αποτελεί μια εξαιρετικά απαιτητική διαδικασία. Η ενσωμάτωση τεχνικών Ενεργής Μάθησης σε προβλήματα Επεξεργασίας Φυσικής Γλώσσας νομικού ενδιαφέροντος θα μπορούσε να συμβάλει στη δημιουργία νομικών βάσεων δεδομένων, να επιταχύνει και να διευκολύνει την ανεύρεση προηγούμενων αποφάσεων και να μειώνει τα λειτουργικά κόστη σε δικηγορικά γραφεία και υπηρεσίες.

Βελτίωση Αλγορίθμων και Νέες Υβριδικές Προσεγγίσεις Η Ενεργή Μάθηση έχει ήδη αποδειχθεί αποτελεσματική στη μείωση των απαιτήσεων για ετικετοποιημένα δεδομένα, ωστόσο η συνεχής βελτίωση των αλγορίθμων της και η αξιοποίηση νέων υβριδικών μοντέλων αποτελούν σημαντικές μελλοντικές κατευθύνσεις σε αναλογία με την όλο και αυξανόμενη πολυπλοκότητα των μοντέλων. Ένα πρώτο βήμα θα μπορούσε να είναι ο συνδυασμός της Ενεργής Μάθησης με μεθόδους Ενισχυτικής Μάθησης. Σε τέτοια σενάρια, ο αλγόριθμος Ενισχυτικής Μάθησης μπορεί να μάθει πώς να επιλέγει αποδοτικότερα τα πιο χρήσιμα δείγματα προς επισήμανση, επιτυγχάνοντας οικονομία στις διαδικασίες επισήμανσης και παράλληλα βελτιώνοντας

την ακρίβεια του μοντέλου. Επιπλέον, η ανάπτυξη meta-learning [279],[41] τεχνικών αποτελεί μια πολλά υποσχόμενη προοπτική καθώς επιτρέπει στο σύστημα να μαθαίνει πώς να μαθαίνει. Με τον τρόπο αυτό, η στρατηγική της Ενεργής Μάθησης θα μπορεί να προσαρμόζεται δυναμικά σε διαφορετικά σύνολα δεδομένων και προβλήματα, χωρίς να απαιτείται εκ νέου επανασχεδιασμός. Παράλληλα, η διερεύνηση αυτοεπιβλεπόμενων μεθόδων σε συνδυασμό με Ενεργή Μάθηση μπορεί να μειώσει περαιτέρω την εξάρτηση από ετικετοποιημένα δεδομένα, αντλώντας πληροφορία από μη επισημασμένα δείγματα. Με αυτό τον τρόπο, το μοντέλο είναι σε θέση να εκμεταλλευτεί μεγάλο όγκο μη ετικετοποιημένων δεδομένων και να βελτιώνει διαρκώς την επίδοσή του σε πραγματικές εφαρμογές.

Χρήση Ενεργής Μάθησης σε streaming δεδομένα Ο χειρισμός μεγάλων ροών δεδομένων (streaming data), όπου τα δεδομένα παράγονται ή αλλάζουν συνεχώς σε πραγματικό χρόνο (π.χ. social media, ειδησεογραφικά websites, IoT συσκευές), αποτελεί μια επιπλέον πρόκληση. Σε τέτοια δυναμικά περιβάλλοντα, η Ενεργή μάθηση θα πρέπει να προσαρμόζεται ταχύτατα, επιλέγοντας δείγματα προς επισήμανση σε πραγματικό χρόνο, με στόχο τη δημιουργία ακριβών μοντέλων που θα μπορούν να παραμείνουν ενημερωμένα. Ένα σημαντικό βήμα προς αυτή την κατεύθυνση είναι η ανάπτυξη τεχνικών Συνεχούς Μάθησης (Continual Learning) [60] σε συνδυασμό με Ενεργή Μάθηση. Με αυτόν τον τρόπο, το μοντέλο διατηρεί όσα έχει μάθει από παλαιότερα δεδομένα, ενώ προσαρμόζεται στα νεότερα, αποφεύγοντας φαινόμενα καταστροφικής λήθης (catastrophic forgetting) [212]. Αυτή η ενσωμάτωση θα μπορούσε να αποδειχθεί ιδιαίτερα χρήσιμη σε σενάρια όπου οι ροές δεδομένων είναι ετερογενείς και οι απαιτήσεις σε χρόνο απόκρισης είναι αυξημένες.

Υπολογιστική αποδοτικότητα και επεκτασιμότητα Τέλος, ένα κρίσιμο ζήτημα για την πρακτική εφαρμογή της Ενεργής Μάθησης σε ευρεία κλίμακα αφορά την υπολογιστική αποδοτικότητα. Πολλά μοντέλα Ενεργής Μάθησης, ειδικά όσα βασίζονται σε τεχνικές Βαθιάς Μάθησης, απαιτούν υψηλό υπολογιστικό κόστος για τη συνεχή επανεκπαίδευση ή την επαναξιολόγηση. Η βελτιστοποίηση αυτών των αλγορίθμων, είτε μέσω πιο αποδοτικών αρχιτεκτονικών είτε μέσω τεχνικών συμπίεσης (model compression) [147] και κβαντοποίησης (quantization) [103], μπορεί να τα καταστήσει πιο προσιτά σε περιβάλλοντα edge computing (π.χ. κινητές συσκευές, IoT κόμβοι). Επιπλέον, η εξερεύνηση τεχνικών παράλληλης επεξεργασίας και κατανεμημένης μάθησης (distributed learning) [297] θα μπορούσε να επιταχύνει σημαντικά την όλη διαδικασία: η επιβάρυνση κατανέμεται σε πολλούς επεξεργαστές ή συστοιχίες υπολογιστών, μειώνοντας τον χρόνο που απαιτείται για την επιλογή δειγμάτων, την επισήμανση και την εκπαίδευση των μοντέλων. Έτσι, εξασφαλίζεται η δυνατότητα εφαρμογής της Ενεργής Μάθησης σε βιομηχανικό ή ερευνητικό επίπεδο για προβλήματα μεγάλης κλίμακας.

Bibliography

[1] Shaza M Abd Elrahman and Ajith Abraham. «A review of class imbalance problem». In: *Journal of Network and Innovative Computing* 1 (2013), pp. 9–9.

[2] Rakesh Agrawal, Tomas Imielinski, and Arun Swami. «Mining association rules between sets of items in large databases». In: *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*. ACM. 1993, pp. 207–216.

[3] Asma Akhtar et al. «COVID-19 detection from CBC using machine learning techniques». In: *International Journal of Technology Innovation and Management (IJTIM)* 1.2 (2021), pp. 65–78.

[4] Muhammad Shoaib Akhtar and Tao Feng. «Malware analysis and detection using machine learning algorithms». In: *Symmetry* 14.11 (2022), p. 2304.

[5] Ali Saleh Alammary. «BERT models for Arabic text classification: a systematic review». In: *Applied Sciences* 12.11 (2022), p. 5720.

[6] Fawaz Khaled Alarfaj et al. «Credit card fraud detection using state-of-the-art machine learning and deep learning algorithms». In: *Ieee Access* 10 (2022), pp. 39700–39715.

[7] Fadi Alharbi and Aleksandar Vakanski. «Machine learning methods for cancer classification using gene expression data: A review». In: *Bioengineering* 10.2 (2023), p. 173.

[8] Moez Ali. *PyCaret: An open source, low-code machine learning library in Python*. PyCaret version 1.0.0. Apr. 2020. URL: `https://www.pycaret.org`.

[9] Constantin Aliferis and Gyorgy Simon. «Overfitting, underfitting and general model overconfidence and under-performance pitfalls and best practices in machine learning and AI». In: *Artificial intelligence and machine learning in health care and medical sciences: Best practices and pitfalls* (2024), pp. 477–524.

[10] Ethem Alpaydin. *Introduction to Machine Learning*. 3rd. Cambridge, MA, USA: MIT Press, 2014.

[11] Ethem Alpaydin. *Introduction to Machine Learning*. 4th. Cambridge, MA, USA: MIT Press, 2020.

