



ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΗΛΕΚΤΡΙΚΩΝ ΒΙΟΜΗΧΑΝΙΚΩΝ ΔΙΑΤΑΞΕΩΝ
ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΑΠΟΦΑΣΕΩΝ

**Ανάλυση και αξιολόγηση καινοτόμων ενεργειών
επιχειρήσεων στους κλάδους της
τεχνολογίας και των χρηματοοικονομικών**

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Κωνσταντίνος Καράμ

Επιβλέπων: Ευάγγελος Σπηλιώτης
Επίκουρος Καθηγητής ΕΜΠ

Υπεύθυνος: Μαρίνα Σολοβιέβα
Υποψήφια Διδάκτωρ

Αθήνα, Ιούλιος 2025



ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΗΛΕΚΤΡΙΚΩΝ ΒΙΟΜΗΧΑΝΙΚΩΝ ΔΙΑΤΑΞΕΩΝ
ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΑΠΟΦΑΣΕΩΝ

**Ανάλυση και αξιολόγηση καινοτόμων ενεργειών
επιχειρήσεων στους κλάδους της
τεχνολογίας και των χρηματοοικονομικών**

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Κωνσταντίνος Καραμ

Επιβλέπων: Ευάγγελος Σπηλιώτης
Επίκουρος Καθηγητής ΕΜΠ

Υπεύθυνος: Μαρίνα Σολοβιέβα
Υποψήφια Διδάκτωρ

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 3^η Ιουλίου 2025.

.....
Ευάγγελος Σπηλιώτης
Επίκουρος Καθηγητής

.....
Ιωάννης Ψαρράς
Καθηγητής

.....
Δημήτρης Ασκούνης
Καθηγητής

Αθήνα, Ιούλιος, 2025

Copyright © Κωνσταντίνος Καράμ, 2025

Με επιφύλαξη παντός δικαιώματος. All rights reserved.

Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν τη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς τον συγγραφέα.

Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τον συγγραφέα και δεν πρέπει να ερμηνευθεί ότι αντιπροσωπεύουν τις επίσημες θέσεις του Εθνικού Μετσόβιου Πολυτεχνείου.

.....
Κωνσταντίνος Καράμ

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Η παρούσα διπλωματική εργασία έχει ως στόχο την αξιολόγηση των καινοτομιών που αναπτύσσονται από μεγάλες εταιρείες και τη χρήση τεχνικών ανάλυσης συναισθήματος των χρηματοοικονομικών τους εγγράφων (SEC filings) για την πρόβλεψη των μελλοντικών αποδόσεων των μετοχών τους. Η ανάλυση επικεντρώθηκε σε τρία κύρια χρηματοοικονομικά έγγραφα, τις αναφορές 10-K, 10-Q και τις δηλώσεις τύπου, από τα οποία εξάχθηκαν συναισθηματικοί δείκτες. Οι συναισθηματικοί δείκτες αυτού του είδους ενσωματώθηκαν σε μοντέλα μηχανικής μάθησης, τα οποία χρησιμοποιήθηκαν για την πρόβλεψη της πορείας της τιμής των μετοχών για τις εταιρείες: Apple (AAPL), Nvidia (NVDA) και Tesla (TSLA).

Η ανάλυση έδειξε ότι η προσθήκη αυτών των δεικτών συναισθήματος ενίσχυσε την ακρίβεια των προβλέψεων, καθώς τα μοντέλα με τα συναισθηματικά δεδομένα παρουσίασαν βελτιωμένες μετρικές ακριβείας σε σχέση με τα βασικά μοντέλα. Η μελέτη επιβεβαιώνει ότι οι συναισθηματικοί δείκτες που προκύπτουν από τα χρηματοοικονομικά έγγραφα των εταιρειών παρέχουν πολύτιμη πληροφορία για την πρόβλεψη των μελλοντικών κινήσεων της αγοράς.

Η ανάλυση των καινοτομιών των μεγάλων εταιρειών αποτελεί ένα κρίσιμο εργαλείο για την κατανόηση της στρατηγικής τους και της πορείας της επιτυχίας τους. Παρά την ποιοτική φύση των δεδομένων που αφορούν τις καινοτομίες, αυτά είναι σημαντικά για την οικονομική ανάλυση, καθώς συχνά αντιπροσωπεύουν σημαντικά στάδια ανάπτυξης μιας εταιρείας. Η κατανόηση της χρονικής ακολουθίας και του τρόπου εισαγωγής καινοτομιών μπορεί να παρέχει πληροφορίες για την επίδρασή τους στην κερδοφορία της εταιρείας και στην τιμή της μετοχής της. Στο πλαίσιο αυτής της μελέτης, οι καινοτομίες των εταιρειών, όπως αναγνωρίστηκαν μέσω των SEC filings, συνδέονται άμεσα με τις οικονομικές επιδόσεις τους, ενώ η ενσωμάτωσή τους σε μοντέλα πρόβλεψης βοηθά στην καλύτερη κατανόηση της στρατηγικής τους.

Λέξεις κλειδιά :

Τεχνικές Προβλέψεων, Μετοχές, Μηχανική Μάθηση, Νευρωνικά δίκτυα, Python.

Abstract

This thesis aims to evaluate the innovations developed by major corporations and the use of sentiment analysis techniques on their financial documents (SEC filings) to predict their future stock performance. The analysis focused on three key financial documents—10-K reports, 10-Q reports, and press releases—from which sentiment indicators were extracted. These sentiment indicators were incorporated into machine learning models used to forecast the stock price direction for the following companies: Apple (AAPL), Nvidia (NVDA), and Tesla (TSLA).

The findings revealed that the inclusion of these sentiment indicators enhanced prediction accuracy, as the sentiment-enriched models outperformed the baseline models across multiple performance metrics. The study confirms that sentiment indicators derived from corporate financial disclosures provide valuable insights for forecasting future market movements.

Moreover, analyzing corporate innovations serves as a critical tool for understanding a company's strategic trajectory and success. Although innovation data is inherently qualitative, it holds substantial relevance for financial analysis, as it often reflects key stages of a company's development. Understanding the timing and nature of innovation introductions can shed light on their impact on corporate profitability and stock valuation. Within the context of this study, company innovations—identified through SEC filings—are directly linked to financial performance, and their integration into predictive models contributes to a deeper understanding of corporate strategy.

Keywords: Forecasting Techniques, Stocks, Machine Learning, Neural Networks, Python.

Περιεχόμενα

1. ΕΙΣΑΓΩΓΗ	15
1.1. ΕΙΣΑΓΩΓΗ ΣΤΙΣ ΧΡΗΜΑΤΟΠΙΣΤΩΤΙΚΕΣ ΑΓΟΡΕΣ	15
1.2. ΜΕΘΟΔΟΙ ΠΡΟΒΛΕΨΗΣ ΧΡΗΜΑΤΙΣΤΗΡΙΑΚΩΝ ΠΡΟΪΟΝΤΩΝ	16
1.3. ΣΧΕΤΙΚΗ ΒΙΒΛΙΟΓΡΑΦΙΑ	19
1.4. ΣΚΟΠΟΣ ΤΗΣ ΕΡΓΑΣΙΑΣ.....	21
1.5. ΟΡΓΑΝΩΣΗ ΤΗΣ ΕΡΓΑΣΙΑΣ.....	23
2. ΘΕΩΡΗΤΙΚΟ ΥΠΟΒΑΘΡΟ	25
2.1. ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΑ ΔΕΔΟΜΕΝΑ ΚΑΙ ΧΡΟΝΟΣΕΙΡΕΣ.....	25
2.1.1. Εισαγωγή στις Χρονοσειρές	25
2.1.2. Ιστορικά δεδομένα τιμών.....	26
2.1.3. Καταθέσεις στην Επιτροπή Κεφαλαιαγοράς των Ηνωμένων Πολιτειών.....	27
2.1.4. Οικονομικό Ημερολόγιο και κέρδη εταιρειών	29
2.1.5. Εισαγωγή στις Χρονοσειρές	31
2.1.6. Ανάλυση Χρονοσειρών.....	32
2.1.7. Χαρακτηριστικά Χρονοσειρών	33
2.1.8. Χρήση Χρονοσειρών στις Χρηματοοικονομικές Αγορές.....	35
2.1.9. Προκλήσεις και Περιορισμοί.....	35
2.2. ΠΡΟΒΛΕΨΕΙΣ.....	36
2.2.1. Εισαγωγή στις προβλέψεις.....	36
2.2.2. Διαδικασία Προβλέψης.....	37
2.2.3. Χρονικός ορίζοντας προβλέψεων.....	38
2.2.4. Μέθοδοι προβλέψεων	39
2.3. ΣΤΑΤΙΣΤΙΚΗ ΑΝΑΛΥΣΗ	43
2.3.1. Εισαγωγή στη Στατιστική Ανάλυση και βασικά μεγέθη.....	43
2.3.2. Μετρικές Ακρίβειας Μοντέλων Μηχανικής Μάθησης	46
3. ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ.....	48
3.1. ΕΙΣΑΓΩΓΗ ΣΤΗ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ.....	48
3.1.1. Ορισμός και στόχοι της μηχανικής μάθησης.....	48
3.1.2. Κατηγορίες μοντέλων.....	49
3.1.3. Ενισχυτική Μάθηση.....	50
3.2. ΑΡΧΙΤΕΚΤΟΝΙΚΕΣ ΝΕΥΡΩΝΙΚΩΝ ΔΙΚΤΥΩΝ	51
3.2.1. Μοντέλο Βιολογικού Νευρώνα	51
3.2.2. Μοντέλο Τεχνητού Νευρώνα (Perceptron).....	53
3.2.3. Συναρτήσεις ενεργοποίησης	54
3.2.4. Εκπαίδευση Νευρωνικών δικτύων.....	56
3.2.5. Πολυεπίπεδο Perceptron (MLP).....	58
3.2.6. Οπίσθια Διάδοση (Backpropagation)	59
3.2.7. Υπερπροσαρμογή (Overfitting)	61
3.3. ΕΠΑΝΑΛΑΜΒΑΝΟΜΕΝΑ ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ (RNN).....	62
3.4. ΜΑΚΡΟΧΡΟΝΙΑ ΜΝΗΜΗ ΚΑΙ ΒΡΑΧΥΠΡΟΘΕΣΜΗ ΜΝΗΜΗ (LSTM)	64
3.4.1. Αρχιτεκτονική των LSTM Δικτύων	64

3.5. ΑΝΑΔΡΟΜΙΚΕΣ ΜΟΝΑΔΕΣ ΜΕ ΠΥΛΕΣ (GRU)	66
3.6. ΕΦΑΡΜΟΓΕΣ GRU ΣΕ ΧΡΟΝΟΣΕΙΡΕΣ	68
3.7. ΠΡΟΚΛΗΣΕΙΣ ΚΑΙ ΠΕΡΙΟΡΙΣΜΟΙ.....	69
4. SENTIMENT ANALYSIS	70
4.1. Η ΕΠΙΔΡΑΣΗ ΤΩΝ ΣΥΝΑΙΣΘΗΜΑΤΩΝ ΣΤΗΝ ΟΙΚΟΝΟΜΙΑ	70
4.1.1. Οικονομικοί κύκλοι.....	70
4.1.2. Φόβος και απληστία.....	72
4.1.3. Το συναίσθημα ως ποιοτικό μέγεθος	73
4.2. ΕΠΕΞΕΡΓΑΣΙΑ ΦΥΣΙΚΗΣ ΓΛΩΣΣΑΣ ΓΙΑ ΕΞΑΓΩΓΗ ΣΥΝΑΙΣΘΗΜΑΤΟΣ	74
4.2.1. Προεπεξεργασία κειμένου	75
4.2.2. Word Embeddings	78
4.3. ΤΟ BERT ΚΑΙ Η ΠΡΟΣΕΓΓΙΣΗ ΤΟΥ ΣΤΗΝ ΚΑΤΑΝΟΗΣΗ ΚΕΙΜΕΝΟΥ	79
4.4. ΤΟ GPT ΚΑΙ Η ΧΡΗΣΗ ΤΟΥ ΣΤΗΝ ΑΝΑΛΥΣΗ ΣΥΝΑΙΣΘΗΜΑΤΟΣ	81
5. ΣΥΛΛΟΓΗ ΚΑΙ ΕΠΕΞΕΡΓΑΣΙΑ ΔΕΔΟΜΕΝΩΝ	83
5.1. ΣΥΛΛΟΓΗ ΔΕΔΟΜΕΝΩΝ.....	83
5.1.1. Επιλογή Μετοχών.....	84
5.1.2. Ιστορικά δεδομένα τιμών.....	84
5.1.3. Ανακοινώσεις Κερδών (Earnings Announcements).....	85
5.1.4. Ρυθμιστικές Καταθέσεις στην Επιτροπή Κεφαλαιαγοράς της Αμερικής	86
5.2. ΕΠΕΞΕΡΓΑΣΙΑ ΔΕΔΟΜΕΝΩΝ	87
5.2.1. Ευθυγράμμιση και Διαχείριση Χρονοσειρών.....	87
5.2.2. Εκκαθάριση και Ανάλυση Ρυθμιστικών Καταθέσεων	87
5.2.3. Οργάνωση και Αποθήκευση Δεδομένων	88
5.3. ΕΞΑΓΩΓΗ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ (FEATURE ENGINEERING).....	88
5.3.1. Ποσοτικά Χαρακτηριστικά.....	88
5.3.2. Ποσοτικοποίηση Ποιοτικών Χαρακτηριστικών	91
5.4. ΕΠΙΛΟΓΗ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ (FEATURE SELECTION)	97
5.4.1. Ανάλυση Συσχέτισης (Correlation analysis).....	97
5.5. ΔΙΕΡΕΥΝΗΤΙΚΗ ΑΝΑΛΥΣΗ (EDA-EXPLORATORY DATA ANALYSIS).....	99
6. ΠΕΙΡΑΜΑΤΙΚΗ ΔΙΑΔΙΚΑΣΙΑ ΚΑΙ ΑΠΟΤΕΛΕΣΜΑΤΑ.....	101
6.1. ΑΡΧΙΤΕΚΤΟΝΙΚΗ ΤΩΝ ΜΟΝΤΕΛΩΝ	101
6.2. ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΑΞΙΟΛΟΓΗΣΗ ΤΩΝ ΜΟΝΤΕΛΩΝ	102
6.3. ΑΠΟΤΕΛΕΣΜΑΤΑ ΑΝΑ ΜΟΝΤΕΛΟ ΓΙΑ ΚΑΘΕ ΜΕΤΟΧΗ	103
6.3.1. Αποτελέσματα Μετοχής AAPL.....	104
6.3.2. Αποτελέσματα Μετοχής NVDA	105
6.3.3. Αποτελέσματα Μετοχής TSLA	106
6.3.4. Σημαντικότητα Αποτελεσμάτων και Ερμηνεία Αποκλίσεων.....	107
7. ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΜΕΛΛΟΝΤΙΚΗ ΕΡΓΑΣΙΑ.....	108
7.1. ΣΗΜΑΝΤΙΚΟΤΗΤΑ ΑΠΟΤΕΛΕΣΜΑΤΩΝ ΚΑΙ ΕΡΜΗΝΕΙΑ ΑΠΟΚΛΙΣΕΩΝ	108
7.2. ΜΕΛΛΟΝΤΙΚΗ ΕΡΓΑΣΙΑ ΚΑΙ ΠΕΡΙΟΡΙΣΜΟΙ.....	109
7.3. ΠΙΘΑΝΕΣ ΧΡΗΣΕΙΣ ΤΗΣ ΜΕΛΕΤΗΣ	110
8. BIBLIOGRAPHY	111

Συντομογραφίες

AI	Artificial Intelligence
ANN	Artificial Neural Network
API	Application Programming Interface
BERT	Bidirectional Encoder Representations from Transformers
BPTT	Backpropagation Through Time
BoW	Bag-of-Words
CLS	Classification Token
ECG	Electrocardiogram
EEG	Electroencephalogram
EPS	Earnings Per Share
ETF	Exchange-Traded Fund
FOMO	Fear of Missing Out
FVG	Fair Value Gap
GPT	Generative Pre-trained Transformer
GRU	Gated Recurrent Unit
IDF	Inverse Document Frequency (μέρος του TF-IDF)
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
ME	Mean Error
ML	Machine Learning
MLP	Multi-Layer Perceptron
MSE	Mean Squared Error
NLP	Natural Language Processing
OHLCV	Open, High, Low, Close, Volume
PNL	Profit and Loss
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Network
SEC	Securities and Exchange Commission
TF	Term Frequency (μέρος του TF-IDF)
TF-IDF	Term Frequency-Inverse Document Frequency

1. Εισαγωγή

1.1. Εισαγωγή στις Χρηματοπιστωτικές Αγορές

Οι χρηματοπιστωτικές αγορές αποτελούν τον πυρήνα της παγκόσμιας οικονομίας, λειτουργώντας ως πλατφόρμες όπου κεφάλαια διατίθενται από επενδυτές σε επιχειρήσεις και κυβερνήσεις. Μέσω αυτών των αγορών, επιτυγχάνεται η αποτελεσματική κατανομή των πόρων, η ενίσχυση της ρευστότητας και η προώθηση της οικονομικής ανάπτυξης. Το χρηματιστήριο, ως κεντρικό στοιχείο των χρηματοπιστωτικών αγορών, παρέχει έναν οργανωμένο χώρο για την αγορά και πώληση χρηματοοικονομικών μέσων, όπως μετοχές και ομόλογα. Η ύπαρξή του επιτρέπει στις επιχειρήσεις να αντλούν κεφάλαια για επενδύσεις και ανάπτυξη, ενώ παράλληλα προσφέρει στους επενδυτές τη δυνατότητα να συμμετέχουν στην οικονομική επιτυχία αυτών των επιχειρήσεων.

Η ιστορία των χρηματιστηρίων ανάγεται στον 17ο αιώνα, με την ίδρυση του πρώτου χρηματιστηρίου στην Ολλανδία. Η Ολλανδική Εταιρεία Ανατολικών Ινδιών (Vereenigde Oostindische Compagnie – VOC) ήταν από τις πρώτες εταιρείες που εξέδωσε μετοχές, επιτρέποντας στο κοινό να επενδύσει στις δραστηριότητές της. Αυτή η καινοτομία έθεσε τα θεμέλια για τη σύγχρονη χρηματιστηριακή αγορά, καθιστώντας δυνατή τη συγκέντρωση κεφαλαίων από ένα ευρύ φάσμα επενδυτών. [1]

Από τον 19ο αιώνα έπειτα, η ανάπτυξη των βιομηχανικών οικονομιών οδήγησε στη δημιουργία μεγαλύτερων και πιο σύνθετων χρηματοοικονομικών αγορών, με την εμφάνιση νέων χρηματιστηρίων σε παγκόσμιο επίπεδο. Οι εξελίξεις στις επικοινωνίες, όπως ο τηλεγράφος και αργότερα το διαδίκτυο, επέτρεψαν τη γρηγορότερη διάδοση των χρηματοοικονομικών πληροφοριών, μειώνοντας τα γεωγραφικά εμπόδια στις συναλλαγές. Παράλληλα, η σταδιακή θεσμική ρύθμιση των αγορών ενίσχυσε τη διαφάνεια και την εμπιστοσύνη των επενδυτών, συμβάλλοντας στην εδραίωση των χρηματιστηρίων ως κεντρικών μηχανισμών συγκέντρωσης κεφαλαίων και κατανομής επενδύσεων.

Η ραγδαία τεχνολογική πρόοδος των τελευταίων δεκαετιών έχει επιφέρει σημαντικές αλλαγές στη λειτουργία των χρηματιστηρίων, καθιστώντας τις συναλλαγές ταχύτερες και πιο προσιτές για τους επενδυτές. Η μετάβαση από τις παραδοσιακές μεθόδους διαπραγμάτευσης σε ηλεκτρονικά συστήματα έχει μειώσει το κόστος συναλλαγών και έχει αυξήσει την αποδοτικότητα των αγορών. Η αυτοματοποίηση των συναλλαγών συνέβαλε στη μείωση του κόστους παροχής υπηρεσιών διαπραγμάτευσης και μετασχημάτισε τη δομή της βιομηχανίας χρηματοοικονομικών υπηρεσιών [2]. Επιπλέον, η ανάπτυξη διαδικτυακών πλατφορμών συναλλαγών έχει διευρύνει την πρόσβαση στις αγορές για ιδιώτες επενδυτές. Αυτές οι πλατφόρμες προσφέρουν φιλικά προς τον χρήστη περιβάλλοντα, πραγματικό χρόνο δεδομένων αγοράς και δυνατότητες άμεσης εκτέλεσης εντολών, επιτρέποντας στους επενδυτές να διαχειρίζονται τα χαρτοφυλάκιά τους με μεγαλύτερη ευκολία και αποτελεσματικότητα. Μελέτη των Chuen et al. (2024) [3] εξετάζει τους κύριους παράγοντες που επηρεάζουν την υιοθέτηση τεχνολογικά προηγμένων πλατφορμών συναλλαγών από ιδιώτες επενδυτές, υπογραμμίζοντας τη σημασία της τεχνολογικής καινοτομίας στη διαμόρφωση των σύγχρονων χρηματοοικονομικών αγορών.

Η τεχνολογική πρόοδος στις χρηματοπιστωτικές αγορές έχει οδηγήσει στην ανάπτυξη νέων μεθόδων πρόβλεψης που αποκλίνουν από τις παραδοσιακές προσεγγίσεις. Σύγχρονες τεχνικές,

όπως η μηχανική μάθηση και η ανάλυση μεγάλων δεδομένων, επιτρέπουν την επεξεργασία τεράστιων όγκων πληροφοριών σε πραγματικό χρόνο, προσφέροντας τη δυνατότητα για πιο ακριβείς και έγκαιρες προβλέψεις. Αυτές οι μέθοδοι μπορούν να ανιχνεύσουν πολύπλοκα μοτίβα και συσχετίσεις που δεν είναι εμφανείς με τις παραδοσιακές στατιστικές τεχνικές, ενισχύοντας έτσι την ικανότητα των επενδυτών να λαμβάνουν ενημερωμένες αποφάσεις. Ωστόσο, η υιοθέτηση αυτών των νέων μεθόδων συνοδεύεται και από προκλήσεις. Η πολυπλοκότητα των αλγορίθμων μπορεί να οδηγήσει σε έλλειψη διαφάνειας, καθιστώντας δύσκολη την κατανόηση του τρόπου με τον οποίο λαμβάνονται οι αποφάσεις [4]. Επιπλέον, η υπερβολική εξάρτηση από αυτοματοποιημένα συστήματα μπορεί να αυξήσει τον κίνδυνο συστημικών αποτυχιών, ιδίως εάν οι αλγόριθμοι βασίζονται σε εσφαλμένα ή μεροληπτικά δεδομένα. Αξίζει επίσης να τονιστεί ότι, παρά τις εξελιγμένες δυνατότητές τους, αυτές οι μέθοδοι δεν είναι σε θέση να προβλέψουν απρόβλεπτα γεγονότα ή φαινόμενα "μαύρου κύκνου", τα οποία μπορεί να επηρεάσουν δραστικά τις αγορές.

Η ενσωμάτωση εναλλακτικών τύπων δεδομένων στις μεθόδους πρόβλεψης των χρηματοπιστωτικών αγορών έχει αναδειχθεί ως σημαντική καινοτομία, επεκτείνοντας τις δυνατότητες ανάλυσης πέρα από τα παραδοσιακά χρηματοοικονομικά δεδομένα. Αυτά τα εναλλακτικά δεδομένα περιλαμβάνουν πληροφορίες από μέσα κοινωνικής δικτύωσης, ειδησεογραφικές ροές, δεδομένα διαδικτυακών αναζητήσεων και δορυφορικές εικόνες, παρέχοντας νέες δυνατότητες για την ανάλυση των αγορών και της επενδυτικής συμπεριφοράς.

Εστιάζοντας στις επενδύσεις, έχει αποδειχθεί ότι τα εναλλακτικά δεδομένα μπορούν να βελτιώσουν την ακρίβεια πρόβλεψης εσόδων, μειώνοντας το μέσο απόλυτο σφάλμα από 88% σε 2,6%. [5]

Παρά τις προκλήσεις που συνοδεύουν την υιοθέτηση νέων μεθόδων πρόβλεψης στις χρηματοπιστωτικές αγορές, οι προοπτικές που διανοίγονται είναι ενθαρρυντικές. Η ενσωμάτωση προηγμένων τεχνολογιών, όπως η μηχανική μάθηση και η ανάλυση μεγάλων δεδομένων, προσφέρει τη δυνατότητα για πιο ακριβείς και έγκαιρες προβλέψεις, ενισχύοντας τη λήψη ενημερωμένων επενδυτικών αποφάσεων. Με την κατάλληλη διαχείριση των κινδύνων και τη συνεχή βελτίωση των αλγορίθμων, αυτές οι καινοτόμες προσεγγίσεις μπορούν να συμβάλουν στη σταθερότητα και την αποδοτικότητα των αγορών, προσφέροντας νέες ευκαιρίες για επενδυτές και φορείς της αγοράς.

1.2. Μέθοδοι πρόβλεψης χρηματιστηριακών προϊόντων

Στην ανάλυση των χρηματοπιστωτικών αγορών, δύο κύριες μεθοδολογίες χρησιμοποιούνται για την αξιολόγηση και πρόβλεψη των τιμών των περιουσιακών στοιχείων: η θεμελιώδης ανάλυση(fundamental analysis) και η τεχνική ανάλυση (technical analysis).

Η πρόβλεψη των τιμών των χρηματιστηριακών προϊόντων αποτελεί θεμελιώδες ζήτημα στη χρηματοοικονομική έρευνα και πρακτική, με στόχο την ενίσχυση της αποτελεσματικότητας των επενδυτικών αποφάσεων και τη μείωση του σχετικού κινδύνου. Οι μεθοδολογίες που έχουν αναπτυχθεί για τον σκοπό αυτό μπορούν να ταξινομηθούν σε τρεις κύριες κατηγορίες: στατιστικές μέθοδοι, τεχνικές μηχανικής μάθησης και ανάλυση συναισθήματος.

Στατιστικές Μέθοδοι

Οι στατιστικές μέθοδοι αποτελούν τις παραδοσιακές προσεγγίσεις στην πρόβλεψη των τιμών των χρηματιστηριακών προϊόντων, βασιζόμενες στην ανάλυση ιστορικών δεδομένων και στη χρήση μαθηματικών μοντέλων για την ανίχνευση τάσεων και μοτίβων. Μία από τις πλέον διαδεδομένες τεχνικές είναι τα αυτοπαλίνδρομα ολοκληρωμένα κινούμενου μέσου όρου μοντέλα (ARIMA - Autoregressive integrated moving average), τα οποία χρησιμοποιούνται για την ανάλυση και πρόβλεψη χρονοσειρών. Τα μοντέλα αυτά συνδυάζουν αυτοπαλίνδρομα (AR-AutoRegressive Models) και κινούμενου μέσου όρου (MA-Moving Average) στοιχεία, επιτρέποντας την ανάλυση των εξαρτήσεων μεταξύ των παρατηρήσεων σε μια χρονοσειρά. Ένα χαρακτηριστικό παράδειγμα εφαρμογής των στατιστικών μεθόδων είναι η χρήση του μοντέλου GARCH (Generalized Autoregressive Conditional Heteroskedasticity) για την πρόβλεψη της μεταβλητότητας των αποδόσεων των μετοχών. Το μοντέλο αυτό επιτρέπει την εκτίμηση της μεταβλητότητας με βάση τις προηγούμενες τιμές της, λαμβάνοντας υπόψη την ετεροσκεδαστικότητα που παρατηρείται στις χρηματοοικονομικές χρονοσειρές [6].

Εκτός από τις παραδοσιακές στατιστικές μεθόδους, αρκετές πιο εξειδικευμένες τεχνικές έχουν εφαρμοστεί στη χρηματοοικονομική ανάλυση και την πρόβλεψη των αγορών. Η εκτίμηση κατάστασης με φίλτρο Kalman (ESM - Exponential Smoothing Model) αποτελεί μία από τις πιο διαδεδομένες μεθόδους πρόβλεψης χρονοσειρών, προσφέροντας δυναμική εκτίμηση των παραμέτρων του μοντέλου με βάση την πιο πρόσφατη διαθέσιμη πληροφορία [7]. Το Γραμμικό και Τετραγωνικό Διακριτό Μοντέλο Ανάλυσης (LDA/QDA - Linear/Quadratic Discriminant Analysis) χρησιμοποιούνται για την ταξινόμηση και διάκριση χρηματοοικονομικών δεδομένων, αξιοποιώντας την κατανομή των χαρακτηριστικών των παρατηρήσεων [8]. Επιπλέον, η Ανάλυση Κύριων Συνιστωσών (PCA - Principal Component Analysis) χρησιμοποιείται ευρέως για τη μείωση της διάστασης και την εξαγωγή των σημαντικότερων πληροφοριών από μεγάλα χρηματοοικονομικά δεδομένα.

Τεχνικές Μηχανικής Μάθησης

Οι ραγδαίες εξελίξεις στην υπολογιστική ισχύ, όπως η ανάπτυξη των GPU και η αύξηση της διαθέσιμης μνήμης, σε συνδυασμό με τη μαζική συσσώρευση χρηματοοικονομικών δεδομένων, έχουν επιτρέψει την εφαρμογή πιο σύνθετων αλγορίθμων μηχανικής μάθησης. Ως αποτέλεσμα, οι τεχνικές αυτές έχουν αποκτήσει εξέχουσα θέση στην πρόβλεψη των τιμών των χρηματιστηριακών προϊόντων, επιτρέποντας τη δημιουργία μοντέλων που μπορούν να ανιχνεύσουν και να αξιοποιήσουν μοτίβα στις αγορές με μεγαλύτερη ακρίβεια. Οι τεχνικές αυτές επιτρέπουν την ανάλυση μεγάλων όγκων δεδομένων και την ανίχνευση σύνθετων μοτίβων που δεν είναι εμφανή με τις παραδοσιακές στατιστικές μεθόδους.

Οι τεχνικές μηχανικής μάθησης διακρίνονται σε επιβλεπόμενες και μη επιβλεπόμενες μεθόδους, καθεμία από τις οποίες εξυπηρετεί διαφορετικούς σκοπούς στην ανάλυση χρηματοοικονομικών δεδομένων. Οι επιβλεπόμενοι αλγόριθμοι, όπως η γραμμική παλινδρόμηση, τα τυχαία δάση και οι υποστηρικτικές διανυσματικές μηχανές (SVM-Support Vector Machines), χρησιμοποιούν ετικετοποιημένα δεδομένα για την πρόβλεψη συγκεκριμένων μεταβλητών-στόχων. Από την άλλη, οι μη επιβλεπόμενες τεχνικές, όπως η ανάλυση κατά συστάδες (clustering) και η ανάλυση κύριων συνιστωσών (PCA - Principal Component Analysis), επιτρέπουν την ανίχνευση υποκείμενων δομών στα δεδομένα χωρίς την ανάγκη προκαθορισμένων ετικετών. Αυτές οι

μέθοδοι έχουν καταστήσει δυνατή την ανάλυση μεγάλων όγκων δεδομένων και την ανίχνευση σύνθετων μοτίβων που δεν είναι εμφανή με τις παραδοσιακές στατιστικές προσεγγίσεις.

Μία από τις πλέον διαδεδομένες τεχνικές στην πρόβλεψη χρηματοοικονομικών δεδομένων είναι τα τεχνητά νευρωνικά δίκτυα (ANN - Artificial Neural Networks), τα οποία εμπνέονται από τη δομή και τη λειτουργία του ανθρώπινου εγκεφάλου. Αυτά τα δίκτυα αποτελούνται από πολλαπλά επίπεδα νευρώνων που προσαρμόζουν τα βάρη τους μέσω της διαδικασίας της εκπαίδευσης, επιτρέποντας την αναγνώριση σύνθετων σχέσεων στα δεδομένα. Ιδιαίτερα στις χρονοσειρές των χρηματοπιστωτικών αγορών, η χρήση νευρωνικών δικτύων μνήμης μακράς και βραχείας διάρκειας (LSTM - Long Short Term Memory) έχει αποδειχθεί εξαιρετικά αποτελεσματική, καθώς μπορούν να διατηρούν πληροφορίες από προηγούμενες χρονικές στιγμές και να ανιχνεύουν μη γραμμικά μοτίβα.

Σε μελέτη των Fischer και Krauss (2018) [8], τα LSTM χρησιμοποιήθηκαν για την πρόβλεψη των τιμών των μετοχών και αποδείχθηκαν η πλέον αποδοτική μέθοδος, επιτυγχάνοντας ετήσιες αποδόσεις 82,29% μετά από κόστη συναλλαγών. Η απόδοση αυτή υπερτερεί σε σύγκριση με άλλες μεθόδους, όπως τυχαία δάση με 67,87%, βαθιά νευρωνικά δίκτυα με 24,60%, λογιστική παλινδρόμηση με 7,11% και την απόδοση της αγοράς στο 9,25%. Αυτό αποδεικνύει τη δυνατότητα των νευρωνικών δικτύων να βελτιώνουν σημαντικά τις επενδυτικές στρατηγικές, αξιοποιώντας προηγμένες τεχνικές ανάλυσης και μεγάλα δεδομένα. [9]

Ανάλυση Συναισθήματος

Η ανάλυση συναισθήματος ανήκει στον ευρύτερο τομέα της μηχανικής μάθησης, καθώς βασίζεται σε αλγόριθμους που επεξεργάζονται και ταξινομούν κείμενα βάσει του συναισθηματικού τους περιεχομένου. Ωστόσο, λόγω της ιδιαιτερότητάς της και της άμεσης σύνδεσής της με την επενδυτική ψυχολογία, εξετάζεται ξεχωριστά. Πρόκειται για μια σχετικά νέα προσέγγιση στην πρόβλεψη των τιμών των χρηματιστηριακών προϊόντων, η οποία αξιοποιεί δεδομένα από ειδήσεις, αναρτήσεις σε μέσα κοινωνικής δικτύωσης, αναφορές αναλυτών και άλλα χρηματοοικονομικά κείμενα. Η βασική της υπόθεση είναι ότι τα συναισθήματα και οι αντιλήψεις των επενδυτών μπορούν να επηρεάσουν τις τιμές των μετοχών, δημιουργώντας μοτίβα που μπορούν να ανιχνευθούν και να αξιοποιηθούν για επενδυτικές αποφάσεις.

Για παράδειγμα, σε μελέτη που πραγματοποιήθηκε από τους Bollen, Mao και Zeng (2011) [10], χρησιμοποιήθηκε ανάλυση συναισθήματος σε δεδομένα από το Twitter για την πρόβλεψη των τιμών των μετοχών. Τα αποτελέσματα έδειξαν ότι η ενσωμάτωση της ανάλυσης συναισθήματος μπορεί να βελτιώσει την ακρίβεια των προβλέψεων, πετυχαίνοντας ακρίβεια 86.7% στην πρόβλεψη των ημερήσιων ανοδικών ή καθοδικών αλλαγών της τιμής κλεισίματος του DJIA (Dow Jones Industrial Average) και μείωση του MAPE κατά 6%, υποδεικνύοντας τη σημασία των συναισθηματικών δεδομένων στην ανάλυση των χρηματοοικονομικών αγορών.

Συνοψίζοντας, οι μέθοδοι πρόβλεψης των τιμών των χρηματιστηριακών προϊόντων έχουν εξελιχθεί σημαντικά, περνώντας από τις παραδοσιακές στατιστικές προσεγγίσεις στις σύγχρονες τεχνικές μηχανικής μάθησης και ανάλυσης συναισθήματος. Αρχικά, η ανάλυση επικεντρωνόταν κυρίως σε δεδομένα από μέσα κοινωνικής δικτύωσης, όπως το Twitter, αλλά πλέον επεκτείνεται σε πιο δομημένες και περιεκτικές πηγές, όπως ειδησεογραφικά άρθρα και αναφορές αναλυτών. Σύμφωνα με πρόσφατη μελέτη των Kirtac and Guido [11], η χρήση μεγάλων γλωσσικών

μοντέλων (LLMs) για ανάλυση συναισθήματος σε ειδήσεις βελτιώνει σημαντικά την ικανότητα πρόβλεψης των αγορών, καθώς επιτυγχάνεται καλύτερη ανάλυση συναισθήματος ενισχύοντας έτσι την πρόβλεψη των χρηματοοικονομικών τιμών. Η επιλογή της κατάλληλης μεθοδολογίας εξαρτάται πλέον από τη φύση των διαθέσιμων δεδομένων, τους επενδυτικούς στόχους και την ικανότητα αξιοποίησης προηγμένων τεχνικών τεχνητής νοημοσύνης.

Θεμελιώδης Ανάλυση:

Η θεμελιώδης ανάλυση επικεντρώνεται στην αξιολόγηση της εσωτερικής αξίας ενός περιουσιακού στοιχείου μέσω της εξέτασης οικονομικών, χρηματοοικονομικών και άλλων σχετικών παραγόντων. Στόχος είναι ο προσδιορισμός της πραγματικής αξίας μιας μετοχής ή άλλου χρηματοοικονομικού προϊόντος, συγκρίνοντάς την με την τρέχουσα τιμή αγοράς για τον εντοπισμό επενδυτικών ευκαιριών. Αυτή η προσέγγιση περιλαμβάνει την ανάλυση οικονομικών καταστάσεων, όπως ο ισολογισμός, η κατάσταση αποτελεσμάτων χρήσης και η κατάσταση ταμειακών ροών, καθώς και την αξιολόγηση μακροοικονομικών δεικτών και κλαδικών τάσεων. Σύμφωνα με τους Yu (2023) [12] και Mandelbrot (1963) [13], η θεμελιώδης ανάλυση αξιολογεί την οικονομική υγεία των εταιρειών μέσω της ανάλυσης χρηματοοικονομικών δεικτών, του υπολογισμού της εσωτερικής αξίας των περιουσιακών στοιχείων και της σύγκρισής τους με τις τιμές της αγοράς, παρέχοντας έτσι τιμές-στόχους για τα περιουσιακά στοιχεία.

Τεχνική Ανάλυση:

Η τεχνική ανάλυση εστιάζει στη μελέτη των ιστορικών δεδομένων τιμών και όγκου συναλλαγών για την πρόβλεψη μελλοντικών κινήσεων των τιμών. Βασίζεται στην υπόθεση ότι όλες οι σχετικές πληροφορίες αντικατοπτρίζονται ήδη στις τιμές της αγοράς και ότι οι τιμές τείνουν να κινούνται σύμφωνα με συγκεκριμένα μοτίβα ή τάσεις. Χρησιμοποιεί εργαλεία όπως διαγράμματα τιμών, τεχνικούς δείκτες (π.χ. κινητούς μέσους όρους, δείκτες σχετικής ισχύος) και μοτίβα σχηματισμών για τον εντοπισμό ευκαιριών συναλλαγών. Η τεχνική ανάλυση περιλαμβάνει τη χρήση διαγραμμάτων και στατιστικών δεικτών για την ανάλυση των τάσεων της αγοράς και τη λήψη επενδυτικών αποφάσεων.

Ενώ η θεμελιώδης ανάλυση προσπαθεί να κατανοήσει το "γιατί" πίσω από τις κινήσεις των τιμών, εξετάζοντας τους υποκείμενους οικονομικούς και χρηματοοικονομικούς παράγοντες, η τεχνική ανάλυση επικεντρώνεται στο "πώς" των κινήσεων, αναλύοντας τα μοτίβα και τις τάσεις που παρουσιάζονται στα διαγράμματα τιμών. Κάθε προσέγγιση έχει τα δικά της πλεονεκτήματα και περιορισμούς, και πολλοί επενδυτές επιλέγουν να συνδυάζουν και τις δύο μεθοδολογίες για μια πιο ολοκληρωμένη ανάλυση της αγοράς. Η συνδυασμένη χρήση θεμελιώδους και τεχνικής ανάλυσης μπορεί να οδηγήσει σε πιο συστηματικές επενδυτικές αποφάσεις στις μετοχικές αγορές [14].

1.3. Σχετική βιβλιογραφία

Η ενότητα αυτή έχει ως στόχο να παρουσιάσει το θεωρητικό πλαίσιο και τη σχετική βιβλιογραφία που υποστηρίζουν τη χρήση προηγμένων μεθόδων πρόβλεψης των χρηματιστηριακών τιμών, όπως αυτές που υλοποιούνται στην παρούσα έρευνα.

Η πρόβλεψη των τιμών των χρηματιστηριακών προϊόντων αποτελεί ένα από τα πλέον μελετημένα πεδία στη χρηματοοικονομική έρευνα, με πληθώρα μοντέλων να έχουν αναπτυχθεί για τον σκοπό αυτό. Παραδοσιακά, οι στατιστικές μέθοδοι, όπως τα αυτοπαλίνδρομα ολοκληρωμένα κινούμενου μέσου όρου μοντέλα (ARIMA) και τα μοντέλα γενικευμένης αυτοπαλίνδρομης ετεροσκεδαστικότητας υπό συνθήκη (GARCH), έχουν χρησιμοποιηθεί ευρέως για την ανάλυση χρονοσειρών και την πρόβλεψη των τιμών των μετοχών. Ωστόσο, η δυναμική και μη γραμμική φύση των χρηματοοικονομικών αγορών παρουσιάζει προκλήσεις για τα παραδοσιακά αυτά μοντέλα, καθώς συχνά αδυνατούν να συλλάβουν τις πολύπλοκες σχέσεις και τις απότομες μεταβολές που χαρακτηρίζουν τις αγορές.

Με την πρόοδο της τεχνολογίας και την αύξηση της διαθεσιμότητας δεδομένων, οι ερευνητές έχουν στραφεί σε πιο προηγμένες μεθόδους, όπως η μηχανική μάθηση και τα νευρωνικά δίκτυα. Μοντέλα όπως τα Long Short-Term Memory (LSTM) νευρωνικά δίκτυα έχουν δείξει υποσχόμενα αποτελέσματα στην πρόβλεψη των τιμών των μετοχών, καθώς μπορούν να διαχειριστούν τις μακροπρόθεσμες εξαρτήσεις και τη μη γραμμικότητα των δεδομένων. Επιπλέον, τεχνικές όπως οι γενετικοί αλγόριθμοι έχουν χρησιμοποιηθεί σε συνδυασμό με νευρωνικά δίκτυα για τη βελτιστοποίηση των προβλέψεων, επιτρέποντας την εξερεύνηση ενός ευρύτερου φάσματος πιθανών λύσεων και τη βελτίωση της ακρίβειας των μοντέλων. [15]

Παρά την πρόοδο αυτή, εξακολουθούν να υπάρχουν προκλήσεις. Ένα από τα κύρια προβλήματα είναι η υπερπροσαρμογή των μοντέλων στα ιστορικά δεδομένα, γεγονός που μπορεί να οδηγήσει σε μειωμένη ικανότητα γενίκευσης σε νέα, αδημοσίευτα δεδομένα. Επιπλέον, η πολυπλοκότητα των αγορών και η επίδραση εξωτερικών παραγόντων, όπως οι μακροοικονομικές συνθήκες και τα συναισθήματα των επενδυτών, καθιστούν δύσκολη την ακριβή πρόβλεψη των μελλοντικών τιμών. Για την αντιμετώπιση αυτών των προκλήσεων, προτείνονται υβριδικά μοντέλα που συνδυάζουν διάφορες τεχνικές, όπως η ενσωμάτωση ανάλυσης συναισθήματος από τα μέσα κοινωνικής δικτύωσης και τις ειδήσεις, προκειμένου να ληφθούν υπόψη οι ψυχολογικοί παράγοντες που επηρεάζουν τις αγορές. Σε μελέτη που πραγματοποιήθηκε από τους Bollen, [16], χρησιμοποιήθηκε ανάλυση συναισθήματος σε δεδομένα από το Twitter για την πρόβλεψη των τιμών των μετοχών, με τα αποτελέσματα να δείχνουν ότι η ενσωμάτωση της ανάλυσης συναισθήματος μπορεί να βελτιώσει την ακρίβεια των προβλέψεων.

Η χρήση της ανάλυσης συναισθήματος και των τεχνικών μηχανικής μάθησης για την πρόβλεψη των τιμών των μετοχών μέσω των καταθέσεων στην Επιτροπή Κεφαλαιαγοράς των ΗΠΑ (SEC) έχει αποτελέσει αντικείμενο σημαντικής έρευνας. Οι καταθέσεις αυτές, όπως οι ετήσιες εκθέσεις 10-K και οι τριμηνιαίες 10-Q, περιέχουν πλούσιες πληροφορίες που μπορούν να αναλυθούν για την εξαγωγή χρήσιμων συμπερασμάτων σχετικά με τη μελλοντική απόδοση των μετοχών.

Μια μελέτη από τους Heeyoung Lee και συνεργάτες [17] παρουσίασε μια προσέγγιση ανάλυσης κειμένου για την κατασκευή προβλεπτικών μοντέλων χρησιμοποιώντας τις αναφορές 8-K των εταιρειών. Η έρευνα αυτή ανέδειξε τη βελτίωση κατά 10% της ακρίβειας μέσω της χρήσης ανάλυσης κειμένου για την πρόβλεψη των μεταβολών στις τιμές των μετοχών με βάση τις πληροφορίες που περιέχονται στις καταθέσεις SEC.