[12] Ahmad Alzu'bi et al. «Masked face recognition using deep learning: A review». In: *Electronics* 10.21 (2021), p. 2666.

[13] Yash Amethiya et al. «Comparative analysis of breast cancer detection using machine learning and biosensors». In: *Intelligent Medicine* 2.2 (2022), pp. 69–81.

[14] Liat Antwarg et al. «Explaining anomalies detected by autoencoders using Shapley Additive Explanations». In: *Expert systems with applications* 186 (2021), p. 115736.

[15] Deniz Araci. «FinBERT: Financial Sentiment Analysis with Pre-trained Language Models». In: *arXiv preprint arXiv:1908.10063* (2019).

[16] M Arshey and KS Angel Viji. «Thwarting cyber crime and phishing attacks with machine learning: a study». In: *2021 7th international conference on advanced computing and communication systems (ICACCS)*. Vol. 1. IEEE. 2021, pp. 353–357.

[17] Jordan T. Ash et al. «Deep Batch Active Learning by Diverse, Uncertain Gradient Embeddings». In: *Proceedings of the 8th International Conference on Learning Representations (ICLR)*. 2020.

[18] Naila Aslam et al. «Sentiment Analysis and Emotion Detection on Cryptocurrency Related Tweets Using Ensemble LSTM-GRU Model». In: *IEEE Access* (Jan. 2022). DOI: 10.1109/access.2022.3165621.

[19] Ali Athar et al. «Sentimental Analysis of Movie Reviews using Soft Voting Ensemble-based Machine Learning». In: *2021 Eighth International Conference on Social Network Analysis, Management and Security (SNAMS)* (2021), pp. 01–05. URL: https://api.semanticscholar.org/CorpusID:247477419.

[20] Mrinal R Bachute and Javed M Subhedar. «Autonomous driving architectures: insights of machine learning and deep learning algorithms». In: *Machine Learning with Applications* 6 (2021), p. 100164.

[21] Shaojie Bai and Vladlen Koltun. «An empirical evaluation of generic convolutional and recurrent networks for sequence modeling». In: *arXiv preprint arXiv:1803.01271* (2018).

[22] Emanuele Bajo and Carlo Raimondo. «Media sentiment and IPO underpricing». In: *Journal of Corporate Finance* 46 (2017), pp. 139–153. ISSN: 0929-1199. DOI: https://doi.org/10.1016/j.jcorpfin.2017.06.003. URL: https://www.sciencedirect.com/science/article/pii/S092911991730370X.

[23] Alexandra Balahur and Marco Turchi. «Improving Sentiment Analysis in Twitter Using Multilingual Machine Translated Data». In: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. 2013, pp. 871–879.

[24] Solon Barocas and Andrew D. Selbst. «Big Data's Disparate Impact». In: *California Law Review* 104.3 (2016), pp. 671–732.

[25] Valerio Basile et al. «SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter». In: *Proceedings of the 13th International Workshop on Semantic Evaluation*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, 2019, pp. 54–63. DOI: 10.18653/v1/S19-2007. URL: https://www.aclweb.org/anthology/S19-2007.

[26] Leonard E. Baum and Ted Petrie. «Statistical Inference for Probabilistic Functions of Finite State Markov Chains». In: *The Annals of Mathematical Statistics* 41.5 (1970), pp. 1554–1563.

[27] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. O'Reilly Media, 2009.

[28] Marouane Birjali, Mohammed Kasri, and Abderrahim Beni-Hssane. «A comprehensive survey on sentiment analysis: Approaches, challenges and trends». In: *Knowledge-Based Systems* 226 (2021), p. 107134.

[29] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.

[30] Barry Boehm and Victor R Basili. «Top 10 list [software development]». In: *Computer* 34.1 (2001), pp. 135–137.

[31] George E. P. Box and Gwilym M. Jenkins. *Time Series Analysis: Forecasting and Control*. Holden-Day, 1970.

[32] George EP Box and David R Cox. «Analysis of transformations». In: *Journal of the Royal Statistical Society: Series B (Methodological)* 26.2 (1964), pp. 211–243.

[33] George EP Box et al. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[34] A. Bozanta et al. «Sentiment Analysis of StockTwits Using Transformer Models». In: *2021 IEEE International Conference on Machine Learning and Applications (ICMLA)*. 2021, pp. 1253–1258.

[35] L. Breiman. «Random Forests». In: *Machine Learning* 45 (2001), pp. 5–32. DOI: 10.1023/A:1010933404324.

[36] L. Breiman et al. *Classification and Regression Trees*. Oxfordshire, UK: Routledge, 2017.

[37] Markus M. Breunig et al. «LOF: Identifying Density-Based Local Outliers». In: *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*. ACM. 2000, pp. 93–104.

[38] Joy Buolamwini and Timnit Gebru. «Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification». In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (FAT*)*. 2018, pp. 77–91.

[39] Erik Cambria et al. «New Avenues in Opinion Mining and Sentiment Analysis». In: *IEEE Intelligent Systems* 28.2 (2017), pp. 15–21.

[40] Gürol Canbek, Tugba Taskaya Temizel, and Seref Sagiroglu. «PToPI: A comprehensive review, analysis, and knowledge representation of binary classification performance measures/metrics». In: *SN Computer Science* 4.1 (2022), p. 13.

[41] Davide Carneiro et al. «Using meta-learning to predict performance metrics in machine learning problems». In: *Expert Systems* 40.1 (2023), e12900.

[42] Joymallya Chakraborty, Suvodeep Majumder, and Tim Menzies. «Bias in machine learning software: Why? how? what to do?» In: *Proceedings of the 29th ACM joint meeting on European software engineering conference and symposium on the foundations of software engineering*. 2021, pp. 429–440.

[43] V. Chandola, A. Banerjee, and V. Kumar. «Anomaly Detection: A Survey». In: *ACM Computing Surveys (CSUR)* 41.3 (2009), pp. 1–58.

[44] Chris Chatfield. *Time-series forecasting*. CRC Press, 2000.

[45] Leiyu Chen et al. «Review of image classification algorithms based on convolutional neural networks». In: *Remote Sensing* 13.22 (2021), p. 4712.

[46] T. Chen et al. *XGBoost: Extreme Gradient Boosting*. R package version 0.4-2. 2015.

[47] Wangmei Chen et al. «Applying machine learning algorithm to optimize personalized education recommendation system». In: *Journal of Theory and Practice of Engineering Science* 4.01 (2024), pp. 101–108.

[48] S. Choi, T. Kim, and W. Yu. «Performance Evaluation of RANSAC». In: *Proceedings of the Conference in London, UK*. 7–10 September 2009. 2009.

[49] Junyoung Chung et al. «Empirical evaluation of gated recurrent neural networks on sequence modeling». In: *arXiv preprint arXiv:1412.3555* (2014).

[50] David A. Cohn, Les Atlas, and Richard E. Ladner. «Improving Generalization with Active Learning». In: *Machine Learning* 15.2 (1994), pp. 201–221.

[51] Enrico Coiera. *Guide to health informatics*. CRC press, 2015.

[52] Stuart Colianni, Stephanie Rosales, and Michael Signorotti. «Algorithmic trading of cryptocurrency based on Twitter sentiment analysis». In: *CS229 Project* 1.5 (2015), pp. 1–4.

[53] Corinna Cortes and Vladimir Vapnik. «Support-vector networks». In: *Machine Learning*. Vol. 20. 3. Springer. 1995, pp. 273–297.

[54] K. Crammer et al. «Online Passive-Aggressive Algorithms». In: *Journal of Machine Learning Research* 7 (2006), pp. 551–585.

[55] M. D'Ambros, M. Lanza, and R. Robbes. «An extensive comparison of bug prediction approaches». In: *Proceedings of MSR 2010 (7th IEEE Working Conference on Mining Software Repositories)*. IEEE CS Press, 2010, pp. 31–41.

[56] X. Dang et al. «The Theil-Sen Estimators in a Multiple Linear Regression Model». In: *Proceedings of the British Machine Vision Conference*. Manuscript, details incomplete. 2024.