Μια πρόσφατη μελέτη των Saleh και Nair [18] παρουσίασε μια μεθοδολογία βασισμένη στη βαθιά μάθηση για την πρόβλεψη των αυξήσεων των μετοχών που προκαλούνται από συγκεκριμένα γεγονότα. Οι ερευνητές εφάρμοσαν νευρωνικές μεθόδους σε ένα υπάρχον σύνολο

δεδομένων, καθώς και σε ένα νέο σύνολο δεδομένων που περιλάμβανε tweets. Συνδύασαν νευρωνικά δίκτυα χρησιμοποιώντας τόσο κειμενικά όσο και μη κειμενικά χαρακτηριστικά για να προβλέψουν τις κινήσεις των μετοχών κατά τις ημέρες υποβολής των αναφορών 8-K. Το καλύτερο μοντέλο τους, χρησιμοποιώντας ενσωματώσεις GloVe με ένα LSTM και πλήρως συνδεδεμένα επίπεδα, πέτυχε ακρίβεια 51% στο αρχικό σύνολο δεδομένων και 52,61% στο νέο σύνολο δεδομένων. Τελικά παρατήρησαν ότι η προσθήκη κειμενικών πληροφοριών βελτίωσε την απόδοση του μοντέλου.

Μια πρόσφατη μελέτη των Satras και Dias [19], διερεύνησε την αποτελεσματικότητα διαφόρων μεθόδων μηχανικής μάθησης στην πρόβλεψη των τιμών των μετοχών τεσσάρων μεγάλων τραπεζών στην Ινδονησία: BRI, BNI, BCA και Mandiri, κατά την περίοδο 2013-2022. Η έρευνα κατέληξε στο συμπέρασμα ότι η μοντελοποίηση ARIMA Box-Jenkins δεν ήταν κατάλληλη για την πρόβλεψη των τιμών των μετοχών αυτών των τραπεζών, ενώ το μοντέλο GRU παρουσίασε την καλύτερη απόδοση στην πρόβλεψη των τιμών των μετοχών τους.

Συνολικά, η ενσωμάτωση της ανάλυσης συναισθήματος και των τεχνικών μηχανικής μάθησης στην ανάλυση των καταθέσεων SEC προσφέρει μια πολλά υποσχόμενη προσέγγιση για την πρόβλεψη των τιμών των μετοχών, επιτρέποντας την εξαγωγή πολύτιμων πληροφοριών από μεγάλα σύνολα δεδομένων κειμένου.

1.4. Σκοπός της εργασίας

Η κατανόηση της στρατηγικής των ισχυρών παικτών στις αγορές αποτελεί κρίσιμο παράγοντα για τη μελέτη της εξέλιξής τους, καθώς και για την εκτίμηση της επίδρασής τους στις χρηματοοικονομικές αγορές. Ένα από τα πιο σημαντικά πεδία αυτής της μελέτης αφορά τις καινοτόμες ενέργειες που αναπτύσσουν οι εταιρείες, οι οποίες συχνά αντικατοπτρίζουν τις στρατηγικές τους επιλογές και τις μελλοντικές προοπτικές τους. Εντούτοις, η ποιοτική φύση των δεδομένων αυτών – όπως οι εταιρικές αναφορές και τα καινοτόμα εγχειρήματα – καθιστά δύσκολη τη συστηματική και αντικειμενική ανάλυση τους. Αυτό έχει ως αποτέλεσμα, τέτοιες κρίσιμες πληροφορίες να μην χρησιμοποιούνται τόσο συχνά σε οικονομικές αναλύσεις, παρά τη δυνατότητά τους να αποκαλύψουν σημαντικά μοτίβα στην κερδοφορία και τη χρηματιστηριακή απόδοση των εταιρειών.

Σκοπός της παρούσας εργασίας είναι να διερευνήσει τη δυνατότητα πρόβλεψης των τιμών των χρηματιστηριακών προϊόντων αξιοποιώντας προηγμένες μεθόδους μηχανικής μάθησης και ανάλυσης συναισθήματος. Ειδικότερα, η μελέτη εστιάζει στην αξιοποίηση ποιοτικών δεδομένων που προέρχονται από εταιρικές αναφορές, όπως οι εκθέσεις 10-K, 10-Q και 8-K. Μέσω της συνδυαστικής αυτής προσέγγισης, στόχος είναι να αναδειχθεί ο βαθμός στον οποίο η ενσωμάτωση ποιοτικών παραγόντων μπορεί να βελτιώσει την ακρίβεια των προβλέψεων και να συμβάλει στην αποτελεσματικότερη λήψη επενδυτικών αποφάσεων. Επιπλέον, η εργασία επιδιώκει να προτείνει ένα μεθοδολογικό πλαίσιο που θα επιτρέπει την αντικειμενική και συστηματική ανάλυση αυτών των ποιοτικών δεδομένων, αντιμετωπίζοντας τις προκλήσεις της ποσοτικοποίησης και ενσωμάτωσής τους σε χρηματοοικονομικά μοντέλα πρόβλεψης.

Η τεχνητή νοημοσύνη και, συγκεκριμένα, η μηχανική μάθηση αποτελούν τον πυρήνα αυτής της προσπάθειας. Μέσω προηγμένων αλγορίθμων, όπως τα νευρωνικά δίκτυα τύπου GRU (Gated

Recurrent Units), επιχειρείται η ανάλυση της σχέσης μεταξύ των εταιρικών δεδομένων και της αγοράς. Αυτά τα μοντέλα προσφέρουν τη δυνατότητα αναγνώρισης μοτίβων σε πολύπλοκα, μη γραμμικά δεδομένα, όπως οι χρηματοοικονομικές χρονοσειρές και τα συναισθήματα των επενδυτών, τα οποία συχνά δεν μπορούν να καταγραφούν εύκολα από παραδοσιακές μεθόδους.

Η συγκεκριμένη μελέτη επεκτείνει τη χρήση της μηχανικής μάθησης στην πρόβλεψη της κατεύθυνσης της τιμής των μετοχών (next day direction) σε καθημερινό επίπεδο. Με την αξιοποίηση ιστορικών δεδομένων τιμών και συναλλαγών, χρηματοοικονομικών καταστάσεων και αναλύσεων συναισθήματος, το αναπτυγμένο μοντέλο στοχεύει στην πρόβλεψη των μελλοντικών διακυμάνσεων της χρηματιστηριακής τιμής των εταιρειών. Για τον σκοπό αυτό, χρησιμοποιούνται πολυεπίπεδα νευρωνικά δίκτυα, τα οποία επιτρέπουν τη σύλληψη μακροπρόθεσμων εξαρτήσεων στα δεδομένα, ενώ παράλληλα προσαρμόζονται στις δυναμικές αλλαγές της αγοράς.

Η εργασία αυτή αντιμετωπίζει επίσης την πρόκληση που προκύπτει από την ποιοτική φύση των δεδομένων. Η ανάγνωση και ανάλυση μεγάλων όγκων εταιρικών εγγράφων από ανθρώπους είναι όχι μόνο χρονοβόρα, αλλά και ευάλωτη σε λάθη ή παραλείψεις. Εδώ έρχεται η μηχανική μάθηση να προσφέρει έναν αντικειμενικό και συστηματικό τρόπο ανάλυσης, ο οποίος μπορεί να εξαλείψει την ανθρώπινη προκατάληψη και να αποκαλύψει κρυφές σχέσεις μεταξύ των δεδομένων.

Σκοπός της παρούσας μελέτης είναι να εντοπίσει και να αξιολογήσει τέτοιες σχέσεις, με έμφαση στη συνεισφορά των καινοτομιών και της συναισθηματικής ανάλυσης (sentiment analysis) στη χρηματιστηριακή απόδοση. Επιπλέον, η μελέτη επικεντρώνεται σε τρεις από τις μεγαλύτερες εταιρείες του αμερικανικού χρηματιστηρίου, αναλύοντας την επίδραση των καινοτομιών τόσο βραχυπρόθεσμα όσο και μακροπρόθεσμα. Οι εταιρείες αυτές αντιπροσωπεύουν διαφορετικούς κλάδους, παρέχοντας ένα ευρύ πλαίσιο για τη μελέτη της σχέσης μεταξύ καινοτομίας και απόδοσης της μετοχής.

Η εργασία δεν περιορίζεται μόνο στη θεωρητική μελέτη, αλλά επιχειρεί να προσφέρει πρακτικές εφαρμογές μέσω της ανάπτυξης ενός εργαλείου που συνδυάζει δεδομένα από πολλαπλές πηγές. Η μελέτη αυτή αποσκοπεί στη βαθύτερη κατανόηση των στρατηγικών εταιρειών και της επίδρασής τους στην αγορά, ενώ παράλληλα να προτείνει βελτιωμένες μεθόδους πρόβλεψης και ανάλυσης για τον χρηματοοικονομικό τομέα.

Τελικά η εργασία στοχεύει στη συμπλήρωση των υφιστάμενων μεθοδολογιών μέσω της ανάπτυξης ενός μοντέλου που συνδυάζει την ανάλυση των εταιρικών αναφορών με προηγμένες τεχνικές μηχανικής μάθησης. Ειδικότερα, εστιάζει στην αξιοποίηση των δημοσιευμένων αναφορών από ισχυρούς παίκτες της αγοράς, όπως τα SEC Filings, για την αποκάλυψη του αντικτύπου τους στη χρηματιστηριακή αξία των εταιρειών. Μέσω της χρήσης GRU νευρωνικών δικτύων, το μοντέλο επιχειρεί να συλλάβει τις μακροπρόθεσμες εξαρτήσεις στα δεδομένα και να προβλέψει τις μελλοντικές διακυμάνσεις των τιμών των μετοχών. Επιπλέον, ενσωματώνει την ανάλυση συναισθήματος από διάφορες πηγές, προκειμένου να ληφθούν υπόψη οι ψυχολογικοί παράγοντες που επηρεάζουν τις αγορές. Με τον τρόπο αυτό, η εργασία φιλοδοξεί να προσφέρει μια πιο ολοκληρωμένη προσέγγιση στην πρόβλεψη των τιμών των χρηματιστηριακών προϊόντων, γεφυρώνοντας το χάσμα μεταξύ των ποιοτικών και ποσοτικών δεδομένων και βελτιώνοντας την ακρίβεια των προβλέψεων.

1.5. Οργάνωση της εργασίας

Η παρούσα διπλωματική εργασία είναι διαρθρωμένη σε επτά κύρια κεφάλαια, τα οποία καλύπτουν τόσο το θεωρητικό υπόβαθρο όσο και την πειραματική μελέτη του θέματος. Η διάρθρωση έχει σχεδιαστεί έτσι ώστε να παρέχει μια σαφή, λογική και ολοκληρωμένη προσέγγιση στο αντικείμενο της ανάλυσης ποιοτικών χρηματοοικονομικών δεδομένων και της εφαρμογής τεχνικών μηχανικής μάθησης.

Στο Κεφάλαιο 1: Εισαγωγή. Στο πρώτο κεφάλαιο, παρουσιάζεται το αντικείμενο της εργασίας, ο σκοπός και η σημασία της. Δίνεται έμφαση στις προκλήσεις που προκύπτουν από την ποιοτική φύση των δεδομένων και στην ανάγκη για συστηματική ανάλυση. Ιδιαίτερη προσοχή δίνεται στη μελέτη της υπάρχουσας βιβλιογραφίας, όπου αναλύονται προηγούμενες προσεγγίσεις στην πρόβλεψη των τιμών χρηματιστηριακών προϊόντων. Επιπλέον, περιγράφεται η δομή της εργασίας και το περιεχόμενό της.

Το Κεφάλαιο 2: Μεθόδοι πρόβλεψης. Το δεύτερο κεφάλαιο εστιάζει στις μεθόδους πρόβλεψης χρηματοοικονομικών δεδομένων. Αναλύονται οι ποσοτικές τεχνικές που χρησιμοποιούνται για την πρόβλεψη, καθώς και οι μέθοδοι αξιολόγησης της ακρίβειας των προβλέψεων. Επιπλέον, γίνεται εισαγωγή στις χρονοσειρές, τα χαρακτηριστικά τους και τις μεθόδους ανάλυσής τους.

Το Κεφάλαιο 3: Νευρωνικά Δίκτυα Στο τρίτο κεφάλαιο, παρουσιάζονται οι βασικές αρχές και η λειτουργία των νευρωνικών δικτύων. Ιδιαίτερη έμφαση δίνεται στα Επαναλαμβανόμενα Νευρωνικά Δίκτυα (RNN) , στα δίκτυα Long Short-Term Memory (LSTM) και Gated Recurrent Units (GRU), με έμφαση στην εφαρμογή τους σε δεδομένα χρονοσειρών. Επιπλέον, συζητούνται στρατηγικές για τη βελτίωση της απόδοσης των μοντέλων και την αποφυγή του overfitting.

Στο Κεφάλαιο 4: Ανάλυση συναισθήματος. Στο τέταρτο κεφάλαιο, εισάγεται η έννοια της ανάλυσης συναισθήματος και εξετάζονται οι βασικές μεθοδολογίες που χρησιμοποιούνται για την εφαρμογή της. Επεξηγείται ο τρόπος με τον οποίο η ανάλυση συναισθήματος αξιοποιείται στα χρηματοοικονομικά, καθώς και οι προκλήσεις που προκύπτουν από την εφαρμογή της σε οικονομικά δεδομένα.

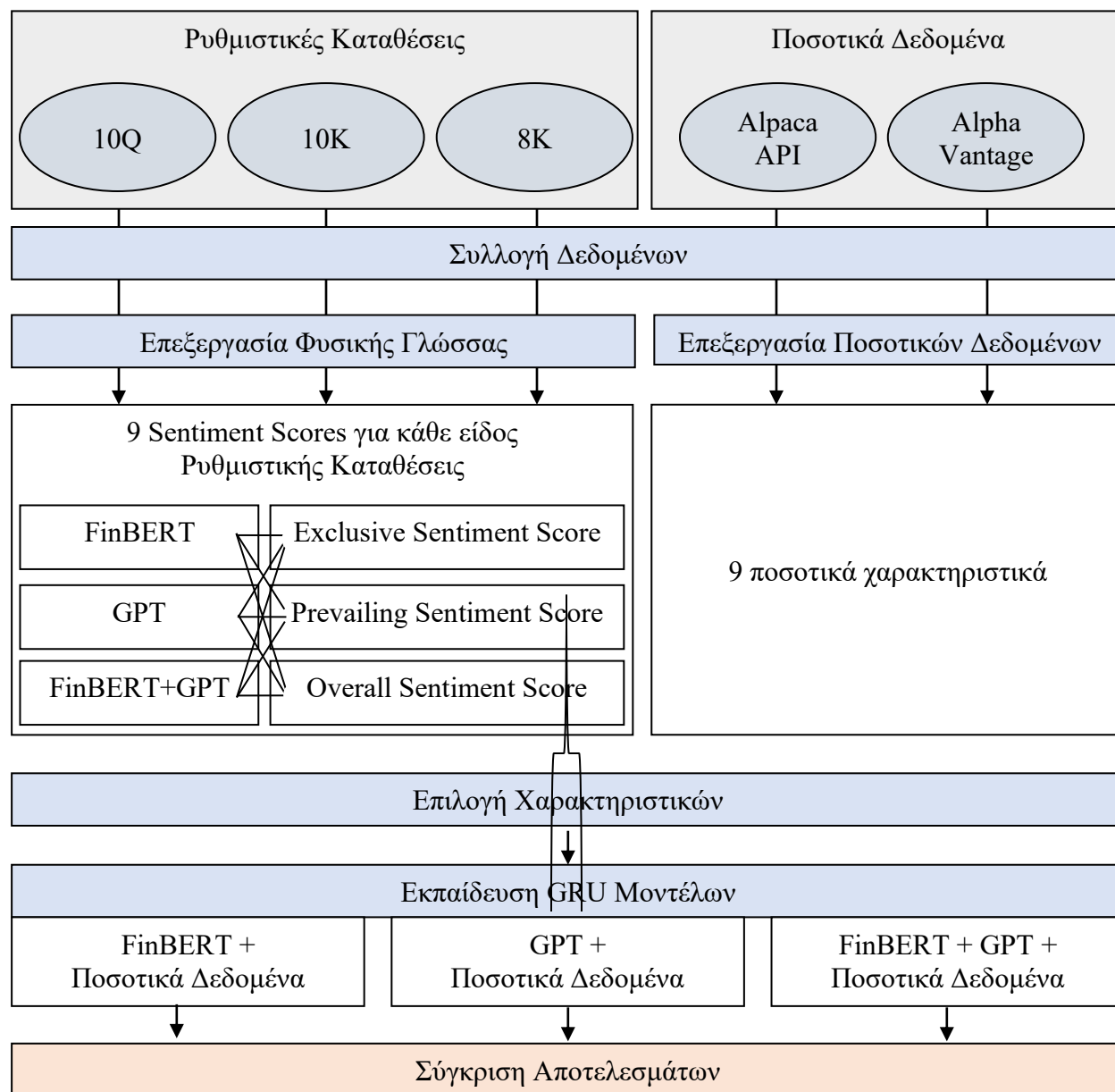
Το Κεφάλαιο 5: Παρουσίαση Δεδομένων και Εξερευνητική Ανάλυση Δεδομένων (EDA). Στο πέμπτο κεφάλαιο, περιγράφονται οι πηγές των δεδομένων και τα API που χρησιμοποιήθηκαν για τη συλλογή τους. Αναλύονται οι διαδικασίες καθαρισμού και προετοιμασίας των δεδομένων για ανάλυση. Παρουσιάζονται γραφήματα και στατιστικά στοιχεία για την κατανόηση των δεδομένων. Επιπλέον, περιγράφεται η μορφή των εταιρικών αναφορών, τα τμήματα που χρησιμοποιούνται για την ανάλυση συναισθήματος και η διαδικασία επεξεργασίας τους.

Στο Κεφάλαιο 6: Πειραματικά Αποτελέσματα. Το έκτο κεφάλαιο επικεντρώνεται στη διαδικασία εκπαίδευσης των μοντέλων, στις μετρικές αξιολόγησης και στις τεχνικές διασταυρούμενης επικύρωσης. Παρουσιάζονται τα αποτελέσματα των μοντέλων, συγκρίνονται με άλλες μεθόδους και αναλύονται τα σφάλματα. Ερμηνεύονται τα αποτελέσματα, γίνεται σύγκριση μεταξύ των εταιρειών και αξιολογείται η επίδραση των χαρακτηριστικών στην απόδοση των προβλέψεων.

Τέλος, το **Κεφάλαιο 7: Συμπεράσματα και Μελλοντική Εργασία** συνοψίζει τα βασικά ευρήματα της έρευνας, αναφέρει τους περιορισμούς που αντιμετωπίστηκαν και προτείνει κατευθύνσεις για μελλοντική έρευνα. Επισημαίνεται η δυνατότητα επέκτασης της μελέτης σε νέες πηγές δεδομένων και πιο σύνθετα μοντέλα.

Η εργασία ολοκληρώνεται με τη **Βιβλιογραφία**, όπου παρατίθενται όλες οι πηγές που χρησιμοποιήθηκαν, ώστε να διασφαλίζεται η ακαδημαϊκή εγκυρότητα και η διαφάνεια της μελέτης.

Σε γενικές γραμμές, η ροή των κυριότερων υπολογιστικών τμημάτων μπορεί να απεικονιστεί ως:



Εικόνα 1: Διάγραμμα ροής διπλωματικής εργασίας

2. Θεωρητικό Υπόβαθρο

2.1. Χρηματοοικονομικά Δεδομένα και Χρονοσειρές

2.1.1. Εισαγωγή στις Χρονοσειρές

Τα χρηματοοικονομικά δεδομένα αποτελούν θεμελιώδη στοιχεία για τη διαδικασία λήψης αποφάσεων στις επιχειρήσεις και στις αγορές, καθώς παρέχουν τόσο ποσοτική όσο και ποιοτική πληροφόρηση σχετικά με την οικονομική κατάσταση και την απόδοση ενός οργανισμού. Αυτά τα δεδομένα περιλαμβάνουν στοιχεία όπως τα έσοδα, τα έξοδα, τα περιουσιακά στοιχεία, οι υποχρεώσεις και οι ταμειακές ροές, προσφέροντας μια ολοκληρωμένη εικόνα της οικονομικής υγείας ενός οργανισμού. Η συλλογή τους πραγματοποιείται είτε περιοδικά (σε καθημερινή, εβδομαδιαία ή μηνιαία βάση) είτε ανάλογα με τις ανάγκες της αγοράς, όπως σε περιπτώσεις σημαντικών οικονομικών συναλλαγών ή γεγονότων. Η ποιότητα αυτών των δεδομένων εξαρτάται από την ακρίβεια, την πληρότητα και την επικαιρότητά τους, καθώς η αξιοπιστία τους είναι κρίσιμη για την υποστήριξη έγκυρων και άμεσων αποφάσεων.

Η σημασία των χρηματοοικονομικών δεδομένων εκτείνεται πέρα από την εσωτερική διαχείριση μιας επιχείρησης. Χρησιμοποιούνται για την ανάλυση τάσεων, την εκτίμηση κινδύνων και τη σύγκριση της απόδοσης μιας εταιρείας σε σχέση με τον ανταγωνισμό και το γενικότερο οικονομικό περιβάλλον. Επιπλέον, διαδραματίζουν καθοριστικό ρόλο στη στρατηγική λήψη αποφάσεων σε τομείς όπως η τιμολόγηση, οι επενδύσεις και η διαχείριση πόρων, καθιστώντας τα ένα αναντικατάστατο εργαλείο για επιχειρήσεις, επενδυτές και ρυθμιστικές αρχές. Η αλληλεπίδραση μεταξύ χρηματοοικονομικών δεδομένων και επενδυτικών στρατηγικών έχει οδηγήσει στην ανάγκη ανάπτυξης μοντέλων πρόβλεψης που επιτρέπουν την καλύτερη κατανόηση των μελλοντικών τάσεων της αγοράς.

Η εφαρμογή προβλέψεων στα χρηματοοικονομικά δεδομένα είναι ζωτικής σημασίας, καθώς επιτρέπει στους επενδυτές, στους διαχειριστές κεφαλαίων και στους οικονομολόγους να λαμβάνουν τεκμηριωμένες αποφάσεις σχετικά με την αγορά, την πώληση ή τη διακράτηση επενδυτικών στοιχείων. Οι προβλέψεις της χρηματιστηριακής αγοράς επιδιώκουν να εντοπίσουν μοτίβα και να εκτιμήσουν πιθανές μελλοντικές διακυμάνσεις στις τιμές των μετοχών, συμβάλλοντας στην αποτελεσματική διαχείριση κινδύνων και στη βελτίωση της αποδοτικότητας των επενδύσεων. Δεδομένου ότι οι αγορές επηρεάζονται από ποικίλους παράγοντες, όπως μακροοικονομικά δεδομένα, ειδήσεις, εταιρικές ανακοινώσεις και συναισθήματα των επενδυτών, η ακριβής πρόβλεψη αποτελεί μία από τις μεγαλύτερες προκλήσεις στον χρηματοοικονομικό τομέα.

Τα θεμελιώδη μεγέθη που θα χρησιμοποιηθούν στην παρούσα εργασία είναι τα ιστορικά δεδομένα της τιμής της μετοχής του εκάστοτε οργανισμού, οι καταθέσεις των αρχείων στο SEC όπως και θεμελιώδη οικονομικά δεδομένα της αγοράς που επηρεάζουν τις αγορές.

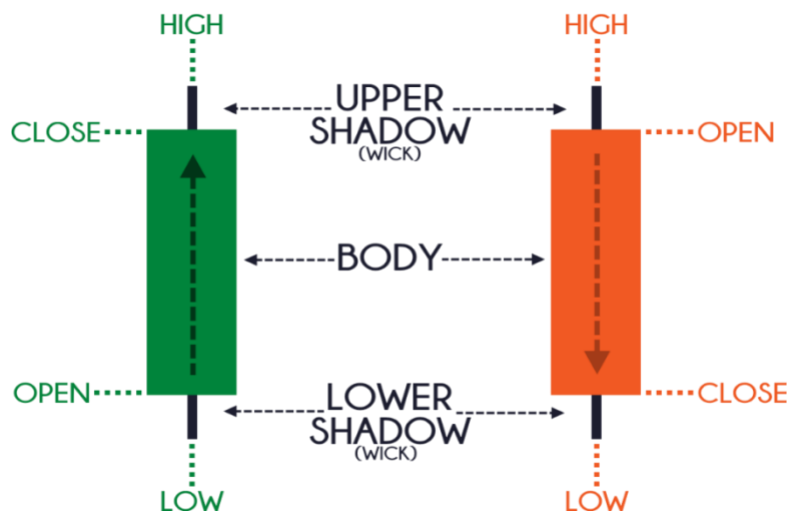
2.1.2. Ιστορικά δεδομένα τιμών

Τα ιστορικά δεδομένα των μετοχών αποτελούν καταγραφές που περιγράφουν τη συμπεριφορά της τιμής και του όγκου συναλλαγών μιας μετοχής σε διαφορετικές χρονικές περιόδους. Αυτά τα δεδομένα περιλαμβάνουν βασικές παραμέτρους, όπως η τιμή ανοίγματος, η μέγιστη και η ελάχιστη τιμή κατά τη διάρκεια της ημέρας, η τιμή κλεισίματος, καθώς και ο συνολικός όγκος συναλλαγών. Τα ιστορικά δεδομένα παρέχουν πολύτιμες πληροφορίες για την απόδοση ενός τίτλου στο παρελθόν, δίνοντας τη δυνατότητα στους επενδυτές να αναλύσουν μακροπρόθεσμες ή βραχυπρόθεσμες τάσεις και να εξάγουν χρήσιμα συμπεράσματα για την πορεία της αγοράς.

Οι παρελθοντικές τιμές μετοχών προέρχονται από εξειδικευμένα χρηματοοικονομικά συστήματα και οργανισμούς που είναι υπεύθυνοι για την καταγραφή και τη διανομή αυτών των δεδομένων. Τα χρηματιστήρια, όπως το Χρηματιστήριο της Νέας Υόρκης (NYSE-New York Stock Exchange) και το NASDAQ, αποτελούν πρωταρχικές πηγές για τη συλλογή των δεδομένων αυτών. Παράλληλα, εξειδικευμένοι πάροχοι δεδομένων, όπως η Bloomberg και η Thomson Reuters, επεξεργάζονται, οργανώνουν και διανέμουν τις τιμές στους χρήστες, συμπεριλαμβανομένων επενδυτών, εταιρειών διαχείρισης κεφαλαίων και ερευνητών. Οι πάροχοι αυτοί διασφαλίζουν ότι τα δεδομένα είναι ακριβή, επικαιροποιημένα και κατάλληλα διαμορφωμένα για αναλύσεις.

Η συλλογή και καταγραφή των δεδομένων γίνεται σε πραγματικό χρόνο κατά τη διάρκεια της διαπραγμάτευσης, ενώ παράλληλα αποθηκεύονται ιστορικές τιμές για μελλοντική χρήση. Αυτά τα δεδομένα περιλαμβάνουν πληροφορίες όπως η τιμή ανοίγματος, η υψηλότερη και η χαμηλότερη τιμή της ημέρας, η τιμή κλεισίματος και ο όγκος συναλλαγών. Σε ορισμένες περιπτώσεις, παρέχονται και προσαρμοσμένες τιμές που λαμβάνουν υπόψη μερίσματα, split μετοχών ή άλλες τροποποιήσεις.

Για την οπτικοποίηση των τιμών σε διαγράμματα, μια από τις πιο δημοφιλείς τεχνικές είναι η χρήση των διαγραμμάτων candlestick (κηροπήγια). Αυτή η μέθοδος παρέχει μια περιεκτική αναπαράσταση της κίνησης της τιμής μιας μετοχής σε συγκεκριμένες χρονικές περιόδους. Κάθε candlestick απεικονίζει την τιμή ανοίγματος, την τιμή κλεισίματος, καθώς και τις μέγιστες και ελάχιστες τιμές, επιτρέποντας στους αναλυτές να εντοπίσουν τάσεις, μοτίβα και αντιστροφές στην αγορά. Επιπλέον, τα candlesticks χρησιμοποιούνται συχνά σε συνδυασμό με τεχνικούς δείκτες για την παροχή πιο εμπεριστατωμένων προβλέψεων.



Εικόνα 2: Οπτική αναπαράσταση Candlestick (Source: Medium)

Ο όγκος συναλλαγών (volume) αποτελεί έναν από τους πιο σημαντικούς δείκτες στην ανάλυση των χρηματοοικονομικών αγορών. Εκφράζει τον συνολικό αριθμό των μετοχών που ανταλλάσσονται κατά τη διάρκεια μιας συγκεκριμένης χρονικής περιόδου, συνήθως ανά ημέρα. Ο όγκος παρέχει κρίσιμες πληροφορίες για τη δυναμική της αγοράς, καθώς αντικατοπτρίζει το επίπεδο ενδιαφέροντος των επενδυτών για μια συγκεκριμένη μετοχή. Υψηλός όγκος συναλλαγών υποδηλώνει συχνά αυξημένη δραστηριότητα και ισχυρό ενδιαφέρον για το εκάστοτε περιουσιακό στοιχείο, ενώ το αντίθετο μπορεί να σημαίνει περιορισμένη συμμετοχή και αβεβαιότητα.

2.1.3. Καταθέσεις στην Επιτροπή Κεφαλαιαγοράς των Ηνωμένων Πολιτειών

Η Επιτροπή Κεφαλαιαγοράς των Ηνωμένων Πολιτειών (SEC - Securities and Exchange Commission) ιδρύθηκε το 1934 μέσω του Securities Exchange Act ως απάντηση στις οικονομικές καταστροφές της Μεγάλης Ύφεσης. Η SEC έχει αναλάβει την ευθύνη να διασφαλίσει την ομαλή λειτουργία των αγορών κεφαλαίου, την προστασία των επενδυτών και τη διατήρηση της εμπιστοσύνης στο χρηματοπιστωτικό σύστημα των Ηνωμένων Πολιτειών. Η αποστολή της συνοψίζεται στη διασφάλιση της πλήρους και δίκαιης δημοσιοποίησης πληροφοριών, την πρόληψη της απάτης και την επιβολή της συμμόρφωσης με τους ομοσπονδιακούς νόμους περί κινητών αξιών [20].

Η SEC εποπτεύει και ρυθμίζει τις δημόσιες εταιρείες, τις χρηματιστηριακές αγορές, τις εταιρείες επενδύσεων και τα αμοιβαία κεφάλαια. Εξασφαλίζει ότι οι συμμετέχοντες στις αγορές παρέχουν διαφανείς και ακριβείς πληροφορίες για την οικονομική τους κατάσταση, τις δραστηριότητές τους και τους κινδύνους που σχετίζονται με τις επενδύσεις τους. Μέσα από τα υποχρεωτικά έγγραφα προς κατάθεση στο αρχείο της, ενισχύει τη λογοδοσία και την αξιοπιστία των δημοσίων εταιρειών [21].

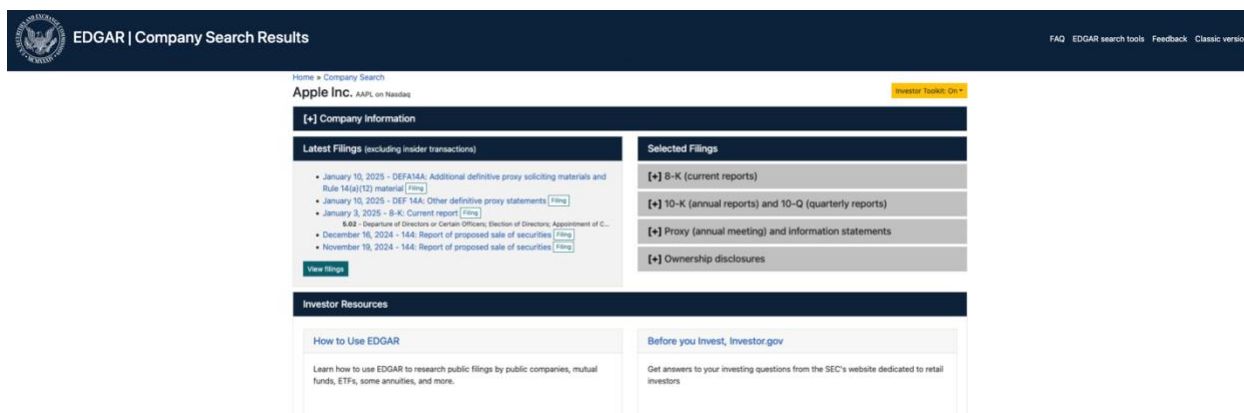
Επιπλέον, η SEC διαδραματίζει ενεργό ρόλο στην προστασία των επενδυτών από παραπλανητικές πρακτικές, όπως η χειραγώγηση της αγοράς ή η διάδοση παραπλανητικών πληροφοριών. Εξοπλισμένη με ισχυρές εξουσίες, μπορεί να διεξάγει έρευνες, να εκδίδει κυρώσεις και να παραπέμπει υποθέσεις στη δικαιοσύνη όταν εντοπίζονται παραβιάσεις [22]. Παράλληλα, επιβλέπει τις δραστηριότητες μεσιτών, εμπόρων και συμβούλων επενδύσεων, διασφαλίζοντας ότι λειτουργούν με τρόπο που εξυπηρετεί τα συμφέροντα των πελατών τους.

Ένας από τους πιο σημαντικούς ρόλους της SEC είναι η προώθηση της καινοτομίας και της εμπιστοσύνης στην αγορά. Μέσα από τη διαχείριση προγραμμάτων όπως ο EDGAR (Electronic Data Gathering, Analysis, and Retrieval system), η SEC διευκολύνει την πρόσβαση του κοινού στις χρηματοοικονομικές πληροφορίες που κατατίθενται από τις δημόσιες εταιρείες. Το EDGAR ενισχύει τη διαφάνεια, επιτρέποντας στους επενδυτές να αξιολογούν ανεξάρτητα τις επενδυτικές επιλογές τους [21].

Η SEC δεν περιορίζεται μόνο στις εγχώριες αγορές αλλά συμμετέχει ενεργά στη διαμόρφωση διεθνών προτύπων ρύθμισης μέσω συνεργασίας με οργανισμούς όπως ο IOSCO (International Organization of Securities Commissions). Έτσι, συνεισφέρει στην παγκόσμια σταθερότητα των χρηματοπιστωτικών αγορών και προωθεί την υιοθέτηση βέλτιστων πρακτικών σε διεθνές επίπεδο [20].

Μέσω της αυστηρής εποπτείας και των πολυδιάστατων πρωτοβουλιών της, η SEC εξασφαλίζει ότι οι χρηματοοικονομικές αγορές των Ηνωμένων Πολιτειών παραμένουν αξιόπιστες, διαφανείς και προσβάσιμες, προστατεύοντας τα συμφέροντα εκατομμυρίων επενδυτών και διατηρώντας την εμπιστοσύνη στο οικονομικό σύστημα [20].

Τα SEC filings είναι έγγραφα που υποβάλλουν δημόσιες εταιρείες και ορισμένοι άλλοι οργανισμοί στην Επιτροπή Κεφαλαιαγοράς των Ηνωμένων Πολιτειών (SEC). Τα έγγραφα αυτά είναι απαραίτητα για τη διαφάνεια και την προστασία των επενδυτών, καθώς παρέχουν λεπτομερείς πληροφορίες σχετικά με την οικονομική κατάσταση, τις δραστηριότητες, τους κινδύνους και τις στρατηγικές των εταιρειών. Οι πιο κοινές κατηγορίες SEC filings περιλαμβάνουν τις 10-K (ετήσιες εκθέσεις), τις 10-Q (τριμηνιαίες εκθέσεις) και τις 8-K (αναφορές σημαντικών γεγονότων) [20] [21].



Εικόνα 3: SEC filings για την APPLE

2.1.3.1. 10-K: Ετήσιες Εκθέσεις

Το αρχείο 10-K είναι μια εκτενής ετήσια έκθεση που καταθέτουν οι δημόσιες εταιρείες στην SEC. Παρέχει πλήρη επισκόπηση της απόδοσης και των δραστηριοτήτων της εταιρείας κατά τη διάρκεια του οικονομικού έτους. Το 10-K πρέπει να κατατίθεται εντός 60 έως 90 ημερών από το τέλος του οικονομικού έτους, ανάλογα με το μέγεθος της εταιρείας.

Περιεχόμενα του 10-K:

1. **Επισκόπηση Επιχειρηματικών Δραστηριοτήτων:** Περιγράφει τη φύση και το εύρος των δραστηριοτήτων της εταιρείας.
2. **Οικονομικές Καταστάσεις:** Αναλυτικές πληροφορίες για τα έσοδα, τα έξοδα, τα περιουσιακά στοιχεία, τις υποχρεώσεις και τις ταμειακές ροές.
3. **Risk Factors:** Περιγράφονται οι σημαντικοί κίνδυνοι που ενδέχεται να επηρεάσουν την επιχείρηση. Αυτή η ενότητα είναι κρίσιμη για την αξιολόγηση των πιθανών κινδύνων που θα μπορούσαν να επηρεάσουν την τιμή της μετοχής.
4. **Management Discussion and Analysis (MD&A):** Αναλύεται η οικονομική απόδοση και οι στρατηγικές της διοίκησης.

5. **Σημαντικές Δικαστικές Διαμάχες:** Αναφέρονται νομικά ζητήματα που ενδέχεται να επηρεάσουν την επιχείρηση.

2.1.3.2. 10-Q: Τριμηνιαίες Εκθέσεις

Το 10-Q είναι μια τριμηνιαία έκθεση που παρέχει ενημερωμένες πληροφορίες για την οικονομική κατάσταση της εταιρείας κατά τη διάρκεια του τριμήνου. Σε αντίθεση με το 10-K, είναι λιγότερο εκτενής, αλλά παραμένει ουσιώδης για τη συνεχή παρακολούθηση των επιδόσεων της εταιρείας. Πρέπει να υποβάλλεται εντός 40 ημερών από το τέλος κάθε τριμήνου, εκτός του τελευταίου, το οποίο καλύπτεται από το 10-K. [20]

Περιεχόμενα του 10-Q:

1. **Οικονομικές Καταστάσεις:** Συνοπτικά οικονομικά δεδομένα για το τρίμηνο.
2. **MD&A:** Ανάλυση της τρέχουσας οικονομικής κατάστασης.
3. **Risk Factors (όπου απαιτείται):** Ενημερώσεις για σημαντικούς κινδύνους που ενδέχεται να έχουν προκύψει από το τελευταίο 10-K.

2.1.3.3. 8-K: Αναφορές Σημαντικών Γεγονότων

Το 8-K υποβάλλεται για να γνωστοποιηθούν σημαντικά γεγονότα που επηρεάζουν την εταιρεία. Δεν έχει τακτική συχνότητα υποβολής, καθώς χρησιμοποιείται για την επικοινωνία έκτακτων και σημαντικών αλλαγών. Η κατάθεση πρέπει να γίνει εντός τεσσάρων εργάσιμων ημερών από τη στιγμή που το γεγονός καθίσταται δημόσιο ή γνωστό στην εταιρεία. [20] [22]

Περιεχόμενα του 8-K:

1. **Press Releases:** Ενημερώσεις για σημαντικές ανακοινώσεις, όπως συγχωνεύσεις, εξαγορές, νέες συνεργασίες ή κυκλοφορία προϊόντων.
2. **Διοικητικές Αλλαγές:** Αναφορές για παραιτήσεις, διορισμούς ή αλλαγές στη διοίκηση.
3. **Εκδόσεις Μετοχών:** Πληροφορίες για νέα έκδοση μετοχών ή αλλαγές στο μετοχικό κεφάλαιο.
4. **Οικονομικές Ανακοινώσεις:** Ενημερώσεις για σημαντικές αλλαγές στις οικονομικές επιδόσεις ή εκτιμήσεις.

2.1.4. Οικονομικό Ημερολόγιο και κέρδη εταιρειών

Το οικονομικό ημερολόγιο αποτελεί ένα απαραίτητο εργαλείο για την παρακολούθηση κρίσιμων οικονομικών γεγονότων, δεικτών και ανακοινώσεων που έχουν τη δυνατότητα να επηρεάσουν τις χρηματοπιστωτικές αγορές. Περιλαμβάνει πληροφορίες για προγραμματισμένες ανακοινώσεις, όπως τις αποφάσεις επιτοκίων από κεντρικές τράπεζες, τα στοιχεία πληθωρισμού, τα επίπεδα ανεργίας και τις εκθέσεις του Ακαθάριστου Εγχώριου Προϊόντος (ΑΕΠ). Αυτά τα γεγονότα, που δημοσιεύονται σε καθορισμένες ημερομηνίες και ώρες, επηρεάζουν σημαντικά τη στρατηγική επενδύσεων, traders και αναλυτών, καθώς προσφέρουν πολύτιμες ενδείξεις για την κατάσταση της παγκόσμιας οικονομίας.

Η χρησιμότητα του οικονομικού ημερολογίου έγκειται στην ικανότητά του να προσφέρει ενόραση σχετικά με τις μελλοντικές κινήσεις της αγοράς. Για παράδειγμα, η ανακοίνωση μιας αύξησης των επιτοκίων μπορεί να ενισχύσει την αξία ενός νομίσματος, ενώ ταυτόχρονα να προκαλέσει πτώση σε αγορές ομολόγων και μετοχών. Παράλληλα, δείκτες όπως ο Δείκτης Τιμών Καταναλωτή (CPI) επηρεάζουν την αγοραστική δύναμη και τις επενδυτικές στρατηγικές. Οι επενδυτές χρησιμοποιούν το οικονομικό ημερολόγιο για να προετοιμάσουν τις θέσεις τους και να προσαρμόσουν τη στρατηγική τους βάσει των αναμενόμενων γεγονότων.

Το οικονομικό ημερολόγιο είναι επίσης ζωτικής σημασίας για τη διαχείριση κινδύνων. Η παρακολούθηση των ημερομηνιών ανακοίνωσης κρίσιμων οικονομικών δεικτών επιτρέπει στους επενδυτές να αποφεύγουν απροσδόκητες αναταράξεις και να εκμεταλλεύονται τη μεταβλητότητα (volatility) που προκαλείται από τα γεγονότα αυτά. Για παράδειγμα, θετικά αποτελέσματα σχετικά με το ΑΕΠ μπορεί να οδηγήσουν σε άνοδο των χρηματιστηρίων, ενώ μια απροσδόκητη αύξηση του πληθωρισμού μπορεί να προκαλέσει πτώση στις αγορές ομολόγων.

Τα κέρδη (earnings) μιας εταιρείας, αποτελούν έναν από τους πιο σημαντικούς δείκτες για την αξιολόγηση της οικονομικής της απόδοσης. Αναφέρονται στα καθαρά κέρδη που καταγράφονται κατά τη διάρκεια μιας συγκεκριμένης χρονικής περιόδου, όπως τριμηνιαίως (quarterly earnings) ή ετησίως (annual earnings). Οι ανακοινώσεις κερδών (earnings reports) περιλαμβάνουν βασικά οικονομικά στοιχεία, όπως τα συνολικά έσοδα, τα καθαρά κέρδη, το κέρδος ανά μετοχή (Earnings Per Share - EPS), καθώς και άλλους δείκτες απόδοσης, όπως το περιθώριο κέρδους. Τα δεδομένα αυτά χρησιμοποιούνται για να μετρήσουν την κερδοφορία μιας εταιρείας και να αξιολογήσουν την ικανότητά της να πετύχει τους στόχους της [23].

Οι ανακοινώσεις των κερδών έχουν σημαντική επίδραση στις χρηματοπιστωτικές αγορές, καθώς επηρεάζουν άμεσα τις προσδοκίες των επενδυτών. Οι επενδυτές και οι αναλυτές συγκρίνουν τα ανακοινωμένα αποτελέσματα με τις προβλέψεις των αναλυτών και τις εκτιμήσεις της αγοράς. Τα αποτελέσματα που ξεπερνούν τις προσδοκίες μπορούν να οδηγήσουν σε άνοδο της τιμής της μετοχής, καθώς ενισχύουν την εμπιστοσύνη στην εταιρεία. Αντίθετα, αρνητικές αποκλίσεις από τις προσδοκίες μπορεί να προκαλέσουν πτώση, καθώς δημιουργούν αβεβαιότητα σχετικά με τη μελλοντική απόδοση της εταιρείας.

Πέρα από τα καθαρά κέρδη, οι ανακοινώσεις κερδών περιλαμβάνουν συχνά προοπτικές για το μέλλον (forward guidance). Η διοίκηση της εταιρείας μπορεί να αναφέρει εκτιμήσεις για τα μελλοντικά έσοδα, τις επενδύσεις ή τις στρατηγικές κινήσεις. Αυτές οι πληροφορίες είναι κρίσιμες για τη διαμόρφωση των προσδοκιών της αγοράς, καθώς παρέχουν ένα πλαίσιο για την αξιολόγηση της ικανότητας της εταιρείας να διαχειρίζεται κινδύνους και να εκμεταλλεύεται ευκαιρίες. Επιπλέον, οι επενδυτές λαμβάνουν υπόψη και το γενικό οικονομικό περιβάλλον, όπως καταγράφεται στο οικονομικό ημερολόγιο, για να αξιολογήσουν τη βιωσιμότητα των προβλέψεων που παρέχονται στις ανακοινώσεις κερδών.

Η διαδικασία ανάλυσης των κερδών περιλαμβάνει επίσης την ερμηνεία δεικτών, όπως οι τάσεις στα έσοδα, η μεταβολή των λειτουργικών εξόδων και η διαχείριση του χρέους. Αυτά τα στοιχεία παρέχουν μια πιο συνολική εικόνα της απόδοσης της εταιρείας και επιτρέπουν στους επενδυτές να αξιολογήσουν εάν η εταιρεία βρίσκεται σε τροχιά ανάπτυξης ή αντιμετωπίζει οικονομικές δυσκολίες. Οι πληροφορίες αυτές είναι κρίσιμες για τη λήψη επενδυτικών αποφάσεων, καθώς προσφέρουν σημαντικά δεδομένα για την πιθανή μελλοντική πορεία της τιμής της μετοχής.

Το οικονομικό ημερολόγιο περιλαμβάνει επίσης ημερομηνίες ανακοινώσεων κερδών από μεγάλες εταιρείες, καθιστώντας το ένα εργαλείο που συνδέει τα μακροοικονομικά δεδομένα με τις μικροοικονομικές επιδόσεις. Οι ανακοινώσεις των earnings επηρεάζονται από εξωτερικούς οικονομικούς παράγοντες, όπως οι αυξήσεις επιτοκίων ή οι μεταβολές στις αγορές εργασίας, που ήδη ανακοινώνονται μέσω του οικονομικού ημερολογίου. Αυτή η σύνδεση επιτρέπει στους επενδυτές να αποκτήσουν μια ολοκληρωμένη εικόνα για τις συνθήκες της αγοράς και να προσαρμόσουν τις στρατηγικές τους ανάλογα.

2.1.5. Εισαγωγή στις Χρονοσειρές

Οι χρονοσειρές αποτελούν ένα από τα πιο κρίσιμα εργαλεία στην ανάλυση δεδομένων, ιδίως όταν πρόκειται για τη μελέτη φαινομένων που εξελίσσονται με την πάροδο του χρόνου. Μια χρονοσειρά είναι μια ακολουθία παρατηρήσεων ή δεδομένων που καταγράφονται σε τακτά χρονικά διαστήματα. Η χρήση τους εκτείνεται σε πολλούς τομείς, όπως τα χρηματοοικονομικά, η μετεωρολογία, η υγεία και οι κοινωνικές επιστήμες, παρέχοντας πολύτιμες πληροφορίες για την κατανόηση και την πρόβλεψη μελλοντικών τάσεων.

Οι χρονοσειρές αποτελούν θεμελιώδες μέσο στην ανάλυση δεδομένων και τη διαμόρφωση προβλέψεων, ιδίως σε πεδία όπως η χρηματοοικονομική ανάλυση, η οικονομία και η επιστήμη των δεδομένων. Η μελέτη των χρονοσειρών επιτρέπει την αναγνώριση επαναλαμβανόμενων προτύπων, εποχικών τάσεων και διαρθρωτικών μεταβολών στα δεδομένα, συμβάλλοντας στην κατανόηση της δυναμικής των συστημάτων. Μέσω της χρήσης στατιστικών και μηχανικών μεθόδων, όπως τα αυτοπαλίνδρομικά μοντέλα (AR), οι κινητοί μέσοι (MA) και οι υβριδικές προσεγγίσεις όπως τα ARIMA και GARCH, οι αναλυτές μπορούν να δημιουργήσουν μοντέλα που περιγράφουν τη συμπεριφορά των χρονοσειρών και να πραγματοποιήσουν αξιόπιστες προβλέψεις.

Η χρησιμότητα των χρονοσειρών επεκτείνεται και στη βελτίωση της διαδικασίας λήψης αποφάσεων, καθώς η πρόβλεψη μελλοντικών τιμών επιτρέπει τη βέλτιστη διαχείριση κινδύνου και την προσαρμογή στρατηγικών. Στις χρηματοπιστωτικές αγορές, για παράδειγμα, η μοντελοποίηση των διακυμάνσεων των τιμών των μετοχών μέσω στοχαστικών διαδικασιών και η ανάλυση της μεταβλητότητας διευκολύνουν την ανάπτυξη στρατηγικών συναλλαγών και διαχείρισης χαρτοφυλακίου. Επιπλέον, ερευνητικά ευρήματα δείχνουν ότι τα σφάλματα πρόβλεψης του ΑΕΠ της Κίνας με χρήση μοντέλων μηχανικής μάθησης είναι κατά μέσο όρο χαμηλότερα σε σύγκριση με αυτά των παραδοσιακών οικονομικών μοντέλων ή των προβλέψεων εμπειρογνομόνων, ιδιαίτερα σε περιόδους οικονομικής σταθερότητας. [24]

Η ανάλυση χρονοσειρών μπορεί να αποκαλύψει πολύτιμες πληροφορίες, όπως τάσεις απόδοσης και ανάπτυξης, βοηθώντας τους οργανισμούς να σχεδιάσουν πιο αποδοτικά την πορεία τους. Επιπλέον, οι χρονοσειρές αποτελούν ένα μέσο για την αξιολόγηση της αποτελεσματικότητας προηγούμενων δράσεων και πολιτικών, καθώς και για την προσαρμογή τους στις μεταβαλλόμενες συνθήκες. Σε έναν κόσμο που διψάει για πληροφορία, τα δεδομένα χρονοσειρών αναδεικνύονται ως ένα από τα πιο πολύτιμα εργαλεία ανάλυσης, προσφέροντας τη βάση για ενημερωμένες και στρατηγικές αποφάσεις που οδηγούν στην ανάπτυξη και την καινοτομία.

2.1.6. Ανάλυση Χρονοσειρών

Η ανάλυση χρονοσειρών αποτελεί ένα κρίσιμο εργαλείο στην ανάλυση δεδομένων που εξελίσσονται χρονικά, επιτρέποντας την εξαγωγή χρήσιμων στατιστικών και χαρακτηριστικών από τα δεδομένα. Οι χρονοσειρές μπορούν να αναφέρονται σε συνεχή αριθμητικά δεδομένα, διακριτά αριθμητικά δεδομένα ή ακόμα και σε συμβολικά δεδομένα, όπως ακολουθίες χαρακτήρων. Η ανάλυση αυτή βρίσκει εφαρμογή σε διάφορους τομείς, όπως η χρηματοοικονομική, η διαχείριση συστημάτων IoT (Internet of Things), και η υγειονομική περίθαλψη, όπου οι οργανισμοί αξιοποιούν τα δεδομένα τους για να βελτιώσουν τις στρατηγικές τους και να προβλέψουν μελλοντικές τάσεις.

Η ανάλυση χρονοσειρών μπορεί να κατηγοριοποιηθεί σε διάφορες μεθόδους και τεχνικές, που επιτρέπουν την κατανόηση και πρόβλεψη των δεδομένων:

Εξερευνητική Ανάλυση Δεδομένων (Exploratory Data Analysis):

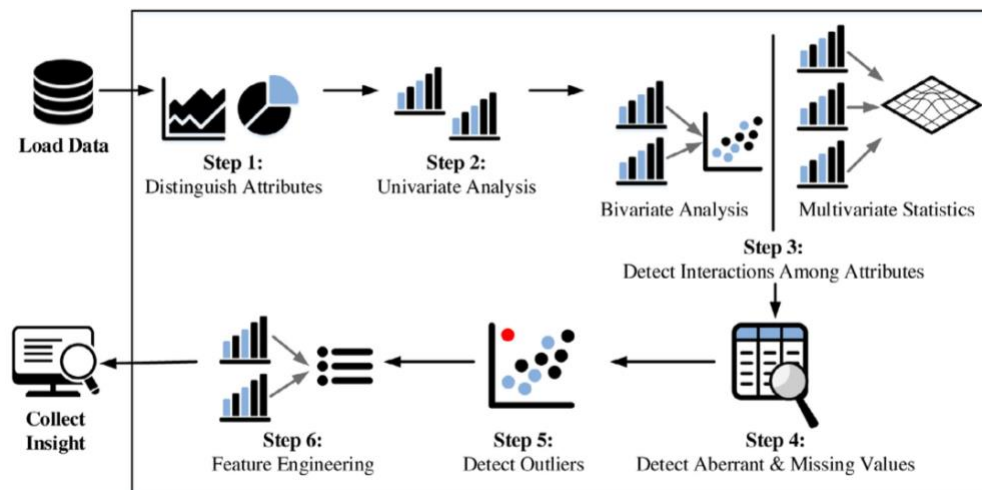
Αυτή η μέθοδος εστιάζει στην ποιοτική κατανόηση της δομής της χρονοσειράς, διαχωρίζοντας την σε στοιχεία όπως η εποχικότητα, η κυκλικότητα, η τάση και οι τυχαίες διακυμάνσεις. Η ανάλυση αυτή βοηθά στην αποσαφήνιση των μοτίβων που παρατηρούνται και των αιτίων που τα προκαλούν, παρέχοντας τη βάση για πιο εξειδικευμένες μελέτες.

Πρόβλεψη (Forecasting):

Η πρόβλεψη αξιοποιεί τις συναρτήσεις που προκύπτουν από την παλινδρόμηση για να εκτιμήσει τη μελλοντική συμπεριφορά μιας χρονοσειράς. Η ικανότητα πρόβλεψης επιτρέπει στους οργανισμούς να σχεδιάζουν στρατηγικές και να λαμβάνουν αποφάσεις βάσει δεδομένων.

Ταξινόμηση (Classification):

Η ταξινόμηση χρονοσειρών δημιουργεί κατηγορίες που περιγράφουν τη συμπεριφορά των δεδομένων, όπως "αύξηση," "κυκλικότητα" ή "σταθερότητα." Αυτή η τεχνική διευκολύνει την ανάλυση και την ερμηνεία των δεδομένων, ενώ παρέχει χρήσιμες ενδείξεις για τη διαχείριση νέων ή άγνωστων δεδομένων.



Εικόνα 4:

Τυπική διαδικασία
EDA.

Πηγή Ghosh et al.
2018, fig. 3.

Η ανάλυση χρονοσειρών έχει ευρύ φάσμα εφαρμογών. Στον χρηματοοικονομικό τομέα, χρησιμοποιείται για την πρόβλεψη τάσεων στις αγορές και τη διαχείριση χαρτοφυλακίων. Στη βιομηχανία IoT (Internet of Things), εφαρμόζεται για την παρακολούθηση συστημάτων σε πραγματικό χρόνο, εξασφαλίζοντας την αποδοτικότητα και την αποφυγή προβλημάτων. Η χρήση της ανάλυσης χρονοσειρών βοηθά επίσης στη μελέτη δημογραφικών αλλαγών, τη διαχείριση πόρων και την πρόβλεψη φυσικών φαινομένων.

2.1.7. Χαρακτηριστικά Χρονοσειρών

Η ανάλυση χρονοσειρών είναι μια μεθοδολογία που αποσκοπεί στην κατανόηση της δομής και της δυναμικής δεδομένων που εξελίσσονται χρονικά. Παρουσιάζονται τα βασικότερα ποιοτικά χαρακτηριστικά που περιγράφουν τη συμπεριφορά των χρονοσειρών, όπως η τάση, η εποχικότητα, η κυκλικότητα, η στασιμότητα, οι τυχαιότητες και οι ασυνέχειες.

Τάση (Trend):

Η τάση είναι το βασικότερο συστατικό μιας χρονοσειράς και περιγράφει τη γενική κατεύθυνση των δεδομένων σε βάθος χρόνου. Αντιπροσωπεύει μακροπρόθεσμες μεταβολές που μπορεί να είναι αυξητικές, μειωτικές ή σταθερές. Για παράδειγμα, οι πωλήσεις μιας επιχείρησης ενδέχεται να παρουσιάζουν ανοδική τάση λόγω αυξανόμενης ζήτησης ή προσαρμογής στις τεχνολογικές εξελίξεις. Η ανάλυση τάσεων είναι κρίσιμη για την κατανόηση της συνολικής δυναμικής των δεδομένων και την πρόβλεψη μελλοντικών τιμών.

Εποχικότητα (Seasonality):

Η εποχικότητα αφορά επαναλαμβανόμενα μοτίβα που συμβαίνουν σε σταθερά χρονικά διαστήματα, όπως μηνιαία, εβδομαδιαία ή ετήσια. Τα χαρακτηριστικά αυτά είναι ιδιαίτερα χρήσιμα για την κατανόηση των κανονικών διακυμάνσεων, όπως η αυξημένη ζήτηση προϊόντων κατά τη διάρκεια των εορτών ή οι μεταβολές στις θερμοκρασίες ανά εποχή. Η ανάλυση εποχικότητας βοηθά στην αναγνώριση των επαναλαμβανόμενων μοτίβων και στη βελτίωση της ακρίβειας προβλέψεων.

Κυκλικότητα (Cyclicity):

Η κυκλικότητα περιγράφει διακυμάνσεις μεγαλύτερης διάρκειας από την εποχικότητα, οι οποίες συνδέονται συνήθως με οικονομικούς ή επιχειρηματικούς κύκλους. Σε αντίθεση με την εποχικότητα, η κυκλικότητα δεν έχει σταθερή διάρκεια ή μέγεθος. Χαρακτηριστικό παράδειγμα κυκλικότητας είναι οι φάσεις ανάπτυξης και ύφεσης σε οικονομικούς κύκλους. Η κατανόηση αυτών των κύκλων είναι απαραίτητη για τη μακροπρόθεσμη ανάλυση και την πρόβλεψη αλλαγών στην οικονομική δραστηριότητα.

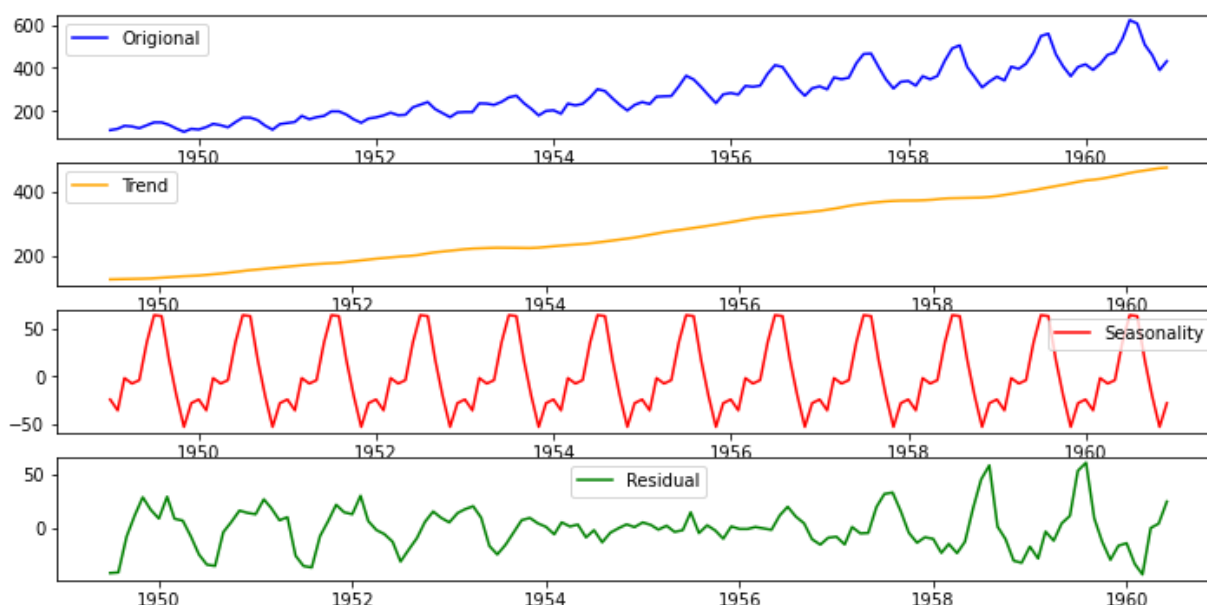
Στασιμότητα (Stationarity):

Η στασιμότητα αναφέρεται στη σταθερότητα των στατιστικών χαρακτηριστικών μιας χρονοσειράς, όπως η μέση τιμή, η διακύμανση και η αυτοσυσχέτιση, σε βάθος χρόνου. Οι στάσιμες χρονοσειρές είναι πιο εύκολες στην ανάλυση και μπορούν να υποβληθούν σε στατιστική επεξεργασία με μεγαλύτερη ακρίβεια. Στην πράξη, πολλές χρονοσειρές δεν είναι

στάσιμες, γεγονός που απαιτεί τεχνικές μετασχηματισμού για την προσαρμογή τους στις αναλυτικές απαιτήσεις.

Τυχαιότητα και Ασυνέχειες (Irregularity and Outliers):

Η τυχαιότητα περιλαμβάνει μη περιοδικές και απρόβλεπτες διακυμάνσεις που δεν εξηγούνται από την τάση, την εποχικότητα ή την κυκλικότητα. Οι ασυνέχειες (outliers) είναι εξαιρετικά ακραίες τιμές που αποκλίνουν σημαντικά από τα υπόλοιπα δεδομένα. Αυτές οι ανωμαλίες μπορούν να προκληθούν από εξωγενείς παράγοντες, όπως φυσικές καταστροφές ή γεωπολιτικά γεγονότα. Ένα ιδιαίτερο παράδειγμα τυχαιότητας είναι το φαινόμενο του random walk, όπου κάθε τιμή μιας χρονοσειράς εξαρτάται από την προηγούμενη συν ένα τυχαίο σφάλμα, καθιστώντας τις μελλοντικές τιμές απρόβλεπτες. Η ανίχνευση και διόρθωση αυτών των στοιχείων είναι απαραίτητη για την ακριβή ανάλυση και πρόβλεψη.



Εικόνα 5: Η παραπάνω εικόνα παρουσιάζει την αποσύνθεση μιας χρονοσειράς στα βασικά της στοιχεία: τάση (trend), εποχικότητα (seasonality) και υπολείμματα (residuals).

- Η αρχική χρονοσειρά (πάνω διάγραμμα) εμφανίζει μια ανοδική τάση, όπως φαίνεται στο δεύτερο διάγραμμα.
- Η εποχικότητα (τρίτο διάγραμμα) δείχνει μια επαναλαμβανόμενη ετήσια διακύμανση, υποδηλώνοντας ότι η χρονοσειρά επηρεάζεται από εποχιακούς παράγοντες.
- Τα υπολείμματα (τελευταίο διάγραμμα) εμφανίζουν τυχαίες διακυμάνσεις, γεγονός που δείχνει ότι η χρονοσειρά περιλαμβάνει και ένα μη προβλέψιμο στοιχείο.

Η αποσύνθεση αυτή είναι ιδιαίτερα χρήσιμη στην ανάλυση χρονοσειρών, καθώς επιτρέπει τον εντοπισμό και τη μεμονωμένη μελέτη των διαφορετικών συστατικών της.

2.1.8. Χρήση Χρονοσειρών στις Χρηματοοικονομικές Αγορές

Οι χρονοσειρές στις χρηματοοικονομικές αγορές χρησιμοποιούνται για την παρακολούθηση και ανάλυση της κίνησης επιλεγμένων δεδομένων, όπως οι τιμές τίτλων, σε συγκεκριμένο χρονικό διάστημα. Τα δεδομένα αυτά καταγράφονται σε τακτά χρονικά διαστήματα, προσφέροντας μια σαφή εικόνα της εξέλιξής τους. Ένα από τα βασικά πλεονεκτήματα των χρονοσειρών είναι ότι δεν υπάρχει συγκεκριμένο όριο στη διάρκεια των δεδομένων που μπορούν να περιληφθούν, επιτρέποντας στους επενδυτές ή αναλυτές να συλλέγουν πληροφορίες με τρόπο που να εξυπηρετεί τους στόχους της ανάλυσής τους. Είτε πρόκειται για βραχυπρόθεσμα δεδομένα, όπως η καταγραφή της τιμής ανά ώρα κατά τη διάρκεια μιας εργάσιμης ημέρας, είτε για μακροπρόθεσμα, όπως η τιμή στο κλείσιμο κάθε μήνα για πέντε χρόνια, οι χρονοσειρές προσαρμόζονται ανάλογα με τις ανάγκες.

Η χρήση χρονοσειρών στις επενδύσεις βοηθά τους αναλυτές να παρακολουθούν τη συμπεριφορά των δεδομένων και να εντοπίζουν μοτίβα, όπως τάσεις, εποχικότητα ή κυκλικότητα. Για παράδειγμα, σε μια αγορά με περιοδικές αυξήσεις πωλήσεων κατά τη διάρκεια συγκεκριμένων περιόδων του έτους, οι χρονοσειρές επιτρέπουν τη μελέτη αυτών των τάσεων και τη χρήση τους για προβλέψεις. Παράλληλα, η δυνατότητα ανάλυσης δεδομένων σε διαφορετικά χρονικά ορίζοντα παρέχει ευελιξία στη δημιουργία στρατηγικών και τη λήψη αποφάσεων.

Οι χρονοσειρές δεν περιορίζονται μόνο στην παρακολούθηση τιμών τίτλων, αλλά μπορούν να περιλαμβάνουν και άλλες χρηματοοικονομικές μεταβλητές, όπως όγκους συναλλαγών, δείκτες αγοράς ή μακροοικονομικούς δείκτες, όπως τα επιτόκια. Η ανάλυσή τους επιτρέπει στους χρηματοοικονομικούς αναλυτές να εξετάζουν τη σχέση μεταξύ διαφόρων παραμέτρων, εντοπίζοντας συνδέσεις και αιτιακές σχέσεις. Αυτές οι πληροφορίες είναι ιδιαίτερα πολύτιμες για τη δημιουργία επενδυτικών στρατηγικών, καθώς προσφέρουν τη δυνατότητα καλύτερης κατανόησης της συμπεριφοράς της αγοράς.

Η χρήση χρονοσειρών δεν περιορίζεται στην ανάλυση του παρελθόντος, αλλά συμβάλλει και στην πρόβλεψη μελλοντικών εξελίξεων. Οι αναλυτές μπορούν να αξιοποιήσουν ιστορικά δεδομένα για τη δημιουργία μοντέλων πρόβλεψης, τα οποία επιτρέπουν την καλύτερη εκτίμηση μελλοντικών τιμών και τάσεων. Σε ένα περιβάλλον όπου η ακρίβεια και η ταχύτητα της πληροφορίας αποτελούν κρίσιμους παράγοντες επιτυχίας, οι χρονοσειρές διαδραματίζουν καθοριστικό ρόλο στη λήψη τεκμηριωμένων αποφάσεων.

2.1.9. Προκλήσεις και Περιορισμοί

Η ανάλυση χρονοσειρών αντιμετωπίζει σημαντικές προκλήσεις, με κυριότερη την παρουσία θορύβου, δηλαδή τυχαίων διακυμάνσεων που μπορούν να συγκαλύψουν τα πραγματικά σήματα των δεδομένων. Η διαχείριση αυτού του θορύβου είναι κρίσιμη, καθώς μπορεί να οδηγήσει σε ανακριβή μοντέλα και εσφαλμένες προβλέψεις. Επιπλέον, η θεωρία του χάους επισημαίνει ότι ορισμένα δυναμικά συστήματα είναι εξαιρετικά ευαίσθητα στις αρχικές συνθήκες, γνωστό ως "φαινόμενο της πεταλούδας", όπου μικρές αλλαγές μπορούν να προκαλέσουν μεγάλες αποκλίσεις στην εξέλιξη του συστήματος. [25] Η παρουσία του φαινομένου αυτού στις χρηματοοικονομικές αγορές αντικατοπτρίζει μια κατάσταση όπου τα ντετερμινιστικά συστήματα, όπως οι κινήσεις της αγοράς, γίνονται εξαιρετικά ευαίσθητα στις αρχικές συνθήκες. Αυτή η ευαισθησία μπορεί να οδηγήσει σε απρόβλεπτα και μη γραμμικά αποτελέσματα, που εκδηλώνονται με σημαντική μεταβλητότητα και αστάθεια. Η θεωρία του χάους στις

χρηματοοικονομικές αγορές υπογραμμίζει ότι, παρόλο που η συμπεριφορά της αγοράς φαίνεται τυχαία, ενδέχεται να εμφανιστούν ορισμένα μοτίβα τάξης μέσα στη φαινομενική αταξία, προσφέροντας δυνατότητες για διερεύνηση των δυναμικών αλληλεπιδράσεων και για τον εντοπισμό υποκείμενων δομών που επηρεάζουν τις τάσεις της αγοράς

2.2. Προβλέψεις

2.2.1. Εισαγωγή στις προβλέψεις

Η πρόβλεψη (forecasting) αποτελεί μία διαδικασία που συνδυάζει επιστημονικές μεθόδους, αναλυτική σκέψη και στατιστική τεχνογνωσία, με στόχο την εκτίμηση μελλοντικών τιμών ενός φαινομένου, βασισμένη σε δεδομένα του παρελθόντος και του παρόντος [26]. Η σημασία της πρόβλεψης αναδεικνύεται στην καθημερινή μας ζωή, από απλές αποφάσεις, όπως η προετοιμασία για τις καιρικές συνθήκες, μέχρι την εφαρμογή της σε πιο σύνθετα περιβάλλοντα, όπως οι χρηματοοικονομικές αγορές, η διαχείριση ενεργειακών πόρων και η χάραξη δημόσιων πολιτικών [27].

Στον πυρήνα της πρόβλεψης βρίσκονται η αβεβαιότητα και ο κίνδυνος, τα οποία διαδραματίζουν καθοριστικό ρόλο στη διαδικασία λήψης αποφάσεων. Οι ειδικοί υποστηρίζουν ότι η πρόβλεψη δεν πρέπει να εκλαμβάνεται ως υποκατάστατο της προφητείας, αλλά ως μία επιστημονική διαδικασία που ενέχει σφάλματα και περιορισμούς [28]. Η ορθότητα των προβλέψεων εξαρτάται από παράγοντες όπως η ποιότητα και η επικαιρότητα των δεδομένων, καθώς και η φύση του φαινομένου που εξετάζεται. Παρόλο που τα μοντέλα πρόβλεψης ενδέχεται να παρουσιάζουν αδυναμίες, παραμένουν ένα από τα πιο σημαντικά εργαλεία για τη διαχείριση της αβεβαιότητας και τη βελτιστοποίηση της λήψης αποφάσεων [29].

Η πρόβλεψη βρίσκει εφαρμογή σε πολλούς τομείς της επιστήμης και της ανθρωπογενούς δραστηριότητας. Στον χρηματοοικονομικό τομέα, οι προβλέψεις των τιμών μετοχών και δεικτών αποτελούν σημαντικό εργαλείο για επενδυτές και αναλυτές. Η ικανότητα εκτίμησης μελλοντικών τάσεων μπορεί να καθορίσει τη στρατηγική διαχείρισης χαρτοφυλακίων και να προσφέρει ανταγωνιστικό πλεονέκτημα στις αγορές [27]. Παράλληλα, στον τομέα της ενέργειας, η πρόβλεψη της ζήτησης παίζει κρίσιμο ρόλο στη διασφάλιση επάρκειας και στη μείωση των εκπομπών αερίων θερμοκηπίου μέσω της βελτιστοποίησης της κατανάλωσης [30].

Στην κοινωνική σφαίρα, η πρόβλεψη της δημογραφικής εξέλιξης και της εξάπλωσης επιδημιών είναι ζωτικής σημασίας για τη χάραξη δημόσιων πολιτικών, όπως η διαχείριση του ασφαλιστικού συστήματος ή η πρόληψη και αντιμετώπιση πανδημιών [26]. Επιπλέον, η σωστή εκτίμηση της τουριστικής κίνησης συμβάλλει στην οικονομική ανάπτυξη χωρών με υψηλή εξάρτηση από τον τουρισμό, όπως η Ελλάδα. Αντίστοιχα, στις μεταφορές, οι προβλέψεις κυκλοφοριακού φόρτου βοηθούν στον σχεδιασμό έργων υποδομής και στη βελτίωση της ασφάλειας στους δρόμους [27].

Παρά τη σημασία της, η διαδικασία πρόβλεψης συνοδεύεται από πολλές προκλήσεις. Ένα από τα μεγαλύτερα εμπόδια είναι η ποιότητα των δεδομένων που χρησιμοποιούνται. Η ύπαρξη επικαιροποιημένων, ακριβών και αξιόπιστων δεδομένων είναι βασική προϋπόθεση για τη δημιουργία αξιόπιστων προβλέψεων. Ωστόσο, συχνά τα δεδομένα δεν είναι διαθέσιμα ή περιέχουν θόρυβο και ανακρίβειες [28]. Επιπλέον, η ίδια η φύση των φαινομένων που

μελετώνται ενδέχεται να είναι επηρεασμένη από απρόβλεπτους εξωτερικούς παράγοντες, όπως νομοθετικές αλλαγές, φυσικές καταστροφές ή γεωπολιτικές εξελίξεις.

Ένα άλλο ζήτημα είναι η επιλογή της κατάλληλης μεθόδου πρόβλεψης. Οι ποιοτικές μέθοδοι, όπως η γνώμη ειδικών, είναι χρήσιμες όταν τα δεδομένα είναι περιορισμένα, αλλά συχνά ενέχουν υποκειμενικότητα. Από την άλλη, οι ποσοτικές μέθοδοι, όπως τα στατιστικά μοντέλα και τα νευρωνικά δίκτυα, απαιτούν μεγάλα σύνολα δεδομένων και ενδέχεται να μην αποδίδουν καλά όταν η φύση του φαινομένου παρουσιάζει μεγάλες μεταβολές [29].

Η ακρίβεια των προβλέψεων επηρεάζεται σημαντικά από εξωτερικούς παράγοντες, όπως οικονομικές συνθήκες, πολιτικές εξελίξεις και κοινωνικές τάσεις [28]. Για παράδειγμα, αλλαγές στη νομοθεσία, φυσικές καταστροφές ή γεωπολιτικές κρίσεις μπορούν να επηρεάσουν σημαντικά την εξέλιξη ενός φαινομένου, καθιστώντας δύσκολη τη σταθερότητα των προβλέψεων. Η ενσωμάτωση αυτών των παραγόντων σε μοντέλα πρόβλεψης αποτελεί πρόκληση, αλλά και αναγκαιότητα για την αύξηση της αξιοπιστίας.

Η χρήση εξωτερικών παραμέτρων όπως η πολιτική σταθερότητα, οι περιβαλλοντικές αλλαγές ή οι κοινωνικές δυναμικές μπορεί να προσφέρει μεγαλύτερη ακρίβεια σε οικονομικά μοντέλα πρόβλεψης, ιδιαίτερα σε αγορές με υψηλή μεταβλητότητα. Η κατανόηση της διασύνδεσης αυτών των παραγόντων με το υπό εξέταση φαινόμενο επιτρέπει τη δημιουργία πιο ευέλικτων και ακριβών μοντέλων.

Η πρόβλεψη διαφέρει από τον σχεδιασμό, αν και οι δύο διαδικασίες είναι στενά συνδεδεμένες. Ενώ η πρόβλεψη επικεντρώνεται στην εκτίμηση του μελλοντικού περιβάλλοντος, ο σχεδιασμός αφορά τη διαμόρφωση στρατηγικών που θα οδηγήσουν στην επιθυμητή κατάσταση [30]. Ο συνδυασμός των δύο εργαλείων είναι απαραίτητος για την αποτελεσματική λήψη αποφάσεων, ειδικά σε ένα διαρκώς μεταβαλλόμενο περιβάλλον όπου η αβεβαιότητα αποτελεί σταθερό παράγοντα.

2.2.2. Διαδικασία Προβλέψης

Η διαδικασία πρόβλεψης αποτελεί μια συστηματική προσέγγιση που περιλαμβάνει μια σειρά από βήματα, τα οποία εφαρμόζονται διαδοχικά για τη διαμόρφωση ακριβών εκτιμήσεων για το μέλλον. Η προσέγγιση αυτή έχει περιγραφεί εκτενώς στη βιβλιογραφία, όπως από τους Makridakis, Wheelwright και Hyndman (1998), και ακολουθείται ευρέως σε πολλούς επιστημονικούς και επαγγελματικούς τομείς.

Το πρώτο βήμα στη διαδικασία είναι ο σαφής καθορισμός του προβλήματος. Στο στάδιο αυτό, είναι απαραίτητο να καθοριστεί ποιο είναι το φαινόμενο που πρέπει να προβλεφθεί, ποια είναι τα δεδομένα που θα χρησιμοποιηθούν και ποιος είναι ο χρονικός ορίζοντας της πρόβλεψης. Για παράδειγμα, σε μια χρηματοοικονομική ανάλυση, μπορεί να χρειάζεται να προβλεφθούν οι τιμές μετοχών για τον επόμενο μήνα ή το επόμενο έτος. Ο καθορισμός του προβλήματος παρέχει μια σαφή κατεύθυνση στη διαδικασία και διασφαλίζει ότι τα υπόλοιπα βήματα είναι κατάλληλα σχεδιασμένα για την αντιμετώπιση του συγκεκριμένου ζητήματος.

Ακολουθεί η συλλογή των απαραίτητων δεδομένων, που αποτελεί ένα από τα πιο κρίσιμα στάδια της διαδικασίας. Τα δεδομένα πρέπει να είναι αξιόπιστα, επικαιροποιημένα και συναφή με το υπό μελέτη πρόβλημα. Αυτά μπορεί να προέρχονται από διαφορετικές πηγές, όπως

ιστορικές βάσεις δεδομένων, στατιστικές καταγραφές ή αισθητήρες. Παράλληλα, πρέπει να λαμβάνονται υπόψη οι γνώσεις και οι εκτιμήσεις ειδικών, καθώς η εμπειρία τους μπορεί να παρέχει κρίσιμη πληροφόρηση που δεν είναι πάντα εμφανής στα αριθμητικά δεδομένα.

Η ανάλυση και η προετοιμασία των δεδομένων είναι το επόμενο στάδιο. Εδώ, τα δεδομένα καθαρίζονται από θόρυβο, ελλείψεις και ανωμαλίες, ενώ πραγματοποιούνται οι απαραίτητοι μετασχηματισμοί για να προσαρμοστούν στις απαιτήσεις του μοντέλου πρόβλεψης. Για παράδειγμα, αν πρόκειται για χρονοσειρές, γίνεται έλεγχος σταθμευτικότητας και, αν απαιτείται, αφαιρείται η εποχικότητα. Αυτό το στάδιο είναι κρίσιμο, καθώς διασφαλίζει ότι τα δεδομένα που εισάγονται στο μοντέλο είναι κατάλληλα για ανάλυση.

Η επιλογή του κατάλληλου μοντέλου πρόβλεψης αποτελεί το επόμενο στάδιο. Η επιλογή αυτή βασίζεται στα χαρακτηριστικά των δεδομένων και στο πρόβλημα που πρέπει να λυθεί. Στο στάδιο αυτό, το μοντέλο προσαρμόζεται στα δεδομένα μέσω διαδικασιών εκπαίδευσης.

Μετά την εφαρμογή του μοντέλου, ακολουθεί η αξιολόγηση της απόδοσής του. Αυτό γίνεται μέσω μετρικών ακρίβειας, όπως το MSE (Mean Squared Error) και το RMSE (Root Mean Squared Error). Οι μετρικές αυτές παρέχουν μια αντικειμενική μέτρηση του βαθμού ακρίβειας της πρόβλεψης και μπορούν να χρησιμοποιηθούν για τη σύγκριση διαφορετικών μοντέλων. Εάν το μοντέλο δεν ικανοποιεί τις απαιτήσεις, μπορεί να τροποποιηθεί ή να αντικατασταθεί από ένα άλλο.

Τέλος, τα αποτελέσματα παρουσιάζονται στους ενδιαφερόμενους φορείς με τρόπο που να είναι κατανοητός και χρήσιμος. Η παρουσίαση αυτή μπορεί να περιλαμβάνει διαγράμματα, πίνακες και στατιστικές αναλύσεις, ενώ συνοδεύεται από εκτιμήσεις για την αβεβαιότητα της πρόβλεψης. Είναι σημαντικό να επισημανθούν οι περιορισμοί του μοντέλου, ώστε οι τελικές αποφάσεις να λαμβάνονται με ρεαλιστική αντίληψη για τις δυνατότητες και τους περιορισμούς της πρόβλεψης.

Συνοψίζοντας, η διαδικασία πρόβλεψης είναι μια πολυδιάστατη διαδικασία που απαιτεί συνδυασμό τεχνικών γνώσεων, αναλυτικών δεξιοτήτων και στρατηγικής σκέψης. Η συστηματική εφαρμογή των παραπάνω βημάτων συμβάλλει στη διαμόρφωση αξιόπιστων προβλέψεων, οι οποίες μπορούν να χρησιμοποιηθούν για τη λήψη τεκμηριωμένων αποφάσεων.

2.2.3. Χρονικός ορίζοντας προβλέψεων

Η πρόβλεψη είναι ένα βασικό εργαλείο που χρησιμοποιείται σε διάφορους τομείς, επιτρέποντας τη διαμόρφωση στρατηγικών αποφάσεων με βάση την εκτίμηση μελλοντικών τάσεων. Οι μέθοδοι που εφαρμόζονται για τον σκοπό αυτό διακρίνονται κυρίως σε ποιοτικές και ποσοτικές, με καθεμία να παρουσιάζει διαφορετικά πλεονεκτήματα και περιορισμούς. Πριν εξεταστούν οι βασικές ποιοτικές μέθοδοι, είναι σημαντικό να αναλυθεί ο χρονικός ορίζοντας της πρόβλεψης, καθώς αυτός επηρεάζει τη μεθοδολογία που θα χρησιμοποιηθεί.

Ο χρονικός ορίζοντας των προβλέψεων αναφέρεται στο διάστημα που καλύπτει η πρόβλεψη και μπορεί να διακριθεί σε βραχυπρόθεσμες, μεσοπρόθεσμες και μακροπρόθεσμες προβλέψεις. Οι βραχυπρόθεσμες προβλέψεις αφορούν χρονικά διαστήματα που εκτείνονται από μερικές ημέρες έως λίγες εβδομάδες. Χρησιμοποιούνται κυρίως για τη βελτιστοποίηση καθημερινών διαδικασιών, όπως ο προγραμματισμός αποθεμάτων, η διαχείριση της παραγωγής και οι

βραχυπρόθεσμες επενδυτικές αποφάσεις. Σε τέτοιες περιπτώσεις, οι διαθέσιμες πληροφορίες είναι συχνά πλούσιες, και η έμφαση δίνεται στην ακριβή ανάλυση ιστορικών δεδομένων.

Οι μεσοπρόθεσμες προβλέψεις καλύπτουν περιόδους που κυμαίνονται από μερικούς μήνες έως περίπου ένα έτος. Χρησιμοποιούνται για την υποστήριξη τακτικών αποφάσεων, όπως η ανάλυση αγοραστικής ζήτησης, ο σχεδιασμός παραγωγής και η ανάπτυξη νέων προϊόντων. Σε αυτό το πλαίσιο, είναι συχνή η χρήση συνδυαστικών μεθόδων, όπου οι ποιοτικές και ποσοτικές τεχνικές αλληλοσυμπληρώνονται, ιδιαίτερα σε περιπτώσεις όπου η δυναμική της αγοράς επηρεάζεται από παράγοντες που δεν αποτυπώνονται εύκολα στα ιστορικά δεδομένα.

Οι μακροπρόθεσμες προβλέψεις, από την άλλη, αφορούν χρονικά διαστήματα που εκτείνονται σε χρόνια ή ακόμη και δεκαετίες. Αυτές είναι απαραίτητες για στρατηγικό σχεδιασμό, επενδύσεις μεγάλης κλίμακας, πολιτικές αποφάσεις και την κατανόηση μακροοικονομικών τάσεων. Δεδομένης της αβεβαιότητας που χαρακτηρίζει μεγάλες χρονικές περιόδους, οι ποιοτικές μέθοδοι αποκτούν ιδιαίτερη σημασία, καθώς επιτρέπουν την ενσωμάτωση της εμπειρίας και της κρίσης ειδικών σε καταστάσεις όπου τα διαθέσιμα δεδομένα είναι ανεπαρκή ή μη αντιπροσωπευτικά.

2.2.4. Μέθοδοι προβλέψεων

Οι ποιοτικές μέθοδοι βασίζονται σε υποκειμενικές εκτιμήσεις, κρίσεις ειδικών και εμπειρικές γνώσεις, καθιστώντας τις ιδιαίτερα χρήσιμες όταν τα ιστορικά δεδομένα είναι ελλιπή ή όταν το υπό εξέταση φαινόμενο είναι καινούριο ή μοναδικό. Ποιοτικές μέθοδοι, όπως η μέθοδος Delphi, η έρευνα αγοράς και η αναλογία ιστορικού κύκλου ζωής, είναι από τις πιο διαδεδομένες τεχνικές πρόβλεψης όταν τα ιστορικά δεδομένα είναι περιορισμένα ή το υπό εξέταση φαινόμενο είναι καινοτόμο. Η μέθοδος Delphi χρησιμοποιείται κυρίως για τη συλλογή και σύνθεση απόψεων ειδικών μέσω διαδοχικών γύρων ερωτήσεων, προσφέροντας μια συστηματική διαδικασία λήψης αποφάσεων. Η έρευνα αγοράς επιτρέπει την ανάλυση προτιμήσεων και αγοραστικών τάσεων μέσα από δεδομένα που προέρχονται απευθείας από τους καταναλωτές, ενώ η αναλογία ιστορικού κύκλου ζωής [26] βοηθά στην πρόβλεψη της πορείας ενός προϊόντος ή μιας τεχνολογίας, συγκρίνοντάς το με παρόμοιες περιπτώσεις του παρελθόντος.

Σε αντίθεση με τις ποσοτικές τεχνικές, οι οποίες στηρίζονται σε μαθηματικά μοντέλα και στατιστικές αναλύσεις, οι ποιοτικές μέθοδοι επιτρέπουν την εκτίμηση μελλοντικών τάσεων μέσω της ανθρώπινης κρίσης. Ωστόσο, λόγω της εγγενούς υποκειμενικότητάς τους, είναι επιρρεπείς σε προκαταλήψεις και ενδέχεται να παρουσιάζουν ασυνέπεια στα αποτελέσματά τους.

2.2.4.1. Ποσοτικές Μέθοδοι Πρόβλεψης

Οι ποσοτικές μέθοδοι αποτελούν ένα ισχυρό εργαλείο ανάλυσης και πρόβλεψης, καθώς βασίζονται στη χρήση μαθηματικών και στατιστικών μοντέλων για την εκτίμηση μελλοντικών τιμών, αξιοποιώντας ιστορικά δεδομένα. Η αποτελεσματικότητά τους εξαρτάται από την ύπαρξη επαρκούς όγκου αριθμητικών δεδομένων και από την υπόθεση ότι τα πρότυπα που έχουν παρατηρηθεί στο παρελθόν θα συνεχιστούν στο μέλλον. Οι μέθοδοι αυτές χρησιμοποιούνται ευρέως σε διάφορους τομείς, όπως η οικονομία, τα χρηματοοικονομικά, η παραγωγή και η διαχείριση εφοδιαστικής αλυσίδας, επιτρέποντας την ανάπτυξη στρατηγικών βασισμένων σε αντικειμενικές μετρήσεις.

Μια από τις πιο απλές ποσοτικές μεθόδους είναι η naïve μέθοδος πρόβλεψης, η οποία υποθέτει ότι η καλύτερη εκτίμηση για την επόμενη χρονική περίοδο είναι η τελευταία διαθέσιμη τιμή. Συμβολικά, αν η τιμή μιας χρονοσειράς στο χρονικό σημείο t είναι X_t , τότε η πρόβλεψη για την επόμενη περίοδο είναι:

$$\widehat{X_{t+1}} = X_t$$

Η προσέγγιση αυτή είναι ιδιαίτερα χρήσιμη όταν δεν υπάρχουν σαφείς τάσεις ή εποχικότητα στα δεδομένα, αλλά δεν είναι κατάλληλη για πιο πολύπλοκες χρονοσειρές.

Για πιο προηγμένες προβλέψεις, χρησιμοποιούνται στατιστικά μοντέλα όπως το ARIMA (AutoRegressive Integrated Moving Average), το οποίο συνδυάζει τρεις βασικές συνιστώσες:

- Αυτοπαλίνδρομο μέρος (AR): Οι προβλέψεις βασίζονται σε παρελθοντικές τιμές της ίδιας μεταβλητής.
- Ολοκλήρωση (I): Η χρονοσειρά μετατρέπεται σε στάσιμη μέσω διαφορών.
- Μέσος κινητός όρος (MA): Η πρόβλεψη βασίζεται σε παρελθοντικά σφάλματα πρόβλεψης.

Το γενικό μοντέλο ARIMA(p,d,q) περιγράφεται από τη σχέση:

$$\left(1 - \sum_{i=1}^p \phi_i L^i\right) (1 - L)^d X_t = \left(1 + \sum_{j=1}^q \theta_j L^j\right) \varepsilon_t$$

όπου:

- L είναι ο τελεστής χρονικής υστέρησης,
- p η τάξη του αυτοπαλίνδρου μέρους,
- d ο αριθμός διαφορών για τη μετατροπή της χρονοσειράς σε στάσιμη,
- q η τάξη του μέσου κινητού όρου,
- ϕ_i και θ_j είναι οι συντελεστές του μοντέλου,
- ε_t είναι ένας όρος λευκού θορύβου.

Το ARIMA χρησιμοποιείται ευρέως σε οικονομικές και χρηματοοικονομικές εφαρμογές, καθώς επιτρέπει την πρόβλεψη τιμών με βάση την εσωτερική δυναμική των δεδομένων. Η επιλογή των παραμέτρων p, d, q γίνεται συχνά με αυτοματοποιημένες διαδικασίες, όπως το κριτήριο πληροφοριών Akaike (AIC) ή το κριτήριο πληροφοριών Bayes (BIC), διασφαλίζοντας τη βέλτιστη προσαρμογή του μοντέλου στα διαθέσιμα δεδομένα.