[57] Jamal I Daoud. «Multicollinearity and regression analysis». In: *journal of physics: Conference series*. Vol. 949. 1. IOP Publishing. 2017, p. 012009.

[58] Sanjoy Dasgupta. «Two Faces of Active Learning». In: *Theoretical Computer Science* 412 (2009), pp. 1767–1781.

[59] Ronald J. Williams David E. Rumelhart Geoffrey E. Hinton. «Learning representations by back-propagating errors». In: *Nature* 323 (1986), pp. 533–536.

[60] Matthias De Lange et al. «A continual learning survey: Defying forgetting in classification tasks». In: *IEEE transactions on pattern analysis and machine intelligence* 44.7 (2021), pp. 3366–3385.

[61] Alysha M De Livera, Rob J Hyndman, and Ralph D Snyder. «Forecasting with TBATS: Trigonometric Box-Cox, ARMA, Trend and Seasonal models». In: *Journal of the American Statistical Association* 106.496 (2011), pp. 1513–1527.

[62] Kim Denecke. «Using SentiWordNet for Multilingual Sentiment Analysis». In: *Proceedings of the 24th International Conference on Data Engineering Workshop*. 2008, pp. 507–512.

[63] Justina Deveikyte et al. «A sentiment analysis approach to the prediction of market volatility». In: *Frontiers in Artificial Intelligence* 5 (2022), p. 836809.

[64] Debashree Devi, Saroj K Biswas, and Biswajit Purkayastha. «A review on solution to class imbalance problem: Undersampling approaches». In: *2020 international conference on computational performance evaluation (ComPE)*. IEEE. 2020, pp. 626–631.

[65] Jacob Devlin et al. «BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding». In: *arXiv preprint arXiv:1810.04805* (2018).

[66] Jacob Devlin et al. «BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding». In: *Proceedings of NAACL-HLT*. 2019.

[67] Jacob Devlin et al. «Bert: Pre-training of deep bidirectional transformers for language understanding». In: *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. 2019, pp. 4171–4186.

[68] L. Devroye et al. «On the Strong Universal Consistency of Nearest Neighbor Regression Function Estimates». In: *Annals of Statistics* 22 (2007), pp. 1371–1385. DOI: 10.1214/009053607000000764.

[69] Tom Dietterich. «Overfitting and undercomputing in machine learning». In: *ACM computing surveys (CSUR)* 27.3 (1995), pp. 326–327.

[70] Kelechukwu Celestine Nosike Dozie and Maxwell Azubuike Ijomah. «A comparative study on additive and mixed models in descriptive time series». In: *American Journal of Mathematical and Computer Modelling* 5.1 (2020), pp. 12–17.

[71] H. Drucker. «Improving Regressors Using Boosting Techniques». In: *Proceedings of the Fourteenth International Conference on Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann, 1997, pp. 107–115.

[72] Kavisha Duggal, Lovi Raj Gupta, and Parminder Singh. «Gamification and machine learning inspired approach for classroom engagement and learning». In: *Mathematical Problems in Engineering* 2021.1 (2021), p. 9922775.

[73] Jannik Dunkelau and Michael Leuschel. «Fairness-aware machine learning». In: *An extensive overview* (2019), pp. 1–60.

[74] Cristina Efremov et al. «Improving syngas yield and quality from biomass/coal co-gasification using cooperative game theory and local interpretable model-agnostic explanations». In: *International Journal of Hydrogen Energy* 96 (2024), pp. 892–907.

[75] B. Efron et al. «Least Angle Regression». In: *Annals of Statistics* 32 (2004), pp. 407–499. DOI: 10.1214/009053604000000067.

[76] Naji Elsayed, Anthony Maida, and Magdy Bayoumi. «Deep gated recurrent and convolutional network hybrid model for univariate time series classification». In: *arXiv preprint arXiv:1812.07683* (2019).

[77] Martin Ester et al. «A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise». In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*. 1996, pp. 226–231.

[78] J. Fan et al. «Light Gradient Boosting Machine: An Efficient Soft Computing Model for Estimating Daily Reference Evapotranspiration with Local and External Meteorological Data». In: *Agricultural Water Management* 225 (2019), p. 105758. DOI: 10.1016/j.agwat.2019.105758.

[79] Zahra Fathali, Zahra Kodia, and Lamjed Ben Said. «Stock market prediction of Nifty 50 index applying machine learning techniques». In: *Applied Artificial Intelligence* 36.1 (2022), p. 2111134.

[80] Mathis Fauvel et al. «XCM: An explainable convolutional neural network for multivariate time series classification». In: *arXiv preprint arXiv:2009.04796* (2021).

[81] Hassan Ismail Fawaz et al. «InceptionTime: Finding AlexNet for time series classification». In: *arXiv preprint arXiv:1909.04939* (2020).

[82] Javad Feizabadi. «Machine learning demand forecasting and supply chain performance». In: *International Journal of Logistics Research and Applications* 25.2 (2022), pp. 119–142.

[83] Peter Filzmoser and Klaus Nordhausen. «Robust linear regression for high-dimensional data: An overview». In: *Wiley Interdisciplinary Reviews: Computational Statistics* 13.4 (2021), e1524.

[84] *Financial Sentiment Analysis — kaggle.com*. https://www.kaggle.com/datasets/sbhatti/financial-sentiment-analysis. [Accessed 29-09-2023].

[85] Chelsea Finn, Pieter Abbeel, and Sergey Levine. «Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks». In: *Proceedings of the 34th International Conference on Machine Learning (ICML)*. 2017, pp. 1126–1135. URL: https://arxiv.org/abs/1703.03400.

[86] Andrew Christian Flores et al. «An Evaluation of SVM and Naive Bayes with SMOTE on Sentiment Analysis Data Set». In: *2018 International Conference on Engineering, Applied Sciences, and Technology (ICEAST)* (2018), pp. 1–4. URL: https://api.semanticscholar.org/CorpusID:52014088.

[87] Yoav Freund et al. «Selective Sampling Using the Query by Committee Algorithm». In: *Machine Learning* 28 (1997), pp. 133–168.

[88] J.H. Friedman. «Greedy Function Approximation: A Gradient Boosting Machine». In: *Annals of Statistics* 29 (2001), pp. 1189–1232. DOI: 10.1214/aos/1013203451.

[89] Yusuke Fujii and Yutaka Matsuo. «Reinforcement Active Learning for Deep Neural Networks». In: *Neural Computation* 34 (2022), pp. 1462–1485.

[90] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. «Deep Bayesian Active Learning with Image Data». In: *Proceedings of the 34th International Conference on Machine Learning (ICML)* (2017).

[91] Bo Gao. «The use of machine learning combined with data mining technology in financial risk prevention». In: *Computational economics* 59.4 (2022), pp. 1385–1405.

[92] Laura Garcia Cuenca et al. «Machine learning techniques for undertaking roundabouts in autonomous driving». In: *Sensors* 19.10 (2019), p. 2386.

[93] Everette S Gardner. «Exponential smoothing: The state of the art». In: *Journal of Forecasting* 4.1 (1985), pp. 1–28.

[94] Brandon George, Samantha Seals, and Inmaculada Aban. «Survival analysis and regression models». In: *Journal of nuclear cardiology* 21.4 (2014), pp. 686–694.

[95] P. Geurts, D. Ernst, and L. Wehenkel. «Extremely Randomized Trees». In: *Machine Learning* 63 (2006), pp. 3–42. DOI: 10.1007/s10994-006-6226-1.

[96] Debanjan Ghosh, Wei Guo, and Smaranda Muresan. «Sarcastic or Not: Word Embeddings to Predict the Literal or Sarcastic Meaning of Words». In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 2015, pp. 1003–1012.

[97] Ben Goertzel. «Human-level artificial general intelligence and the possibility of a technological singularity». In: *Journal of Consciousness Studies* 13.1–2 (2007), pp. 67–84.

[98] Wenceslao González-Manteiga and Rosa M Crujeiras. «An updated review of goodness-of-fit tests for regression models». In: *Test* 22 (2013), pp. 361–411.

[99] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[100] Ian Goodfellow et al. «Generative adversarial nets». In: *Advances in Neural Information Processing Systems*. 2014, pp. 2672–2680.

[101] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. «Explaining and Harnessing Adversarial Examples». In: *International Conference on Learning Representations (ICLR)* (2015). URL: https://arxiv.org/abs/1412.6572.

[102] Anjana Gosain and Saanchi Sardana. «Handling class imbalance problem using oversampling techniques: A review». In: *2017 international conference on advances in computing, communications and informatics (ICACCI)*. IEEE. 2017, pp. 79–85.

[103] Mitul Goswami, Suneeta Mohanty, and Prasant Kumar Pattnaik. «Optimization of machine learning models through quantization and data bit reduction in healthcare datasets». In: *Franklin Open* 8 (2024), p. 100136.

[104] A. Graves, S. Fernandez, and J. Schmidhuber. «Bidirectional LSTM Networks for Improved Phoneme Classification and Recognition». In: *Proceedings of the International Conference on Artificial Neural Networks*. Berlin, Germany: Springer, 2005.

[105] Riccardo Guidotti et al. «A survey of methods for explaining black box models». In: *ACM Computing Surveys (CSUR)* 51.5 (2018), 93:1–93:42.

[106] A. Gupta and V.K. Tayal. «Analysis of Twitter Sentiment to Predict Financial Trends». In: *2023 International Conference on Artificial Intelligence and Smart Communication (AISC)*. 2023, pp. 1027–1031.

[107] Harsha Gurulingappa et al. «Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports». In: *Journal of Biomedical Informatics* 45.5 (2012). Text Mining and Natural Language Processing in Pharmacogenomics, pp. 885–892. ISSN: 1532-0464. DOI: https://doi.org/10.1016/j.jbi.2012.04.008. URL: http://www.sciencedirect.com/science/article/pii/S1532046412000615.

[108] Curt Hagquist and Magnus Stenbeck. «Goodness of fit in regression analysis–R 2 and G 2 reconsidered». In: *Quality and Quantity* 32.3 (1998), pp. 229–245.

[109] F.R. Hampel et al. *Robust Statistics: The Approach Based on Influence Functions*. Vol. 196. Hoboken, NJ, USA: John Wiley & Sons, 2011.

[110] Jiawei Han, Jian Pei, and Yi Yin. «Mining frequent patterns without candidate generation». In: *ACM SIGMOD Record*. Vol. 29. 2. ACM. 2000, pp. 1–12.

[111] Frank E Harrell Jr et al. «Regression models for prognostic prediction: advantages, problems, and suggested solutions.» In: *Cancer treatment reports* 69.10 (1985), pp. 1071–1077.

[112] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd. New York, NY, USA: Springer, 2009.

[113] A. Hazourli. *FinancialBERT - A Pretrained Language Model for Financial Text Mining*. Tech. rep. Technical Report, 2022.

[114] A. Hazourli. *ahmedrachid/FinancialBERT*. [Accessed: 29-09-2023]. URL: https://huggingface.co/ahmedrachid/FinancialBERT.

[115] Maleika Heenaye-Mamode Khan et al. «Multi-class classification of breast cancer abnormalities using Deep Convolutional Neural Network (CNN)». In: *Plos one* 16.8 (2021), e0256500.

[116] Richard M Heiberger et al. «Polynomial regression». In: *R through excel: a spreadsheet interface for statistics, data analysis, and graphics* (2009), pp. 269–284.

[117] Arash Heidari et al. «Machine learning applications for COVID-19 outbreak management». In: *Neural Computing and Applications* 34.18 (2022), pp. 15313–15348.

[118] S. Hochreiter and J. Schmidhuber. «Long Short-Term Memory». In: *Neural Computation* 9 (1997), pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.

[119] Sepp Hochreiter and Jürgen Schmidhuber. «Long short-term memory». In: *Neural Computation* 9.8 (1997), pp. 1735–1780.

[120] Steven C. H. Hoi et al. «Batch Mode Active Learning and Its Application to Medical Image Classification». In: *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)* (2012), pp. 435–443.

[121] Charles C Holt. «Forecasting seasonal and trends by exponentially weighted moving averages». In: *ONR Memorandum, Carnegie Institute of Technology* 52 (1957).

[122] Minqing Hu and Bing Liu. «Mining and Summarizing Customer Reviews». In: *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2004, pp. 168–177.

[123] A. Huang, H. Wang, and Y. Yang. «FinBERT: A Large Language Model for Extracting Information from Financial Text». In: *Contemporary Accounting Research* (2022).

[124] Clayton J. Hutto and Eric Gilbert. «VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text». In: *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM)*. 2014, pp. 216–225.

[125] Alexander I Iliev and Amruth Anand. «Huber Loss and Neural Networks Application in Property Price Prediction». In: *Future of Information and Communication Conference*. Springer. 2023, pp. 242–256.

[126] Shantanu Jain et al. «A Survey on Face Recognition Techniques in Machine Learning». In: *International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT)* 8.6 (2022), pp. 50–66.

[127] Nathalie Japkowicz and Shaju Stephen. «The class imbalance problem: A systematic study». In: *Intelligent data analysis* 6.5 (2002), pp. 429–449.

[128] Adel Javanmard and Andrea Montanari. «Confidence intervals and hypothesis testing for high-dimensional regression». In: *The Journal of Machine Learning Research* 15.1 (2014), pp. 2869–2909.

[129] Zhao Jianqiang, Gui Xiaolin, and Zhang Xuejun. «Deep Convolution Neural Networks for Twitter Sentiment Analysis». In: *IEEE Access* 6 (2018), pp. 23253–23260. DOI: 10.1109/ACCESS.2017.2776930.

[130] Long Jing and Ying Tian. «Self-supervised learning for graph neural networks». In: *arXiv preprint arXiv:2006.14837* (2020).

[131] Anna Jobin, Marcello Ienca, and Effy Vayena. «The global landscape of AI ethics guidelines». In: *Nature Machine Intelligence* 1 (2019), pp. 389–399.

[132] Vivek Justus and GR Kanagachidambaresan. «Machine learning based fault-oriented predictive maintenance in industry 4.0». In: *International Journal of System Assurance Engineering and Management* 15.1 (2024), pp. 462–474.

[133] Gozde Karatas, Onder Demir, and Ozgur Koray Sahingoz. «Deep Learning in Intrusion Detection Systems». In: *2018 International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism (IBIGDELFT)*. 2018, pp. 113–116. DOI: 10.1109/IBIGDELFT.2018.8625278.

[134] Fazle Karim et al. «LSTM fully convolutional networks for time series classification». In: *IEEE Access* 6 (2018), pp. 1662–1669.

[135] Benjamin Kedem and Konstantinos Fokianos. *Regression models for time series analysis*. John Wiley & Sons, 2005.

[136] D.P. Kingma and J. Ba. «Adam: A Method for Stochastic Optimization». In: *arXiv preprint arXiv:1412.6980* (2014).

[137] Jens Kober, Jonathan A. Bagnell, and Jan Peters. «Reinforcement Learning in Robotics: A Survey». In: *The International Journal of Robotics Research* 32.11 (2013), pp. 1238–1274.

[138] Andrei Kolmogorov. «Analytical methods in probability theory». In: *Mathematische Annalen* 104 (1931), pp. 415–458.

[139] Olivier Kraaijeveld and Johannes De Smedt. «The predictive power of public Twitter sentiment for forecasting cryptocurrency prices». In: *Journal of International Financial Markets, Institutions and Money* 65 (2020), p. 101188. ISSN: 1042-4431. DOI: https://doi.org/10.1016/j.intfin.2020.101188. URL: https://www.sciencedirect.com/science/article/pii/S104244312030072X.

[140] Danylo Krasovytskyi and Andriy Stavytskyy. «Predicting mortgage loan defaults using machine learning techniques». In: *Ekonomika* 103.2 (2024), pp. 140–160.

[141] Naik Kshirasagar and Tripathy Priyadarshi. «Software testing and quality assurance: theory and practice». In: *Scopus* (2008), pp. 539–598.