Για την ανάλυση της μεταβλητότητας στις χρηματοοικονομικές χρονοσειρές, συχνά χρησιμοποιούνται τα μοντέλα GARCH (Generalized Autoregressive Conditional Heteroskedasticity), τα οποία επιτρέπουν τη μοντελοποίηση της ετεροσκεδαστικότητας, δηλαδή της χρονικά μεταβαλλόμενης διακύμανσης μιας σειράς δεδομένων. Το βασικό μοντέλο GARCH(p,q) περιγράφεται από τις εξισώσεις:

$$r_t = \mu + \varepsilon_t, \quad \varepsilon_t = \sigma_t z_t, \quad z_t \sim N(0,1)$$

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^p \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^q \beta_j \sigma_{t-j}^2$$

όπου:

- r_t είναι η απόδοση της χρονοσειράς στο χρόνο t ,
- μ είναι η μέση απόδοση,
- ε_t είναι το σφάλμα πρόβλεψης, το οποίο ακολουθεί κανονική κατανομή με μηδενική μέση τιμή και διακύμανση σ_t^2 ,
- σ_t^2 είναι η υπό όρους διακύμανση της απόδοσης,
- α_0 είναι ένας σταθερός όρος,
- α_i είναι οι συντελεστές των παρελθοντικών τετραγώνων σφαλμάτων,
- β_j είναι οι συντελεστές των παρελθοντικών διακυμάνσεων.

Το GARCH αποτελεί επέκταση του βασικού μοντέλου ARCH (Autoregressive Conditional Heteroskedasticity), το οποίο προτάθηκε από τον Engle (1982) και στη συνέχεια γενικεύτηκε από τον Bollerslev (1986). Σε αντίθεση με το ARIMA, το οποίο επικεντρώνεται στη μοντελοποίηση των τιμών, το GARCH είναι σχεδιασμένο για την πρόβλεψη της διακύμανσης, επιτρέποντας την ανάλυση της αστάθειας στις χρηματοπιστωτικές αγορές.

Η επιλογή των βέλτιστων παραμέτρων ppp και qqq γίνεται συχνά μέσω κριτηρίων επιλογής μοντέλων, όπως το Akaike Information Criterion (AIC) και το Bayesian Information Criterion (BIC), προκειμένου να εξασφαλιστεί η κατάλληλη προσαρμογή του μοντέλου στα δεδομένα. Επεκτάσεις του βασικού GARCH, όπως το EGARCH (Exponential GARCH) και το TGARCH (Threshold GARCH), χρησιμοποιούνται για να μοντελοποιήσουν επιπλέον χαρακτηριστικά της αγοράς, όπως η ασυμμετρία της μεταβλητότητας (volatility clustering).

2.2.4.2. Χρήση Μηχανικής Μάθησης στις Προβλέψεις

Η αυξανόμενη διαθεσιμότητα δεδομένων και η βελτίωση των υπολογιστικών δυνατοτήτων έχουν καταστήσει δυνατή την εφαρμογή προηγμένων αλγορίθμων πρόβλεψης σε ένα ευρύ φάσμα τομέων. Οι παραδοσιακές στατιστικές μέθοδοι, όπως τα αυτοπαλίνδρομα μοντέλα και οι κινητοί μέσοι, εξακολουθούν να διαδραματίζουν κεντρικό ρόλο στην ανάλυση χρονοσειρών, παρέχοντας ερμηνεύσιμες προβλέψεις και στατιστικά τεκμηριωμένες εκτιμήσεις. Ωστόσο, η ενσωμάτωση τεχνικών μηχανικής μάθησης έχει ανοίξει νέες προοπτικές, επιτρέποντας τη μοντελοποίηση πιο περίπλοκων σχέσεων στα δεδομένα και τη βελτίωση της απόδοσης των προβλέψεων.

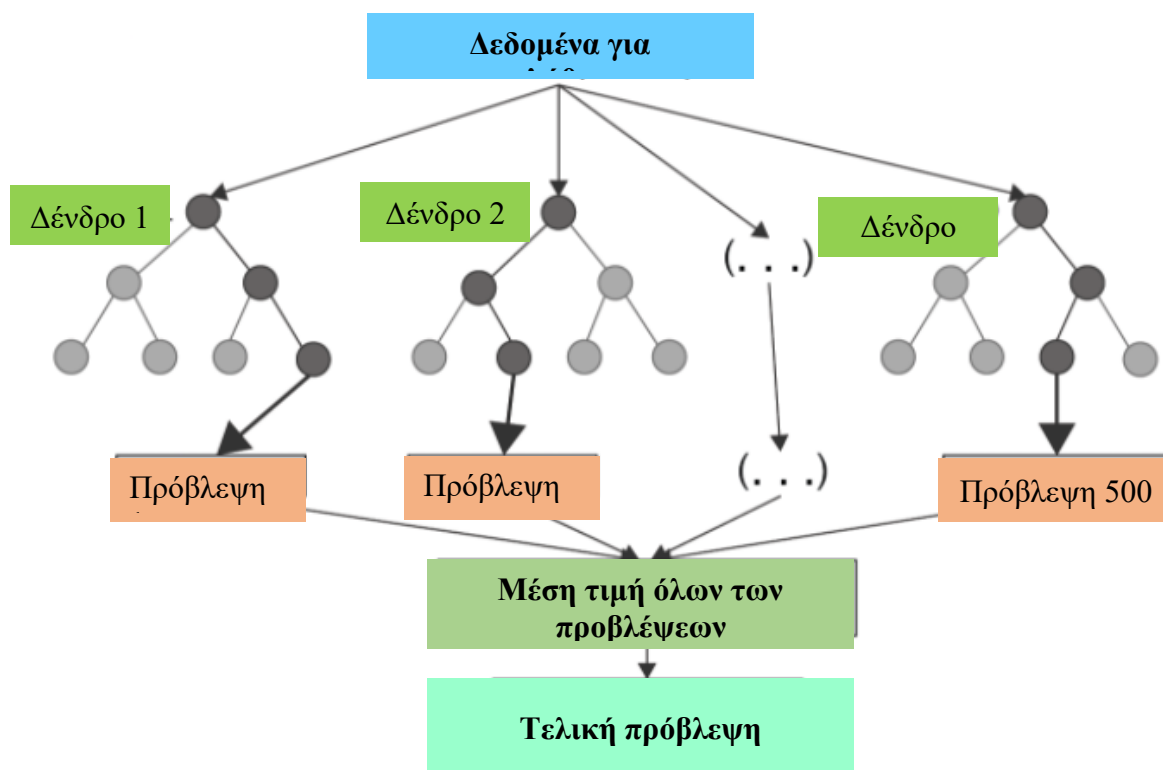
Η μηχανική μάθηση έχει αναδειχθεί ως ισχυρό εργαλείο στην πρόβλεψη, ιδιαίτερα σε δεδομένα χρονοσειρών. Αλγόριθμοι όπως τα Νευρωνικά Δίκτυα Μακράς Βραχύχρονης Μνήμης (LSTM) και οι Γκαουσιανές Διαδικασίες (GPs- Gaussian Processes) έχουν χρησιμοποιηθεί με επιτυχία σε προβλήματα πρόβλεψης, επιτυγχάνοντας υψηλή ακρίβεια [31]. Επιπλέον, η ενσωμάτωση ανάλυσης συναισθημάτων από κοινωνικά δίκτυα, όπως το Twitter, σε συνδυασμό με τεχνικές

βαθιάς μάθησης, έχει δείξει υποσχόμενα αποτελέσματα στην πρόβλεψη τιμών κρυπτονομισμάτων [32].

Δύο από τις πλέον διαδεδομένες τεχνικές στον χώρο αυτό είναι τα Τυχαία Δάση (Random Forests) και η Ενίσχυση Κλίσης (XGBoosting), οι οποίες βασίζονται στη δημιουργία και συνδυασμό πολλαπλών δέντρων απόφασης για τη βελτίωση της ακρίβειας των προβλέψεων.

Τα Random Forests αποτελεί μία από τις πλέον ισχυρές τεχνικές επιβλεπόμενης μάθησης και ανήκει στην κατηγορία των ensemble learning models. Η βασική του αρχή στηρίζεται στη δημιουργία ενός μεγάλου αριθμού δέντρων απόφασης (decision trees), τα οποία εκπαιδεύονται σε τυχαία υποσύνολα των διαθέσιμων δεδομένων και στη συνέχεια συνδυάζονται για την παραγωγή μιας βελτιστοποιημένης πρόβλεψης. Η μέθοδος αυτή, γνωστή ως bagging (bootstrap aggregation), μειώνει την ευαισθησία στα ακραία δεδομένα και περιορίζει το πρόβλημα της υπερπροσαρμογής (overfitting) που συναντάται συχνά σε μεμονωμένα δέντρα απόφασης.

Στον τομέα των χρηματοοικονομικών προβλέψεων, το Random Forests χρησιμοποιείται για την ανάλυση και πρόβλεψη διαφόρων μετρικών, όπως η κατεύθυνση των τιμών των μετοχών, η διακύμανση των αγορών και η ανίχνευση παραγόντων που επηρεάζουν την απόδοση των επενδυτικών χαρτοφυλακίων. Η μέθοδος διακρίνεται για τη ρομποτική της προσέγγιση στην επιλογή χαρακτηριστικών (feature selection), καθώς μπορεί να αναγνωρίσει ποιοι παράγοντες έχουν τη μεγαλύτερη σημασία στις προβλέψεις, κάτι που καθιστά τη χρήση της ιδιαίτερα αποδοτική σε πολύπλοκα χρηματοοικονομικά δεδομένα.



Εικόνα 6: Διαδικασία Πρόβλεψης με χρήση Random Forest

Το XGBoost (Extreme Gradient Boosting) ανήκει στην κατηγορία των boosting algorithms, οι οποίοι διαφοροποιούνται από το bagging καθώς βασίζονται σε μία αλληλουχία δέντρων απόφασης που εκπαιδεύονται διαδοχικά. Η κεντρική ιδέα πίσω από την προσέγγιση boosting είναι ότι κάθε νέο δέντρο απόφασης μαθαίνει από τα λάθη του προηγούμενου, διορθώνοντας τα σφάλματα πρόβλεψης και ελαχιστοποιώντας τη συνολική απώλεια.

Το XGBoost αποτελεί μία από τις πιο αποδοτικές τεχνικές μηχανικής μάθησης στον τομέα των χρηματοοικονομικών προβλέψεων, καθώς συνδυάζει:

1. **Αυξημένη Υπολογιστική Απόδοση:** Χρησιμοποιεί παράλληλη επεξεργασία (parallel computing) και βελτιστοποιήσεις στη χρήση μνήμης, γεγονός που του επιτρέπει να επεξεργάζεται μεγάλα σύνολα δεδομένων με εξαιρετικά υψηλή ταχύτητα.
2. **Μείωση της Μεροληψίας και της Διακύμανσης:** Σε αντίθεση με το Random Forests, το XGBoost ενσωματώνει regularization terms (L1 & L2 penalties) για την αποφυγή της υπερπροσαρμογής, εξασφαλίζοντας έτσι μεγαλύτερη γενικευσιμότητα των προβλέψεων.
3. **Προσαρμοστικότητα στα Δεδομένα:** Η μέθοδος αυτή μπορεί να ενσωματώσει μη γραμμικές σχέσεις και αλληλεπιδράσεις μεταξύ των χαρακτηριστικών, καθιστώντας την ιδιαίτερα κατάλληλη για την πρόβλεψη χρηματιστηριακών τιμών και άλλων σύνθετων οικονομικών δεικτών.

Σύμφωνα με τον Zhang (2023) [33], η μέθοδος αυτή υπερτερεί των παραδοσιακών στατιστικών μοντέλων, όπως το CAPM και το Fama-French, ενσωματώνοντας νέους χρηματοοικονομικούς δείκτες και αξιοποιώντας τεχνικές βελτιστοποίησης. Χάρη στη χρήση δέντρων απόφασης και τεχνικών ενίσχυσης κλίσης, το XGBoost μπορεί να μειώσει το σφάλμα πρόβλεψης και να προσφέρει πιο ακριβείς εκτιμήσεις για την πορεία των μετοχών.

Μια πιο λεπτομερής ανάλυση των τεχνικών μηχανικής μάθησης, με έμφαση στα νευρωνικά δίκτυα και τη συμβολή τους στην πρόβλεψη χρηματοοικονομικών δεδομένων, θα παρουσιαστεί στο Κεφάλαιο 3.

2.3. Στατιστική Ανάλυση

2.3.1. Εισαγωγή στη Στατιστική Ανάλυση και βασικά μεγέθη

Η στατιστική ανάλυση αποτελεί τον κλάδο της επιστήμης που ασχολείται με τη συστηματική συλλογή, επεξεργασία, ανάλυση και ερμηνεία δεδομένων, με στόχο την εξαγωγή ουσιαστικών πληροφοριών και τη λήψη τεκμηριωμένων αποφάσεων. Ενσωματώνοντας στοιχεία της θεωρίας πιθανοτήτων και των εφαρμοσμένων μαθηματικών, η στατιστική αντιμετωπίζει την αβεβαιότητα και επιτρέπει την εξαγωγή συμπερασμάτων από μεγάλα και συχνά πολύπλοκα σύνολα δεδομένων.

Η στατιστική ανάλυση διακρίνεται σε δύο κύριες κατηγορίες: την περιγραφική και την επαγωγική στατιστική. Η περιγραφική στατιστική εστιάζει στην οργάνωση και παρουσίαση δεδομένων μέσω μέτρων κεντρικής τάσης, όπως ο μέσος όρος και η διάμεσος, καθώς και μέτρων διασποράς, όπως η διακύμανση και η τυπική απόκλιση. Από την άλλη πλευρά, η επαγωγική

στατιστική χρησιμοποιεί δείγματα δεδομένων για να εξαγάγει συμπεράσματα σχετικά με έναν πληθυσμό, υπολογίζοντας πιθανότητες και εκτιμήσεις για την ακρίβεια των αποτελεσμάτων.

Η σημασία της στατιστικής ανάλυσης είναι ιδιαίτερα έντονη σε ένα ευρύ φάσμα εφαρμογών, από την επιστημονική έρευνα και τις επιχειρηματικές αποφάσεις έως τη δημόσια πολιτική. Μέσα από τη χρήση κατάλληλων μεθόδων, η στατιστική ανάλυση επιτρέπει την αναγνώριση τάσεων, τη μέτρηση σχέσεων μεταξύ μεταβλητών, και την πρόβλεψη μελλοντικών γεγονότων, διαδραματίζοντας καίριο ρόλο στη διαχείριση της αβεβαιότητας και την επίλυση σύνθετων προβλημάτων.

Παρακάτω παρατίθενται τα βασικά στατιστικά μεγέθη που χρησιμοποιήθηκαν κατά τη διεξαγωγή του πειραματικού μέρους της παρούσας διπλωματικής εργασίας:

Μέση τιμή (Mean)

Η μέση τιμή (mean) είναι ένα από τα πιο βασικά μέτρα κεντρικής τάσης, το οποίο υπολογίζεται ως το άθροισμα όλων των τιμών σε ένα δείγμα ή πληθυσμό, διαιρεμένο με το πλήθος τους. Αν για ένα σύνολο δεδομένων

$$x_1, x_2, \dots, x_n$$

το πλήθος των τιμών είναι n , η μέση τιμή δίνεται από τη σχέση:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

Η μέση τιμή προσφέρει μια συνολική εικόνα για την τοποθεσία των δεδομένων στον άξονα των τιμών και είναι ευαίσθητη σε ακραίες τιμές (outliers).

Τυπική απόκλιση (Standard Deviation)

Η τυπική απόκλιση (standard deviation) είναι ένα μέτρο διασποράς που δείχνει το πόσο οι τιμές ενός δείγματος αποκλίνουν από τη μέση τιμή τους. Υπολογίζεται ως η τετραγωνική ρίζα της διακύμανσης και δίνεται από τη σχέση:

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}}$$

Η τυπική απόκλιση εκφράζεται στην ίδια μονάδα μέτρησης με τα δεδομένα και χρησιμοποιείται ευρέως για την αξιολόγηση της μεταβλητότητας ενός συνόλου δεδομένων.

Διακύμανση (Variance)

Η διακύμανση (variance) είναι το τετράγωνο της τυπικής απόκλισης και μετρά τη μέση τιμή των τετραγωνικών αποκλίσεων από τη μέση τιμή. Υπολογίζεται ως:

$$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$$

Αντιπροσωπεύει το επίπεδο μεταβλητότητας των δεδομένων. Μια χαμηλή τυπική απόκλιση υποδηλώνει ότι τα δείγματα έχουν την τάση να βρίσκονται κοντά στον μέσο όρο, ενώ υψηλή τυπική απόκλιση υποδηλώνει ότι τα στοιχεία απλώνονται σε ένα ευρύτερο φάσμα τιμών.

Συνδιακύμανση (Covariance)

Η συνδιακύμανση (covariance) είναι ένα μέτρο που αξιολογεί τη γραμμική σχέση μεταξύ δύο μεταβλητών. Για δύο μεταβλητές X και Y , υπολογίζεται ως:

$$\text{Cov}(X, Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n}$$

Θετική συνδιακύμανση υποδηλώνει ότι οι δύο μεταβλητές αυξάνονται ή μειώνονται ταυτόχρονα, ενώ αρνητική υποδηλώνει αντίθετη κατεύθυνση.

Συντελεστής γραμμικής συσχέτισης (Linear Correlation Coefficient)

Ο συντελεστής γραμμικής συσχέτισης (linear correlation coefficient), γνωστός και ως συντελεστής Pearson, μετρά τη δύναμη και την κατεύθυνση της γραμμικής σχέσης μεταξύ δύο μεταβλητών. Δίνεται από τη σχέση:

$$\rho = \frac{\text{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y}$$

Οι τιμές του κυμαίνονται από -1 έως 1. Μία τιμή κοντά στο 1 υποδηλώνει ισχυρή θετική συσχέτιση, ενώ μία τιμή κοντά στο -1 υποδηλώνει ισχυρή αρνητική συσχέτιση.

Συντελεστής αυτοσυσχέτισης (Autocorrelation Coefficient)

Ο συντελεστής αυτοσυσχέτισης (autocorrelation coefficient) μετρά τη γραμμική σχέση μεταξύ των τιμών μιας χρονοσειράς και των καθυστερημένων εκδοχών της (lags). Για μία χρονοσειρά X_t , ο συντελεστής αυτοσυσχέτισης για καθυστέρηση k ορίζεται ως:

$$\rho_k = \frac{\sum_{t=1}^{n-k} (X_t - \bar{X})(X_{t+k} - \bar{X})}{\sum_{t=1}^n (X_t - \bar{X})^2}$$

Ο συντελεστής αυτοσυσχέτισης μετρά τη συσχέτιση μιας χρονοσειράς με τον εαυτό της σε διαφορετικές χρονικές καθυστερήσεις (lags). Όταν χρησιμοποιείται για την ανίχνευση μη τυχαιότητας, επικεντρώνεται κυρίως στο πρώτο lag, ενώ σε ανάλυση μοντέλων χρονοσειρών, υπολογίζεται για πολλαπλά lags και απεικονίζεται σε διαγράμματα αυτοσυσχέτισης (ACF).

2.3.2. Μετρικές Ακρίβειας Μοντέλων Μηχανικής Μάθησης

Η ανάλυση των σφαλμάτων πρόβλεψης αποτελεί βασικό κριτήριο για την αξιολόγηση και επιλογή μοντέλων μηχανικής μάθησης. Το σφάλμα πρόβλεψης ορίζεται ως η διαφορά μεταξύ των πραγματικών τιμών (Y_i) και των προβλεπόμενων τιμών ($\hat{Y}_i$):

$$\varepsilon_i = Y_i - \hat{Y}_i$$

Οι δείκτες σφαλμάτων διαχωρίζονται σε δύο κατηγορίες:

1. **In-sample errors:** Τα σφάλματα αυτά υπολογίζονται από το ίδιο σύνολο δεδομένων που χρησιμοποιήθηκε για την εκπαίδευση του μοντέλου. Συνήθως, είναι μικρότερα από τα σφάλματα νέων δεδομένων.
2. **Out-of-sample errors:** Αφορούν νέα σύνολα δεδομένων που δεν χρησιμοποιήθηκαν κατά την εκπαίδευση. Αποτελούν αντιπροσωπευτικό δείκτη της γενικής ακρίβειας του μοντέλου.

Η ακρίβεια του μοντέλου μπορεί να εκτιμηθεί με τη βοήθεια των παρακάτω στατιστικών δεικτών:

Μέση Τιμή Σφάλματος (Mean Error, ME)

Ο μέσος όρος των σφαλμάτων, ο οποίος σπάνια χρησιμοποιείται λόγω εξουδετέρωσης θετικών και αρνητικών σφαλμάτων:

$$ME = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)$$

Μέσο Απόλυτο Σφάλμα (Mean Absolute Error, MAE)

Η μέση τιμή των απόλυτων διαφορών μεταξύ πραγματικών και προβλεπόμενων τιμών. Είναι εύκολα ερμηνεύσιμος και εκφράζεται στις ίδιες μονάδες μέτρησης:

$$MAE = \frac{1}{n} \sum_{i=1}^n |Y_i - \hat{Y}_i|$$

Μέσο Τετραγωνικό Σφάλμα (Mean Squared Error, MSE)

Ο μέσος όρος των τετραγώνων των διαφορών, όπου μεγαλύτερη έμφαση δίνεται στα μεγάλα σφάλματα:

$$MSE = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

Ρίζα Μέσου Τετραγωνικού Σφάλματος (Root Mean Squared Error, RMSE)

Η τετραγωνική ρίζα του MSE. Χρησιμοποιείται συχνά για τη μέτρηση της ακρίβειας κατά τη διάρκεια περιόδων υψηλής μεταβλητότητας:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2}$$

Μέσο Απόλυτο Ποσοστιαίο Σφάλμα (Mean Absolute Percentage Error, MAPE)

Ο μέσος όρος των απόλυτων ποσοστιαίων σφαλμάτων:

$$MAPE = \frac{1}{n} \sum_{i=1}^n \frac{|Y_i - \hat{Y}_i|}{Y_i} \times 100\%$$

Συντελεστής Προσδιορισμού (R²)

Ο συντελεστής R² εκφράζει το ποσοστό της διακύμανσης της εξαρτημένης μεταβλητής που εξηγείται από το μοντέλο:

$$R^2 = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

όπου:

- R²=1: Το μοντέλο εξηγεί πλήρως τη διακύμανση.
- R²=0: Το μοντέλο δεν εξηγεί καμία διακύμανση.

3. Νευρωνικά Δίκτυα

3.1. Εισαγωγή στη Μηχανική Μάθηση

3.1.1. Ορισμός και στόχοι της μηχανικής μάθησης.

Η Μηχανική Μάθηση (Machine Learning) αποτελεί έναν από τους πιο δυναμικούς κλάδους της Τεχνητής Νοημοσύνης (Artificial Intelligence), με ευρεία εφαρμογή σε πολλούς τομείς της κοινωνίας και της οικονομίας. Στο πυρήνα της, η Μηχανική Μάθηση αφορά τη δημιουργία αλγορίθμων που επιτρέπουν στα συστήματα να "μαθαίνουν" από δεδομένα και να βελτιώνουν τις επιδόσεις τους χωρίς να απαιτείται άμεσος προγραμματισμός από τον άνθρωπο. Σε αντίθεση με τις παραδοσιακές προσεγγίσεις, όπου οι κανόνες και οι διαδικασίες καθορίζονται εξ' αρχής, η Μηχανική Μάθηση βασίζεται στην ικανότητα των συστημάτων να αναγνωρίζουν μοτίβα στα δεδομένα και να προσαρμόζονται ανάλογα ([34]; [35]).

Η εποχή που ζούμε χαρακτηρίζεται από τεράστιες ποσότητες δεδομένων που παράγονται καθημερινά. Τα δεδομένα αυτά προέρχονται από ποικίλες πηγές, όπως το Διαδίκτυο των Πραγμάτων (IoT), τα κοινωνικά μέσα, τις επιχειρηματικές διαδικασίες, τα smartphones, τα συστήματα υγείας και πολλά άλλα. Αυτά τα δεδομένα μπορεί να είναι δομημένα, ημιδομημένα ή αδόμητα και αυξάνονται συνεχώς με ταχείς ρυθμούς. Η ανάλυσή τους μπορεί να προσφέρει πολύτιμες πληροφορίες, οι οποίες χρησιμοποιούνται για την ανάπτυξη έξυπνων εφαρμογών σε διάφορους τομείς. Για παράδειγμα, τα δεδομένα από αισθητήρες μπορούν να χρησιμοποιηθούν για την πρόβλεψη της συντήρησης μηχανημάτων, ενώ τα δεδομένα από τις πωλήσεις μπορούν να υποστηρίξουν τη βελτιστοποίηση αποθεμάτων ([35]; [36]).

Η Μηχανική Μάθηση διαδραματίζει επίσης καίριο ρόλο στη Βιομηχανία 4.0, που αποτελεί την τέταρτη βιομηχανική επανάσταση. Η Βιομηχανία 4.0 χαρακτηρίζεται από την αυτοματοποίηση των παραδοσιακών βιομηχανικών διαδικασιών μέσω της χρήσης προηγμένων τεχνολογιών, όπως η Μηχανική Μάθηση. Η ενσωμάτωση αυτών των τεχνολογιών επιτρέπει τη δημιουργία ευφών συστημάτων που μπορούν να ανταποκρίνονται δυναμικά στις αλλαγές του περιβάλλοντος. Για παράδειγμα, στη διαχείριση εφοδιαστικής αλυσίδας, η Μηχανική Μάθηση χρησιμοποιείται για την πρόβλεψη της ζήτησης προϊόντων, βελτιώνοντας την αποδοτικότητα και μειώνοντας το κόστος ([35]).

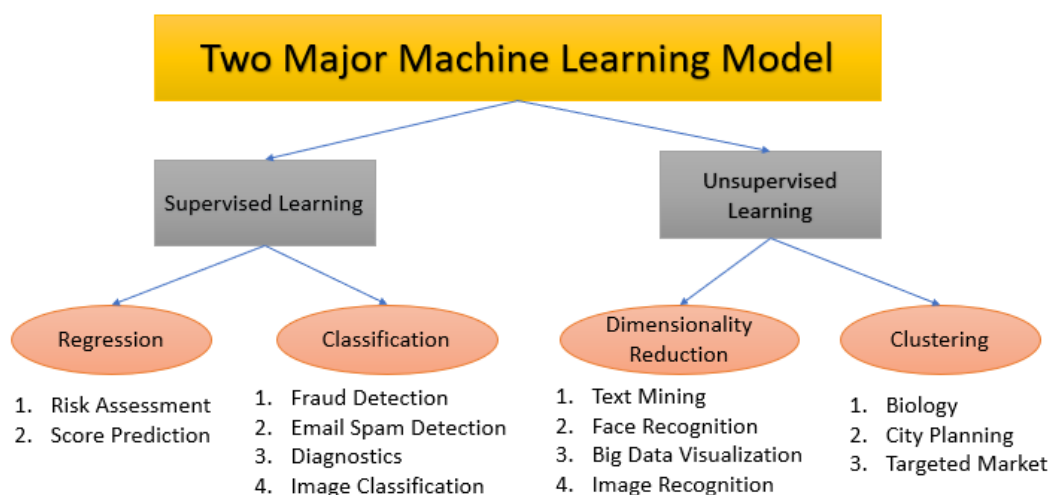
Η εφαρμογή της Μηχανικής Μάθησης περιλαμβάνει την ανάπτυξη αλγορίθμων που κατηγοριοποιούνται σε διάφορους τύπους, όπως η ταξινόμηση, η παλινδρόμηση, η συσταδοποίηση, και η μάθηση ενισχυτών. Επιπλέον, οι τεχνικές της χρησιμοποιούνται ευρέως για την ανάλυση δεδομένων, την πρόβλεψη μελλοντικών τάσεων, και τη δημιουργία έξυπνων συστημάτων σε τομείς όπως η κυβερνοασφάλεια, οι έξυπνες πόλεις και η υγεία ([36]).

Η Βιομηχανία 4.0 και η Μηχανική Μάθηση βρίσκονται σε άμεση αλληλεπίδραση, καθώς η χρήση δεδομένων αποτελεί τη βάση για την ανάπτυξη εφαρμογών με σημαντικό αντίκτυπο στην παραγωγικότητα και την καινοτομία. Η επιτυχία των εφαρμογών αυτών εξαρτάται από την ποιότητα των δεδομένων, καθώς και από την επιλογή κατάλληλων αλγορίθμων που να ανταποκρίνονται στις ανάγκες κάθε εφαρμογής. Η πρόοδος στις τεχνολογίες υπολογιστών υψηλής απόδοσης και στις τεχνικές Μηχανικής Μάθησης έχει ενισχύσει την ικανότητά τους να

αντιμετωπίζουν σύνθετες προκλήσεις, καθιστώντας τη Μηχανική Μάθηση αναπόσπαστο μέρος της σύγχρονης τεχνολογικής εξέλιξης.

3.1.2. Κατηγορίες μοντέλων:

Η μηχανική μάθηση αποτελεί ένα πεδίο της τεχνητής νοημοσύνης που βασίζεται στην εκπαίδευση μοντέλων για την ανάλυση δεδομένων και τη λήψη αποφάσεων, χωρίς να απαιτείται άμεσος προγραμματισμός. Υπάρχουν διαφορετικές κατηγορίες μοντέλων μηχανικής μάθησης, καθεμία με συγκεκριμένα χαρακτηριστικά και περιοχές εφαρμογής. Οι κύριες κατηγορίες περιλαμβάνουν την επιβλεπόμενη μάθηση (supervised learning), τη μη επιβλεπόμενη μάθηση (unsupervised learning) και την ενισχυτική μάθηση (reinforcement learning). Επιπλέον θα γίνει λόγος για τα προεκπαιδευμένα μοντέλα (pre-trained models). Αυτές οι προσεγγίσεις διαφοροποιούνται ανάλογα με τον τρόπο εκπαίδευσης, τα δεδομένα που απαιτούνται και τους στόχους που εξυπηρετούν.



Εικόνα 7: Κατηγοριοποίηση τεχνικών Μηχανικής Μάθησης σε επιβλεπόμενη και μη-επιβλεπόμενη (Πηγή: Medium)

3.1.2.1. Supervised Learning

Η supervised learning αποτελεί μία από τις πιο δημοφιλείς τεχνικές μηχανικής μάθησης, με ευρύ φάσμα εφαρμογών. Η βασική της ιδέα στηρίζεται στην εκμάθηση ενός μοντέλου που περιγράφει τη σχέση μεταξύ ενός συνόλου εισόδων (inputs) και των αντίστοιχων επισημασμένων εξόδων (outputs). Αυτή η σχέση επιτυγχάνεται μέσω της εκπαίδευσης του μοντέλου σε ένα σύνολο δεδομένων που περιλαμβάνει γνωστά ζεύγη εισόδων-εξόδων ([34]).

Τα καθήκοντα που επιτελούνται μέσω της supervised learning περιλαμβάνουν την κατηγοριοποίηση (classification) και την παλινδρόμηση (regression). Στην κατηγοριοποίηση, οι έξοδοι ανήκουν σε διακριτές κατηγορίες, όπως για παράδειγμα η πρόβλεψη του φύλου (άνδρας/γυναίκα) ή η ανάλυση συναισθήματος (θετικό/αρνητικό/ουδέτερο). Αντίθετα, στην παλινδρόμηση, οι έξοδοι είναι συνεχείς αριθμητικές τιμές, όπως η πρόβλεψη της θερμοκρασίας ή του μισθού.

Η διαδικασία εκπαίδευσης ενός μοντέλου supervised learning βασίζεται στη διαίρεση του συνόλου δεδομένων σε δύο μέρη: σύνολο εκπαίδευσης (training set) και σύνολο δοκιμών (testing set). Το σύνολο εκπαίδευσης χρησιμοποιείται για την εκμάθηση των χαρακτηριστικών του μοντέλου, ενώ το σύνολο δοκιμών αξιοποιείται για τη μέτρηση της απόδοσης του μοντέλου, εξασφαλίζοντας ότι δεν παρουσιάζει υπερπροσαρμογή (overfitting).

Ωστόσο, η supervised learning έχει σημαντικές απαιτήσεις σε δεδομένα. Η δημιουργία επισημασμένων συνόλων δεδομένων είναι συχνά δαπανηρή και χρονοβόρα, ενώ απαιτείται σαφής και αξιόπιστη συσχέτιση μεταξύ των εισόδων και των εξόδων. Παρόλα αυτά, η supervised learning έχει αποδειχθεί ιδιαίτερα επιτυχημένη, με χαμηλά ποσοστά σφάλματος σε πολλές εφαρμογές. Δημοφιλείς αλγόριθμοι περιλαμβάνουν τα artificial neural networks, logistic regression, linear regression, decision trees, support vector machines και K-nearest neighbor.

3.1.2.2. Unsupervised Learning

Η unsupervised learning αποτελεί μια πιο διερευνητική προσέγγιση στη μηχανική μάθηση, καθώς δεν απαιτεί την ύπαρξη επισημασμένων δεδομένων. Σε αυτό το μοντέλο, τα δεδομένα χρησιμοποιούνται όπως έχουν, χωρίς την ανάγκη για προηγούμενη γνώση ή ανθρώπινη παρέμβαση, επιτρέποντας την αυτόματη ανακάλυψη προτύπων, δομών και σχέσεων.

Μία από τις κύριες εφαρμογές της unsupervised learning είναι η ομαδοποίηση (clustering), όπου τα δεδομένα κατηγοριοποιούνται σε ομάδες με βάση ομοιότητες ή διαφορές. Τέτοιες εφαρμογές βρίσκουν χρήση σε διάφορους τομείς, όπως η τμηματοποίηση πελατών στο μάρκετινγκ, η ανίχνευση κοινοτήτων σε κοινωνικά δίκτυα και η βελτιστοποίηση ροών εργασίας. Άλλες κοινές εφαρμογές περιλαμβάνουν την εκτίμηση πυκνότητας (density estimation), τη μείωση διαστάσεων (dimensionality reduction), και την ανίχνευση ανωμαλιών (anomaly detection).

Η απουσία επισημασμένων δεδομένων καθιστά την unsupervised learning ιδιαίτερα χρήσιμη σε περιπτώσεις όπου η προεπεξεργασία δεδομένων είναι δύσκολη ή δαπανηρή. Ωστόσο, η ερμηνεία των αποτελεσμάτων της unsupervised learning είναι συχνά πιο περίπλοκη, καθώς τα μοντέλα δεν παρέχουν άμεσα εξηγήσεις για τις ανακαλύψεις τους. Επίσης, η αξιολόγηση της ποιότητας των αποτελεσμάτων αποτελεί πρόκληση, δεδομένου ότι δεν υπάρχει προφανής «σωστή» απάντηση.

Δημοφιλείς αλγόριθμοι unsupervised learning περιλαμβάνουν το k-means, το hierarchical clustering, τον αλγόριθμο DBSCAN και τη μέθοδο principal component analysis (PCA). Αυτοί οι αλγόριθμοι επιτρέπουν τη μείωση της πολυπλοκότητας των δεδομένων και τη βελτίωση της κατανόησης για τη δομή τους.

3.1.3. Ενισχυτική Μάθηση

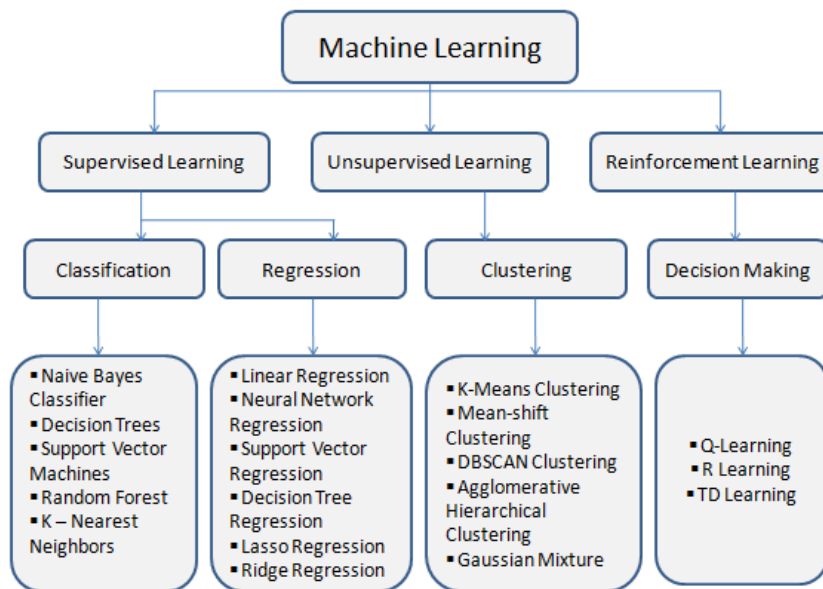
3.1.3.1. Προεκπαιδευμένα Μοντέλα και Μεταφορά Μάθησης

Η κατηγορία των pre-trained models συνδέεται στενά με την τεχνική του transfer learning, που βασίζεται στη μεταφορά γνώσης από μια πηγή (source task) σε έναν νέο στόχο (target task). Η διαδικασία αυτή περιλαμβάνει τη χρήση προκαταρκτικής εκπαίδευσης (pre-training) σε μεγάλα σύνολα δεδομένων ώστε να ενσωματωθεί σημαντική γνώση στις παραμέτρους του μοντέλου.

Στη συνέχεια, το μοντέλο μπορεί να εξειδικευτεί μέσω fine-tuning για την επίλυση συγκεκριμένων προβλημάτων.

Δύο βασικές προσεγγίσεις pre-training είναι η μεταφορά χαρακτηριστικών (feature transfer) και η μεταφορά παραμέτρων (parameter transfer). Η πρώτη μέθοδος επικεντρώνεται στην ανάπτυξη γενικών αναπαραστάσεων που μπορούν να ενσωματωθούν σε διαφορετικές εφαρμογές ([37]). Η δεύτερη βασίζεται στην κοινή χρήση παραμέτρων μεταξύ εργασιών.

Ειδικότερα, η ανάπτυξη deep neural networks, όπως το ResNet ([38]), αποτέλεσε σημείο καμπής για την επίτευξη υψηλής ακρίβειας σε σύνθετα προβλήματα. Η προκαταρκτική εκπαίδευση σε μεγάλα σύνολα δεδομένων, όπως το ImageNet ([39]), έχει συμβάλει σημαντικά στην πρόοδο εφαρμογών στον τομέα της όρασης υπολογιστών (CV) και της επεξεργασίας φυσικής γλώσσας (NLP). Για παράδειγμα, το BERT ([40]) αποτελεί πρωτοποριακό παράδειγμα pre-trained μοντέλου που αξιοποιεί τη μέθοδο του transfer learning για τη βελτίωση της επίδοσης.



Εικόνα 8: Κατηγορίες Μηχανικής Μάθησης [67]

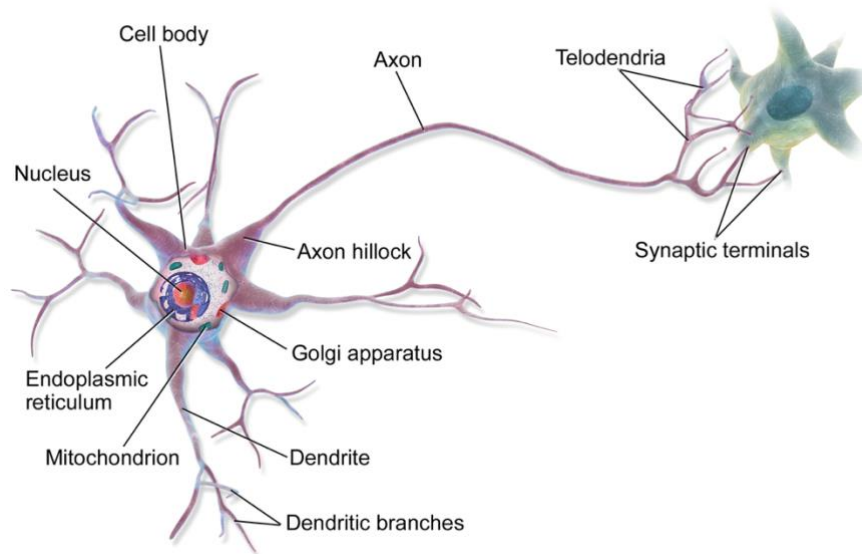
3.2. Αρχιτεκτονικές Νευρωνικών Δικτύων

3.2.1. Μοντέλο Βιολογικού Νευρώνα

Ο βιολογικός νευρώνας αποτελεί τη θεμελιώδη μονάδα του ανθρώπινου εγκεφάλου και παίζει καθοριστικό ρόλο στην επεξεργασία και μετάδοση πληροφοριών. Ο ανθρώπινος εγκέφαλος περιλαμβάνει περίπου 100 δισεκατομμύρια νευρώνες, οι οποίοι συνδέονται μέσω περίπου 100 τρισεκατομμυρίων συνάψεων. Αυτές οι συνδέσεις επιτρέπουν την ανταλλαγή πληροφοριών με εξαιρετική ταχύτητα, αποτελώντας τη βάση για τη μάθηση, τη μνήμη και τη λήψη αποφάσεων.

Ένας τυπικός νευρώνας αποτελείται από τα εξής μέρη:

1. **Δενδρίτες:** Λειτουργούν ως η είσοδος του νευρώνα, λαμβάνοντας ηλεκτρικά σήματα από άλλους νευρώνες. Οι δενδρίτες διαθέτουν διακλαδώσεις που αυξάνουν την επιφάνεια επικοινωνίας, επιτρέποντας στον νευρώνα να λαμβάνει σήματα από πολλές διαφορετικές πηγές.
2. **Κυτταρικό σώμα :** Εδώ βρίσκεται ο πυρήνας του νευρώνα, ο οποίος επεξεργάζεται τα σήματα που λαμβάνονται από τους δενδρίτες. Στο κυτταρικό σώμα συντελείται η άθροιση όλων των εισερχόμενων σημάτων και λαμβάνεται η απόφαση για τη μετάδοση του σήματος.
3. **Άξονας (Axon):** Ο άξονας λειτουργεί ως η έξοδος του νευρώνα, μεταφέροντας το επεξεργασμένο σήμα προς άλλους νευρώνες. Το μήκος του άξονα μπορεί να κυμαίνεται από μερικά χιλιοστά έως πάνω από 1 μέτρο.
4. **Συνάψεις:** Τα σημεία σύνδεσης μεταξύ του άξονα ενός νευρώνα και των δενδριτών άλλων νευρώνων. Μέσω των συναπτικών συνδέσεων πραγματοποιείται η μετάδοση του σήματος, είτε με ενισχυτικό (excitatory) είτε με ανασταλτικό (inhibitory) τρόπο.



Εικόνα 9: Αναπαράσταση Βιολογικού Νευρώνα (Πηγή:Wikipedia)

Η λειτουργία του νευρώνα βασίζεται στη μετάδοση ηλεκτρικών και χημικών σημάτων. Τα σήματα που λαμβάνονται από τους δενδρίτες συνδυάζονται στο κυτταρικό σώμα, όπου πραγματοποιείται η άθροισή τους. Εάν το συνολικό σήμα υπερβεί μια ορισμένη τιμή κατωφλίου, τότε δημιουργείται μια δυναμική δράση (action potential). Αυτή η δυναμική δράση διαδίδεται κατά μήκος του άξονα του νευρώνα και στη συνέχεια μεταδίδεται σε άλλους νευρώνες μέσω των συνάψεων. Επιπλέον, η ικανότητα του νευρώνα να μαθαίνει και να προσαρμόζεται εξαρτάται από την πλαστικότητα των συνάψεων, δηλαδή από τη δυνατότητά τους να μεταβάλλουν την αγωγιμότητά τους με την πάροδο του χρόνου.

3.2.2. Μοντέλο Τεχνητού Νευρώνα (Perceptron)

Το Perceptron, το οποίο εισήχθη από τον Frank Rosenblatt το 1957, αποτελεί τη βάση για τη σύγχρονη κατανόηση των τεχνητών νευρωνικών δικτύων. Ο σχεδιασμός του βασίζεται στο μοντέλο McCulloch-Pitts (MCP), που παρουσιάστηκε το 1943 από τους Warren McCulloch και Walter Pitts. Το μοντέλο MCP εισήγαγε έναν τρόπο αναπαράστασης των βιολογικών νευρώνων μέσω μαθηματικών συναρτήσεων, όπου οι εισόδοι πολλαπλασιάζονται με βάρη και περνούν από μια συνάρτηση ενεργοποίησης για την παραγωγή εξόδου.

Το Perceptron αποτελεί έναν απλό τεχνητό νευρώνα που λειτουργεί ως γραμμικός δυαδικός ταξινομητής. Δέχεται πολλαπλές εισόδους $x_1, x_2, \dots, x_n$ οι οποίες πολλαπλασιάζονται με αντίστοιχα βάρη $w_1, w_2, \dots, w_n$. Το αποτέλεσμα αυτών των πολλαπλασιασμών προστίθεται σε έναν όρο πόλωσης b (bias), και το συνολικό άθροισμα περνά από μια συνάρτηση ενεργοποίησης που καθορίζει την έξοδο.