[142] Akshi Kumar and Arunima Jaiswal. «Systematic literature review of sentiment analysis on Twitter using soft computing techniques». In: *Concurrency and Computation: Practice and Experience* 32.1 (2020). e5107 CPE-18-1167.R1, e5107. DOI: https://doi.org/10.1002/cpe.5107. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpe.5107. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/cpe.5107.

[143] Chia-Chen Lee, Zhen Gao, and Chia-Lin Tsai. «BERT-Based Stock Market Sentiment Analysis». In: *2020 IEEE International Conference on Consumer Electronics–Taiwan (ICCE-Taiwan)*. 2020, pp. 1–2.

[144] Adrien-Marie Legendre. *Nouvelles méthodes pour la détermination des orbites des comètes*. F. Didot, 1805.

[145] *Lexis Nexis Database*. https://www.lexisnexis.com/en-us/home.page. [Accessed 29-09-2023].

[146] Xuhong Li et al. «Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond». In: *Knowledge and Information Systems* 64.12 (2022), pp. 3197–3234.

[147] Zhuo Li, Hengyi Li, and Lin Meng. «Model compression for deep neural networks: A survey». In: *Computers* 12.3 (2023), p. 60.

[148] Bing Liu. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers, 2012.

[149] Lanbiao Liu, Chen Chen, and Bo Wang. «Predicting financial crises with machine learning methods». In: *Journal of Forecasting* 41.5 (2022), pp. 871–910.

[150] Yi Liu, Lei Chen, and Zerui Yao. «The application of artificial intelligence assistant to deep learning in teachers' teaching and students' learning processes». In: *Frontiers in Psychology* 13 (2022), p. 929175.

[151] Yinhan Liu et al. «RoBERTa: A Robustly Optimized BERT Pretraining Approach». In: *arXiv preprint arXiv:1907.11692* (2019).

[152] Steven Loria. *TextBlob Documentation*. `https://textblob.readthedocs.io/en/dev/`. Accessed: 2025-03-02. 2018.

[153] Dengsheng Lu and Qihao Weng. «A survey of image classification methods and techniques for improving classification performance». In: *International journal of Remote sensing* 28.5 (2007), pp. 823–870.

[154] Scott M. Lundberg and Su-In Lee. «A Unified Approach to Interpreting Model Predictions». In: *Advances in Neural Information Processing Systems (NIPS)* 30 (2017), pp. 4765–4774.

[155] Andrew L. Maas et al. «Learning Word Vectors for Sentiment Analysis». In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, June 2011, pp. 142–150. URL: `http://www.aclweb.org/anthology/P11-1015`.

[156] J. MacQueen. «Some Methods for Classification and Analysis of Multivariate Observations». In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. 1967, pp. 281–297.

[157] Macedo Maia et al. «WWW'18 Open Challenge: Financial Opinion Mining and Question Answering». In: *Companion Proceedings of the The Web Conference 2018* (2018). URL: `https://api.semanticscholar.org/CorpusID:13866508`.

[158] Anastasia Makarova et al. «Overfitting in Bayesian optimization: an empirical study and early-stopping solution». In: *2nd Workshop on Neural Architecture Search (NAS 2021)@ ICLR 2021*. NAS 2021. 2021.

[159] Spyros Makridakis et al. *NeuralProphet: Explainable Forecasting at Scale*. `https://github.com/ourownstory/neural_prophet`. 2021.

[160] Lorenzo Malandri et al. «Public Mood–Driven Asset Allocation: the Importance of Financial Sentiment in Portfolio Management». In: *Cognitive Computation* 10 (2018), pp. 1167–1176. URL: `https://api.semanticscholar.org/CorpusID:53795790`.

[161] Pekka Malo et al. *FinancialPhraseBank-v1.0*. July 2013.

[162] Keith A Marill. «Advanced statistics: linear regression, part I: simple linear regression». In: *Academic emergency medicine* 11.1 (2004), pp. 87–93.

[163] D.W. Marquardt and R.D. Snee. «Ridge Regression in Practice». In: *The American Statistician* 29 (1975), pp. 3–20. DOI: 10.1080/00031305.1975.10479232.

[164] Michael L Martini et al. «Application of cooperative game theory principles to interpret machine learning models of nonhome discharge following spine surgery». In: *Spine* 46.12 (2021), pp. 803–812.

[165] Ann Neethu Mathew, V Rohini, and Joy Paulose. «NLP-based personal learning assistant for school education». In: *Int. J. Electr. Comput. Eng* 11.5 (2021), pp. 4522–4530.

[166] Norman Matloff. *Statistical regression and classification: from linear models to machine learning*. Chapman and Hall/CRC, 2017.

[167] Andrew McCallum and Kamal Nigam. «Efficient Clustering of High-Dimensional Data Sets with Application to Reference Matching». In: *Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining (KDD)* (2003).

[168] JS McCarley, Rishav Chakravarti, and Avirup Sil. «Structured pruning of a bert-based question answering model». In: *arXiv preprint arXiv:1910.06360* (2019).

[169] Gary C McDonald. «Ridge regression». In: *Wiley Interdisciplinary Reviews: Computational Statistics* 1.1 (2009), pp. 93–100.

[170] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. «Sentiment Analysis Algorithms and Applications: A Survey». In: *Ain Shams Engineering Journal* 5.4 (2014), pp. 1093–1113.

[171] Abolfazl Mehbodniya et al. «[Retracted] Financial Fraud Detection in Healthcare Using Machine Learning and Deep Learning Techniques». In: *Security and Communication Networks* 2021.1 (2021), p. 9293877.

[172] Ninareh Mehrabi et al. «A Survey on Bias and Fairness in Machine Learning». In: *ACM Comput. Surv.* 54.6 (July 2021). ISSN: 0360-0300. DOI: 10.1145/3457607. URL: https://doi.org/10.1145/3457607.

[173] Tom M. Mitchell. *Machine Learning*. New York, NY, USA: McGraw-Hill, 1997.

[174] Saif Mohammad et al. «Semeval-2018 task 1: Affect in tweets». In: *Proceedings of the 12th international workshop on semantic evaluation*. 2018, pp. 1–17.

[175] Senthilkumar Mohan, Chandrasegar Thirumalai, and Gautam Srivastava. «Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques». In: *IEEE Access* 7 (2019), pp. 81542–81554.

[176] Javier Montalt-Tordera et al. «Machine learning in magnetic resonance imaging: image reconstruction». In: *Physica Medica* 83 (2021), pp. 79–87.

[177] Douglas C Montgomery, Cheryl L Jennings, and Murat Kulahci. *Introduction to time series analysis and forecasting*. John Wiley & Sons, 2015.

[178] Eduardo Mosqueira-Rey et al. «Human-in-the-loop machine learning: a state of the art». In: *Artificial Intelligence Review* 56.4 (2023), pp. 3005–3054.

[179] Khadijeh Moulaei et al. «Comparing machine learning algorithms for predicting COVID-19 mortality». In: *BMC medical informatics and decision making* 22.1 (2022), p. 2.

[180] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. Cambridge, MA, USA: MIT Press, 2012.

[181] F. Murtagh. «Multilayer Perceptrons for Classification and Regression». In: *Neurocomputing* 2 (1991), pp. 183–197. DOI: 10.1016/0925-2312(91)90002-P.

[182] Andrew Y Ng et al. «Preventing" overfitting" of cross-validation data». In: *ICML*. Vol. 97. Citeseer. 1997, pp. 245–253.

[183] Jovial Niyogisubizo et al. «Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization». In: *Computers and Education: Artificial Intelligence* 3 (2022), p. 100066.

[184] NLP Town. *nlptown/bert-base-multilingual-uncased-sentiment*. https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment. Accessed: 2024-07-28. 2024.

[185] Shadrack Obeng et al. «Utilizing machine learning algorithms to prevent financial fraud and ensure transaction security». In: *World Journal of Advanced Research and Reviews* 23.1 (2024), pp. 1972–1980.

[186] I. Oguiza. *tsAI Models: FCNPlus*. https://timeseriesai.github.io/tsai/models.fcnplus.html. Accessed on 7 October 2021.

[187] I. Oguiza. *tsAI Models: InceptionTimePlus*. https://timeseriesai.github.io/tsai/models.inceptiontimeplus.html. Accessed on 7 October 2021.

[188] I. Oguiza. *tsAI Models: ResNetPlus*. https://timeseriesai.github.io/tsai/models.resnetplus.html. Accessed on 7 November 2022.

[189] I. Oguiza. *tsAI Models: RNN_FCN*. https://timeseriesai.github.io/tsai/models.rnn_fcn.html. Accessed on 7 November 2022.

[190] I. Oguiza. *tsAI Models: RNN_FCNPlus*. https://timeseriesai.github.io/tsai/models.rnn_fcnplus.html. Accessed on 7 November 2022.

[191] I. Oguiza. *tsAI Models: RNNS*. https://timeseriesai.github.io/tsai/models.rnn.html. Accessed on 7 November 2022.

[192] I. Oguiza. *tsAI Models: Transformermodel*. https://timeseriesai.github.io/tsai/models.transformermodel.html. Accessed on 7 November 2022.

[193] I. Oguiza. *tsAI Models: TSIT*. https://timeseriesai.github.io/tsai/models.tsitplus.html. Accessed on 7 November 2022.

[194] I. Oguiza. *tsAI Models: TSTPlus*. https://timeseriesai.github.io/tsai/models.tstplus.html. Accessed on 7 November 2022.

[195] I. Oguiza. *tsAI Models: XceptionTimePlus*. https://timeseriesai.github.io/tsai/models.xceptiontimeplus.html. Accessed on 7 November 2022.

[196] I. Oguiza. *tsAI Models: XCMPlus*. https://timeseriesai.github.io/tsai/models.xcmplus.html. Accessed on 7 November 2022.

[197] OpenAI. *GPT-4 Technical Report*. https://cdn.openai.com/papers/GPT-4.pdf. Accessed: 2023-03-14. 2023.

[198] Boris Oreshkin et al. «N-BEATS: Neural basis expansion analysis for interpretable time series forecasting». In: *Advances in Neural Information Processing Systems*. 2019, pp. 893–903.

[199] Boris N Oreshkin et al. «N-BEATS: Neural basis expansion analysis for interpretable time series forecasting». In: *arXiv preprint arXiv:1905.10437* (2019).

[200] Alvaro Ortigosa, José M. Martín, and Rosa M. Carro. «Sentiment analysis in Facebook and its application to e-learning». In: *Computers in Human Behavior* 31 (2014), pp. 527–541. ISSN: 0747-5632. DOI: https://doi.org/10.1016/j.chb.2013.05.024. URL: https://www.sciencedirect.com/science/article/pii/S0747563213001751.

[201] Kağan Özgün et al. «Malfunction detection on production line using machine learning: case study in wood industry». In: *Intelligent and Fuzzy Techniques: Smart and Innovative Solutions: Proceedings of the INFUS 2020 Conference, Istanbul, Turkey, July 21-23, 2020*. Springer. 2021, pp. 1116–1124.

[202] Venkata Sasank Pagolu et al. «Sentiment analysis of Twitter data for predicting stock market movements». In: *2016 international conference on signal processing, communication, power and embedded system (SCOPES)*. IEEE. 2016, pp. 1345–1350.

[203] Prajyot Palimkar, Rabindra Nath Shaw, and Ankush Ghosh. «Machine learning technique to prognosis diabetes disease: Random forest classifier approach». In: *Advanced computing and intelligent technologies: proceedings of ICACIT 2021*. Springer, 2021, pp. 219–244.

[204] Bo Pang and Lillian Lee. «Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales». In: *Proceedings of the ACL*. 2005.

[205] Bo Pang and Lillian Lee. «Opinion Mining and Sentiment Analysis». In: *Foundations and Trends in Information Retrieval* 2.1–2 (2008), pp. 1–135.

[206] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. «Thumbs Up?: Sentiment Classification Using Machine Learning Techniques». In: *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*. 2002, pp. 79–86.

[207] Panagiotis Papaioannou et al. «Can social microblogging be used to forecast intraday exchange rates?» In: *NETNOMICS: Economic Research and Electronic Networking* 14 (2013), pp. 47–68. URL: https://api.semanticscholar.org/CorpusID:2516894.

[208] Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, 1988.

[209] Francisco José García Peñalvo et al. «Sustainable stock market prediction framework using machine learning models». In: *International Journal of Software Science and Computational Intelligence (IJSSCI)* 14.1 (2022), pp. 1–15.

[210] Juan Manuel Pérez, Juan Carlos Giudici, and Franco Luque. *pysentimiento: A Python Toolkit for Sentiment Analysis and SocialNLP tasks*. 2021. arXiv: 2106.09462 [cs.CL].

[211] Emily R Pfaff et al. «Identifying who has long COVID in the USA: a machine learning approach using N3C data». In: *The Lancet Digital Health* 4.7 (2022), e532–e541.

[212] Benedikt Pfülb et al. «Catastrophic forgetting: still a problem for DNNs». In: *Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part I 27*. Springer. 2018, pp. 487–497.

[213] Soujanya Poria et al. «A Review of Affective Computing: From Unimodal Analysis to Multimodal Fusion». In: *Information Fusion*. Vol. 37. 2017, pp. 98–125.

[214] L. Prokhorenkova et al. «CatBoost: Unbiased Boosting with Categorical Features». In: *arXiv preprint arXiv:1706.09516* (2019).

[215] *ProsusAI FinBERT Model*. [Accessed: 29-09-2023]. URL: `https://huggingface.co/ProsusAI/finbert`.

[216] Wisam A Qader, Musa M Ameen, and Bilal I Ahmed. «An overview of bag of words; importance, implementation, applications, and challenges». In: *2019 international engineering conference (IEC)*. IEEE. 2019, pp. 200–204.

[217] Atika Qazi et al. «Machine Learning-Based Opinion Spam Detection: A Systematic Literature Review». In: *IEEE Access* (2024).

[218] Ruoxi Qin et al. «Weighted focal loss: An effective loss function to overcome unbalance problem of chest X-ray14». In: *IOP conference series: Materials science and engineering*. Vol. 428. 1. IOP Publishing. 2018, p. 012022.

[219] Qiang Qiu and Zichen Song. «A nonuniform weighted loss function for imbalanced image classification». In: *Proceedings of the 2018 international conference on image and graphics processing*. 2018, pp. 78–82.

[220] Ehsan Rahimian et al. «XceptionTime: A novel deep architecture based on depthwise separable convolutions for hand gesture classification». In: *arXiv preprint arXiv:1911.03803* (2019).

[221] Juan Ramos et al. «Using tf-idf to determine word relevance in document queries». In: *Proceedings of the first instructional conference on machine learning*. Vol. 242. 1. Citeseer. 2003, pp. 29–48.

[222] Jonas Ranstam and Jonathan A Cook. «LASSO regression». In: *Journal of British Surgery* 105.10 (2018), pp. 1348–1348.

[223] Carl Edward Rasmussen and Christopher K.I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[224] Rui Ren, Desheng Dash Wu, and Tianxiang Liu. «Forecasting Stock Market Movement Direction Using Sentiment Analysis and Support Vector Machine». In: *IEEE Systems Journal* 13 (2019), pp. 760–770. URL: `https://api.semanticscholar.org/CorpusID:67870584`.

[225] Juha Reunanen. «Overfitting in making comparisons between variable selection methods». In: *Journal of Machine Learning Research* 3.Mar (2003), pp. 1371–1382.

[226] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. «"Why Should I Trust You?": Explaining the Predictions of Any Classifier». In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. 2016, pp. 1135–1144.

[227] Rebecca Roelofs et al. «A meta-analysis of overfitting in machine learning». In: *Advances in neural information processing systems* 32 (2019).

[228] Lavinia Rognone, Stuart Hyde, and S. Sarah Zhang. «News sentiment in the cryptocurrency market: An empirical comparison with Forex». In: *International Review of Financial Analysis* 69.C (2020). DOI: 10.1016/j.irfa.2020.10146.

[229] Frank Rosenblatt. «The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain». In: *Psychological Review* 65.6 (1958), pp. 386–408.

[230] Frank Rosenblatt. «The perceptron: A probabilistic model for information storage and organization in the brain». In: *Psychological Review* 65.6 (1958), pp. 386–408.