Η έξοδος του Perceptron υπολογίζεται από την εξίσωση:

$$f(x) = \begin{cases} 1 & \text{if } w \cdot x + b > 0 \\ 0 & \text{if } w \cdot x \leq 0 \end{cases}$$

όπου:

- $w \cdot x = \sum_{i=1}^n w_i x_i$ είναι το εσωτερικό γινόμενο των εισόδων και των βαρών.
- b είναι η πόλωση που μετατοπίζει το όριο απόφασης.

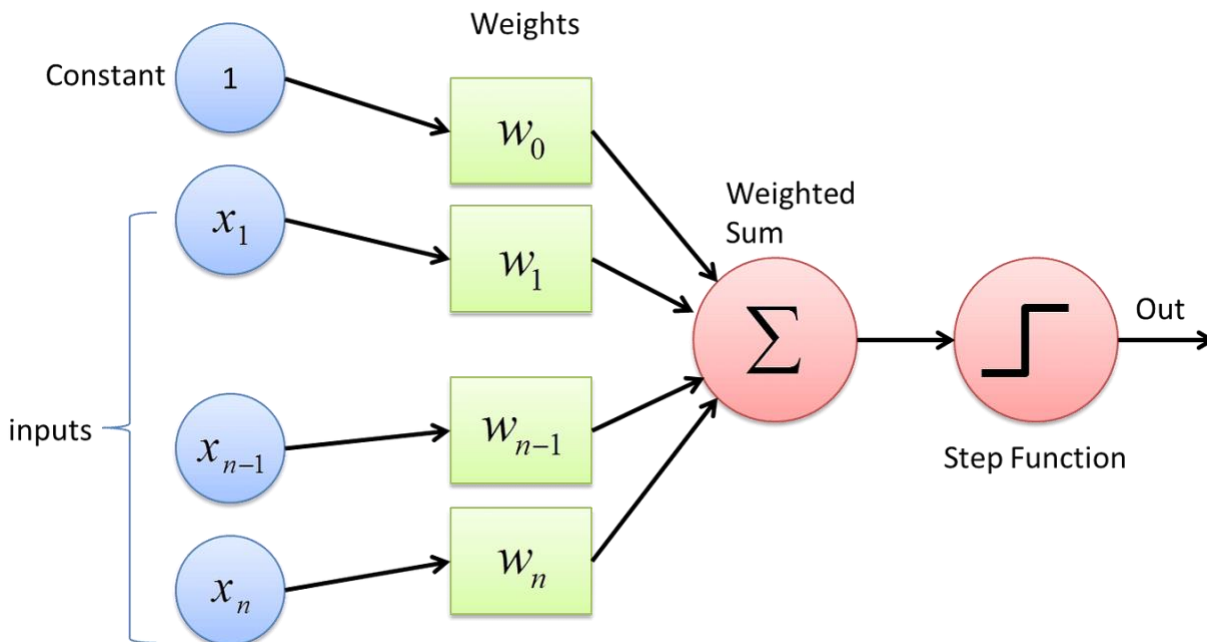
Το Perceptron χρησιμοποιείται σε προβλήματα επιβλεπόμενης μάθησης και μπορεί να κατηγοριοποιήσει δεδομένα σε δύο γραμμικά διαχωρίσιμες κλάσεις.

Οι εισόδοι, συμβολιζόμενες ως $x_1, x_2, \dots, x_n$, αντιπροσωπεύουν τα δεδομένα που παρέχονται στο μοντέλο. Αυτές μπορεί να είναι αριθμητικές τιμές που περιγράφουν χαρακτηριστικά ενός δείγματος. Ο αριθμός των εισόδων (n) καθορίζεται από τη διάσταση του προβλήματος. Κάθε εισόδος συνοδεύεται από ένα αντίστοιχο βάρος ($w_1, w_2, \dots, w_n$), που προσδιορίζει τη σημασία της συγκεκριμένης εισόδου στη διαδικασία λήψης απόφασης. Τα βάρη είναι πραγματικοί αριθμοί, που μπορούν να είναι είτε θετικά, ενισχύοντας τη συνεισφορά της εισόδου, είτε αρνητικά, μειώνοντας τη σημασία της. Κατά τη διαδικασία μάθησης, τα βάρη προσαρμόζονται ώστε να μειώνεται το σφάλμα πρόβλεψης.

Μία σημαντική συνιστώσα του Perceptron είναι η πόλωση (b), η οποία είναι ένας σταθερός όρος που προστίθεται στο ζυγισμένο άθροισμα των εισόδων. Η πόλωση εξυπηρετεί τη μετακίνηση της ευθείας απόφασης (decision boundary) μακριά από την αρχή των αξόνων, παρέχοντας στο μοντέλο μεγαλύτερη ευελιξία. Ο υπολογισμός του ζυγισμένου αθροίσματος των εισόδων και των βαρών εκφράζεται μαθηματικά ως:

Η συνάρτηση ενεργοποίησης παίζει καθοριστικό ρόλο στο Perceptron, καθώς αποφασίζει την τελική έξοδο του μοντέλου. Η λειτουργία της βασίζεται στο αποτέλεσμα του ζυγισμένου αθροίσματος. Στο κλασικό Perceptron, χρησιμοποιείται μια δυαδική συνάρτηση ενεργοποίησης, όπως ορίστηκε προηγουμένως.

Αυτή η συνάρτηση επιτρέπει στο Perceptron να παράγει δυαδική έξοδο, η οποία μπορεί να χρησιμοποιηθεί για την ταξινόμηση δειγμάτων σε δύο κατηγορίες. Η συνάρτηση ενεργοποίησης μπορεί να μεταβληθεί ανάλογα με τις ανάγκες του προβλήματος, με τη ReLU (Rectified Linear Unit) να αποτελεί μια πιο σύγχρονη επιλογή.



Εικόνα 10: Αναπαράσταση διάταξης Perceptron

3.2.3. Συναρτήσεις ενεργοποίησης

Οι συναρτήσεις ενεργοποίησης αποτελούν έναν κρίσιμο παράγοντα στη σχεδίαση και λειτουργία των νευρωνικών δικτύων, καθορίζοντας τον τρόπο με τον οποίο τα δεδομένα περνούν από τους κόμβους του δικτύου και συμβάλλοντας στη μη γραμμική μοντελοποίηση σύνθετων σχέσεων. Η επιλογή της κατάλληλης συνάρτησης εξαρτάται από το είδος του προβλήματος που αντιμετωπίζεται, όπως ταξινόμηση, παλινδρόμηση ή πρόβλεψη πιθανοτήτων. [41]

Σε κάθε νευρώνα, το εισερχόμενο σήμα πολλαπλασιάζεται με τα αντίστοιχα βάρη και στη συνέχεια αθροίζεται. Το αποτέλεσμα αυτού του υπολογισμού ονομάζεται τιμή προ-ενεργοποίησης ($h = W \cdot Xh$), ενώ η έξοδος μετά την εφαρμογή της συνάρτησης ενεργοποίησης ονομάζεται τιμή μετα-ενεργοποίησης ($a = \Phi(h)$). Η τιμή μετα-ενεργοποίησης είναι αυτή που περνά στον επόμενο νευρώνα ή χρησιμοποιείται ως τελική έξοδος του δικτύου.

3.2.3.1. Γραμμική Συνάρτηση Ενεργοποίησης

Η πιο απλή συνάρτηση ενεργοποίησης είναι η γραμμική ($\Phi(v) = v$), η οποία χρησιμοποιείται συνήθως στον τελικό κόμβο ενός δικτύου όταν η έξοδος είναι πραγματική τιμή. Παρόλο που είναι χρήσιμη για απλές παλινδρομήσεις, η γραμμική συνάρτηση δεν επιτρέπει στο δίκτυο να μάθει μη γραμμικές σχέσεις.

3.2.3.2. Μη Γραμμικές Συναρτήσεις Ενεργοποίησης

Οι μη γραμμικές συναρτήσεις ενεργοποίησης είναι απαραίτητες για την αύξηση της υπολογιστικής δύναμης των νευρωνικών δικτύων. Ορισμένες από τις πιο κλασικές επιλογές περιλαμβάνουν:

Sign Function:

$$\Phi(v) = \text{sign}(v)$$

Χρησιμοποιείται κυρίως για δυαδική ταξινόμηση. Παρά τη χρησιμότητά της, δεν είναι διαφορίσιμη, γεγονός που την καθιστά λιγότερο κατάλληλη για εκπαίδευση μέσω οπισθοδιάδοσης.

Sigmoid Function:

$$\Phi(v) = 1 / (1 + e^{-v})$$

Παράγει τιμές στο εύρος (0,1) και χρησιμοποιείται για την πρόβλεψη πιθανοτήτων σε ταξινομητές. Η χρήση της στη διαμόρφωση πιθανοτήτων και στη μέγιστη πιθανοφάνεια την καθιστά δημοφιλή, αν και ο κορεσμός των τιμών της για μεγάλα απόλυτα μεγέθη εισόδων μπορεί να μειώσει την ταχύτητα εκπαίδευσης.

Tanh Function:

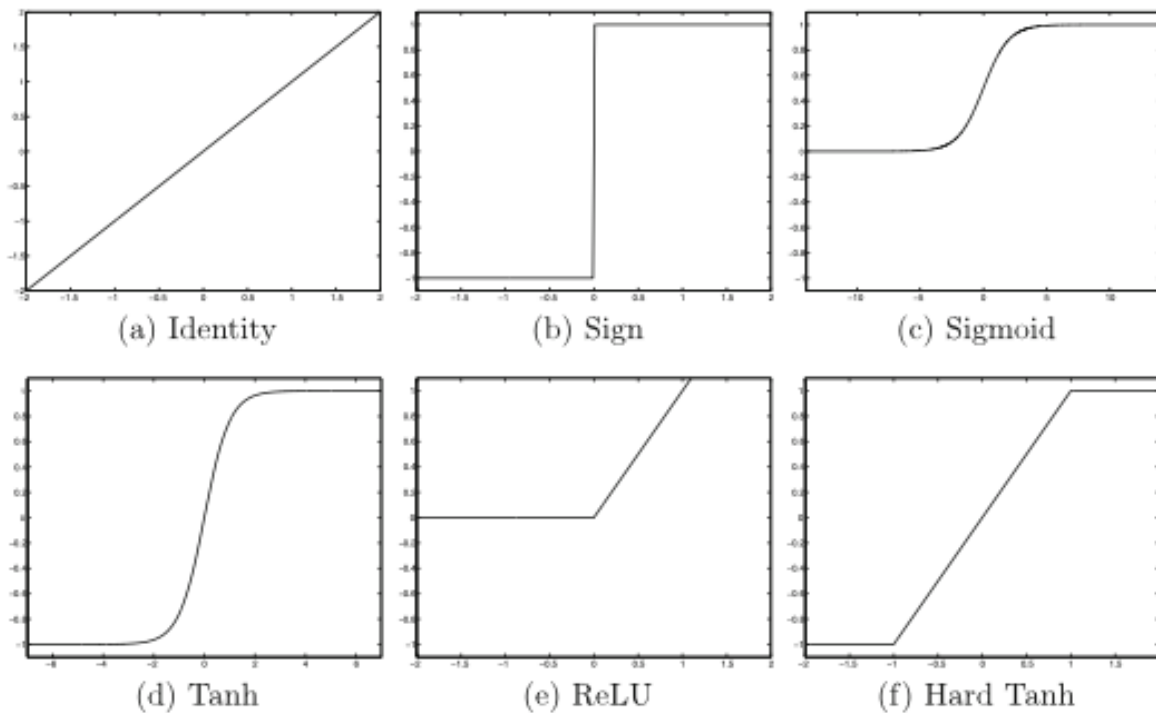
$$\Phi(v) = (e^{2v} - 1) / (e^{2v} + 1)$$

Παράγει τιμές στο εύρος [-1,1] και είναι κατάλληλη για περιπτώσεις όπου απαιτείται μέση κεντροποίηση της εξόδου. Σε σχέση με τη Sigmoid, διαθέτει μεγαλύτερο εύρος τιμών και μεγαλύτερη κλίση, γεγονός που βελτιώνει την εκπαίδευση.

Rectified Linear Unit (ReLU):

$$\Phi(v) = \max(0, v)$$

Η ReLU είναι σήμερα η πιο δημοφιλής συνάρτηση ενεργοποίησης σε πολυεπίπεδα νευρωνικά δίκτυα. Επιτρέπει την εκπαίδευση βαθύτερων δικτύων λόγω της απλότητας και της αποτελεσματικότητάς της. Ωστόσο, μπορεί να οδηγήσει σε «νεκρούς νευρώνες» όταν οι τιμές των βαρών καθίστανται τέτοιες ώστε οι τιμές προ-ενεργοποίησης να είναι συνεχώς αρνητικές.



Εικόνα 11: Διάφορες συναρτήσεις ενεργοποίησης [13]

3.2.4. Εκπαίδευση Νευρωνικών δικτύων

Η μάθηση στα τεχνητά νευρωνικά δίκτυα (ΤΝΔ) αποτελεί έναν θεμελιώδη μηχανισμό που επιτρέπει στο δίκτυο να προσαρμόζεται και να αποδίδει ακριβείς προβλέψεις. Η διαδικασία της μάθησης συνίσταται στην τροποποίηση των βαρών του δικτύου, έτσι ώστε οι έξοδοι που παράγονται για δεδομένα εισόδου να προσεγγίζουν τις επιθυμητές τιμές. Ο στόχος της μάθησης είναι η βελτιστοποίηση της λειτουργίας του δικτύου, ώστε να εξάγονται οι σωστές λύσεις για τα προβλήματα που αντιμετωπίζονται.

Η εκπαίδευση ενός ΤΝΔ ξεκινά με την αρχικοποίηση των βαρών σε τυχαίες τιμές. Κατά την εκπαίδευση, το δίκτυο λαμβάνει δεδομένα εισόδου, τα επεξεργάζεται και παράγει μια έξοδο. Αυτή η έξοδος συγκρίνεται με την επιθυμητή τιμή, και η διαφορά που προκύπτει ονομάζεται σφάλμα (error). Το σφάλμα αυτό χρησιμοποιείται ως βάση για την αναπροσαρμογή των βαρών, με σκοπό τη μείωση του και την επίτευξη μεγαλύτερης ακρίβειας στις προβλέψεις.

Η διαδικασία της μάθησης μπορεί να κατηγοριοποιηθεί σε τρία είδη: τη μάθηση με επίβλεψη (supervised learning), τη μάθηση χωρίς επίβλεψη (unsupervised learning) και τη βαθμολογημένη μάθηση (graded learning). Στη μάθηση με επίβλεψη, τα δεδομένα περιλαμβάνουν ζεύγη εισόδου-εξόδου, όπου το δίκτυο μαθαίνει να συσχετίζει τις εισόδους με τις επιθυμητές εξόδους. Στη μάθηση χωρίς επίβλεψη, το δίκτυο αυτό-οργανώνεται με βάση τις εισόδους, ενώ στη βαθμολογημένη μάθηση οι έξοδοι αξιολογούνται με βάση μια κλίμακα και τα βάρη προσαρμόζονται ανάλογα.

Η εκπαίδευση ενός νευρωνικού δικτύου περιλαμβάνει την εφαρμογή επαναληπτικών αλγορίθμων για την ελαχιστοποίηση του σφάλματος. Ένας από τους πιο κλασικούς αλγορίθμους

είναι ο κανόνας του Perceptron. Ο συγκεκριμένος κανόνας βασίζεται στην αρχική σύγκριση των παραγόμενων εξόδων με τις ιδανικές τιμές. Μετά τη σύγκριση, τα βάρη τροποποιούνται, ώστε να μειωθεί η διαφορά μεταξύ των δύο. Αυτή η επαναληπτική διαδικασία συνεχίζεται έως ότου το σφάλμα φτάσει σε ένα αποδεκτό επίπεδο.

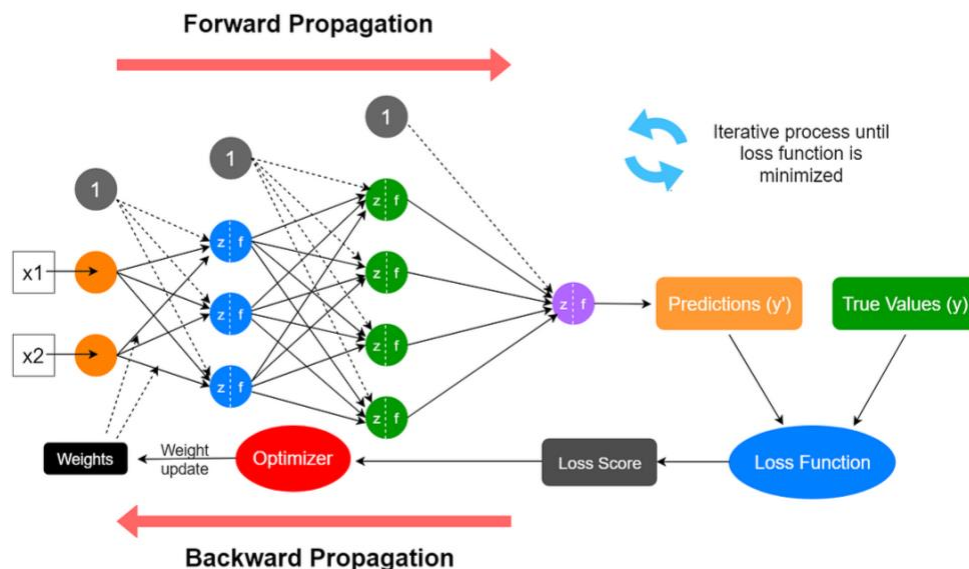
Ο μαθηματικός τύπος που περιγράφει την προσαρμογή των βαρών στο Perceptron βασίζεται στον κανόνα ενημέρωσης των βαρών:

$$w_i^{(t+1)} = w_i^{(t)} + \eta \cdot (y - \hat{y}) \cdot x_i$$

- $w_i^{(t)}$: είναι το βάρος της σύνδεσης του i-οστού χαρακτηριστικού στη χρονική στιγμή t
- η : είναι ο ρυθμός εκμάθησης (learning rate)
- y : είναι η επιθυμητή έξοδος
- $\hat{y}$: είναι η έξοδος που προβλέφθηκε από το μοντέλο
- x_i : είναι η είσοδος που αντιστοιχεί στο βάρος w_i

Στα πολυεπίπεδα δίκτυα, όπου η σχέση μεταξύ των βαρών και του σφάλματος γίνεται πιο περίπλοκη, χρησιμοποιείται ο αλγόριθμος πίσω διάδοσης σφάλματος (backpropagation). Ο αλγόριθμος αυτός βασίζεται στον κανόνα της αλυσίδας του διαφορικού λογισμού για τον υπολογισμό των παραγώνων του σφάλματος ως προς τα βάρη. Η διαδικασία περιλαμβάνει δύο φάσεις: την προώθηση (forward phase) και την οπισθοχώρηση (backward phase). Στην προώθηση, οι είσοδοι περνούν μέσα από το δίκτυο και παράγονται οι έξοδοι, ενώ στην οπισθοχώρηση οι παράγωγοι του σφάλματος υπολογίζονται για κάθε επίπεδο και τα βάρη ενημερώνονται.

Η απόδοση ενός ΤΝΔ επηρεάζεται από διάφορους παράγοντες. Ένας από αυτούς είναι η αρχική τιμή των βαρών, η οποία μπορεί να επηρεάσει την ταχύτητα σύγκλισης του αλγορίθμου. Ο ρυθμός εκμάθησης (learning rate) είναι επίσης κρίσιμος, καθώς καθορίζει το μέγεθος των αλλαγών στα βάρη σε κάθε επανάληψη. Επιπλέον, το μέγεθος των batch δεδομένων εκπαίδευσης επηρεάζει τη σταθερότητα και την ακρίβεια της διαδικασίας. Τέλος, η επιλογή της συνάρτησης απώλειας (loss function) παίζει καθοριστικό ρόλο στην ποιότητα της εκπαίδευσης.



Εικόνα 12:
Διαδικασία
προώθησης και
οπισθοχώρησης

3.2.5. Πολυεπίπεδο Perceptron (MLP)

Το πολυεπίπεδο perceptron (MLP) αποτελεί έναν από τους πιο σημαντικούς τύπους τεχνητών νευρωνικών δικτύων, καθώς έχει βρει ευρεία εφαρμογή σε πολλά πεδία, από την επεξεργασία εικόνων και την αναγνώριση προτύπων έως την ανάλυση χρηματοοικονομικών δεδομένων. Η κύρια διαφορά του MLP από το απλό perceptron είναι η προσθήκη ενός ή περισσότερων κρυφών επιπέδων (hidden layers), γεγονός που του επιτρέπει να μάθει και να αναπαραστήσει μη γραμμικές σχέσεις μεταξύ των εισόδων και των εξόδων του.

Το MLP αποτελείται από τρία βασικά είδη επιπέδων:

1. **Το επίπεδο εισόδου (input layer)**, στο οποίο εισάγονται τα δεδομένα.
2. **Τα κρυφά επίπεδα (hidden layers)**, όπου εφαρμόζονται οι μη γραμμικοί μετασχηματισμοί.
3. **Το επίπεδο εξόδου (output layer)**, το οποίο παρέχει την τελική πρόβλεψη του μοντέλου.

Κάθε νευρώνας ενός επιπέδου συνδέεται με όλους τους νευρώνες του επόμενου μέσω βαρών σύνδεσης. Η πληροφορία ρέει από την είσοδο προς την έξοδο μέσω μιας διαδικασίας προώθησης προς τα εμπρός (feedforward propagation).

Τα βήματα που αφορούν την ροή των δεδομένων κατά την διαδικασία της ταξινόμησης στα perceptrons είναι τα παρακάτω [42]:

1. Παρουσιάζεται το διάνυσμα εισόδου $x_1, x_2, \dots, x_n$
2. Η έξοδος κάθε νευρώνα υπολογίζεται ως εξής:

$$h_j = \frac{1}{1 + \exp -w_{0j} + \sum_{i=1}^l w_{ij}x_i}, \quad 1 \leq j \leq m$$

όπου:

- w_{ij} είναι το βάρος μεταξύ του εισόδου x_i και του νευρώνα h_j ,
 - w_{0j} είναι η πόλωση (bias) του νευρώνα h_j ,
 - l είναι ο αριθμός των εισόδων,
 - m είναι ο αριθμός των κρυφών νευρώνων.
3. Η συνάρτηση ενεργοποίησης που χρησιμοποιείται στο MLP είναι συνήθως η σιγμοειδής:

$$f(x) = \frac{1}{1 + e^{-x}}$$

η οποία βοηθά το δίκτυο να μάθει πολύπλοκες σχέσεις μεταξύ των δεδομένων εισόδου και εξόδου.

4. Η τελική έξοδος του δικτύου δίνεται από:

$$o_k = \frac{1}{1 + \exp(v_{0k} + \sum_{j=1}^m v_{jk} h_j)}, \quad 1 \leq k \leq n$$

Η κλάση $c = [c_1, c_2, \dots, c_n]$, στην οποία το perceptron ταξινομεί το πρότυπο εισόδου p , πρέπει να είναι ένας δυαδικός διανυσματικός δείκτης. Αυτό σημαίνει ότι η έξοδος o_k πρέπει να υπερβαίνει ένα κατώφλι τ που ορίζεται ανάλογα με την εκάστοτε εφαρμογή.

5. Η διαδικασία επαναλαμβάνεται ακολουθώντας τα βήματα 1, 2, 3 και 4 για κάθε δεδομένο πρότυπο εισόδου.

Τα πολυεπίπεδα perceptrons (MLPs) έχουν εφαρμοστεί αποτελεσματικά σε διάφορους επιστημονικούς κλάδους, όπως η ιατρική και η γεωλογία. Στον τομέα της ιατρικής, τα MLPs έχουν χρησιμοποιηθεί για την πρόβλεψη καρδιακών παθήσεων, επιτυγχάνοντας ακρίβεια 88,10% στην αναγνώριση καρδιακής ανεπάρκειας με βάση δεδομένα ασθενών. Η χρήση των MLPs σε αυτό το πλαίσιο αποδεικνύει την ικανότητά τους να συμβάλλουν στην έγκαιρη διάγνωση και παρέμβαση, οδηγώντας σε βελτιωμένα θεραπευτικά αποτελέσματα και μείωση των δαπανών στον τομέα της υγείας [43]. Γεωλογία, τα MLPs έχουν αξιοποιηθεί για τη γεωλογική μοντελοποίηση κατηγοριοποιημένων μεταβλητών. Σε μια σχετική μελέτη, χρησιμοποιήθηκε ένα Τεχνητό Νευρωνικό Δίκτυο τύπου Πολυεπίπεδου Perceptron (ANN-MLP) για την ταξινόμηση τύπων πετρωμάτων, βασιζόμενο σε μεταβλητές όπως γεωγραφικές συντεταγμένες και συγκεντρώσεις ορυκτών. Τα αποτελέσματα κατέδειξαν την αποτελεσματικότητα και την ακρίβεια των ANN-MLP στην κατηγοριοποίηση γεωλογικών δεδομένων, γεγονός που υπογραμμίζει τη συμβολή της τεχνητής νοημοσύνης στη γεωλογική ανάλυση και μοντελοποίηση [44].

3.2.6. Οπίσθια Διάδοση (Backpropagation)

Τα πολυεπίπεδα perceptrons (MLPs) έχουν χρησιμοποιηθεί με επιτυχία για την επίλυση σύνθετων προβλημάτων ταξινόμησης. Η εκπαίδευση του δικτύου πραγματοποιείται μέσω μιας επιβλεπόμενης τεχνικής μάθησης, η οποία απαιτεί σύνολα δεδομένων εισόδου-εξόδου. Η συγκεκριμένη τεχνική εκπαίδευσης είναι γνωστή ως αλγόριθμος οπισθοδιάδοσης σφάλματος (backpropagation) και αποτελεί μια αμφίδρομη διαδικασία που περιλαμβάνει δύο στάδια: τη διάδοση προς τα εμπρός (forward propagation) και τη διάδοση προς τα πίσω (backward propagation). Στο πρώτο στάδιο, ένας διάνυσμα εκπαίδευσης παρουσιάζεται στο δίκτυο και ταξινομείται. Στη συνέχεια, στο στάδιο της οπισθοδιάδοσης, τα βάρη του δικτύου ενημερώνονται αναδρομικά σε όλα τα επίπεδα με βάση το υπολογισμένο σφάλμα. Τα σφάλματα των νευρώνων στην έξοδο είναι συνάρτηση της επιθυμητής κλάσης c_t και της εξόδου του perceptron $(o_1, o_2, \dots, o_n)$. [42]

Η αλγοριθμική περιγραφή του αλγορίθμου οπισθοδιάδοσης έχει ως εξής:

1. **Αρχικοποίηση:** Τα βάρη του perceptron αρχικοποιούνται τυχαία με αριθμούς στο διάστημα $[-0.1, 0.1]$, δηλαδή:

$$w_{ij} = \text{random}([-0.1, 0.1]), \quad 0 \leq i \leq l, \quad 1 \leq j \leq m$$

$$v_{jk} = \text{random}([-0.1, 0.1]), \quad 0 \leq j \leq m, \quad 1 \leq k \leq n$$

2. **Παρουσίαση των προτύπων εκπαίδευσης:** Το διάνυσμα εκπαίδευσης $p_t = [p_1^t, p_2^t, \dots, p_l^t]$ από το ζεύγος εκπαίδευσης (p_t, c_t) παρουσιάζεται στο perceptron και εφαρμόζονται τα αντίστοιχα βήματα από τον αλγόριθμο ταξινόμησης του perceptron.
3. **Υπολογισμός εμπρός (επίπεδο εξόδου):** Τα σφάλματα των νευρώνων στην έξοδο υπολογίζονται ως εξής:

$$\delta_{o_k} = o_k(1 - o_k)(c_k^t - o_k), \quad 1 \leq k \leq n$$

όπου $ct = [c_1^t, c_2^t, \dots, c_n^t]$ αντιπροσωπεύει τη σωστή κλάση του προτύπου p^t και το διάνυσμα $(o_1, o_2, \dots, o_n)$ περιγράφει την έξοδο του perceptron.

4. **Υπολογισμός εμπρός (κρυφό επίπεδο):** Τα σφάλματα των νευρώνων στο κρυφό επίπεδο υπολογίζονται ως εξής:

$$\delta_{h_j} = h_j(1 - h_j) \sum_{k=1}^n \delta_{o_k} v_{jk}, \quad 1 \leq j \leq m$$

5. **Οπισθοδιάδοση (ενημέρωση βαρών εξόδου):** Τα βάρη μεταξύ του κρυφού επιπέδου και του επιπέδου εξόδου ενημερώνονται σύμφωνα με τον τύπο:

$$v_{jk}(t) = v_{jk}(t-1) + \eta \delta_{o_k} h_j, \quad 0 \leq j \leq m, \quad 1 \leq k \leq n$$

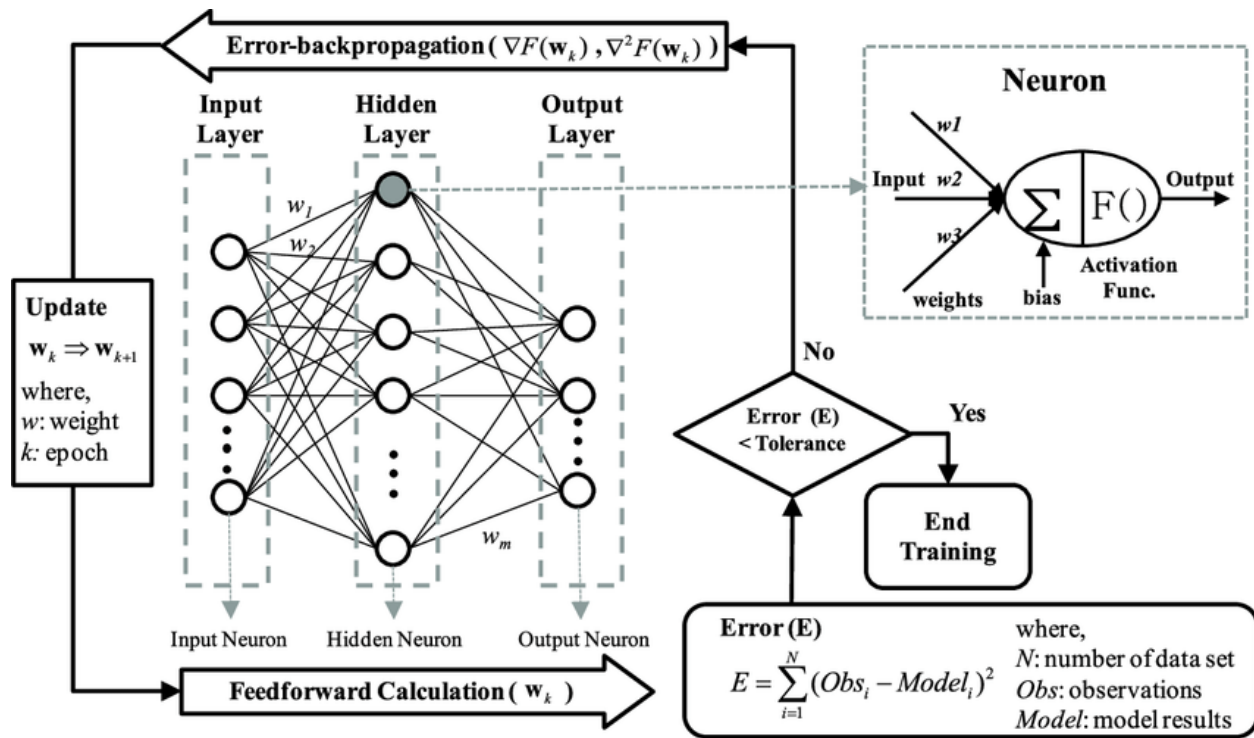
όπου η παράμετρος η είναι ο ρυθμός μάθησης και λαμβάνει τιμές $0 \leq \eta \leq 1$.

6. **Οπισθοδιάδοση (ενημέρωση βαρών κρυφού επιπέδου):** Τα βάρη μεταξύ του επιπέδου εισόδου και του κρυφού επιπέδου ενημερώνονται με τον τύπο:

$$w_{ij}(t) = w_{ij}(t-1) + \eta \delta_{h_j} p_i^t, \quad 0 \leq i \leq l, \quad 1 \leq j \leq m$$

7. **Επανάληψη:** Τα βήματα 2–6 επαναλαμβάνονται για κάθε πρότυπο στο σύνολο δεδομένων εκπαίδευσης. Κάθε πλήρης διέλευση μέσα από το σύνολο εκπαίδευσης αποτελεί μία επανάληψη του αλγορίθμου.

Ο αλγόριθμος οπισθοδιάδοσης αποτελεί τη θεμελιώδη μέθοδο εκπαίδευσης των πολυεπίπεδων perceptrons και επιτρέπει στο δίκτυο να βελτιώνει σταδιακά τις προβλέψεις του μέσω της προσαρμογής των βαρών με βάση το παραγόμενο σφάλμα.



Εικόνα 13: Αναπαράσταση διαδικασίας backpropagation

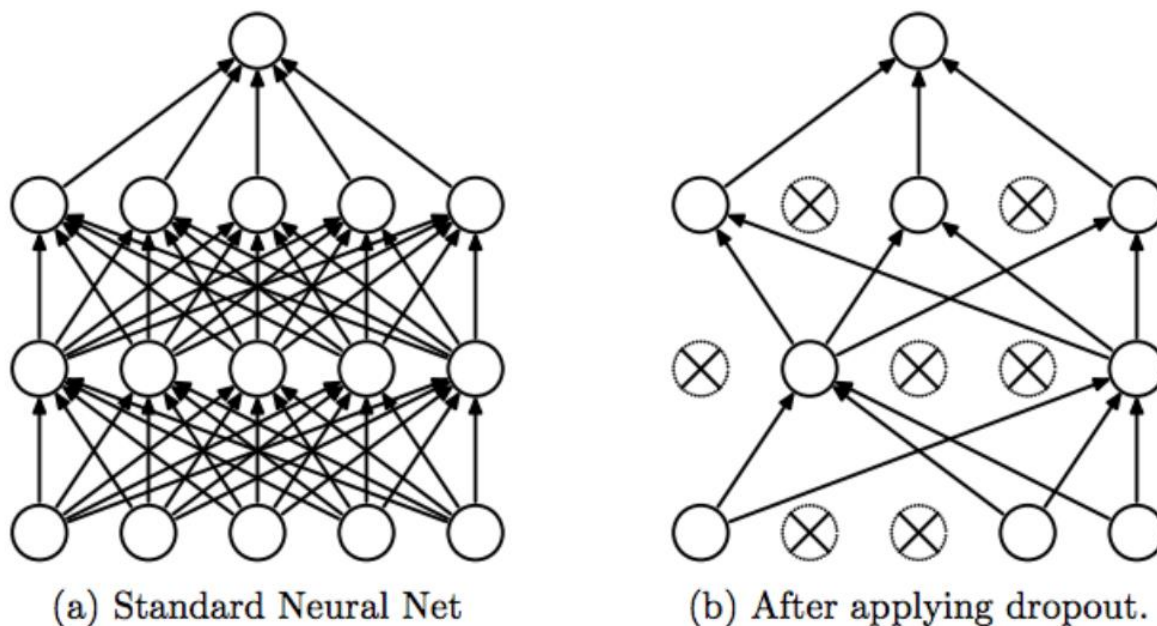
3.2.7. Υπερπροσαρμογή (Overfitting)

Το overfitting αποτελεί μία από τις βασικότερες προκλήσεις στη χρήση τεχνικών μηχανικής μάθησης, ειδικά σε προβλήματα πρόβλεψης χρηματοοικονομικών δεδομένων. Πρόκειται για το φαινόμενο κατά το οποίο ένα μοντέλο προσαρμόζεται υπερβολικά στα δεδομένα εκπαίδευσης, καταγράφοντας ακόμα και τον θόρυβο αντί για τα ουσιώδη μοτίβα. Ως αποτέλεσμα, το μοντέλο παρουσιάζει πολύ υψηλή ακρίβεια στα δεδομένα εκπαίδευσης, αλλά αδυνατεί να γενικεύσει σωστά σε νέα, άγνωστα δεδομένα, οδηγώντας σε χαμηλή προβλεπτική ικανότητα.

Για την αντιμετώπιση του overfitting, εφαρμόζονται τεχνικές όπως η διασταυρούμενη επικύρωση (cross-validation), η κανονικοποίηση (regularization) μέσω όρων όπως οι L1 και L2, καθώς και η χρήση μεθόδων όπως το dropout στα νευρωνικά δίκτυα. Επιπλέον, η επιλογή του κατάλληλου αριθμού παραμέτρων και η χρήση περισσότερων δεδομένων εκπαίδευσης μπορούν να συμβάλουν στη βελτίωση της γενικευσιμότητας του μοντέλου.

Η τεχνική dropout αποτελεί μία από τις πιο αποτελεσματικές μεθόδους για την αντιμετώπιση του overfitting στα νευρωνικά δίκτυα. Στην ουσία, το dropout λειτουργεί ως μια μορφή κανονικοποίησης (regularization), κατά την οποία, σε κάθε επανάληψη της εκπαίδευσης, ορισμένοι νευρώνες του δικτύου απενεργοποιούνται τυχαία, δηλαδή δεν συμμετέχουν στην πρόβλεψη και στην ενημέρωση των βαρών. Αυτό αποτρέπει το δίκτυο από το να εξαρτάται υπερβολικά από συγκεκριμένους νευρώνες και συμβάλλει στην ανάπτυξη ενός πιο ανθεκτικού και γενικεύσιμου μοντέλου. Στη φάση της πρόβλεψης, όλοι οι νευρώνες ενεργοποιούνται ξανά, αλλά οι τιμές τους προσαρμόζονται κατάλληλα ώστε να διατηρηθεί η κλίμακα των εξόδων.

σταθερή. Η προσθήκη dropout σε κρυφά στρώματα ενός νευρωνικού δικτύου έχει αποδειχθεί ιδιαίτερα χρήσιμη σε προβλήματα χρηματοοικονομικής πρόβλεψης, όπου η υψηλή μεταβλητότητα των δεδομένων μπορεί να οδηγήσει σε υπερπροσαρμογή αν δεν εφαρμοστούν κατάλληλες τεχνικές κανονικοποίησης.



Εικόνα 14: Τεχνική Dropout στην Μηχανική Μάθηση

3.3. Επαναλαμβανόμενα Νευρωνικά Δίκτυα (RNN)

Τα Αναδρομικά Νευρωνικά Δίκτυα (Recurrent Neural Networks - RNNs) αποτελούν ένα ιδιαίτερο είδος τεχνητών νευρωνικών δικτύων, που έχει τη δυνατότητα να επεξεργάζεται ακολουθιακά δεδομένα, αξιοποιώντας πληροφορίες από προηγούμενες χρονικές στιγμές. Αυτή η ιδιότητά τους τα καθιστά ιδανικά για περιπτώσεις όπου η σχέση μεταξύ των δεδομένων βασίζεται στη χρονική διαδοχή, όπως συμβαίνει στην ανάλυση φυσικής γλώσσας, στην αναγνώριση φωνής, στην πρόβλεψη χρονοσειρών και στην ανάλυση ιατρικών δεδομένων.

Η βασική αρχιτεκτονική των RNNs περιλαμβάνει ένα σύνολο νευρώνων, όπου κάθε κόμβος συνδέεται τόσο με την είσοδο του δικτύου όσο και με τις προηγούμενες καταστάσεις του, επιτρέποντας την αποθήκευση πληροφοριών για μικρά ή μεγάλα χρονικά διαστήματα. Σε αντίθεση με τα κλασικά πολυεπίπεδα νευρωνικά δίκτυα (MLPs), στα οποία η ροή της πληροφορίας κατευθύνεται μονοσήμαντα από την είσοδο προς την έξοδο, τα RNNs περιλαμβάνουν αναδρομικές συνδέσεις, μέσω των οποίων μπορούν να μεταφέρουν πληροφορίες από παλαιότερα χρονικά σημεία.

Τα Αναδρομικά Νευρωνικά Δίκτυα διαφέρουν από τα συμβατικά πολυεπίπεδα δίκτυα λόγω της ικανότητάς τους να αποθηκεύουν πληροφορίες από προηγούμενα χρονικά βήματα. Αυτό επιτυγχάνεται μέσω μιας αναδρομικής εξίσωσης, που συνδέει την τρέχουσα κατάσταση του κρυφού επιπέδου με την προηγούμενη κατάστασή του, δημιουργώντας μια μορφή εσωτερικής μνήμης.

Η κατάσταση ενός RNN σε ένα συγκεκριμένο χρονικό βήμα t εξαρτάται τόσο από την τρέχουσα είσοδο x_t όσο και από την κατάσταση του προηγούμενου χρονικού βήματος h_{t-1} :

$$h_t = \phi(W_{xh}x_t + W_{hh}h_{t-1} + b_h)$$

όπου:

- W_{xh} : αντιπροσωπεύει τα βάρη που συνδέουν την είσοδο με την κρυφή κατάσταση,
- W_{hh} : είναι ο πίνακας βαρών που ελέγχει τη μεταφορά πληροφορίας από το προηγούμενο χρονικό βήμα στο τρέχον,
- b_h : είναι ένας όρος μετατόπισης (bias),
- $\phi(\cdot)$: είναι η συνάρτηση ενεργοποίησης, που συνήθως είναι η $\tanh$ ή η ReLU .

Η έξοδος του δικτύου σε κάθε χρονική στιγμή προσδιορίζεται από τη σχέση:

$$y_t = W_{hy}h_t + b_y$$

όπου:

- W_{hy} : αντιπροσωπεύει τα βάρη που συνδέουν την είσοδο με την κρυφή κατάσταση,
- b_y : είναι ο πίνακας βαρών που ελέγχει τη μεταφορά πληροφορίας από το προηγούμενο χρονικό βήμα στο τρέχον,

Η εκπαίδευση των RNNs γίνεται μέσω της διαδικασίας της οπισθοδιάδοσης στο χρόνο (Backpropagation Through Time - BPTT), η οποία αποτελεί μια γενικευμένη μορφή του κλασικού αλγορίθμου οπισθοδιάδοσης. Ωστόσο, η αναδρομική φύση αυτών των δικτύων δημιουργεί ορισμένες δυσκολίες. Το πιο σύνηθες πρόβλημα είναι η σταδιακή μείωση ή η εκθετική αύξηση των κλίσεων των συναρτήσεων ενεργοποίησης κατά τη διάρκεια της εκπαίδευσης, γνωστό ως φαινόμενο της εξαφάνισης και της έκρηξης των κλίσεων (vanishing and exploding gradients).

Για την αντιμετώπιση αυτών των προκλήσεων, έχουν αναπτυχθεί διάφορες τεχνικές, όπως η κανονικοποίηση των βαρών και η χρήση εναλλακτικών συναρτήσεων ενεργοποίησης όπως η ReLU . Επιπλέον, πιο εξελιγμένες αρχιτεκτονικές, όπως τα Long Short-Term Memory (LSTM) και Gated Recurrent Units (GRU), έχουν σχεδιαστεί ώστε να διαχειρίζονται καλύτερα μακροχρόνιες εξαρτήσεις.

Για να ξεπεραστούν οι περιορισμοί των κλασικών RNNs, έχουν αναπτυχθεί διάφορες βελτιωμένες εκδοχές τους. Μία από τις σημαντικότερες είναι το Long Short-Term Memory (LSTM), το οποίο εισάγει μηχανισμούς διαχείρισης της μνήμης μέσω πύλων που ελέγχουν ποια πληροφορία θα αποθηκευτεί και ποια θα απορριφθεί. Αυτό καθιστά τα LSTMs ιδιαίτερα αποτελεσματικά στην κατανόηση μακροχρόνιων σχέσεων σε δεδομένα.

Μια πιο απλοποιημένη εκδοχή του LSTM είναι το Gated Recurrent Unit (GRU), το οποίο λειτουργεί παρόμοια αλλά με λιγότερες παραμέτρους, καθιστώντας το πιο αποδοτικό υπολογιστικά. Άλλη σημαντική παραλλαγή είναι τα Bidirectional RNNs, τα οποία επεξεργάζονται δεδομένα τόσο προς την κατεύθυνση του χρόνου όσο και αντίστροφα, παρέχοντας πληρέστερη πληροφόρηση για το υπό ανάλυση πρόβλημα.

Τα Αναδρομικά Νευρωνικά Δίκτυα έχουν βρει εκτεταμένη χρήση σε διάφορους επιστημονικούς και τεχνολογικούς τομείς. Στην επεξεργασία φυσικής γλώσσας, χρησιμοποιούνται για τη μηχανική μετάφραση, την ανάλυση συναισθήματος και τη δημιουργία αυτόματων περιλήψεων κειμένου. Στην αναγνώριση φωνής, αξιοποιούνται από εφαρμογές όπως οι ψηφιακοί βοηθοί Siri και Google Assistant.

Στη χρηματοοικονομική ανάλυση, τα RNNs είναι χρήσιμα για την πρόβλεψη τιμών μετοχών και άλλων χρηματοοικονομικών δεικτών, ενώ στον τομέα της ιατρικής χρησιμοποιούνται για την ανάλυση ηλεκτροκαρδιογραφικών σημάτων και τη διάγνωση παθήσεων μέσω της επεξεργασίας ιατρικών εικόνων. Παράλληλα, στην ανάλυση χρονοσειρών βρίσκουν εφαρμογή σε πεδία όπως η πρόβλεψη καιρού και η ανίχνευση ανωμαλιών σε βιομηχανικά δεδομένα.