[231] Sara Rosenthal, Noura Farra, and Preslav Nakov. «SemEval-2017 task 4: Sentiment analysis in Twitter». In: *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*. 2017, pp. 502–518.

[232] Nicholas Roy and Andrew McCallum. «Toward Optimal Active Learning through Sampling Estimation of Error Reduction». In: *Proceedings of the 18th International Conference on Machine Learning (ICML)* (2001), pp. 441–448.

[233] R. Rubinstein, M. Zibulevsky, and M. Elad. «Efficient Implementation of the KSVD Algorithm Using Batch Orthogonal Matching Pursuit». In: *Proceedings of the Computer Science Department, Technion*. Haifa, Israel, 2008, pp. 1–15.

[234] Cynthia Rudin et al. «Interpretable machine learning: Fundamental principles and 10 grand challenges». In: *Statistics Surveys* 16.none (2022), pp. 1–85. DOI: 10.1214/21-SS133. URL: https://doi.org/10.1214/21-SS133.

[235] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. 4th. New York, NY, USA: Pearson, 2020.

[236] Emrullah ŞAHiN, Naciye Nur Arslan, and Durmuş Özdemir. «Unlocking the black box: an in-depth review on interpretability, explainability, and reliability in deep learning». In: *Neural Computing and Applications* (2024), pp. 1–107.

[237] Haidara Saleh and JA Layous. «Machine Learning–Regression». In: *Suriye Üniversitesi, Yüksek Uygulamalı Bilimler ve Teknoloji Enstitüsü, Elektronik ve Mekanik Sistemler Anabilim Dalı* 23 (2022).

[238] David Salinas et al. «DeepAR: Probabilistic forecasting with autoregressive recurrent networks». In: *International Journal of Forecasting*. Elsevier, 2019.

[239] David Salinas et al. «DeepAR: Probabilistic forecasting with autoregressive recurrent networks». In: *International Journal of Forecasting* 36.3 (2020), pp. 1181–1191. ISSN: 0169-2070. DOI: https://doi.org/10.1016/j.ijforecast.2019.07.001. URL: https://www.sciencedirect.com/science/article/pii/S0169207019301888.

[240] Arthur L. Samuel. «Some studies in machine learning using the game of Checkers». In: *IBM Journal of Research and Development* 3.3 (1959), pp. 210–229.

[241] Victor Sanh et al. *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. 2020. arXiv: 1910.01108 [cs.CL]. URL: https://arxiv.org/abs/1910.01108.

[242] Claudio Filipi Gonçalves Dos Santos and João Paulo Papa. «Avoiding overfitting: A survey on regularization methods for convolutional neural networks». In: *ACM Computing Surveys (Csur)* 54.10s (2022), pp. 1–25.

[243] Elvis Saravia et al. «CARER: Contextualized Affect Representations for Emotion Recognition». In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 3687–3697. DOI: 10.18653/v1/D18-1404. URL: https://www.aclweb.org/anthology/D18-1404.

[244] Widi Satriaji and Retno Kusumaningrum. «Effect of Synthetic Minority Oversampling Technique (SMOTE), Feature Representation, and Classification Algorithm on Imbalanced Sentiment Analysis». In: *2018 2nd International Conference on Informatics and Computational Sciences (ICICoS)* (2018), pp. 1–5. URL: https://api.semanticscholar.org/CorpusID:59231725.

[245] Fabrizio Sebastiani. «Machine learning in automated text categorization». In: *ACM Computing Surveys* 34.1 (2002), pp. 1–47.

[246] G.A.F. Seber and A.J. Lee. *Linear Regression Analysis*. Vol. 329. Hoboken, NJ, USA: John Wiley & Sons, 2012.

[247] Ozan Sener and Silvio Savarese. «Active Learning for Convolutional Neural Networks: A Core-Set Approach». In: *Proceedings of the International Conference on Learning Representations (ICLR)*. 2018.

[248] Burr Settles. *Active Learning*. Morgan & Claypool Publishers, 2012.

[249] Burr Settles and Mark Craven. «An Analysis of Active Learning Strategies for Sequence Labeling Tasks». In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2007).

[250] Emily Sheng and David Uthus. *Investigating Societal Biases in a Poetry Composition System*. 2020. arXiv: 2011.02686 [cs.CL].

[251] J. Sayyad Shirabad and T. Menzies. *The PROMISE Repository of Software Engineering Databases*. School of Information Technology and Engineering, University of Ottawa, Canada. http://promise.site.uottawa.ca/SERepository. 2005.

[252] Wesam Shishah. «An Efficient Early Stage Heart Disease Risk Detection Using Machine Learning Techniques». In: *2022 Second International Conference on Power, Control and Computing Technologies (ICPC2T)*. IEEE. 2022, pp. 1–6.

[253] Reza Shokri et al. «Membership Inference Attacks Against Machine Learning Models». In: *2017 IEEE Symposium on Security and Privacy (SP)*. 2017, pp. 3–18.

[254] David Silver et al. «Mastering the game of Go with deep neural networks and tree search». In: *Nature* 529.7587 (2016), pp. 484–489.

[255] A. Sinha et al. «SEntFiN 10: Entity-aware Sentiment Analysis for Financial News». In: *Journal of the Association for Information Science and Technology* 73 (2022), pp. 1314–1335.

[256] A.J. Smola and B. Schölkopf. «A Tutorial on Support Vector Regression». In: *Statistics and Computing* 14 (2004), pp. 199–222. DOI: 10.1023/B:STCO.0000035301.49549.88.

[257] Derek Snow. *AtsPy: Automated Time Series Models in Python.* Version 1.15. Feb. 17, 2020. URL: https://github.com/firmai/atspy/.

[258] Stefan Stieglitz and Linh Dang-Xuan. «Emotions and Information Diffusion in Social Media—Sentiment of Microblogs and Sharing Behavior». In: *Journal of Management Information Systems* 29.4 (2013), pp. 217–248.

[259] Stephen M Stigler. «Gauss and the invention of least squares». In: *the Annals of Statistics* (1981), pp. 465–474.

[260] Emma Strubell, Ananya Ganesh, and Andrew McCallum. «Energy and Policy Considerations for Deep Learning in NLP». In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*. 2019, pp. 3645–3650.

[261] František Štulajter. *Predictions in time series using regression models.* Springer, 2002.

[262] Harry Surden. «Machine learning and law: An overview». In: *Research handbook on big data law* (2021), pp. 171–184.

[263] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction.* 2nd. Cambridge, MA, USA: MIT Press, 2018.

[264] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction.* 2nd. MIT Press, 2018.

[265] Kyle Swanson et al. «From patterns to patients: Advances in clinical machine learning for cancer diagnosis, prognosis, and treatment». In: *Cell* 186.8 (2023), pp. 1772–1791.

[266] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining.* Addison-Wesley, 2006.

[267] Duyu Tang, Bing Qin, and Ting Liu. «Document Modeling with Gated Recurrent Neural Network for Sentiment Classification». In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing.* 2015, pp. 1422–1432.

[268] Wenjie Tang et al. «Omni-Scale CNNs: A simple and effective kernel size configuration for time series classification». In: *arXiv preprint arXiv:2002.10061* (2022).

[269] Adane Nega Tarekegn, Mario Giacobini, and Krzysztof Michalak. «A review of methods for imbalanced multi-label classification». In: *Pattern Recognition* 118 (2021), p. 107965.

[270] Sean J Taylor and Benjamin Letham. «Forecasting at scale». In: *The American Statistician* 72.1 (2018), pp. 37–45.

[271] Mike A Thelwall. «Social media analytics for YouTube comments: potential and limitations». In: *International Journal of Social Research Methodology* 21 (2018), pp. 303–316. URL: https://api.semanticscholar.org/CorpusID:148591270.

[272] R. Tibshirani. «Regression Shrinkage and Selection via the Lasso». In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58 (1996), pp. 267–288. DOI: 10.2307/2346178.

[273] Simon Tong and Daphne Koller. «Support Vector Machine Active Learning with Applications to Text Classification». In: *Proceedings of the 17th International Conference on Machine Learning (ICML)*. 2001, pp. 999–1006.

[274] Mark Tranmer and Mark Elliot. «Multiple linear regression». In: *The Cathie Marsh Centre for Census and Survey Research (CCSR)* 5.5 (2008), pp. 1–5.

[275] Evanthia E. Tripoliti et al. «Heart Failure: Diagnosis, Severity Estimation and Prediction of Adverse Events Through Machine Learning Techniques». In: *Computational and Structural Biotechnology Journal* 15 (2017), pp. 26–47. ISSN: 2001-0370. DOI: https://doi.org/10.1016/j.csbj.2016.11.001. URL: https://www.sciencedirect.com/science/article/pii/S2001037016300460.

[276] Alan Turing. «Computing Machinery and Intelligence». In: *Mind* LIX.236 (1950), pp. 433–460.

[277] Alan M. Turing. «Computing Machinery and Intelligence». In: *Mind* LIX (1950), pp. 433–460.

[278] *Tweet Sentiment Airlines Dataset*. https://huggingface.co/datasets/jos-ger/tweet-sentiment-airlines. Accessed: 28-07-2024.

[279] Joaquin Vanschoren. «Meta-learning». In: *Automated machine learning: methods, systems, challenges* (2019), pp. 35–61.

[280] Ashish Vaswani et al. «Attention is all you need». In: *Advances in neural information processing systems* 30 (2017).

[281] Sahil Verma, Michael Ernst, and Rene Just. «Removing biased data to improve fairness and accuracy». In: *arXiv preprint arXiv:2102.03054* (2021).

[282] V. Vovk. «Kernel Ridge Regression». In: *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*. Ed. by B. Scholkopf, Z. Luo, and V. Vovk. Berlin, Germany: Springer, 2013, pp. 105–116.

[283] Romi Satria Wahono. «A systematic literature review of software defect prediction». In: *Journal of software engineering* 1.1 (2015), pp. 1–16.

[284] Yun Wan and Qigang Gao. «An Ensemble Sentiment Classification System of Twitter Data for Airline Services Analysis». In: *2015 IEEE International Conference on Data Mining Workshop (ICDMW)* (2015), pp. 1318–1325. URL: https://api.semanticscholar.org/CorpusID:21442.

[285] Ke Wang and Dit-Yan Yeung. «Batch Mode Active Learning for Deep Neural Networks». In: *Proceedings of the 22nd International Conference on Machine Learning (ICML)* (2015).

[286] Qi Wang et al. «A comprehensive survey of loss functions in machine learning». In: *Annals of Data Science* (2020), pp. 1–26.

[287] Rongguang Wang, Pratik Chaudhari, and Christos Davatzikos. «Bias in machine learning models can be significantly mitigated by careful training: Evidence from neuroimaging studies». In: *Proceedings of the National Academy of Sciences* 120.6 (2023), e2211613120.

[288] Zhiguang Wang, Weizhong Yan, and Tim Oates. «Time series classification from scratch with deep neural networks: A strong baseline». In: *Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2017, pp. 1578–1585.

[289] J. H. Ward. «Hierarchical Grouping to Optimize an Objective Function». In: *Journal of the American Statistical Association* 58.301 (1963), pp. 236–244.

[290] Pengcheng Wei et al. «Research on sound classification based on SVM». In: *Neural Computing and Applications* 32 (2020), pp. 1593–1607.

[291] Sanford Weisberg. *Applied linear regression*. Vol. 528. John Wiley & Sons, 2005.

[292] Norbert Wiener. *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. MIT Press, 1949.

[293] Peter R Winters. «Forecasting sales by exponentially weighted moving averages». In: *Management Science* 6.3 (1960), pp. 324–342.

[294] D. Wipf and S. Nagarajan. «A New View of Automatic Relevance Determination». In: *Advances in Neural Information Processing Systems*. Vol. 20. Red Hook, NY, USA: Curran Associates, Inc., 2008.

[295] Xueqing Wu et al. «A study of BERT for context-aware neural machine translation». In: *Machine Learning* 111.3 (2022), pp. 917–935.

[296] Zonghan Wu et al. «A comprehensive survey on graph neural networks». In: *IEEE Transactions on Neural Networks and Learning Systems* 32.1 (2020), pp. 4–24.

[297] Eric P Xing et al. «Strategies and principles of distributed machine learning on big data». In: *Engineering* 2.2 (2016), pp. 179–195.

[298] Frank Xing, E. Cambria, and Roy E. Welsch. «Natural language based financial forecasting: a survey». In: *Artificial Intelligence Review* 50 (2017), pp. 49–73. URL: https://api.semanticscholar.org/CorpusID:207079655.

[299] Frank Xing et al. «Financial Sentiment Analysis: An Investigation into Common Mistakes and Silver Bullets». In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 978–987. DOI: 10.18653/v1/2020.coling-main.85. URL: https://aclanthology.org/2020.coling-main.85.

[300] Frank Z. Xing, Erik Cambria, and Roy E. Welsch. «Intelligent Asset Allocation via Market Sentiment Views». In: *IEEE Computational Intelligence Magazine* 13.4 (2018), pp. 25–34. DOI: 10.1109/MCI.2018.2866727.

[301] Rui Xu and Delwyn Wunsch. «Survey of Clustering Algorithms». In: *IEEE Transactions on Neural Networks* 16.3 (2005), pp. 645–678.

[302] Avanish Vijaybahadur Yadav, Sanket Saheb Verma, and Deepak Dinesh Singh. «Virtual assistant for blind people». In: *International Journal* 6.5 (2021).

[303] Ram Narayan Yadav, Prem Kumar Kalra, and Joseph John. «Time series prediction with single multiplicative neuron model». In: *Applied soft computing* 7.4 (2007), pp. 1157–1163.

[304] Dong Yu and Lin Deng. *Automatic speech recognition*. Vol. 1. Springer, 2016.

[305] G. Udny Yule. «On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers». In: *Philosophical Transactions of the Royal Society of London. Series A* 226 (1927), pp. 267–298.

[306] Muhammad Rehman Zafar and Naimul Khan. «Deterministic local interpretable model-agnostic explanations for stable explainability». In: *Machine Learning and Knowledge Extraction* 3.3 (2021), pp. 525–541.

[307] George Zerveas et al. «A Transformer-based Framework for Multivariate Time Series Representation Learning». In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. Singapore, 2021, pp. 2114–2124.

[308] Bo Zhang, Huiping Shi, and Hongtao Wang. «Machine learning and AI in cancer prognosis, prediction, and treatment selection: a critical approach». In: *Journal of multidisciplinary healthcare* (2023), pp. 1779–1791.

[309] Dailing Zhang et al. «Can Sentiment Analysis Help Mimic Decision-Making Process of Loan Granting? A Novel Credit Risk Evaluation Approach Using GMKL Model». In: *2015 48th Hawaii International Conference on System Sciences* (2015), pp. 949–958. URL: https://api.semanticscholar.org/CorpusID:17733609.

[310] Longfei Zhou, Lin Zhang, and Nicholas Konz. «Computer vision techniques in manufacturing». In: *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 53.1 (2022), pp. 105–117.

[311] David Zimbra, Ahmed Abbasi, and Daniel Zeng. «The State-of-the-Art in Twitter Sentiment Analysis: A Review and Benchmark Evaluation». In: *ACM Transactions on Management Information Systems* xx, No. x (May 2018). DOI: 10.1145/3185045.

[312] Indre Zliobaite. «Fairness-aware machine learning: a perspective». In: *arXiv preprint arXiv:1708.00754* (2017).

[313] H. Zou and T. Hastie. «Regularization and Variable Selection via the Elastic Net». In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67 (2005), pp. 301–320. DOI: 10.1111/j.1467-9868.2005.00503.x.

[314] Hui Zou and Trevor Hastie. «Regularization and variable selection via the elastic net». In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 67.2 (2005), pp. 301–320.

[315] Xiaowu Zou et al. «Integration of residual network and convolutional neural network along with various activation functions and global pooling for time series classification». In: *Neurocomputing* 367 (2019), pp. 39–45.