3.4. Μακροχρόνια Μνήμη και Βραχυπρόθεσμη Μνήμη (LSTM)

Τα Long Short-Term Memory (LSTM) δίκτυα αποτελούν μια εξειδικευμένη αρχιτεκτονική αναδρομικών νευρωνικών δικτύων (RNNs) που έχει σχεδιαστεί για την αντιμετώπιση του προβλήματος της εξασθένησης και της έκρηξης των κλίσεων (vanishing and exploding gradients). Αυτά τα προβλήματα εμφανίζονται κατά την εκπαίδευση των παραδοσιακών RNNs, καθιστώντας δύσκολη την εκμάθηση εξαρτήσεων που εκτείνονται σε μεγάλες χρονικές ακολουθίες. Τα LSTM δίκτυα προτάθηκαν από τους Hochreiter & Schmidhuber (1997) και έχουν αποδειχθεί εξαιρετικά αποτελεσματικά στην ανάλυση χρονοσειρών, την αναγνώριση φωνής και την επεξεργασία φυσικής γλώσσας.

3.4.1. Αρχιτεκτονική των LSTM Δικτύων

Η κύρια διαφορά μεταξύ των LSTM και των συμβατικών RNNs είναι η παρουσία μνημονικών κυττάρων (memory cells), τα οποία επιτρέπουν τη μακροπρόθεσμη αποθήκευση πληροφοριών. Κάθε LSTM μονάδα περιλαμβάνει τρεις κύριες πύλες ελέγχου (gates):

1. **Πύλη Εισόδου (Input Gate):** Ελέγχει ποια πληροφορία από την τρέχουσα είσοδο και την προηγούμενη κρυφή κατάσταση θα αποθηκευτεί στο memory cell.
2. **Πύλη Λήθης (Forget Gate):** Καθορίζει ποιο τμήμα της προηγούμενης κατάστασης μνήμης θα ξεχαστεί ή θα διατηρηθεί.
3. **Πύλη Εξόδου (Output Gate):** Ελέγχει ποιο μέρος της μνήμης θα χρησιμοποιηθεί για την παραγωγή της εξόδου και τη νέα κρυφή κατάσταση.

Η κατάσταση μνήμης c_t ενημερώνεται μέσω μιας ισορροπημένης διαδικασίας αποθήκευσης και διαγραφής πληροφοριών, επιτρέποντας στο LSTM να διατηρεί χρήσιμα δεδομένα για εκτεταμένες χρονικές περιόδους.

Η λειτουργία ενός LSTM κυττάρου περιγράφεται από τις παρακάτω εξισώσεις:

1. **Πύλη Λήθης (Forget Gate):**

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

2. **Πύλη Εισόδου (Input Gate):**

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

3. **Κυψέλη Μνήμης (Memory Cell Update):**

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

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

4. **Πύλη Εξόδου (Output Gate):**

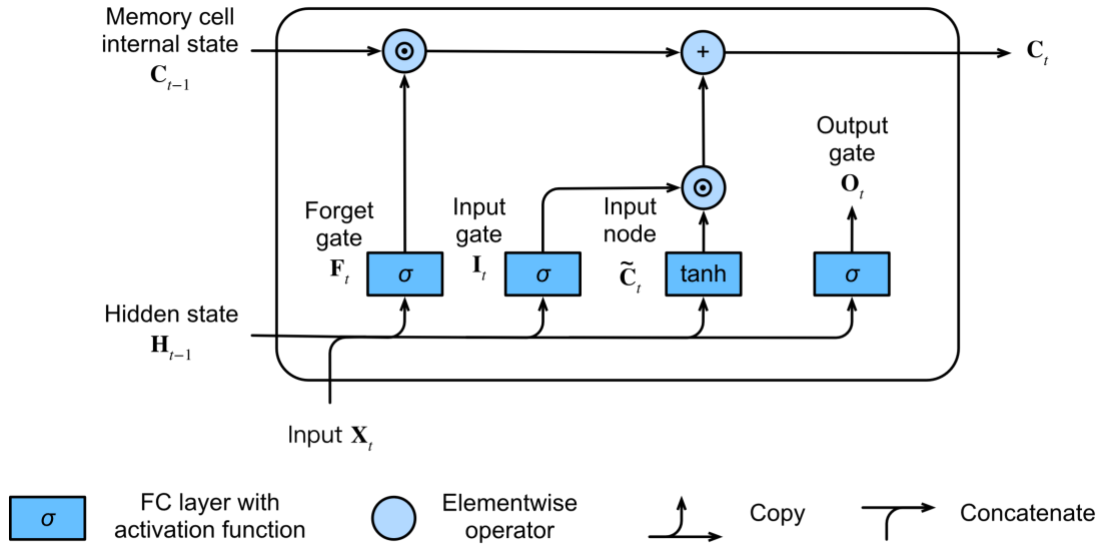
$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o)$$

5. **Κρυφή Κατάσταση (Hidden State Update):**

$$h_t = o_t \odot \tanh(c_t)$$

όπου:

- x_t είναι το διάνυσμα εισόδου,
- h_t η κρυφή κατάσταση,
- c_t η κατάσταση μνήμης,
- W_f, W_i, W_c, W_o είναι τα αντίστοιχα βάρη,
- b_f, b_i, b_c, b_o είναι οι όροι μετατόπισης (bias),
- σ είναι η σιγμοειδής συνάρτηση (sigmoid function), και
- $\odot$ συμβολίζει το στοιχειομετρικό γινόμενο (element-wise multiplication).



Εικόνα 15:Απεικόνιση LSTM κόμβου

Στη λειτουργία ενός κόμβου LSTM, οι πύλες ελέγχουν τη ροή της πληροφορίας, επιτρέποντας στο δίκτυο να διαχειρίζεται τη μνήμη αποτελεσματικά. Η πύλη εισόδου λαμβάνει την τρέχουσα είσοδο x_t και την προηγούμενη κρυφή κατάσταση h_{t-1} , τις οποίες μετασχηματίζει μέσω σιγμοειδούς συνάρτησης σε βάρη που καθορίζουν τη σημασία της νέας πληροφορίας. Παράλληλα, μέσω της $\tanh$, δημιουργείται ένας υποψήφιος νέος καταχωρητής μνήμης $\tilde{c}_t$, ο οποίος συνδυάζεται με το διάνυσμα εισόδου i_t για να ενημερώσει το κελί μνήμης c_t . Η πύλη λήθης ελέγχει ποιες πληροφορίες από την προηγούμενη κατάσταση μνήμης c_{t-1} θα διατηρηθούν, επιτρέποντας την προσαρμογή της μακροπρόθεσμης μνήμης. Τέλος, η πύλη εξόδου χρησιμοποιεί τη νέα κατάσταση μνήμης για να υπολογίσει την κρυφή κατάσταση h_t , η οποία προκύπτει από τον σημειακό πολλαπλασιασμό της εξόδου της $\tanh$ με το διάνυσμα εξόδου o_t . Οι τιμές c_t και h_t , μεταφέρονται στο επόμενο χρονικό βήμα, επιτρέποντας στο δίκτυο να διατηρεί μακροχρόνιες εξαρτήσεις στη ροή της πληροφορίας.

Παρόλο που τα LSTM παρέχουν καλύτερη διαχείριση της μνήμης σε σχέση με τα κλασικά RNNs, έχουν υψηλές υπολογιστικές απαιτήσεις. Μια ελαφρύτερη παραλλαγή τους, τα Gated Recurrent Units (GRUs), προτάθηκε από τους Cho et al. (2014). Τα GRUs διαφέρουν από τα LSTM στο γεγονός ότι συγχωνεύουν τις πύλες εισόδου και λήθης σε μία ενιαία πύλη ενημέρωσης (update gate), μειώνοντας έτσι τον αριθμό των απαιτούμενων παραμέτρων και αυξάνοντας την υπολογιστική τους αποδοτικότητα.

3.5. Αναδρομικές Μονάδες με Πύλες (GRU)

Το Gated Recurrent Unit (GRU) περιλαμβάνει δύο βασικές πύλες: την πύλη ενημέρωσης (Update Gate) και την πύλη εξόδου (Reset Gate). Οι εξισώσεις για το GRU είναι οι εξής:

Πύλη Ενημέρωσης (Update Gate):

Η πύλη ενημέρωσης καθορίζει πόσο από την προηγούμενη κατάσταση h_{t-1} θα διατηρηθεί στην τρέχουσα κατάσταση h_t και πόσο από την νέα πληροφορία πρέπει να ενσωματωθεί στην τρέχουσα κατάσταση.

$$z_t = \sigma(W_z[h_{t-1}, x_t] + b_z)$$

Όπου:

- z_t είναι η έξοδος της πύλης ενημέρωσης.
- W_z είναι τα βάρη της πύλης ενημέρωσης.
- b_z είναι ο όρος μετατόπισης (bias).
- σ είναι η σιγμοειδής συνάρτηση (sigmoid), που εξασφαλίζει ότι η έξοδος της πύλης είναι στο διάστημα $[0, 1]$.

Πύλη Εξόδου (Reset Gate):

Η πύλη εξόδου καθορίζει πόσο από την προηγούμενη κατάσταση h_{t-1} θα επηρεάσει τη νέα κατάσταση μνήμης και πόσο από αυτήν θα χρησιμοποιηθεί για να υπολογιστεί η τρέχουσα έξοδος.

$$r_t = \sigma(W_r[h_{t-1}, x_t] + b_r)$$

Όπου:

- r_t είναι η έξοδος της πύλης εξόδου.
- W_r είναι τα βάρη της πύλης εξόδου.
- b_r είναι ο όρος μετατόπισης (bias).

Υποψήφια Κρυφή Κατάσταση (Candidate Hidden State):

Η υποψήφια κρυφή κατάσταση $\tilde{h}_t$ υπολογίζεται χρησιμοποιώντας την προηγούμενη κατάσταση h_{t-1} και την τρέχουσα είσοδο x_t , υπό την επίδραση της πύλης εξόδου r_t .

$$\tilde{h}_t = \tanh(W_h[r_t \odot h_{t-1}, x_t] + b_h)$$

Όπου:

- $\tilde{h}_t$ είναι η υποψήφια νέα κρυφή κατάσταση.
- $\tanh$ είναι η συνάρτηση υπερβολικής εφαπτομένης (hyperbolic tangent), η οποία διασφαλίζει ότι η έξοδος της υποψήφιας κατάστασης είναι στο διάστημα $[-1, 1]$.
- $\odot$ συμβολίζει το στοιχειομετρικό γινόμενο (element-wise multiplication).

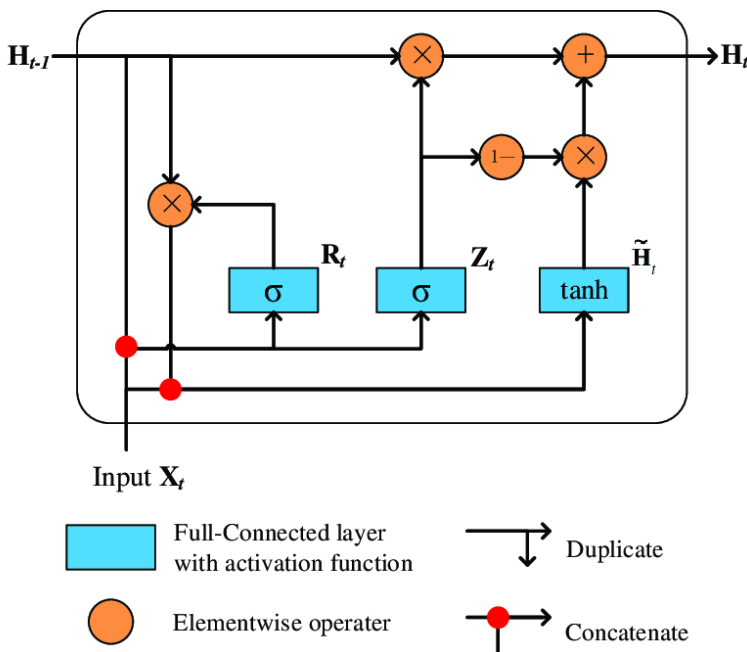
Ενημέρωση της Κρυφής Κατάστασης:

Η τελική κρυφή κατάσταση h_t υπολογίζεται μέσω του στοιχειομετρικού πολλαπλασιασμού της πύλης ενημέρωσης z_t και της υποψήφιας κατάστασης $\tilde{h}_t$, καθώς και της διατήρησης της προηγούμενης κατάστασης h_{t-1} .

$$h_t = (1 - z_t) \odot \tilde{h}_t + z_t \odot h_{t-1}$$

Όπου:

- h_t είναι η νέα κρυφή κατάσταση.
- z_t είναι η έξοδος της πύλης ενημέρωσης.
- $\tilde{h}_t$ είναι η υποψήφια κρυφή κατάσταση.
- $1 - z_t$ επιτρέπει την αποθήκευση της προηγούμενης κατάστασης όταν το z_t είναι μικρό, ενώ το z_t καθοδηγεί πόσο να ανανεωθεί η κρυφή κατάσταση με την υποψήφια κατάσταση



Εικόνα 16: Απεικόνιση GRU κόμβου

3.6. Εφαρμογές GRU σε χρονοσειρές

Τα GRU έχουν αποδειχθεί εξαιρετικά χρήσιμα σε πολλές εφαρμογές, ειδικά όταν πρόκειται για την κατανόηση και πρόβλεψη δεδομένων χρονοσειρών. Η ευελιξία τους τα καθιστά κατάλληλα σε διάφορους τομείς:

- **Χρηματοοικονομικές Αγορές:** Τα GRU χρησιμοποιούνται ευρέως για την πρόβλεψη χρηματιστηριακών τιμών, επιτοκίων και συναλλαγματικών ισοτιμιών. Η δυνατότητά τους να «θυμούνται» παλιότερα μοτίβα τα κάνει ιδιαίτερα κατάλληλα για τη μελέτη της δυναμικής των αγορών και την εκτίμηση μελλοντικών μεταβολών.

- **Ενέργεια και Διαχείριση Πόρων:** Στον τομέα της ενέργειας, τα GRU συμβάλλουν στην πρόβλεψη της κατανάλωσης ρεύματος και της ζήτησης. Τέτοιες προβλέψεις βοηθούν τους παρόχους να προγραμματίσουν καλύτερα την παραγωγή και να διαχειριστούν αποδοτικότερα τους διαθέσιμους πόρους.
- **Ιατρική και Βιοπληροφορική:** Στην υγειονομική περίθαλψη, μοντέλα GRU χρησιμοποιούνται για την ανάλυση σημάτων όπως τα ηλεκτροκαρδιογραφήματα (ECG) και τα εγκεφαλογραφήματα (EEG). Μέσω της συνεχούς παρακολούθησης, μπορούν να εντοπιστούν πρώιμα σημάδια καρδιολογικών ή νευρολογικών διαταραχών.
- **Μετεωρολογία και Κλίμα:** Οι μονάδες GRU είναι χρήσιμες στην πρόβλεψη καιρικών φαινομένων, όπως μεταβολές στη θερμοκρασία ή τα επίπεδα βροχής, καθώς και στη μοντελοποίηση πιο σύνθετων κλιματικών εξελίξεων.
- **Επεξεργασία Φυσικής Γλώσσας και Ομιλίας:** Στον τομέα της γλώσσας, τα GRU αποτελούν βασικά εργαλεία σε εφαρμογές όπως η αυτόματη μετάφραση, η φωνητική αναγνώριση και η δημιουργία λεζάντων για εικόνες, δείχνοντας την ικανότητά τους να κατανοούν και να παράγουν φυσική γλώσσα.

3.7. Προκλήσεις και περιορισμοί

Αν και τα GRU αποτελούν ισχυρά εργαλεία για την ανάλυση χρονοσειρών, η εφαρμογή τους συνοδεύεται από ορισμένες προκλήσεις που δεν πρέπει να αγνοούνται. Μερικές από αυτές είναι:

- **Θόρυβος στα Δεδομένα:** Τα πραγματικά δεδομένα συχνά περιέχουν θόρυβο, δηλαδή ακανόνιστες ή άσχετες τιμές, που μπορεί να αποπροσανατολίσουν το μοντέλο. Αυτό καθιστά δυσκολότερη την εκμάθηση των πραγματικών σχέσεων που κρύβονται στα δεδομένα, οδηγώντας σε ανακριβείς προβλέψεις.
- **Overfitting:** Παρόλο που τα GRU έχουν λιγότερες παραμέτρους από τα LSTM, εξακολουθούν να είναι ευαίσθητα στην υπερπροσαρμογή, ειδικά όταν δουλεύουμε με μικρά σύνολα δεδομένων. Το μοντέλο μπορεί να «μάθει» πολύ καλά τα δεδομένα εκπαίδευσης, αλλά να αποτύχει να γενικεύσει σε νέα, άγνωστα δεδομένα.
- **Υπολογιστικοί Πόροι:** Τα GRU είναι πιο «ελαφριά» από άλλα επαναλαμβανόμενα δίκτυα, αλλά όταν μεγαλώνει το μέγεθος των δεδομένων ή η πολυπλοκότητα του μοντέλου, οι απαιτήσεις σε μνήμη και επεξεργαστική ισχύ αυξάνονται σημαντικά. Η εκπαίδευση μεγάλων μοντέλων μπορεί να γίνει ιδιαίτερα απαιτητική χωρίς την κατάλληλη υποδομή.

Για την αντιμετώπιση αυτών των προκλήσεων, χρησιμοποιούνται διάφορες τεχνικές όπως: το dropout μειώνει την πιθανότητα overfitting, το batch normalization βελτιώνει τη σταθερότητα της εκπαίδευσης, ενώ το early stopping σταματά την εκπαίδευση πριν το μοντέλο αρχίσει να μαθαίνει «θόρυβο». Επιπλέον, η χρήση υβριδικών μοντέλων ή η μείωση των παραμέτρων μπορεί να οδηγήσει σε πιο αποδοτικές λύσεις για εφαρμογές μεγάλης κλίμακας.

4. Sentiment Analysis

4.1. Η επίδραση των συναισθημάτων στην οικονομία

Τα συναισθήματα διαδραματίζουν κρίσιμο ρόλο στις οικονομικές αγορές, επηρεάζοντας την ψυχολογία της αγοράς, τη λήψη επενδυτικών αποφάσεων και τη συνολική δυναμική των οικονομικών κύκλων. Η έννοια της "συναισθηματικής οικονομικής συμπεριφοράς" αναδεικνύει πώς οι ανθρώπινες αντιδράσεις σε συνθήκες αβεβαιότητας ή κινδύνου μπορούν να επηρεάσουν τις οικονομικές δραστηριότητες. Σε μια αγορά όπου κυριαρχούν οι ανθρώπινες αποφάσεις, οι συναισθηματικές αντιδράσεις μπορούν να υπερσχύσουν της λογικής, δημιουργώντας φαινόμενα όπως η υπεραγορά και η υπερπώληση.

Η ψυχολογία της αγοράς αναφέρεται στη συλλογική στάση και συμπεριφορά των συμμετεχόντων σε μια δεδομένη χρονική στιγμή. Οι συμμετέχοντες, είτε είναι επενδυτές είτε καταναλωτές, δεν δρουν πάντα ορθολογικά. Αντίθετα, καθοδηγούνται από συναισθηματικούς παράγοντες, όπως ο φόβος, η απληστία και οι προσδοκίες. Για παράδειγμα, σε περιόδους ευφορίας, η αισιοδοξία μπορεί να ενισχύσει την υπερτίμηση των περιουσιακών στοιχείων, ενώ σε περιόδους κρίσης, ο φόβος οδηγεί σε μαζικές πωλήσεις και πτώση τιμών.

Συχνά, τα συναισθήματα ενισχύονται από εξωτερικούς παράγοντες, όπως οι οικονομικές ειδήσεις, οι γεωπολιτικές εξελίξεις και οι κοινωνικές τάσεις. Ο "CNN effect," όπως αποκαλείται, περιγράφει πώς η συνεχής ροή ειδήσεων επηρεάζει τη συλλογική ψυχολογία, εντείνοντας τις συναισθηματικές αντιδράσεις της αγοράς. Αυτό το φαινόμενο μπορεί να οδηγήσει σε αυτοτροφοδοτούμενους κύκλους, όπου τα συναισθήματα ενισχύουν τις μεταβολές των αγορών πέρα από τα θεμελιώδη δεδομένα.

Ο John Maynard Keynes εισήγαγε τον όρο «animal spirits» για να περιγράψει τον ρόλο των ανθρώπινων συναισθημάτων και ενστίκτων στις οικονομικές αποφάσεις, ιδιαίτερα στις επενδύσεις. Με τον όρο αυτό αναφερόταν στην αυθόρμητη αισιοδοξία και παρόρμηση για δράση που ωθεί τους επιχειρηματίες και επενδυτές να προχωρούν, πέρα από ό,τι θα υπαγόρευε μια καθαρά λογική ανάλυση. Αν αυτή η αυθόρμητη αισιοδοξία «σβήσει», οι επενδυτικές πρωτοβουλίες επίσης ατονούν, υπογραμμίζοντας πόσο κρίσιμο είναι το ψυχολογικό κλίμα για την οικονομική δραστηριότητα [45].

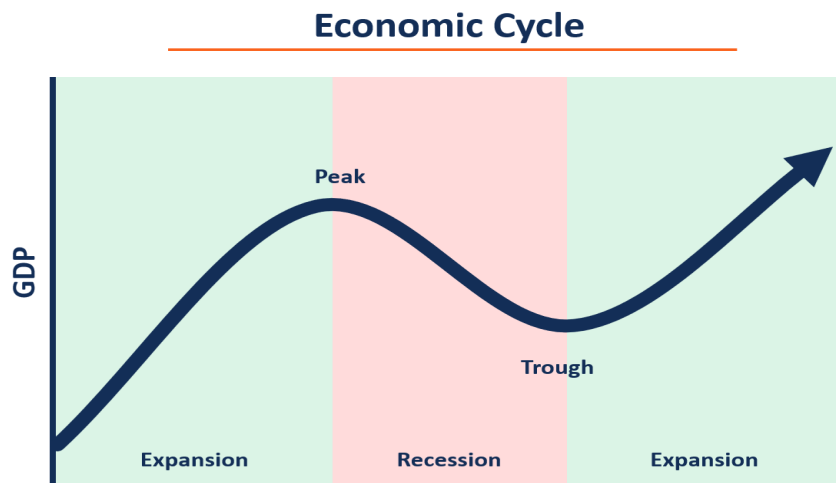
Τέλος, επίδραση των συναισθημάτων στην οικονομία και τις αγορές συχνά ξεπερνά τη λογική και τις θεμελιώδεις αρχές που διέπουν τις αγορές. Σε ορισμένες περιπτώσεις, η παρορμητική συμπεριφορά των συμμετεχόντων, καθοδηγούμενη από φόβο ή απληστία, οδηγεί σε αποφάσεις που δεν ευθυγραμμίζονται με τους βασικούς κανόνες της προσφοράς και της ζήτησης. Αυτή η συμπεριφορά μπορεί να προκαλέσει έντονες αποκλίσεις από την πραγματική αξία των περιουσιακών στοιχείων, δημιουργώντας είτε υπερτιμήσεις είτε υποτιμήσεις στις αγορές.

4.1.1. Οικονομικοί κύκλοι

Οι οικονομικοί κύκλοι αποτελούν έναν από τους βασικούς μηχανισμούς που καθορίζουν τη συνολική πορεία της οικονομικής δραστηριότητας και επηρεάζουν άμεσα την ευημερία των ατόμων και των επιχειρήσεων. Αυτοί οι κύκλοι αντικατοπτρίζουν την αλληλεπίδραση μεταξύ

ζήτησης και προσφοράς, καθώς και τη δυναμική που αναπτύσσεται μέσα από τις επενδύσεις, την κατανάλωση, την παραγωγή και την απασχόληση. Κατά τη διάρκεια της φάσης επέκτασης, παρατηρείται αύξηση της οικονομικής δραστηριότητας, με τους καταναλωτές να δαπανούν περισσότερο, τις επιχειρήσεις να επενδύουν σε κεφαλαιουχικό εξοπλισμό και τις αγορές να αναπτύσσονται. Αυτό συχνά συνοδεύεται από αύξηση των εισοδημάτων, μείωση της ανεργίας και γενικότερα αισιοδοξία στις αγορές.

Οι οικονομικοί κύκλοι (ακμή και ύφεση) δεν καθορίζονται μόνο από θεμελιώδη μεγέθη, αλλά και από τη συλλογική ψυχολογία. Σε περιόδους ανόδου, η ευρέως διαδεδομένη αισιοδοξία και η αγελαία συμπεριφορά μπορεί να φουσκώσουν «φούσκες», ενώ σε περιόδους καθόδου, ο φόβος και η απαισιοδοξία μπορούν να επιτείνουν την ύφεση. Η σύγχρονη έρευνα αναγνωρίζει όλο και περισσότερο ότι οι μακροοικονομικές διακυμάνσεις έχουν ψυχολογικά αίτια: οι μεταβολές στην εμπιστοσύνη και το συλλογικό συναίσθημα επηρεάζουν σημαντικά τις επενδύσεις, την κατανάλωση και συνεπώς τη διαμόρφωση των φάσεων του οικονομικού κύκλου [46].



Εικόνα 17:
Αναπαράσταση τυπικού
οικονομικού κύκλου

Ωστόσο, οι οικονομικοί κύκλοι δεν ακολουθούν πάντα μια ομαλή πορεία, καθώς η φάση της κορύφωσης συχνά οδηγεί σε υπερθέρμανση της οικονομίας, με αυξημένες πληθωριστικές πιέσεις και αυξανόμενους κινδύνους οικονομικών ανισορροπιών. Καθώς η οικονομία περνά στη φάση συρρίκνωσης, οι επενδύσεις μειώνονται, η παραγωγή συρρικνώνεται, και η ανεργία αυξάνεται, δημιουργώντας έναν φαύλο κύκλο μειωμένης κατανάλωσης και ζήτησης. Στο χαμηλότερο σημείο του κύκλου, τη φάση του πυθμένα, η οικονομική δραστηριότητα φτάνει στο ελάχιστο, αλλά δημιουργούνται προϋποθέσεις για ανάκαμψη μέσω ενισχυμένων μέτρων νομισματικής και δημοσιονομικής πολιτικής.

Αυτές οι διακυμάνσεις συνδέονται στενά με τα συναισθήματα των οικονομικών παραγόντων, όπως ο φόβος και η απληστία, τα οποία συχνά μεγεθύνουν τις επιπτώσεις των κύκλων. Για παράδειγμα, σε περιόδους κρίσης, η έλλειψη εμπιστοσύνης μπορεί να επιδεινώσει τη συρρίκνωση, ενώ σε περιόδους ευφορίας, η υπεραισιοδοξία μπορεί να οδηγήσει σε υπερβολικές επενδύσεις και φούσκες στις αγορές. Επιπλέον, οι εξωτερικοί παράγοντες, όπως γεωπολιτικές εξελίξεις, φυσικές καταστροφές ή σημαντικές τεχνολογικές καινοτομίες, μπορούν να επηρεάσουν την ένταση και τη διάρκεια κάθε φάσης. Κατανοώντας τους μηχανισμούς αυτούς, οι επενδυτές, οι επιχειρήσεις και οι κυβερνήσεις μπορούν να προετοιμαστούν καλύτερα για τις

επιπτώσεις των οικονομικών κύκλων και να αξιοποιήσουν τις ευκαιρίες που προσφέρουν οι διαφορετικές φάσεις τους.

Οι οικονομικοί κύκλοι δεν είναι ομοιογενείς ως προς τη διάρκειά τους, καθώς μπορούν να εκτυλίσσονται σε διάφορους χρονικούς ορίζοντες, από βραχυπρόθεσμες διακυμάνσεις μέχρι μακροχρόνιες τάσεις. Οι βραχυπρόθεσμοι οικονομικοί κύκλοι, όπως οι κύκλοι των επιχειρήσεων (business cycles), διαρκούν συνήθως μερικά χρόνια και αντικατοπτρίζουν μεταβολές που σχετίζονται με την προσφορά και τη ζήτηση, τη νομισματική πολιτική και άλλους εξωτερικούς παράγοντες. Από την άλλη πλευρά, οι μακροχρόνιοι κύκλοι διαρκούν δεκαετίες και συνδέονται με δομικές αλλαγές στην οικονομία, όπως οι τεχνολογικές επαναστάσεις ή οι μεταβολές στις αγορές εργασίας.

Επιπλέον, είναι σημαντικό να αναφερθεί ότι μέσα σε έναν κύριο οικονομικό κύκλο μπορεί να δημιουργηθούν υπο-κύκλοι ή μικρότεροι κύκλοι. Για παράδειγμα, κατά τη διάρκεια μιας φάσης επέκτασης σε έναν μακροχρόνιο κύκλο, ενδέχεται να εμφανιστούν βραχυχρόνιες περίοδοι συρρίκνωσης που αντικατοπτρίζουν προσωρινές διαταραχές ή προσαρμογές στις αγορές. Αυτή η έννοια των «κύκλων μέσα σε κύκλους» προσθέτει ένα επίπεδο πολυπλοκότητας στην ανάλυση των οικονομικών φαινομένων, καθιστώντας απαραίτητη τη λεπτομερή παρακολούθηση και κατανόηση των δεδομένων.

4.1.2. Φόβος και απληστία

Ο φόβος και η απληστία είναι δύο από τις πιο σημαντικές συναισθηματικές δυνάμεις που καθοδηγούν τις χρηματοπιστωτικές αγορές. Αυτά τα συναισθήματα, αν και αντικρουόμενα, συχνά αλληλοσυμπληρώνονται και δημιουργούν τις έντονες διακυμάνσεις που χαρακτηρίζουν τις αγορές.

Ο φόβος κυριαρχεί σε περιόδους αβεβαιότητας και κρίσεων. Όταν οι επενδυτές ανησυχούν για την απώλεια κεφαλαίου, συχνά λαμβάνουν βιαστικές αποφάσεις, όπως η πώληση περιουσιακών στοιχείων. Αυτές οι μαζικές πωλήσεις δημιουργούν πίεση στις τιμές, οδηγώντας σε περαιτέρω πτώση. Το φαινόμενο αυτό ενισχύεται όταν η πτώση προκαλείται από εξωτερικά γεγονότα, όπως γεωπολιτικές εντάσεις ή μακροοικονομικές αλλαγές. Επιπλέον, ο φόβος της "χαμένης ευκαιρίας" (fear of missing out - FOMO) μπορεί να ωθήσει τους επενδυτές να εισέλθουν σε αγορές που βρίσκονται σε υπερτίμηση, οδηγώντας σε φαινόμενα φούσκας

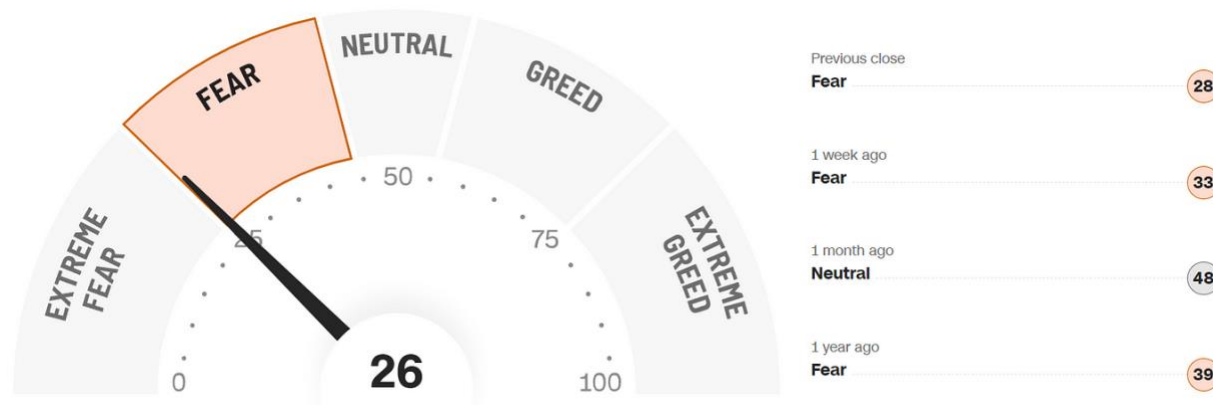
Αντίθετα, η απληστία οδηγεί τους επενδυτές να επιδιώκουν υψηλότερες αποδόσεις, ακόμη και όταν αυτό απαιτεί ανάληψη υπερβολικού ρίσκου. Σε περιόδους ευφορίας, η αισιοδοξία και οι υψηλές προσδοκίες ωθούν τις αγορές σε ανοδικές τάσεις. Ωστόσο, αυτή η υπεραισιοδοξία συχνά καταλήγει σε φαινόμενα υπερτίμησης, καθώς οι επενδυτές αγνοούν τους κινδύνους. Το αποτέλεσμα είναι η δημιουργία χρηματοοικονομικών φουσκών, οι οποίες τελικά καταρρέουν όταν οι τιμές επιστρέφουν στα θεμελιώδη επίπεδα

Το δίπολο φόβου και απληστίας θεωρείται ευρέως ότι βρίσκεται πίσω από τις διακυμάνσεις των αγορών. Οι οικονομικοί ιστορικοί όπως ο Hyman Minsky και ο Charles Kindleberger έχουν επισημάνει ότι σε φάσεις χρηματοπιστωτικής μανίας επικρατεί μια μεταδοτική κερδοσκοπική ευφορία (απληστία) που ωθεί τους επενδυτές σε υπερβολικό ρίσκο. Αντίθετα, κατά την

κατάρρευση η διάθεση μεταστρέφεται σε υπερβολική απαισιοδοξία και αποστροφή κινδύνου (φόβος), επιταχύνοντας την πτώση [47].

Ο συνδυασμός φόβου και απληστίας δημιουργεί έναν φαύλο κύκλο, όπου η εναλλαγή των συναισθημάτων εντείνει τη μεταβλητότητα των αγορών. Αυτή η ψυχολογική δυναμική αποτελεί σημαντικό παράγοντα που πρέπει να λαμβάνουν υπόψη οι επενδυτές και οι διαχειριστές κεφαλαίων.

Οι επενδυτές κατά τις οικονομικές αναλύσεις τους χρησιμοποιούν συνήθως τον δείκτη “fear and greed index” ώστε να καταλάβουν την κατάσταση φόβου/απληστίας στην οποία βρίσκονται οι συμμετέχοντες στις αγορές. Με βάση την τιμή του δείκτη αυτού μπορούν να προσαρμόσουν τις κινήσεις τους, είτε επιτρέποντας παραπάνω ρίσκο είτε κρατώντας μια πιο συντηρητική στάση.



Εικόνα 18: Δείκτης απληστίας-φόβου

4.1.3. Το συναίσθημα ως ποιοτικό μέγεθος

Το συναίσθημα αποτελεί ένα κατεξοχήν ποιοτικό μέγεθος, το οποίο αντανακλά την υποκειμενική αντίληψη, τις προσδοκίες και τις συναισθηματικές αντιδράσεις των συμμετεχόντων στις αγορές. Η φύση του συναίσθηματος, ως υποκειμενική και συχνά άμεσα μεταβαλλόμενη, δημιουργεί σημαντικές δυσκολίες στην ποσοτικοποίησή του. Ενώ εργαλεία όπως η ανάλυση κειμένων από οικονομικές ειδήσεις, αναρτήσεις στα κοινωνικά δίκτυα ή ακόμα και τα σχόλια σε επενδυτικά φόρουμ μπορούν να παρέχουν ενδείξεις για το γενικό κλίμα της αγοράς, τα αποτελέσματα επηρεάζονται από τη μεροληψία (bias) που μπορεί να εισάγεται από την ίδια τη μέθοδο ανάλυσης. Επιπλέον, το συναίσθημα δεν είναι στατικό· μεταβάλλεται δυναμικά με βάση τα εξωτερικά γεγονότα και τις προσωπικές ερμηνείες των συμμετεχόντων.

Το επενδυτικό συναίσθημα αποτελεί μια ποιοτική διάσταση – ουσιαστικά τη διάθεση και την εμπιστοσύνη των συμμετεχόντων – και επομένως είναι δύσκολο να αποτυπωθεί με συγκεκριμένους αριθμούς. Σε αντίθεση με χρηματοοικονομικούς δείκτες ή άλλα μετρήσιμα μεγέθη, το συναίσθημα είναι μια ψυχολογική κατάσταση, η οποία συνήθως εκτιμάται εμμέσως μέσω δεικτών κλίματος, ερευνών εμπιστοσύνης ή άλλων έμμεσων μετρήσεων όπως ο δείκτης μεταβλητότητας και ο όγκος συναλλαγών. Η δυσκολία ποσοτικοποίησης αυτού του παράγοντα είναι εμφανής: το ζητούμενο πλέον δεν είναι αν το συναίσθημα επηρεάζει τις τιμές, αλλά πώς μπορεί να μετρηθεί και να αποτιμηθεί ποσοτικά η επίδρασή του [48].

Οι μέθοδοι που χρησιμοποιούνται για την ανάλυση συναισθήματος, όπως η επεξεργασία φυσικής γλώσσας (Natural Language Processing - NLP) ή η ανάλυση δεδομένων από κοινωνικά δίκτυα, προσφέρουν χρήσιμα εργαλεία για τη διερεύνηση του γενικού κλίματος στις αγορές. Ωστόσο, τα δεδομένα αυτά συχνά περιέχουν θόρυβο, όπως υπερβολικές ή μεροληπτικές απόψεις, που μπορεί να οδηγήσουν σε ανακριβείς εκτιμήσεις. Επιπλέον, το συναίσθημα δεν είναι στατικό μέγεθος· εξελίσσεται δυναμικά και συχνά μεταβάλλεται απότομα ανάλογα με τις συνθήκες που επικρατούν στις αγορές ή τα απρόβλεπτα γεγονότα. Για παράδειγμα, μια ξαφνική ανακοίνωση κερδών που ξεπερνά τις προσδοκίες μπορεί να μετατρέψει την αβεβαιότητα σε αισιοδοξία, ενώ μια αρνητική είδηση μπορεί να προκαλέσει απότομη στροφή προς τον φόβο.

Η δυσκολία στην ποσοτικοποίηση του συναισθήματος απορρέει επίσης από το γεγονός ότι συχνά εκφράζεται μέσα από διαφορετικά κανάλια και τρόπους, όπως οι δηλώσεις των αναλυτών, οι αναρτήσεις στα μέσα κοινωνικής δικτύωσης ή οι συναλλακτικές συμπεριφορές στις αγορές. Κάθε μία από αυτές τις πηγές εμπεριέχει το δικό της βαθμό αβεβαιότητας και υποκειμενικότητας, γεγονός που καθιστά την ερμηνεία και τη στατιστική μοντελοποίηση ιδιαίτερα απαιτητική. Παρόλα αυτά, η σημασία της ανάλυσης συναισθήματος παραμένει τεράστια, καθώς παρέχει πολύτιμες πληροφορίες για τη διάθεση των επενδυτών και μπορεί να συμβάλει στην πρόβλεψη τάσεων και μεταβολών στις αγορές. Η ανάπτυξη πιο εξελιγμένων αλγορίθμων και μεθοδολογιών που συνδυάζουν ποιοτικά και ποσοτικά δεδομένα είναι απαραίτητη για την καλύτερη ενσωμάτωση του συναισθήματος στα χρηματοοικονομικά μοντέλα και τη βελτίωση της λήψης αποφάσεων.

4.2. Επεξεργασία Φυσικής Γλώσσας για εξαγωγή συναισθήματος

Η Επεξεργασία Φυσικής Γλώσσας (Natural Language Processing - NLP) αποτελεί έναν από τους σημαντικότερους τομείς της Τεχνητής Νοημοσύνης και έχει αναδειχθεί ως κρίσιμο εργαλείο για την εξαγωγή χρήσιμων πληροφοριών από αδόμητο κείμενο. Στο πλαίσιο των χρηματοοικονομικών αναλύσεων, το NLP επιτρέπει την αυτοματοποιημένη επεξεργασία μεγάλου όγκου εταιρικών εγγράφων, ειδήσεων, αναφορών και σχολίων επενδυτών, με στόχο την εξαγωγή συναισθηματικής πληροφορίας (sentiment analysis) [49]. Η ανάλυση συναισθήματος είναι μια διαδικασία που στοχεύει στον εντοπισμό και την ποσοτικοποίηση του συναισθηματικού τόνου ενός κειμένου, είτε ως θετικού, αρνητικού ή ουδέτερου, είτε μέσω πιο λεπτομερών μοντέλων που συλλαμβάνουν την ένταση και τη δυναμική των συναισθημάτων.

Στα χρηματοοικονομικά, η συναισθηματική ανάλυση έχει ιδιαίτερη σημασία καθώς οι αντιλήψεις και τα συναισθήματα των επενδυτών μπορούν να επηρεάσουν τις αγορές και να οδηγήσουν σε διακυμάνσεις των τιμών των μετοχών. Με τη χρήση προηγμένων μοντέλων NLP, όπως το BERT (Bidirectional Encoder Representations from Transformers) και εξειδικευμένων παραλλαγών του όπως το FinBERT, οι ερευνητές μπορούν να αναλύσουν εταιρικές ανακοινώσεις, οικονομικές εκθέσεις και καταθέσεις ρυθμιστικών αρχών (SEC filings), εξάγοντας δομημένα sentiment scores. Αυτά τα μοντέλα, εκπαιδευμένα σε χρηματοοικονομικά δεδομένα, επιτρέπουν τη λεπτομερή αξιολόγηση του τρόπου με τον οποίο οι λεκτικές διατυπώσεις και οι αφηγηματικές στρατηγικές των εταιρειών αντικατοπτρίζουν τη μελλοντική τους πορεία.

Η διαδικασία εξαγωγής συναισθήματος προϋποθέτει συγκεκριμένα στάδια, από την προεπεξεργασία του κειμένου (tokenization, stopword removal, stemming/lemmatization) έως

την εφαρμογή τεχνικών βαθιάς μάθησης. Σε αυτή την ενότητα, παρουσιάζεται η μεθοδολογία που ακολουθήθηκε για την εξαγωγή συναισθήματος από τις SEC καταθέσεις, καθώς και η θεωρητική βάση που υποστηρίζει τις εφαρμογές του NLP στην ανάλυση χρηματοοικονομικών εγγράφων.

4.2.1. Προεπεξεργασία κειμένου

Η προεπεξεργασία αποτελεί κρίσιμο στάδιο της ανάλυσης συναισθήματος, καθώς εξαλείφει περιττά ή θορυβώδη στοιχεία του κειμένου, βελτιώνοντας τη σημασιολογική συνοχή και μειώνοντας την πολυπλοκότητα της γλωσσικής αναπαράστασης (Manning et al., 2008). Οι κύριες τεχνικές που εφαρμόζονται περιλαμβάνουν:

4.2.1.1. Λεξικολογική Ανάλυση

Η λεξικολογική ανάλυση (Lexical Analysis), γνωστή και ως τοκενοποίηση (tokenization), είναι η διαδικασία διαχωρισμού ενός κειμένου σε διακριτές μικρές μονάδες, τα λεγόμενα tokens. Συνήθως αυτά τα tokens αντιστοιχούν σε λέξεις (αλλά μπορεί επίσης να είναι αριθμοί, σημεία στίξης ή άλλα σύμβολα) τα οποία μπορούν στη συνέχεια να αναλυθούν περαιτέρω από αλγορίθμους μηχανικής μάθησης. Η σωστή αναγνώριση των ορίων μεταξύ λέξεων και συμβόλων επιτρέπει στο σύστημα να κατανοήσει τη δομή του κειμένου και να εξάγει χρήσιμα χαρακτηριστικά (features) από αυτό. Η τοκενοποίηση αποτελεί θεμελιώδες προκαταρκτικό βήμα σε σχεδόν όλες τις εφαρμογές επεξεργασίας φυσικής γλώσσας – συμπεριλαμβανομένης της ανάλυσης συναισθήματος – καθώς χωρίς αυτήν δεν είναι εφικτή η αποτελεσματική επεξεργασία του κειμένου από τον υπολογιστή [50]. Με άλλα λόγια, η λεξικολογική ανάλυση παρέχει τη βάση πάνω στην οποία θα εφαρμοστούν οι επόμενες διαδικασίες φιλτραρίσματος και μετασχηματισμού του κειμένου.

Stocks	declined	,	and	investors	remained	cautious	throughout	the	crisis	.
--------	----------	---	-----	-----------	----------	----------	------------	-----	--------	---

Εικόνα 19: Διάσπαση πρότασης σε tokens.

4.2.1.2. Αφαίρεση Τετριμμένων Λέξεων

Μετά την τοκενοποίηση, ένα σύνηθες βήμα είναι η αφαίρεση των λεγόμενων τετριμμένων λέξεων (Stopwords Removal) από το κείμενο. Οι λέξεις αυτές είναι συνήθως πολύ κοινές (π.χ. άρθρα, σύνδεσμοι, αντωνυμίες) και εμφανίζονται με μεγάλη συχνότητα, αλλά μεταφέρουν ελάχιστη σημασιολογική πληροφορία ή συναισθηματικό περιεχόμενο. Για παράδειγμα, λέξεις όπως «και», «το», «είναι» στα ελληνικά (ή αντίστοιχα “and”, “the”, “is” στα αγγλικά) σπάνια συμβάλλουν στη διάκριση θετικού ή αρνητικού συναισθήματος σε μια πρόταση. Η απομάκρυνσή τους από το κείμενο μπορεί να μειώσει τον «θόρυβο» στα δεδομένα και να βελτιώσει την απόδοση μοντέλων ταξινόμησης συναισθήματος, καθώς το μοντέλο επικεντρώνεται σε πιο σημαντικές λέξεις-φορείς συναισθήματος. Η αφαίρεση των stopwords βελτιώνει την απόδοση των μοντέλων, μειώνοντας τη διάσταση του συνόλου δεδομένων και αυξάνοντας την ταχύτητα επεξεργασίας (Bird et al., 2009). Συνεπώς, η αφαίρεση τετριμμένων λέξεων θεωρείται ενδεδειγμένη ως ένα απλό προληπτικό βήμα καθαρισμού, αν και η συμβολή της πρέπει να αξιολογείται κατά περίπτωση ανάλογα με το μοντέλο ανάλυσης συναισθήματος που χρησιμοποιείται.

4.2.1.3. Αποκατάληξη και Λημματοποίηση

Οι γλώσσες χαρακτηρίζονται από μορφολογικές παραλλαγές – για παράδειγμα, μια ρίζα λέξης μπορεί να εμφανίζεται με διαφορετικές καταλήξεις (χρόνοι ρημάτων, πληθυντικοί ουσιαστικών, παράγωγα επίθετα κ.ο.κ.). Οι τεχνικές αποκατάληξης και λημματοποίησης (Stemming και Lemmatization) στοχεύουν στο να μειώσουν αυτές τις παραλλαγές, μετατρέποντας τις λέξεις στις βασικές τους μορφές.

Η αποκατάληξη (stemming) είναι η αυτοματοποιημένη διαδικασία με την οποία μια λέξη κόβεται στη ρίζα της (τον stem) μέσω αφαίρεσης επιθημάτων (καταλήξεων) – και μερικές φορές προθημάτων – βάσει προκαθορισμένων κανόνων. Ο σκοπός είναι διαφορετικές κλίσεις ή παράγωγα μιας λέξης να καταλήγουν σε μια κοινή μορφή, ώστε να θεωρούνται ισοδύναμες κατά την επεξεργασία. Για παράδειγμα, οι αγγλικές λέξεις “playing”, “played” και “plays” μπορούν όλες να μειωθούν στο ίδιο stem “play”. Οι πιο διαδεδομένοι αλγόριθμοι stemming, όπως ο Porter Stemmer, χρησιμοποιούν σύνολα κανόνων για την αφαίρεση τυπικών καταλήξεων και άλλων μορφολογικών στοιχείων. Η διαδικασία αυτή είναι απολύτως μηχανιστική – αφαιρεί χαρακτήρες από το τέλος (ή την αρχή) της λέξης χωρίς γνώση της σημασίας – και έτσι το αποτέλεσμα (stem) ενδέχεται να μην αποτελεί έγκυρη λέξη. Για παράδειγμα, εφαρμόζοντας stemming στη λέξη “producing” μπορεί να προκύψει το stem “produc”, το οποίο δεν είναι αυτόνομη λέξη. Παρόλα αυτά, η αποκατάληξη είναι υπολογιστικά αποδοτική και βοηθά στη μείωση της διασποράς των χαρακτηριστικών, καθώς λέξεις με κοινή ρίζα αντιμετωπίζονται ως ένα στοιχείο από το μοντέλο.

Η λημματοποίηση (lemmatization), από την άλλη πλευρά, είναι μια πιο σύνθετη διαδικασία που αξιοποιεί λεξικογραφική και μορφολογική γνώση προκειμένου να μετατρέψει μια δοσμένη λέξη στην κανονική λεξική μορφή της – δηλαδή στο λήμμα όπως θα εμφανιζόταν σε ένα λεξικό. Η λημματοποίηση λαμβάνει υπόψη το μέρος του λόγου (π.χ. αν μια λέξη είναι ουσιαστικό, ρήμα, επίθετο κλπ.) και χρησιμοποιεί πόρους όπως μορφολογικά λεξικά ή βάσεις γνώσεων (π.χ. το WordNet στα αγγλικά) ώστε να εντοπίσει τη σωστή βασική μορφή. Για παράδειγμα, οι αγγλικές λέξεις “plays”, “played”, “playing” αναγνωρίζονται όλες με λημματοποίηση ως “play” – τη μορφή ενεστώτα του ρήματος to play. Σε αντίθεση με την αποκατάληξη, η οποία μπορεί να δώσει μη πραγματικές ρίζες, η λημματοποίηση εγγυάται ότι η προκύπτουσα μορφή είναι έγκυρη λέξη (lemma). Αυτό επιτυγχάνεται διότι η μέθοδος αυτή ουσιαστικά κατανοεί τη λέξη στο γλωσσικό της πλαίσιο (π.χ. αναγνωρίζοντας ότι “better” ως επίρρημα αντιστοιχεί στο lemma “well”, ενώ ως επίθετο στο lemma “good”). Η λημματοποίηση είναι υπολογιστικά ακριβότερη από την αποκατάληξη, αλλά συχνά πιο ακριβής. Μελέτες έχουν δείξει ότι για την αγγλική γλώσσα η λημματοποίηση τείνει να υπερτερεί της απλής αποκατάληξης σε ανεπίσημα κείμενα (π.χ. tweets), διότι το επιπλέον πλαίσιο που λαμβάνει υπόψη (γραμματική κατηγορία) τη βοηθά να χειριστεί καλύτερα τις μορφολογικές παραλλαγές. Γενικότερα, η εφαρμογή είτε stemming είτε lemmatization (ή συνδυασμού τους) μειώνει την πολυπλοκότητα του λεξιλογίου που πρέπει να μάθει ένα μοντέλο, ομαδοποιώντας τις διαφορετικές μορφές μιας λέξης σε μία αντιπροσώπευση. Αυτή η μείωση της διαστασιμότητας των δεδομένων εισόδου μπορεί να βελτιώσει την ακρίβεια και την αποδοτικότητα μοντέλων μηχανικής μάθησης σε εργασίες ανάλυσης κειμένου– συμπεριλαμβανομένης της ανάλυσης συναισθήματος [51]. Με απλά λόγια, με το stemming/lemmatization το μοντέλο “βλέπει” λιγότερα διακριτά χαρακτηριστικά, χωρίς ωστόσο να χάνει ουσιαστική πληροφορία, αφού λέξεις με την ίδια ρίζα/λήμμα αντιμετωπίζονται ως ισοδύναμες.

4.2.1.4. Αλγόριθμος Τμηματοποίησης WordPiece

Οι παραπάνω τεχνικές προεπεξεργασίας δρουν σε επίπεδο ολόκληρων λέξεων. Μια πιο πρόσφατη προσέγγιση στην ανάλυση κειμένου – που έχει καθιερωθεί στα σύγχρονα μοντέλα βαθιάς μάθησης – είναι η τοκενοποίηση υπο-λέξεων (subword tokenization). Ο αλγόριθμος WordPiece αποτελεί μια τέτοια μέθοδο, όπου οι λέξεις δεν αντιμετωπίζονται ως αδιαίρετες μονάδες αλλά διασπώνται σε μικρότερα τμήματα (υπο-λέξεις ή wordpieces) βάσει της στατιστικής συχνότητας εμφάνισης υπο-μορίων στο σώμα κειμένων εκπαίδευσης [50].

Τα υπολεξιακά αυτά tokens μπορούν να θεωρηθούν ανάλογα των μορφημάτων της γλώσσας, καθώς διατηρούν σε μεγάλο βαθμό σημασιολογική πληροφορία (π.χ. η ρίζα “happy” στο “unhappiness” δηλώνει θετικό συναίσθημα παρά την προσθήκη αρνητικού προθέματος) και ταυτόχρονα αντιμετωπίζουν αποτελεσματικά το πρόβλημα των άγνωστων ή σπάνιων λέξεων. Επειδή μια άγνωστη λέξη μπορεί να αναλυθεί στα επιμέρους γνωστά τμήματά της, μειώνονται οι περιπτώσεις out-of-vocabulary (λέξεων εκτός λεξιλογίου) ακόμη και με σχετικά μικρό μέγεθος λεξιλογίου υπο-λέξεων. Για παράδειγμα, η αγγλική λέξη “unhappiness” μπορεί να διασπαστεί στα μέρη “un”, “happy”, “ness”, έτσι ώστε αν η λέξη ως σύνολο δεν έχει εμφανιστεί στην εκπαίδευση, το σύστημα να μπορεί να αναγνωρίσει τη συμβολή του προθέματος “un” (άρνηση), της ρίζας “happy” (χαρά) και της κατάληξης “-ness” (ουσιαστικοποίηση) στη συνολική σημασία/συναίσθημα.

Η υπο-λεξιακή τοκενοποίηση του WordPiece έχει καταστεί συνηθισμένη πρακτική σε πολλά σύγχρονα μοντέλα, ιδιαίτερα σε μοντέλα βασισμένα σε Transformers που επιτυγχάνουν υψηλές επιδόσεις σε εργασίες ανάλυσης κειμένου (μεταξύ των οποίων και η ανάλυση συναισθήματος). Χαρακτηριστικό παράδειγμα αποτελεί το BERT (Bidirectional Encoder Representations from Transformers): το μοντέλο αυτό χρησιμοποιεί tokenization WordPiece με λεξιλόγιο ~30.000 υπο-λέξεων για την αναπαράσταση των εισόδων του. Κάθε ακολουθία κειμένου κωδικοποιείται σε tokens WordPiece, ξεκινώντας πάντα με ένα ειδικό token ταξινόμησης ([CLS]) και χωρίζοντας τις λέξεις στις αντίστοιχες υπο-λέξεις του λεξιλογίου. Η υιοθέτηση του WordPiece επιτρέπει στο BERT και σε παρόμοια μοντέλα να χειρίζονται αποτελεσματικά σπάνιες ή σύνθετες λέξεις, εφόσον ακόμη και άγνωστες λέξεις μπορούν να αναλυθούν στα δομικά τους στοιχεία. Με αυτόν τον τρόπο, μεγιστοποιείται η κάλυψη του λεξιλογίου (vocabulary coverage) χωρίς να απαιτείται ένα τεράστιο λεξικό όλων των πιθανών λέξεων, επιτυγχάνοντας μια ισορροπία μεταξύ λεξιλογικού μεγέθους και εκφραστικής δύναμης του μοντέλου.



Εικόνα 20: Παράδειγμα δημιουργίας διακριτικών στο WordPiece. Πηγή: Xinying Song

4.2.2. Word Embeddings

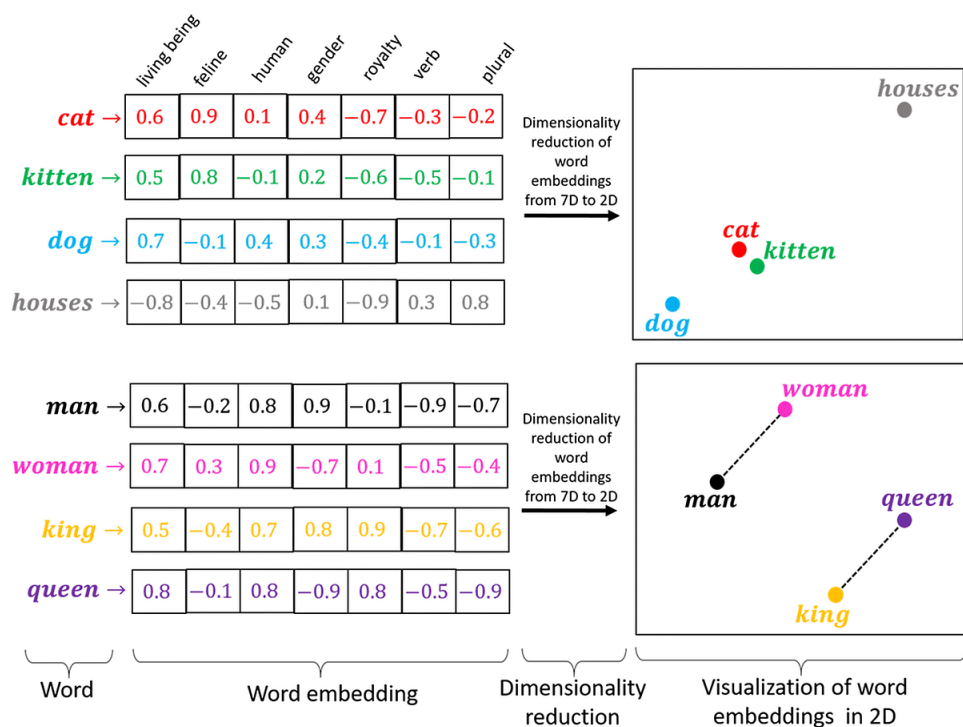
Η αναπαράσταση των λέξεων σε έναν πολυδιάστατο χώρο μέσω των Word Embeddings αποτελεί θεμελιώδη τεχνική στην επεξεργασία φυσικής γλώσσας (Natural Language Processing - NLP), καθώς επιτρέπει τη μοντελοποίηση της σημασιολογίας και της συντακτικής σχέσης μεταξύ των λέξεων. Σε αντίθεση με τις παραδοσιακές τεχνικές αναπαράστασης, όπως τα Bag-of-Words (BoW) ή TF-IDF, οι τεχνικές Word Embeddings επιτρέπουν τη σύλληψη των σημασιολογικών ομοιοτήτων μεταξύ των λέξεων, προβάλλοντας τις λέξεις σε συνεχείς διανυσματικούς χώρους όπου παρόμοιες έννοιες βρίσκονται εγγύτερα μεταξύ τους [52].

Οι πιο διαδεδομένοι αλγόριθμοι που έχουν χρησιμοποιηθεί για την εκπαίδευση Word Embeddings περιλαμβάνουν:

- **Word2Vec:** Αναπτύχθηκε από τους Mikolov et al. [53] και βασίζεται σε δύο μοντέλα, το Continuous Bag of Words (CBOW) και το Skip-Gram. Το CBOW προβλέπει τη λέξη-στόχο με βάση το περιβάλλον της, ενώ το Skip-Gram μαθαίνει να προβλέπει το περιβάλλον μιας λέξης-στόχου.
- **GloVe (Global Vectors for Word Representation):** Παρουσιάστηκε από τους Pennington et al. [54] και χρησιμοποιεί μια προσέγγιση βασισμένη σε πίνακες κατανομής συχνοτήτων των λέξεων, παρέχοντας πιο σταθερές σημασιολογικές αναπαραστάσεις σε σχέση με το Word2Vec.
- **FastText:** Μια βελτιωμένη έκδοχή του Word2Vec, που εισήγαγαν οι Bojanowski et al. [55], η οποία λαμβάνει υπόψη υπολέξεις (subword embeddings) και είναι ιδιαίτερα χρήσιμη για γλώσσες με σύνθετη μορφολογία ή περιορισμένα δεδομένα εκπαίδευσης.
- **BERT (Bidirectional Encoder Representations from Transformers):** Αναπτύχθηκε από τους Devlin et al. [56] και διαφέρει από τα προηγούμενα μοντέλα, καθώς χρησιμοποιεί αμφίδρομη μάθηση (bidirectional learning) μέσω Transformers, λαμβάνοντας υπόψη τόσο τις προηγούμενες όσο και τις επόμενες λέξεις κατά την εκπαίδευση.

Τα Word Embeddings έχουν αποδειχθεί εξαιρετικά αποτελεσματικά σε εφαρμογές NLP, όπως ανάλυση συναισθήματος (sentiment analysis), αναγνώριση οντοτήτων (Named Entity Recognition - NER), και μηχανική μετάφραση (machine translation) [57]. Η ικανότητά τους να διατηρούν σημασιολογικές σχέσεις μεταξύ λέξεων είναι ιδιαίτερα χρήσιμη στη χρηματοοικονομική ανάλυση, όπου μικρές διαφορές στη διατύπωση μπορούν να επηρεάσουν σημαντικά τη συνολική ερμηνεία ενός κειμένου [49].

Η επιλογή του κατάλληλου Word Embedding εξαρτάται από τα δεδομένα και το εκάστοτε πρόβλημα. Τα στατικά embeddings, όπως το Word2Vec και το GloVe, είναι αποδοτικά σε κλασικές εφαρμογές NLP, ενώ τα δυναμικά embeddings, όπως το BERT, είναι προτιμότερα όταν απαιτείται κατανόηση του πλαισίου σε προτάσεις με περίπλοκη σύνταξη [58].



Εικόνα 21: Υπολογισμός και οπτική αναπαράσταση των word embeddings

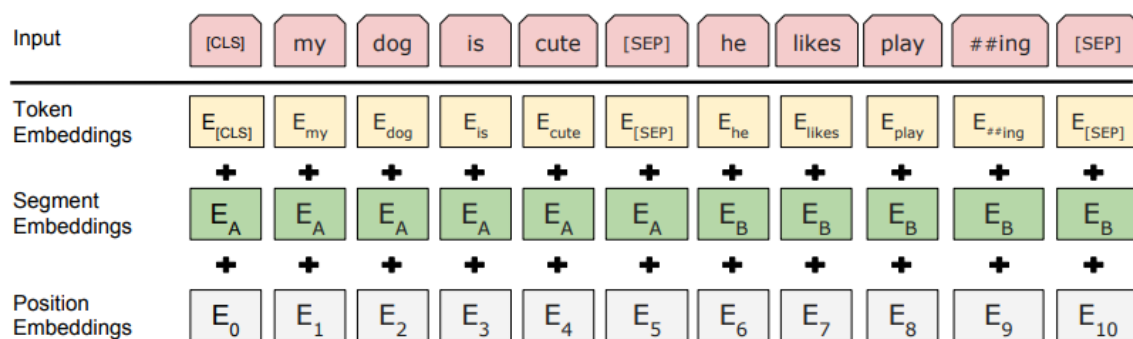
4.3. Το BERT και η προσέγγισή του στην κατανόηση κειμένου

Το BERT (Bidirectional Encoder Representations from Transformers) είναι ένα μοντέλο γλωσσικής αναπαράστασης που έφερε επανάσταση στην κατανόηση κειμένου. Σε αντίθεση με προηγούμενα μοντέλα όπως εκείνο του Radford et al. (2018), τα οποία διάβαζαν το κείμενο μονοδρομικά, το BERT αξιοποιεί αμφίδρομη κωδικοποίηση – δηλαδή λαμβάνει υπόψη τα συμφραζόμενα τόσο από τα αριστερά όσο και από τα δεξιά κάθε λέξης [59]. Μέσω αυτής της διαδικασίας, το BERT παράγει πλαισιοεξαρτώμενες ενσωματώσεις λέξεων (contextualized word embeddings), όπου η σημασιολογική αναπαράσταση κάθε λέξης προσαρμόζεται δυναμικά ανάλογα με το περιβάλλον της. Αυτό του δίνει τη δυνατότητα να συλλαμβάνει λεπτές εννοιολογικές διαφορές, π.χ. να αποδίδει διαφορετική σημασία στη λέξη “τράπεζα” όταν συνοδεύεται από οικονομικούς όρους έναντι γεωγραφικών όρων. Σημαντικό χαρακτηριστικό του BERT είναι ότι προεκπαιδεύεται (pre-trained) σε τεράστια μη επισημασμένα corpora (όπως το Wikipedia και βιβλία) με δύο αυτοεπιβλεπόμενους στόχους: την μάσκαρισμένη μοντελοποίηση γλώσσας (Masked Language Modeling - πρόβλεψη τυχαία καλυμμένων λέξεων) και την πρόβλεψη επόμενης πρότασης (Next Sentence Prediction) [60]. Αυτή η προεκπαίδευση επιτρέπει στο BERT να μάθει γενικές ιδιότητες της γλώσσας, αποθηκεύοντας γνώση που μπορεί να μεταφερθεί σε ποικιλία εργασιών κατανόησης.

Κατά τη φάση του fine-tuning, το προεκπαιδευμένο BERT προσαρμόζεται εύκολα σε συγκεκριμένες εργασίες NLP προσθέτοντας μόνο ένα επιπλέον επίπεδο εξόδου. Η ευελιξία αυτή, σε συνδυασμό με την ισχυρή του κατανόηση συμφραζομένων, τού επέτρεψε να πετύχει αποτελέσματα αιχμής σε πληθώρα δοκιμασιών. Για παράδειγμα, το BERT σημείωσε νέα κορυφαία ακρίβεια σε 11 διαφορετικές εργασίες γλωσσικής κατανόησης, συμπεριλαμβανομένων

δοκιμασιών όπως το GLUE και το SQuAD. Είναι αξιοσημείωτο ότι στο σύνολο GLUE (General Language Understanding Evaluation) – το οποίο περιλαμβάνει και την ανάλυση συναισθήματος SST-2 – το BERT επέφερε σημαντική βελτίωση της συνολικής βαθμολογίας (80.5, αύξηση 7.7 μονάδων) σε σχέση με προηγούμενα μοντέλα [59]. Η ικανότητά του να κατανοεί αποχρώσεις στη σημασία και στο συναισθήμα των λέξεων το κάνει ιδιαίτερα αποτελεσματικό για εργασίες ταξινόμησης κειμένου, συμπεριλαμβανομένης της ανάλυσης συναισθήματος. Πράγματι, με κατάλληλη προσαρμογή, το BERT μπορεί να εκτιμήσει τον συναισθηματικό τόνο ενός κειμένου (θετικό, αρνητικό, ουδέτερο) λαμβάνοντας υπόψη ολόκληρο το περιεχόμενο της πρότασης ή παραγράφου, αντί απλής λεξικής ανάλυσης.

Η αρχιτεκτονική του BERT μπορεί να προσαρμοστεί σε εξειδικευμένους τομείς, όπως τα χρηματοοικονομικά. Το FinBERT, που απεικονίζεται ενδεικτικά παραπάνω, αποτελεί ένα παράδειγμα προσαρμογής του BERT για ανάλυση χρηματοοικονομικών κειμένων. Συγκεκριμένα, το FinBERT είναι ένα μοντέλο βασισμένο στο BERT που προπονήθηκε περαιτέρω σε μεγάλου όγκου χρηματοοικονομικά δεδομένα κειμένου, ώστε να κατανοεί την εξειδικευμένη ορολογία και φρασεολογία του χώρου. Η ανάγκη για το FinBERT προέκυψε επειδή τα γενικά μοντέλα γλώσσας συχνά αποτυγχάνουν να αποδώσουν με ακρίβεια το νόημα σε κείμενα οικονομικών ειδήσεων ή αναφορών, όπου χρησιμοποιείται ιδιαίτερη ορολογία και στυλ. Με την περαιτέρω προεκπαίδευση στον συγκεκριμένο τομέα, το FinBERT κατορθώνει να συλλάβει τα χρηματοοικονομικά συμφραζόμενα καλύτερα από το γενικό BERT. Σε συγκριτικές μελέτες, το FinBERT πέτυχε βελτιωμένες επιδόσεις σε σύνολα δεδομένων ανάλυσης συναισθήματος χρηματοοικονομικών ειδήσεων, υπερβαίνοντας κάθε προηγούμενο αποτέλεσμα σε μέτρα απόδοσης [61]. Αξιοσημείωτο είναι ότι ακόμη και με μικρότερο σύνολο εκπαιδευτικών δεδομένων και προσαρμόζοντας μόνο μέρος των παραμέτρων του μοντέλου, το FinBERT ξεπέρασε παραδοσιακές μεθόδους μηχανικής μάθησης στην ταξινόμηση του συναισθήματος οικονομικών κειμένων. Αυτό υπογραμμίζει τη σημασία των προσαρμοσμένων γλωσσικών μοντέλων σε ειδικούς τομείς γνώσης: επεκτείνοντας την αρχιτεκτονική του BERT με εξειδικευμένα δεδομένα, μπορούμε να πετύχουμε ακριβέστερη κατανόηση και ανάλυση συναισθήματος σε τομείς όπως οι χρηματοπιστωτικές αγορές.



Εικόνα 20: Απεικόνιση της διαδικασίας ενσωμάτωσης εισόδου στο BERT, όπου κάθε λέξη μετατρέπεται σε ένα διάνυσμα αναπαράστασης μέσω της άθροισης τριών συνιστωσών: ενσωματώσεις λέξεων (Token Embeddings), ενσωματώσεις τμημάτων (Segment Embeddings) και ενσωματώσεις θέσης (Position Embeddings)

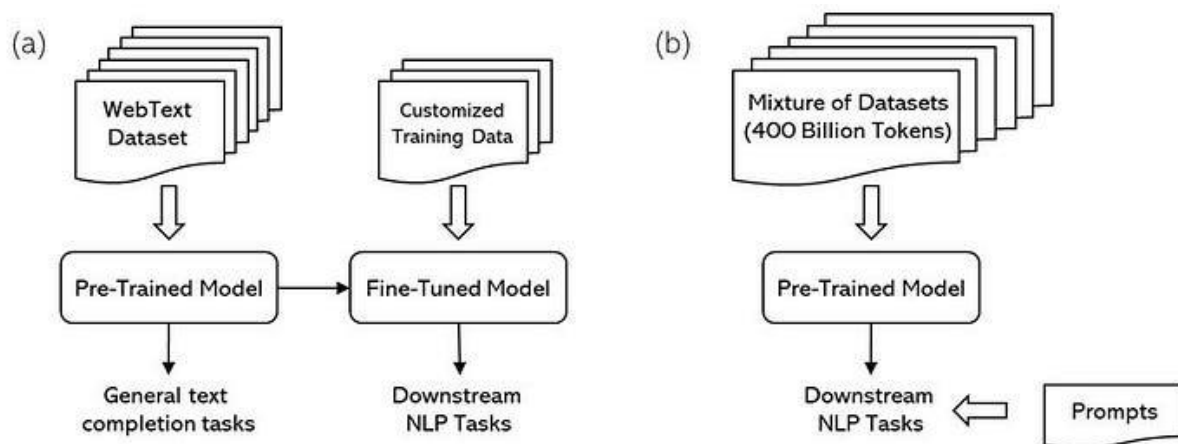
4.4. Το GPT και η χρήση του στην Ανάλυση Συναισθήματος

Το GPT (Generative Pre-Trained Transformer) εισάγει μια διαφορετική προσέγγιση στην επεξεργασία φυσικής γλώσσας, επικεντρωμένη κυρίως στην γενετική παραγωγή κειμένου. Ενώ το BERT είναι μοντέλο *encoder* που απορροφά πλήρως τα συμφραζόμενα εισόδου, το GPT βασίζεται αποκλειστικά στον *decoder* του μετασχηματιστή, γεγονός που το καθιστά αυτοπαλινδρομικό μοντέλο γλώσσας (autoregressive) με μονοκατευθυντική κωδικοποίηση [60]. Αυτό σημαίνει ότι το GPT διαβάζει το κείμενο μόνο προς μία κατεύθυνση – συνήθως από αριστερά προς τα δεξιά – και σε κάθε θέση προβλέπει την επόμενη λέξη λαμβάνοντας υπόψη μόνο το ήδη παραχθέν προηγούμενο κείμενο (π.χ. τις προηγούμενες λέξεις μιας πρότασης). Η εκπαίδευσή του γίνεται με στόχο την *πρόβλεψη της επόμενης λέξης* (objective causal language modeling), αξιοποιώντας τεράστιο όγκο μη επισημασμένου κειμένου για να μάθει στατιστικές συσχετίσεις της γλώσσας[49] . Με άλλα λόγια, το GPT μαθαίνει να συνεχίζει ένα δοσμένο κείμενο με τρόπο που είναι συντακτικά και νοηματικά συνεπής με όσα έχουν προηγηθεί. Η μονοδιάστατη αυτή προεκπαίδευση το καθιστά εξαιρετικά ικανό στη δημιουργία συνεκτικού κειμένου, όπως σύνταξη παραγράφων, άρθρων ή διαλόγων, όπου κάθε νέα λέξη παράγεται με βάση το ιστορικό των προηγούμενων. Ωστόσο, σε αντίθεση με το BERT, δεν έχει εγγενή πληροφόρηση για τις “επόμενες” λέξεις κατά την εκπαίδευση, κάτι που σημαίνει ότι μπορεί να δυσκολεύεται περισσότερο σε ορισμένες καθαρά ερμηνευτικές εργασίες που απαιτούν πλήρη κατανόηση ολόκληρης της πρότασης εκ των προτέρων. Παρά τους περιορισμούς αυτούς, η αρχιτεκτονική GPT έχει αποδειχθεί εξαιρετικά αποτελεσματική, ιδίως όταν το μοντέλο είναι πολύ μεγάλο σε κλίμακα. Χαρακτηριστικό παράδειγμα είναι το GPT-3 με 175 δισεκατομμύρια παραμέτρους, το οποίο μπορεί να παράγει κείμενο υψηλής ποιότητας και να προσαρμόζεται σε πλήθος εργασιών με ελάχιστη καθοδήγηση [62].

Παρότι το GPT σχεδιάστηκε πρωτίστως για παραγωγή κειμένου, μπορεί να αξιοποιηθεί και για την εξαγωγή συναισθηματικών χαρακτηριστικών από μεγάλα κείμενα, είτε μέσω προσαρμογής (fine-tuning) είτε μέσω κατάλληλης καθοδήγησης εισόδου (prompting). Στην περίπτωση της επιβλεπόμενης προσαρμογής, ένα προεκπαιδευμένο μοντέλο GPT μπορεί να εκπαιδευτεί περαιτέρω σε δεδομένα με ετικέτες συναισθήματος (π.χ. κριτικές ταινιών χαρακτηριζόμενες ως θετικές/αρνητικές) ώστε να μάθει να προβλέπει το συναίσθημα ενός δοσμένου κειμένου. Οι Radford et al. (2018) έδειξαν ότι ένα GPT μοντέλο, αφού προσαρμοστεί σε δεδομένα ταξινόμησης, μπορεί να πετύχει ακρίβεια ~91% στο σύνολο SST-2 της ανάλυσης συναισθήματος – επίδοση συγκρίσιμη με την καλύτερη τότε διαθέσιμη [63]. Αυτό είναι αξιοσημείωτο διότι καταδεικνύει ότι ένα γενετικό μοντέλο, παρότι δεν σχεδιάστηκε αρχικά για κατανόηση, μπορεί να ανταγωνιστεί εξειδικευμένα δίκτυα κατανόησης όταν εκπαιδευτεί καταλλήλως.

Εναλλακτικά, τα σύγχρονα μεγάλα γλωσσικά μοντέλα τύπου GPT μπορούν να εκτελέσουν ανάλυση συναισθήματος χωρίς περαιτέρω εκπαίδευση, εκμεταλλευόμενα τις γενικές γνώσεις που έχουν αποκτήσει. Μέσω καθοδηγούμενης εισόδου (prompts), το μοντέλο μπορεί να παρακινηθεί να παράξει ενδείξεις συναισθήματος. Για παράδειγμα, δίνοντας στο GPT-3 μια μεγάλη παράγραφο κειμένου και ζητώντας του να προσδιορίσει αν το ύφος είναι θετικό ή αρνητικό, το μοντέλο αξιοποιεί τις συσσωρευμένες γλωσσικές του γνώσεις για να συλλάβει τα συναισθηματικά συμφραζόμενα και να δώσει απάντηση. Έρευνες έχουν δείξει ότι τα μεγάλα προεκπαιδευμένα μοντέλα μπορούν να εντοπίσουν αυτόματα συναισθηματικές αποχρώσεις σε

κείμενο μεγάλης κλίμακας, μαθαίνοντας από τα πρότυπα της γλώσσας ακόμα και χωρίς ρητή επισήμανση κάθε περίπτωσης. Αυτό σημαίνει ότι το GPT-3 και τα συναφή μοντέλα καταφέρνουν να αφογκραστούν το συναίσθημα μέσα από τα συμφραζόμενα: για παράδειγμα, κατανοούν πότε μια κριτική προϊόντος εκφράζει δυσαρέσκεια ή ενθουσιασμό, βασιζόμενα στις λέξεις-κλειδιά και στον τόνο. Επιπλέον, η μεγάλη χωρητικότητα σε συμφραζόμενα που διαθέτουν (π.χ. δυνατότητα επεξεργασίας κειμένου με χιλιάδες λέξεις) επιτρέπει στο GPT να αναλύει μεγαλύτερα κείμενα σε βάθος, συσχετίζοντας πληροφορίες που εμφανίζονται διάσπαρτες σε όλο το έγγραφο για μια συνολική εκτίμηση του συναισθήματος.



Εικόνα 21: Μέθοδος εκπαίδευσης GPT μοντέλων

5. Συλλογή και Επεξεργασία Δεδομένων

5.1. Συλλογή Δεδομένων

Για την υλοποίηση του πειραματικού σκέλους της παρούσας διπλωματικής εργασίας, κρίθηκε απαραίτητη η επιλογή εταιρειών με υψηλή κεφαλαιοποίηση, προκειμένου να διασφαλιστεί η αντιπροσωπευτικότητα και η αξιοπιστία των δεδομένων. Συνεπώς, επιλέχθηκαν τρεις από τις μεγαλύτερες και πλέον επιδραστικές εταιρείες που διαπραγματεύονται στο Αμερικανικό Χρηματιστήριο, οι οποίες χαρακτηρίζονται από σημαντική εμπορευσιμότητα και επενδυτικό ενδιαφέρον.

Οι εταιρείες που εξετάζονται στο πλαίσιο της μελέτης είναι οι ακόλουθες:

1. Η **Apple Inc.** αποτελεί μία από τις πιο σταθερές και αξιόπιστες εταιρείες στο χρηματιστήριο, χαρακτηριστικό που την καθιστά ιδανική για μελέτες συναισθηματικής ανάλυσης και αποδόσεων. Ως μία από τις μεγαλύτερες πολυεθνικές εταιρείες στον κόσμο, η Apple διακρίνεται για την ισχυρή της κερδοφορία, τη συνεχή ανάπτυξη και το ισχυρό brand loyalty των καταναλωτών της. Η σταθερή βάση εσόδων της προέρχεται κυρίως από πωλήσεις hardware (iPhone, Mac, iPad), λογισμικού και υπηρεσιών (Apple Music, iCloud, App Store). Η επιλογή της Apple στη συγκεκριμένη ανάλυση έγινε λόγω της ανθεκτικότητάς της στις οικονομικές διακυμάνσεις και της σταθερής της απόδοσης. Σε αντίθεση με πιο ευμετάβλητες εταιρείες, όπως οι νεοφυείς επιχειρήσεις ή οι εταιρείες με έντονη κυκλικότητα στη ζήτηση, η Apple έχει αποδείξει ότι διατηρεί υψηλή κερδοφορία ανεξαρτήτως μακροοικονομικών συνθηκών.
2. Η **Tesla, Inc.** είναι ένας από τους πιο ασταθείς τίτλους της Wall Street, καθώς η τιμή της επηρεάζεται έντονα από τις ειδήσεις, τις ανακοινώσεις και τις απόψεις των επενδυτών. Σε αντίθεση με πιο σταθερές επιχειρήσεις όπως η Apple, η Tesla κινείται στον χώρο της ηλεκτροκίνησης, της καινοτομίας και της τεχνητής νοημοσύνης, τομείς που χαρακτηρίζονται από υψηλό βαθμό αβεβαιότητας. Οι ειδήσεις σχετικά με τις πωλήσεις, την παραγωγή, τις νέες τεχνολογίες ή ακόμα και οι δηλώσεις του CEO Elon Musk συχνά προκαλούν έντονες διακυμάνσεις στην τιμή της μετοχής. Η επιλογή της Tesla βασίστηκε στη μοναδική της ικανότητα να επηρεάζεται από το sentiment της αγοράς, καθιστώντας την ιδανικό παράδειγμα για τη διερεύνηση της σχέσης μεταξύ SEC filings και χρηματιστηριακών αποδόσεων. Η Tesla ανήκει σε μια κατηγορία εταιρειών που βασίζονται στην τεχνολογική καινοτομία και την αλλαγή παραδείγματος στη βιομηχανία τους.
3. Η **NVIDIA Corporation** είναι ένα από τα πλέον χαρακτηριστικά παραδείγματα μιας εταιρείας που γνώρισε ραγδαία ανάπτυξη λόγω τεχνολογικών εξελίξεων και μακροπρόθεσμων τάσεων στην αγορά. Από μια εταιρεία που αρχικά επικεντρωνόταν στην παραγωγή καρτών γραφικών για gaming, η NVIDIA έχει μετατραπεί σε παγκόσμιο ηγέτη στους τομείς της τεχνητής νοημοσύνης, του cloud computing και της επεξεργασίας δεδομένων. Η ζήτηση για τις GPUs της εκτοξεύτηκε χάρη στις εφαρμογές τους στο machine learning, τα data centers και το metaverse. Η επιλογή της NVIDIA βασίστηκε στο γεγονός ότι αντιπροσωπεύει εταιρείες με εκθετική ανάπτυξη, οι οποίες συχνά

εμφανίζουν μεγάλη συσχέτιση μεταξύ των χρηματοοικονομικών ανακοινώσεων και των αποδόσεων τους. Οι εταιρείες αυτού του τύπου μπορεί να έχουν έντονες διακυμάνσεις στις τιμές των μετοχών τους, ειδικά όταν ανακοινώνουν νέα προϊόντα ή οικονομικά αποτελέσματα. Παρόμοιες εταιρείες περιλαμβάνουν την AMD (Advanced Micro Devices), την TSMC (Taiwan Semiconductor Manufacturing Company) και την Intel (INTC) στον χώρο των ημιαγωγών.

Αυτές οι τρεις εταιρείες επιλέχθηκαν για την ανάλυση, καθώς καλύπτουν διαφορετικούς τύπους επιχειρηματικών μοντέλων και στρατηγικών ανάπτυξης: η Apple ως σταθερός τεχνολογικός γίγαντας, η Tesla ως εταιρεία που επηρεάζεται από το sentiment και τις ειδήσεις, και η NVIDIA ως παράδειγμα ραγδαίας ανάπτυξης. Αυτή η διαφοροποίηση επιτρέπει μια πιο ολοκληρωμένη ανάλυση της σχέσης μεταξύ των ενεργειών και αποδόσεων των εταιρειών.

5.1.1. Επιλογή Μετοχών

Για την επιλογή των μετοχών που χρησιμοποιούνται στη μελέτη, αξιοποιήθηκε το Alpaca API, το οποίο είναι μια σύγχρονη χρηματιστηριακή πλατφόρμα που παρέχει πρόσβαση σε δεδομένα αγοράς και επιτρέπει την εκτέλεση συναλλαγών μέσω ενός RESTful API. Σχεδιάστηκε κυρίως για αλγοριθμική διαπραγμάτευση και ανάλυση δεδομένων, δίνοντας στους χρήστες τη δυνατότητα να ανακτούν ιστορικά και real-time δεδομένα, να διαχειρίζονται επενδυτικά χαρτοφυλάκια και να εκτελούν συναλλαγές μετοχών και ETFs

Αρχικά, μέσω της `list_assets`, ανακτήθηκαν όλα τα tickers των διαπραγματεύσιμων μετοχών, μαζί με τις βασικές πληροφορίες τους, όπως η κατάσταση διαπραγμάτευσης (ενεργή ή μη), ο τύπος του asset (μετοχή, ETF κ.λπ.) και το χρηματιστήριο στο οποίο είναι εισηγμένες. Από αυτά, επιλέχθηκαν οι μετοχές που θα συμπεριληφθούν στο πειραματικό μέρος της εργασίας, λαμβάνοντας υπόψη:

1. **Ενεργές μετοχές** (status = 'active').
2. **Μετοχές εισηγμένες σε μεγάλα χρηματιστήρια** (NYSE, NASDAQ).
3. **Κοινές μετοχές** (Asset class = 'us_equity').

Με αυτόν τον τρόπο, συγκροτήθηκε ένα αντιπροσωπευτικό δείγμα μετοχών που ανήκουν στον ίδιο κλάδο ή έχουν παρόμοιο μέγεθος κεφαλαιοποίησης, ώστε να διατηρηθεί μια συνεκτική βάση δεδομένων για την ανάλυση.

Από τα δεδομένα που επιστράφηκαν από το API, διατηρήθηκαν μόνο τα ticker symbols των επιλεγμένων μετοχών, καθώς αυτά ήταν απαραίτητα για την περαιτέρω συλλογή χρηματιστηριακών δεδομένων (τιμές, όγκοι κ.λπ.) μέσω άλλων κλήσεων API.

5.1.2. Ιστορικά δεδομένα τιμών

Για την απόκτηση ιστορικών δεδομένων τιμών των μετοχών, χρησιμοποιήθηκε το Alpha Vantage API, το οποίο παρέχει μη επεξεργασμένα ημερήσια δεδομένα (raw as-traded data) για χρηματιστηριακές αξίες παγκοσμίως.

Για τη συλλογή των δεδομένων, χρησιμοποιήθηκε η συνάρτηση `TIME_SERIES_DAILY`, η οποία επιστρέφει ημερήσια δεδομένα χρονοσειρών για μια συγκεκριμένη μετοχή. Τα δεδομένα

περιλαμβάνουν τις ακόλουθες μεταβλητές OHLCV, γνωστές και ως "candles" στη χρηματοοικονομική βιβλιογραφία:

1. **Date:** Ημερομηνία της συναλλαγής.
2. **Open:** Τιμή ανοίγματος της ημέρας.
3. **High:** Υψηλότερη τιμή κατά τη διάρκεια της ημέρας.
4. **Low:** Χαμηλότερη τιμή κατά τη διάρκεια της ημέρας.
5. **Close:** Τιμή κλεισίματος της συνεδρίασης.
6. **Volume:** Συνολικός όγκος συναλλαγών.

Κατά την ανάκτηση των δεδομένων, ορίστηκαν οι εξής παράμετροι στο API:

1. **symbol:** Το ticker της μετοχής που μελετάται.
2. **outputsize:** Ορίστηκε σε full για τη λήψη πλήρους ιστορικού δεδομένων.
3. **datatype:** Ορίστηκε σε json για τη μετατροπή των δεδομένων σε δομή που επιτρέπει ευκολότερη επεξεργασία.

Τα δεδομένα αυτά αποτέλεσαν τη βάση για την ανάλυση των αγορών, την εφαρμογή τεχνικών δεικτών και τη διαμόρφωση χρονοσειρών. Για τη λήψη δεδομένων που περιλαμβάνουν προσαρμοσμένες τιμές (adjusted close) λόγω μερισμάτων ή splits, το Alpha Vantage API προσφέρει την εναλλακτική επιλογή TIME_SERIES_DAILY_ADJUSTED, η οποία δεν χρησιμοποιήθηκε στην παρούσα μελέτη.

5.1.3. Ανακοινώσεις Κερδών (Earnings Announcements)

Για την απόκτηση δεδομένων κερδοφορίας των εταιρειών, χρησιμοποιήθηκε το Alpha Vantage API και συγκεκριμένα η συνάρτηση EARNINGS, η οποία επιστρέφει ετήσια και τετραμηνιαία δεδομένα κερδών ανά μετοχή (EPS). Εκτός από τα πραγματικά αποτελέσματα, το API παρέχει και εκτιμήσεις αναλυτών, καθώς και τον βαθμό απόκλισης μεταξύ εκτιμήσεων και πραγματικών κερδών (surprise metrics).

Αν και το API επιστρέφει στοιχεία όπως τα ετήσια EPS, τετραμηνιαία EPS και εκτιμήσεις αναλυτών, στην παρούσα μελέτη χρησιμοποιήθηκαν αποκλειστικά τα surprise metrics, δηλαδή η απόκλιση μεταξύ των αναμενόμενων και των πραγματικών κερδών ανά μετοχή. Αυτή η μεταβλητή θεωρείται κρίσιμη, καθώς μπορεί να επηρεάζει τη συμπεριφορά των επενδυτών και τις διακυμάνσεις των μετοχών μετά την ανακοίνωση των οικονομικών αποτελεσμάτων.

Η συλλογή των surprise metrics πραγματοποιήθηκε με τις εξής παραμέτρους API:

1. **function:** earnings, για την ανάκτηση των αποτελεσμάτων κερδοφορίας.
2. **symbol:** Το ticker της εταιρείας.

Η ενσωμάτωση αυτής της πληροφορίας στην ανάλυση επέτρεψε τη μελέτη της σχέσης μεταξύ μη αναμενόμενων χρηματοοικονομικών εξελίξεων και των αντίστοιχων μεταβολών στις τιμές των μετοχών.

5.1.4. Ρυθμιστικές Καταθέσεις στην Επιτροπή Κεφαλαιαγοράς της Αμερικής

Για τη συλλογή και ανάλυση των ρυθμιστικών καταθέσεων (filings) των εισηγμένων εταιρειών στο χρηματιστήριο, χρησιμοποιήθηκε η Python βιβλιοθήκη edgar-tools. Η συγκεκριμένη βιβλιοθήκη επιτρέπει την αυτοματοποιημένη ανάκτηση εγγράφων από τη βάση δεδομένων SEC EDGAR (Electronic Data Gathering, Analysis, and Retrieval System), η οποία περιλαμβάνει εταιρικές αναφορές, οικονομικές καταστάσεις και λοιπά κανονιστικά αρχεία.

Τα αρχεία που ανακτήθηκαν μέσω του edgar-tools είναι δομημένα HTML, TXT ή XML αρχεία, ανάλογα με τη μορφή στην οποία έχουν αποθηκευτεί στο EDGAR. Η βιβλιοθήκη επιτρέπει:

1. Την αναζήτηση ανά εταιρεία ή ανά τύπο εγγράφου (π.χ., 10-K, 10-Q, 8-K).
2. Τη λήψη των εγγράφων σε ακατέργαστη μορφή από τον SEC EDGAR server.
3. Τη μετατροπή των εγγράφων σε κειμενική μορφή για περαιτέρω ανάλυση.

Μετά την ανάκτηση, τα κείμενα καθαρίστηκαν από HTML tags, αριθμούς και ειδικούς χαρακτήρες, ώστε να καταστεί δυνατή η ανάλυση συναισθήματος (sentiment analysis).

Για την ανάλυση συναισθήματος, επιλέχθηκαν συγκεκριμένα τμήματα των ρυθμιστικών καταθέσεων, τα οποία περιέχουν κρίσιμες πληροφορίες για τις στρατηγικές, τους κινδύνους και τις επίσημες ανακοινώσεις των εταιρειών.

1. Risk Factors (Παράγοντες Κινδύνου) – 10-K & 10-Q Filings

Αυτή η ενότητα περιέχει πληροφορίες σχετικά με τους επιχειρηματικούς και χρηματοοικονομικούς κινδύνους που αντιμετωπίζει η εταιρεία. Εφαρμόστηκε ανάλυση συναισθηματικής φόρτισης, με έμφαση σε διατυπώσεις που αφορούν αβεβαιότητα, νομικούς κινδύνους, ρυθμιστικά εμπόδια, ανάπτυξη και επεκτατική πολιτική.

2. Management Discussion and Analysis (Συζήτηση Διοίκησης) – 10-K & 10-Q Filings

Περιλαμβάνει την αξιολόγηση της διοίκησης σχετικά με τις οικονομικές επιδόσεις και τις μελλοντικές στρατηγικές. Η ανάλυση επικεντρώθηκε στον τόνο της διοίκησης, καταγράφοντας αν είναι θετικός, αρνητικός ή ουδέτερος, προκειμένου να εκτιμηθεί η στάση της απέναντι στην εταιρική ανάπτυξη.

3. Press Releases (Δελτία Τύπου) – 8-K Filings

Τα 8-K filings περιλαμβάνουν επίσημες ανακοινώσεις, όπως στρατηγικές επενδύσεις, εξαγορές και οικονομικά αποτελέσματα. Εφαρμόστηκε ανάλυση συναισθήματος για τον εντοπισμό λέξεων και φράσεων με θετικό ή αρνητικό φορτίο, προκειμένου να αξιολογηθεί η επίδρασή τους στη χρηματιστηριακή αγορά.

Η συλλογή και επεξεργασία αυτών των τμημάτων επέτρεψε την κατασκευή ενός δομημένου συναισθηματικού dataset, το οποίο χρησιμοποιήθηκε για τη διερεύνηση της σχέσης μεταξύ των ρυθμιστικών ανακοινώσεων και των διακυμάνσεων των τιμών των μετοχών.

5.2. Επεξεργασία δεδομένων

Μετά τη συλλογή των δεδομένων από τις παραπάνω πηγές, πραγματοποιήθηκε η διαδικασία της προεπεξεργασίας των δεδομένων, ώστε να διασφαλιστεί η συνοχή και η ορθότητα των δεδομένων πριν τη χρήση τους στο πειραματικό μέρος της μελέτης.

5.2.1. Ευθυγράμμιση και Διαχείριση Χρονοσειρών

Για τη σωστή διαχείριση των χρηματοοικονομικών δεδομένων, πραγματοποιήθηκε ευθυγράμμιση των χρονοσειρών ώστε να διασφαλιστεί η ακρίβεια στην αντιστοίχιση των οικονομικών στοιχείων με τις αντίστοιχες χρονικές στιγμές.

Συγκεκριμένα:

1. **Εναρμόνιση Ημερομηνιών:** Οι ημερομηνίες των OHLCV δεδομένων (Open, High, Low, Close, Volume) συνδέθηκαν με τις ημερομηνίες των εταιρικών ανακοινώσεων (όπως earnings reports) και των ρυθμιστικών καταθέσεων (SEC filings) ώστε να αποφεύγονται χρονικές ανακολουθίες.
2. **Διαχείριση Ημερολογιακών Ασυνεχειών:** Οι ημέρες όπου η αγορά ήταν κλειστή (π.χ. Σαββατοκύριακα, αργίες, ή τεχνικά ζητήματα) εντοπίστηκαν και αφαιρέθηκαν από το σύνολο δεδομένων, αποτρέποντας τη χρήση μη συναλλακτικών ημερών.
3. **Αντιστοίχιση Χρονοσειρών:** Οι καταχωρήσεις των SEC filings και των εταιρικών ανακοινώσεων ευθυγραμμίστηκαν με τις αντίστοιχες ημερομηνίες χρηματιστηριακών δεδομένων, ώστε κάθε γεγονός να συσχετίζεται με τις κατάλληλες τιμές μετοχών.
4. **Χειρισμός Λείπουσας Πληροφορίας:** Σε περιπτώσεις όπου υπήρχαν ελλείποντα δεδομένα OHLCV λόγω χαμηλής ρευστότητας ή απουσίας συναλλαγών, εφαρμόστηκαν τεχνικές απαλοιφής.

5.2.2. Εκκαθάριση και Ανάλυση Ρυθμιστικών Καταθέσεων

Τα δεδομένα των ρυθμιστικών καταθέσεων (SEC filings) εξήχθησαν σε αδόμητη μορφή (raw text) μέσω της βιβλιοθήκης edgar-tools. Προκειμένου να χρησιμοποιηθούν στην ανάλυση, πραγματοποιήθηκε προγραμματιστικός καθαρισμός των δεδομένων μέσω τεχνικών επεξεργασίας κειμένου.

Συγκεκριμένα:

1. Αφαιρέθηκαν ειδικοί χαρακτήρες και μεταδεδομένα που δεν ήταν απαραίτητα για την ανάλυση.
2. Ομαλοποιήθηκε η μορφοποίηση, όπως η αφαίρεση πολλαπλών κενών χαρακτήρων.
3. Αφαιρέθηκαν πίνακες και μη σχετικές ενότητες, ώστε να διατηρηθεί μόνο το κείμενο των οικονομικών αναφορών.
4. Για περαιτέρω επεξεργασία και ανάλυση, εφαρμόστηκαν τεχνικές tokenization, ώστε να διευκολυνθεί η εξαγωγή χρηστικών προτάσεων από τα κείμενα.

5.2.3. Οργάνωση και Αποθήκευση Δεδομένων

Για τη γρήγορη και αποδοτική προσπέλαση των δεδομένων, όλες οι πληροφορίες αποθηκεύτηκαν σε βάση δεδομένων SQLite. Η επιλογή της SQLite έγινε λόγω της ευκολίας ενσωμάτωσης σε πειραματικά περιβάλλοντα, της χαμηλής κατανάλωσης πόρων και της υποστήριξης δομημένων ερωτημάτων SQL.

Συγκεκριμένα, δημιουργήθηκαν ξεχωριστοί πίνακες για:

1. **Stock_Price**: Αποθηκεύει τις ημερήσιες τιμές μετοχών σε μορφή OHLCV (Open, High, Low, Close, Volume). Αυτός ο πίνακας επιτρέπει τη γρήγορη πρόσβαση και επεξεργασία ιστορικών δεδομένων τιμών.
2. **Quarterly_Earnings**: Περιέχει δεδομένα ανακοινώσεων κερδών, όπως το αναφερόμενο κέρδος ανά μετοχή (EPS), το εκτιμώμενο EPS και την απόκλιση (surprise) μεταξύ τους.
3. **TenK και TenQ**: Περιλαμβάνουν δεδομένα από τις ρυθμιστικές καταθέσεις 10-K και 10-Q. Αποθηκεύονται τόσο το ακατέργαστο κείμενο, το πλήρες καθαρισμένο κείμενο όσο και αναλύσεις συναισθήματος (sentiment analysis) για συγκεκριμένα τμήματα, όπως τα "Risk Factors" και "Management Discussion."
4. **EightK**: Περιέχει δεδομένα από αναφορές 8-K, που εστιάζουν σε σημαντικές εταιρικές ανακοινώσεις. Αποθηκεύονται τόσο τα κείμενα (ακατέργαστα και μη) όσο και τα συναισθήματα (sentiment scores) που έχουν εξαχθεί από τις ανακοινώσεις.
5. **Stock**: Περιέχει βασικές πληροφορίες για κάθε μετοχή, όπως το σύμβολο (symbol) και το όνομα της εταιρείας (company), επιτρέποντας τη συσχέτιση μεταξύ των διαφορετικών πινάκων.

Επιπλέον δημιουργήθηκαν οι θυγατρικοί των TenK , TenQ, EightK(TenK_GPT , TenQ_GPT, EightK_GPT) πίνακες οι οποίοι περιέχουν τα sentiment scores που δόθηκαν από το ChatGPT (αναφορά στην επόμενη ενότητα). Η αποθήκευση των δεδομένων σε SQLite επέτρεψε την εύκολη εκτέλεση ερωτημάτων και την ενσωμάτωσή τους στα επόμενα στάδια ανάλυσης και εκπαίδευσης μοντέλων.

5.3. Εξαγωγή Χαρακτηριστικών (Feature Engineering)

5.3.1. Ποσοτικά Χαρακτηριστικά

Η διαδικασία του Feature Engineering αποτελεί κρίσιμο στάδιο στη δημιουργία ενός επιτυχημένου μοντέλου μηχανικής μάθησης, καθώς επιτρέπει την εξαγωγή πρόσθετων χαρακτηριστικών από τα δεδομένα, τα οποία μπορούν να βελτιώσουν την απόδοση του μοντέλου. Στην παρούσα διπλωματική εργασία, υλοποιήθηκαν τα παρακάτω χαρακτηριστικά:

1. Surprise

Το χαρακτηριστικό Surprise αντιπροσωπεύει την "έκπληξη" που προκαλεί η ανακοίνωση κερδών μιας εταιρείας. Υπολογίζεται ως η διαφορά μεταξύ των αναμενόμενων και ανακοινωθέντων κερδών ανά μετοχή (EPS). Αυτό το χαρακτηριστικό είναι ιδιαίτερα χρήσιμο, καθώς ενσωματώνει πληροφορίες για τις αποκλίσεις από τις προβλέψεις των αναλυτών, οι

οποίες συχνά επηρεάζουν σημαντικά τις χρηματιστηριακές τιμές. Το surprise των earnings δίνεται από τον τύπο:

$$\text{Earnings Surprise} = \frac{\text{Actual EPS} - \text{Expected EPS}}{|\text{Expected EPS}|}$$

όπου:

- **Actual EPS:** Τα πραγματικά κέρδη ανά μετοχή που ανακοινώθηκαν από την εταιρεία.
- **Expected EPS:** Οι εκτιμήσεις των αναλυτών για τα κέρδη ανά μετοχή πριν την ανακοίνωση.
- **|Expected EPS|:** Η απόλυτη τιμή των εκτιμώμενων κερδών χρησιμοποιείται ώστε το αποτέλεσμα να είναι κανονικοποιημένο.

Αν το Earnings Surprise είναι θετικό, σημαίνει ότι η εταιρεία ανακοίνωσε καλύτερα από τα αναμενόμενα κέρδη, ενώ αν είναι αρνητικό, η εταιρεία ανακοίνωσε χειρότερα αποτελέσματα από τις εκτιμήσεις.

2. Ημερήσια κατεύθυνση (Intraday Direction)

Δυαδικός δείκτης που λαμβάνει τιμή 1 αν η μετοχή έκλεισε υψηλότερα από το άνοιγμά της, αλλιώς 0:

$$\text{intraday direction}_t = \begin{cases} 1, & \text{αν } \text{close}_t > \text{open}_t \\ 0, & \text{αλλιώς} \end{cases}$$

3. Επόμενη Ημερήσια Απόδοση (Προηγούμενο Στάδιο)

Ποσοστιαία μεταβολή της τιμής κλεισίματος σε σχέση με την προηγούμενη ημέρα, μετατοπισμένη κατά μια ημέρα μπροστά ώστε να χρησιμοποιηθεί ως στόχος:

$$\text{next_day_return}_t = \frac{\text{close}_{t+1} - \text{close}_t}{\text{close}_t}$$

4. Δυαδικός Στόχος (Binary Target)

Δυαδικός δείκτης που λαμβάνει τιμή 1 αν η μετοχή έκλεισε υψηλότερα από το άνοιγμά της, αλλιώς 0:

$$\text{next - day direction}_t = \begin{cases} 1, & \text{αν } \text{close}_t > \text{open}_t \\ 0, & \text{αλλιώς} \end{cases}$$

5. Ημερήσια Μεταβλητότητα (Daily Volatility)

Τυπική απόκλιση του intraday_direction σε κινούμενο παράθυρο 7 ημερών, υπολογιζόμενη ανά μετοχή.

$$\text{daily_volatility}_t = \sqrt{\frac{1}{6} \sum_{i=t-6}^t (\text{intraday_direction}_i - \overline{\text{intraday_direction}_{t-6:t}})^2}$$

Δυναδικός δείκτης που λαμβάνει τιμή 1 αν η επόμενη ημερήσια απόδοση είναι θετική, αλλιώς 0:

6. Ποσοστιαία Κατάταξη Τιμής Κλεισίματος (Percentile Rank)

Η θέση της σημερινής τιμής κλεισίματος ως ποσοστό έναντι όλων των προηγούμενων κλεισιμάτων:

$$\text{percentile_rank}_t = \frac{\# \{ \text{close}_i - \text{close}_t \}}{N_t} \times 100\%$$

Όπου N_t ο αριθμός των διαθέσιμων ημερών έως t .

7. 1-μηνιαίο Σωρευτικό Επιστροφών (Rolling 1-month Cumulative Return)

Ποσοστιαία μεταβολή της τιμής κλεισίματος σε διάστημα 21 ημερών:

$$\text{rolling_rets_1}_t = \frac{\text{close}_t}{\text{close}_{t-21}} - 1$$

8. Απόσταση από τον Κινούμενο Μέσο Όρο (Distance from 7-day Rolling Mean)

Διαφορά της σημερινής τιμής κλεισίματος από τον 7-ήμερο κινούμενο μέσο όρο της ίδιας τιμής.

$$\text{distance_from_mean}_t = \frac{1}{7} \sum_{i=t-6}^t \text{close}_i$$

9. Δείκτης TRIX (15-day Triple Exponential Moving Average)

Υπολογίζεται ως το ποσοστιαίο ποσοστό αλλαγής του τριπλού εκθετικού κινούμενου μέσου όρου.

$$EMA_t^{(1)} = \alpha \text{close}_t + (1 - \alpha) EMA_{t-1}^{(1)}$$

$$EMA_t^{(2)} = \alpha EMA_t^{(1)} + (1 - \alpha) EMA_{t-1}^{(2)}$$

$$EMA_t^{(3)} = \alpha EMA_t^{(2)} + (1 - \alpha) EMA_{t-1}^{(3)}$$

$$TRIX_t = \left(\frac{EMA_t^{(3)} - EMA_{t-1}^{(3)}}{EMA_{t-1}^{(3)}} \right) \times 100\%$$

Όπου $\alpha = \frac{2}{\text{window}+1}$ για $\text{window} = 15$

Η διαδικασία επιλογής και κατασκευής χαρακτηριστικών στηρίχθηκε στην ανάγκη για αντιπροσωπευτική αποτύπωση της δυναμικής των χρηματοοικονομικών χρονοσειρών, με στόχο τη βελτίωση της ακρίβειας πρόβλεψης και της γενικευσιμότητας των μοντέλων. Η έμφαση δόθηκε στη δημιουργία χαρακτηριστικών που συνδυάζουν πληροφορία κατεύθυνσης,

μεταβλητότητας και σχετικής θέσης, μέσω υπολογισμών που βασίζονται σε ιστορικά δεδομένα τιμών και αποτελεσμάτων εταιρειών.

Ιδιαίτερη προσοχή δόθηκε στην επεξεργασία των αποδόσεων μέσω μετασχηματισμών που ενισχύουν τη στασιμότητα, στοιχείο κρίσιμο για την επιτυχημένη εφαρμογή αλγορίθμων μηχανικής μάθησης. Χρησιμοποιήθηκαν τόσο απλές ποσοστιαίες μεταβολές όσο και σωρευτικές επιστροφές σε κινούμενα χρονικά παράθυρα, με στόχο την αποτύπωση των τάσεων και των σημείων αλλαγής στην κατεύθυνση της αγοράς.

Επιπλέον, υλοποιήθηκαν δείκτες όπως ο TRIX (Triple Exponential Moving Average) και η ποσοστιαία κατάταξη της τιμής κλεισίματος, που παρέχουν πληροφορία για την ορμή (momentum) και τη σχετική θέση της τιμής εντός του ιστορικού εύρους. Η προσθήκη χαρακτηριστικών όπως η ημερήσια μεταβλητότητα, υπολογιζόμενη σε κινούμενα παράθυρα, επιτρέπει την αποτύπωση των εναλλαγών έντασης της αγοράς.

Η επεξεργασία βασίστηκε σε σταθερά βήματα υπολογισμού, ώστε να διασφαλίζεται η συστηματική μεταχείριση των δεδομένων και η συνέπεια μεταξύ των μεταβλητών. Κάθε χαρακτηριστικό εντάσσεται σε μία από τις ακόλουθες κατηγορίες:

- Κατεύθυνση (Direction-based): όπως το Intraday Direction και ο Binary Target.
- Απόδοση (Return-based): όπως η Επόμενη Ημερήσια Απόδοση και το Rolling 1-month Return.
- Θέση/Κατάταξη (Position-based): όπως το Percentile Rank και η Απόσταση από τον Κινούμενο Μέσο Όρο.
- Μεταβλητότητα (Volatility-based): όπως η Ημερήσια Μεταβλητότητα.
- Τεχνικοί Δείκτες (Technical indicators): όπως ο TRIX και το Earnings Surprise.

Η επιλογή των παραπάνω χαρακτηριστικών έγινε με βάση την εμπειρική τους χρησιμότητα στη χρηματοοικονομική ανάλυση και την τεκμηριωμένη συμβολή τους στη βελτίωση της προβλεπτικής ισχύος των μοντέλων.

5.3.2. Ποσοτικοποίηση Ποιοτικών Χαρακτηριστικών

Για την ποσοτικοποίηση των ποιοτικών χαρακτηριστικών που περιέχονται στα SEC filings, όπως οι αναφορές σε στρατηγικές κινήσεις, καινοτομίες και οικονομικές προβλέψεις, χρησιμοποιήθηκε η μέθοδος της συναισθηματικής ανάλυσης. Αυτή η προσέγγιση επέτρεψε τη μετατροπή του κειμένου σε ποσοτικά δεδομένα, τα οποία μπορούν να ενσωματωθούν στα μοντέλα πρόβλεψης της μελέτης. Για την εξαγωγή των συναισθηματικών χαρακτηριστικών, αξιοποιήθηκαν τρεις διαφορετικές μεθοδολογίες, ώστε να εξασφαλιστεί μεγαλύτερη ακρίβεια και πληρότητα στα αποτελέσματα.

5.3.2.1. Επιλογή μεθόδων ποσοτικοποίησης

1η Προσέγγιση: Χρήση FinBERT για Συναισθηματική Ανάλυση Προτάσεων

Η πρώτη μεθοδολογία περιλάμβανε την εξαγωγή προτάσεων από τα κείμενα των SEC filings μετά από καθαρισμό τους. Ο καθαρισμός περιελάμβανε την απομάκρυνση μη σχετικών

τμημάτων, όπως μεταδεδομένα, επικεφαλίδες ή κοινότητες φράσεις που εμφανίζονται επαναλαμβανόμενα στα έγγραφα. Οι καθαρές προτάσεις εισήχθησαν στο προεκπαιδευμένο γλωσσικό μοντέλο FinBERT, το οποίο έχει σχεδιαστεί ειδικά για την ανάλυση χρηματοοικονομικών δεδομένων.

Το FinBERT ανέλυσε κάθε πρόταση και παρήγαγε τρία διαφορετικά sentiment scores: positive, negative και neutral, με βάση το περιεχόμενο της πρότασης. Αυτή η μέθοδος επέτρεψε τη λεπτομερή καταγραφή του συναισθήματος σε επίπεδο πρότασης, παρέχοντας μια ακριβή εκτίμηση της συναισθηματικής φόρτισης που περιέχεται σε κάθε τμήμα των εγγράφων.

2η Προσέγγιση: Χρήση API ChatGPT για Εξαγωγή Σημαντικών Προτάσεων

Η δεύτερη μέθοδος βασίστηκε στο API του ChatGPT, το οποίο χρησιμοποιήθηκε για την εξαγωγή προτάσεων που περιείχαν σαφώς θετικά ή αρνητικά συναισθήματα. Ο καθαρισμός των κειμένων και η εξαγωγή των προτάσεων πραγματοποιήθηκαν βάσει ενός προκαθορισμένου prompt, το οποίο σχεδιάστηκε για να εξασφαλίσει τη διατήρηση μόνο των σχετικών πληροφοριών.

Το prompt καθοδηγούσε το ChatGPT ώστε να:

- Εντοπίζει προτάσεις που αντικατοπτρίζουν θετικό ή αρνητικό συναίσθημα, όπως ανάπτυξη, κίνδυνο, εμπιστοσύνη ή προκλήσεις.
- Εξαιρεί προτάσεις με ουδέτερο ή περιγραφικό περιεχόμενο, όπως ημερομηνίες, νομικές δηλώσεις ή επαναλαμβανόμενη γλώσσα.
- Παράγει μόνο πλήρεις, ανεξάρτητες προτάσεις που περιέχουν σημαντικές πληροφορίες.
- Να αναλύσει το κείμενο και να παράξει ένα sentiment score

Το αποτέλεσμα αυτής της διαδικασίας ήταν μια συλλογή από σημαντικές προτάσεις, οι οποίες αποτελούσαν ένα καθαρό υποσύνολο των εγγράφων, κατάλληλο για περαιτέρω ανάλυση. Παράλληλα, χρησιμοποιήθηκε το sentiment score που παρήγαγε το ChatGPT για κάθε έγγραφο αυτοτελώς ως χαρακτηριστικό για τη συναισθηματική ανάλυση των εγγράφων.

Παρακάτω παρατίθεται το prompt που δόθηκε στο ChatGPT:

You are an advanced text extraction assistant. Your task is to analyze raw text from SEC filings (10-K, 10-Q, 8-K) and extract sentences that meet the following criteria:

- 1. Include sentences reflecting positive or negative sentiment (e.g., confidence, growth, challenges, risks, etc.).*
- 2. Exclude sentences that:*
 - Are repetitive phrases or boilerplate language commonly found in SEC filings.*
 - Contain neutral or purely descriptive content without any sentiment (e.g., dates, legal disclaimers).*
- 3. Do not paraphrase the sentences. Retain them exactly as they appear in the text.*
- 4. Ensure each extracted sentence is a standalone, full, grammatically correct sentence.*
- 5. Ignore headers, section titles, tables, metadata, and any content unrelated to financial statements or future plans.*
- 6. Do not format the output as a list or include numbering of any kind.*
- 7. Output only the extracted sentences as plain text, each on a new line.*
- 8. After extracting the sentences, provide an overall sentiment score for the text on a scale from -1 to 1:*
 - 1: Strongly Positive*
 - 0: Neutral*
 - -1: Strongly Negative*
 - Base this score on the tone, content, context and sentiment on profitability of the extracted sentences.*

Provide the extracted sentences as plain text followed by the sentiment score. Output should be JSON.

3η Προσέγγιση: Συναισθηματική Ανάλυση των Εξαγόμενων Προτάσεων από ChatGPT μέσω FinBERT

Οι προτάσεις που εξήχθησαν μέσω του ChatGPT χρησιμοποιήθηκαν στη συνέχεια ως είσοδος στο μοντέλο FinBERT. Η διαδικασία ήταν πανομοιότυπη με την 1η προσέγγιση, με τη διαφορά ότι αυτή τη φορά το FinBERT ανέλυσε αποκλειστικά τις προτάσεις που είχαν ήδη προεπιλεγεί από το ChatGPT. Το FinBERT παρήγαγε sentiment scores (positive, negative, neutral) για κάθε πρόταση, διατηρώντας τη λεπτομερή καταγραφή του συναισθήματος, αλλά βασιζόμενο σε μια ήδη φιλτραρισμένη και επιμελημένη συλλογή προτάσεων.

Με αυτόν τον τρόπο, η προσέγγιση αυτή συνδύασε τα πλεονεκτήματα της ακρίβειας του FinBERT με την ικανότητα του ChatGPT να απομονώνει σχετικό περιεχόμενο, παρέχοντας ένα ακόμη επίπεδο ανάλυσης για τα SEC filings.

5.3.2.2. Υπολογισμός sentiment scores

Για την ποσοτικοποίηση των συναισθηματικών χαρακτηριστικών των SEC filings, αναπτύχθηκε μια συστηματική διαδικασία βασισμένη σε προηγμένα εργαλεία επεξεργασίας φυσικής γλώσσας. Συγκεκριμένα, χρησιμοποιήθηκε το FinBERT, ένα γλωσσικό μοντέλο ειδικά σχεδιασμένο για ανάλυση συναισθημάτων σε χρηματοοικονομικά κείμενα. Το FinBERT εφαρμόστηκε στις επιλεγμένες προτάσεις που εξήχθησαν από τα SEC filings για την παραγωγή τριών βασικών sentiment scores: overall sentiment score, exclusive sentiment score, και prevailing sentiment score. Η ανάλυση πραγματοποιήθηκε σε διαφορετικές ενότητες των εγγράφων, όπως τα Risk Factors, Management Discussion, και οι Press Releases.

Η συνολική διαδικασία περιλάμβανε τα εξής βήματα:

1. **Καθαρισμός Κειμένων:** Η διαδικασία καθαρισμού εξασφάλισε ότι οι προτάσεις ήταν καλά δομημένες και περιείχαν τουλάχιστον επτά σημαντικές λέξεις (3 ρήματα και 4 ουσιαστικά), ώστε να είναι κατάλληλες για ανάλυση.
2. **Διαχωρισμός σε Παρτίδες:** Δεδομένου του περιορισμού των 512 tokens του FinBERT, το κείμενο χωρίστηκε σε παρτίδες. Κάθε παρτίδα περιείχε προτάσεις που μπορούσαν να αναλυθούν αποτελεσματικά χωρίς να υπερβαίνουν αυτόν τον περιορισμό.
3. **Ανάλυση Συναισθημάτων με FinBERT:** Οι παρτίδες αναλύθηκαν χρησιμοποιώντας το FinBERT. Το αποτέλεσμα ήταν ένα σύνολο προβλέψεων με ετικέτες Positive, Negative, και Neutral, συνοδευόμενο από τις αντίστοιχες πιθανότητες.
4. **Υπολογισμός Συνολικών και Ειδικών Δεικτών:** Δεδομένου ότι σε πολλές περιπτώσεις τα αποτελέσματα της συναισθηματικής ανάλυσης περιλάμβαναν μεγάλο αριθμό neutral προβλέψεων, η δημιουργία διαφορετικών sentiment scores κρίθηκε απαραίτητη για την πληρέστερη κατανόηση των συναισθηματικών τάσεων.

a. Overall Sentiment Score:

Υπολογίστηκε ως ο σταθμισμένος μέσος όρος των συναισθηματικών ετικετών, όπου οι ετικέτες Positive, Negative, και Neutral αντιστοιχούσαν σε τιμές +1, -1, και 0 αντίστοιχα.

$$\text{Overall Sentiment Score} = \frac{\sum_{i=1}^N \text{label}_i \times \text{score}_i}{N}$$

όπου label_i είναι οι τιμές (+1 για Positive, -1 για Negative, 0 για Neutral) και score_i οι πιθανότητες.

b. Exclusive Sentiment Score:

Προέκυψε ως ο μέσος όρος των πιθανών τιμών Positive και Negative, εξαιρώντας την ουδέτερη ετικέτα.

$$\text{Exclusive Sentiment Score} = \frac{\sum \text{Positive Scores} + \sum \text{Negative Scores}}{N}$$

c. Prevailing Sentiment Score:

Υπολόγισε τον επικρατέστερο τόνο, βασισμένο στον αριθμό των θετικών ή αρνητικών προτάσεων. Αν υπερείχαν οι θετικές προτάσεις, το αποτέλεσμα ήταν ο μέσος όρος των Positive scores, και το αντίθετο για τις αρνητικές.

$$\text{Prevailing Sentiment Score} = \begin{cases} \frac{\sum \text{Positive Scores}}{N}, & \text{Positive} > \text{Negative} \\ \frac{\sum \text{Negative Scores}}{N}, & \text{Positive} < \text{Negative} \\ 0, & \text{Positive} = \text{Negative} \end{cases}$$

Το overall sentiment score αποτύπωνε τη συνολική συναισθηματική φόρτιση λαμβάνοντας υπόψη όλες τις ετικέτες, ενώ το exclusive sentiment score εξέταζε αποκλειστικά τις θετικές και αρνητικές προβλέψεις, αποφεύγοντας την επίδραση των ουδέτερων τιμών. Παράλληλα, το prevailing sentiment score εστίαζε στην κυρίαρχη συναισθηματική κατεύθυνση, αποτυπώνοντας με μεγαλύτερη σαφήνεια τον κυρίαρχο τόνο κάθε εγγράφου. Με αυτόν τον τρόπο, διασφαλίστηκε ότι η ποσοτικοποίηση των SEC filings δεν επηρεαζόταν υπερβολικά από τις ουδέτερες ετικέτες, επιτρέποντας πιο ακριβή ερμηνεία των δεδομένων στις χρηματοοικονομικές αναλύσεις.

Για την εξαγωγή ενός συνολικού δείκτη συναισθήματος (sentiment score) για τις κατηγορίες 10-K και 10-Q, σχεδιάστηκε και υλοποιήθηκε μια προσαρμοσμένη μέθοδος συνδυασμού των επιμέρους ενότητων των καταχωρίσεων τύπου της κάθε μιας, όπως περιγράφεται αναλυτικά στην ενότητα 5.1.4. Η μέθοδος αυτή βασίστηκε στη στάθμιση δύο βασικών θεματικών πεδίων των SEC filings: των παραγόντων ρίσκου (Risk Factors) και των δηλώσεων της διοίκησης (Management Discussion and Analysis).

Η συναισθηματική πληροφορία από τις δύο αυτές ενότητες δεν συνδυάστηκε απλώς μέσω ενός σταθμισμένου αθροίσματος, αλλά λήφθηκε υπόψη η μεταξύ τους συμφωνία ή αντίθεση ως προς το πρόσημο της συναισθηματικής έκφρασης. Συγκεκριμένα, στην περίπτωση όπου και οι δύο τιμές φέρουν το ίδιο πρόσημο (και οι δύο θετικές ή και οι δύο αρνητικές), τότε το τελικό sentiment score προκύπτει μέσω ενίσχυσης της έντασης, δηλαδή του γινομένου των επιμέρους scores με τους αντίστοιχους συντελεστές στάθμισης. Αντίθετα, όταν τα πρόσημα διαφέρουν, η τελική τιμή προκύπτει από το σταθμισμένο άθροισμά τους, αντανakλώντας μια πιο ισορροπημένη εικόνα.

Ο αλγόριθμος ορίζεται ως εξής:

$$\text{Combined Sentiment Score} = \begin{cases} a * x + b * y, & \text{αν } x * y < 0 \\ a * x * b * y, & \text{αν } x < 0 \wedge y < 0 \\ -(a * x * b * y), & \text{αν } x < 0 \vee y < 0 \\ x \text{ ή } y, & \text{αν η μια τιμή είναι κενή} \end{cases}$$

όπου:

- ο x είναι το sentiment score της ενότητας Risk Factors,
- ο y το αντίστοιχο score της ενότητας Management Discussion,
- ο $a=0.4$ και $b=0.6$ είναι οι συντελεστές στάθμισης.

Η επιλογή αυτών των συντελεστών δεν είναι αυθαίρετη. Δόθηκε μεγαλύτερη βαρύτητα στη διοικητική ενότητα, καθώς θεωρείται ότι περιέχει στρατηγικές εκτιμήσεις και μελλοντικές προβλέψεις που αντικατοπτρίζουν την πρόθεση και τη θέση της διοίκησης, στοιχεία ιδιαίτερα κρίσιμα για την εκτίμηση της μελλοντικής πορείας μιας εταιρείας.

Εφαρμόζοντας αυτή τη μεθοδολογία για κάθε τύπο filing και για κάθε κατηγορία βαθμολόγησης δημιουργήθηκε το Combined Sentiment score κάθε μιας. Η κατηγορία EightK δεν υπέστη κάποια επεξεργασία καθώς αποτελείται από ένα και μοναδικό score.

Για κάθε τύπο καταχώρισης (filing) προέκυψαν τα αντίστοιχα χαρακτηριστικά συναισθήματος, τα οποία αποτυπώθηκαν ως εξής:

1. **TenK_Overall_Combined_Sentiment_Score,**
TenQ_Overall_Combined_Sentiment_Score,
EightK_Press_Release_Overall_Sentiment_Score
2. **TenK_Exclusive_Combined_Sentiment_Score,**
TenQ_Exclusive_Combined_Sentiment_Score,
EightK_Press_Release_Exclusive_Sentiment_Score
3. **TenK_Prevaling_Combined_Sentiment_Score,**
TenQ_Prevaling_Combined_Sentiment_Score,
EightK_Press_Release_Prevaling_Sentiment_Score.

Οι μεταβλητές αυτές επέτρεψαν την ξεχωριστή αποτίμηση και περαιτέρω ανάλυση της συναισθηματικής πληροφορίας για κάθε κατηγορία εγγράφου, ενισχύοντας την ακρίβεια της ποσοτικής ανάλυσης και διευκολύνοντας τη συσχέτιση των αποτελεσμάτων με τις μεταβολές στη συμπεριφορά της αγοράς.

5. **Αποθήκευση των Αποτελεσμάτων:** Τα αποτελέσματα ενημέρωσαν τις σχετικές στήλες στη βάση δεδομένων SQLite για κάθε κατηγορία εγγράφων

5.4. Επιλογή Χαρακτηριστικών (Feature Selection)

5.4.1. Ανάλυση Συσχέτισης (Correlation analysis)

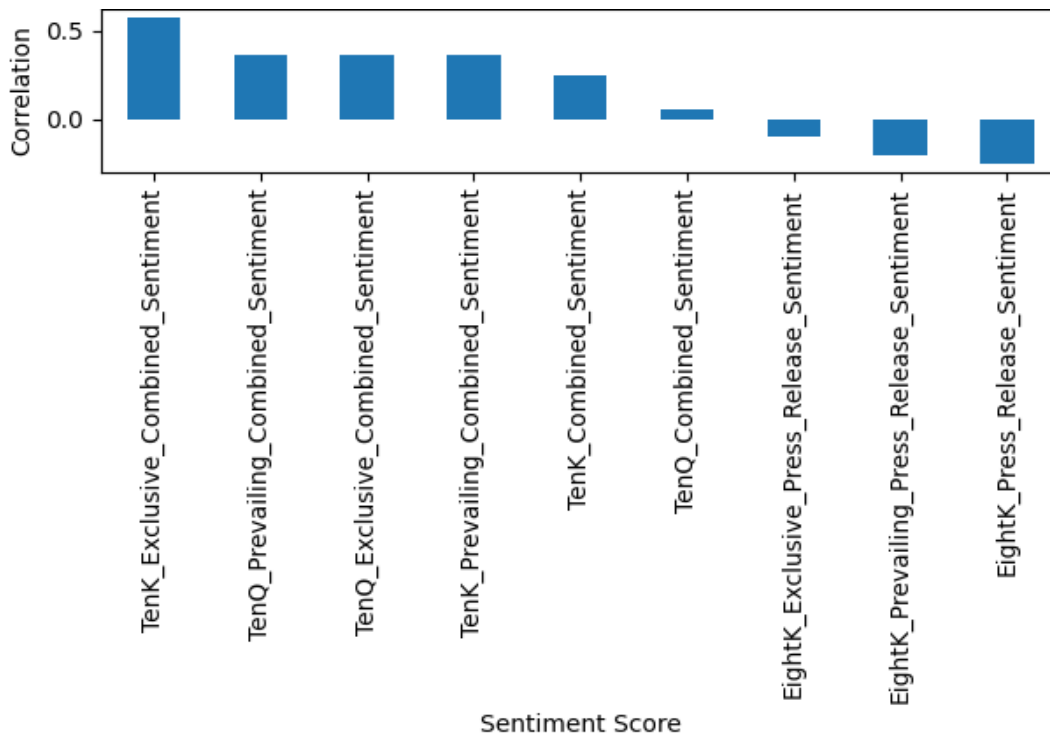
Το Correlation Analysis διεξήχθη με στόχο την αξιολόγηση της σχέσης μεταξύ των sentiment scores και των cumulative returns της επόμενης ημέρας για κάθε μετοχή. Η ανάλυση αποσκοπούσε στον προσδιορισμό των πιο σημαντικών sentiment scores, τα οποία θα μπορούσαν να χρησιμοποιηθούν για την ανάπτυξη ενός ακριβούς και αξιόπιστου προβλεπτικού μοντέλου.

Αρχικά, εξετάστηκαν τρεις διαφορετικές προσεγγίσεις για την εξαγωγή των sentiment scores:

1. **Finbert Manual Scores:** Τα scores εξήχθησαν αποκλειστικά μέσω του FinBERT από τα πρωτογενή δεδομένα των SEC filings.
2. **Finbert GPT Scores:** Οι προτάσεις των SEC filings φιλτραρίστηκαν πρώτα μέσω GPT, και στη συνέχεια αναλύθηκαν από το FinBERT.
3. **GPT Scores:** Τα scores δημιουργήθηκαν αποκλειστικά από το API του ChatGPT.

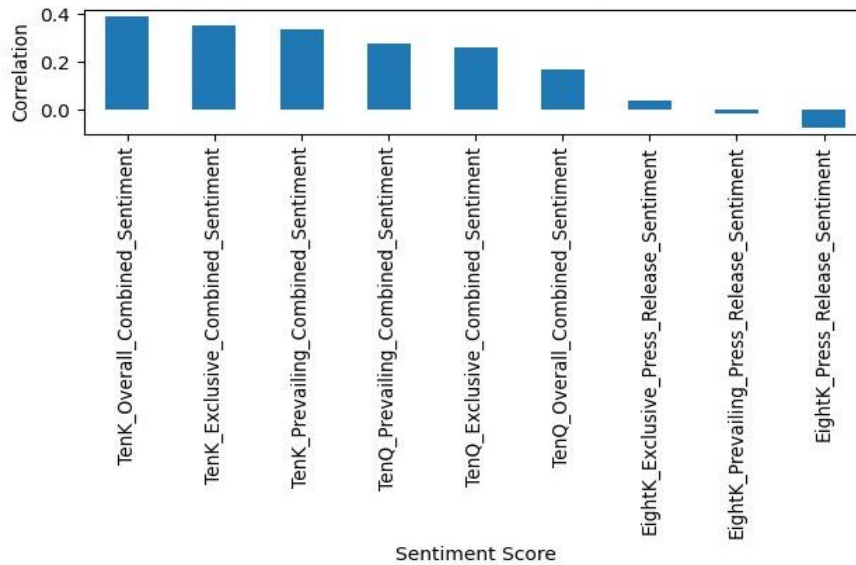
5.4.1.1. Finbert Manual Scores:

Η ανάλυση των συσχετίσεων έδειξε ότι τα TenK Exclusive Sentiment Scores και TenQ Exclusive Sentiment Scores είχαν τη μεγαλύτερη συσχέτιση με τις cumulative returns της επόμενης ημέρας. Αυτό υποδεικνύει ότι οι συγκεκριμένες ενότητες των SEC filings περιέχουν κρίσιμες πληροφορίες για την απόδοση των μετοχών.



Εικόνα 24: Διάγραμμα συσχέτισης 1^{ης} μεθόδου (FinBERT)

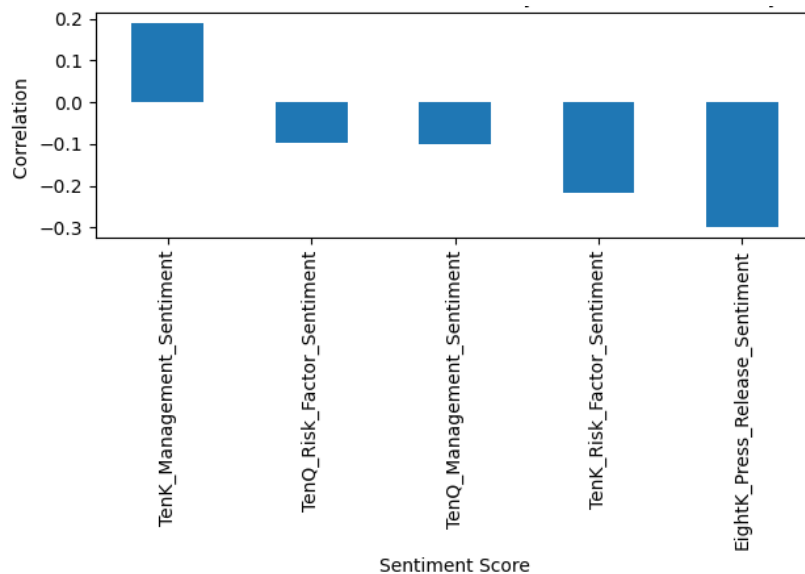
5.4.1.2. Finbert GPT Scores:



Όταν οι προτάσεις φιλτράρονταν μέσω GPT και στη συνέχεια αναλύονταν με το FinBERT, τα αποτελέσματα παρουσίασαν υψηλή συσχέτιση, ιδιαίτερα για τις ενότητες TenK Exclusive και TenQ Exclusive. Ωστόσο η τιμές της συσχέτισης ήταν μειωμένες σε σχέση με την πρώτη προσέγγιση.

Εικόνα 22: Διάγραμμα συσχέτισης 2^{ης} μεθόδου. (FinBERT + GPT)

5.4.1.3. GPT Scores:



Τα scores που δημιουργήθηκαν αποκλειστικά μέσω GPT είχαν μικρότερη συσχέτιση σε σχέση με τις δύο προηγούμενες μεθόδους, γεγονός που υποδεικνύει ότι η αυτόνομη χρήση του GPT για sentiment ανάλυση δεν είναι εξίσου αποτελεσματική στη συγκεκριμένη περίπτωση.

Εικόνα 23: Διάγραμμα συσχέτισης 3^{ης} μεθόδου.

5.4.1.4. Συγκεντρωτικά αποτελέσματα

Βάσει των αποτελεσμάτων της ανάλυσης, για κάθε μοντέλο επιλέχθηκαν ως sentiment features οι διπλές sentiment που παρουσίαζαν τη μεγαλύτερη συσχέτιση με τα cumulative returns.

Αξίζει να σημειωθεί ότι, όπως παρουσιάζεται και παρακάτω, η μεταβλητή στόχος του μοντέλου είναι η δυαδική κατεύθυνση της απόδοσης της επόμενης ημέρας (binary next day direction), δηλαδή αν η τιμή κλεισίματος της επόμενης ημέρας είναι υψηλότερη ή χαμηλότερη. Επιπλέον, τα sentiment scores που χρησιμοποιούνται ως χαρακτηριστικά είναι αριθμητικές (numeric) μεταβλητές, ενώ η target μεταβλητή είναι διακριτή δυαδική (binary discrete).

Η διερεύνηση της συσχέτισης με τις cumulative returns δεν μειώνει την εγκυρότητα της ανάλυσης, καθώς τόσο οι σωρευτικές αποδόσεις όσο και η δυαδική κατεύθυνση βασίζονται στην ίδια υποκείμενη δυναμική τιμών και διατηρούν ισχυρή μαθηματική σύνδεση. Η εκτίμηση της συσχέτισης με τις cumulative returns παρέχει πολύτιμες ενδείξεις για τη γενική τάση και το μέγεθος του αντίκτυπου των ποιοτικών παραμέτρων (sentiment), ενώ η binary κατεύθυνση αξιοποιείται τελικά ως στόχος για την απλούστερη και πιο πρακτική πρόβλεψη της μελλοντικής κατεύθυνσης τιμών.

Κατά συνέπεια, η προσέγγιση αυτή θεωρείται μεθοδολογικά ισοδύναμη και επιστημονικά τεκμηριωμένη, διατηρώντας τη συνοχή μεταξύ της διερευνητικής ανάλυσης και της προβλεπτικής εφαρμογής.

5.5. Διερευνητική ανάλυση (EDA-Exploratory Data Analysis)

Κατά την προεπεξεργασία των δεδομένων, πραγματοποιήθηκε διερευνητική ανάλυση (EDA) στα χαρακτηριστικά εισόδου των μοντέλων πρόβλεψης με στόχο την κατανόηση της κατανομής, της μεταβλητότητας και των μεταξύ τους συσχετίσεων. Η διαδικασία αυτή αποσκοπεί στον εντοπισμό πιθανών προβλημάτων στα δεδομένα, όπως ακραίες τιμές ή ελλείψεις, αλλά και στην εξαγωγή πληροφοριών που μπορούν να αξιοποιηθούν για τη βελτίωση της επιλογής χαρακτηριστικών και της συνολικής απόδοσης των μοντέλων.

Για τα χαρακτηριστικά TenK και TenQ, η διερευνητική ανάλυση πραγματοποιήθηκε αποκλειστικά για τις περιπτώσεις όπου οι μεταβλητές αυτές έφεραν μη μηδενική τιμή. Η επιλογή αυτή βασίστηκε στο γεγονός ότι οι συγκεκριμένοι δείκτες συναισθήματος είναι διαθέσιμοι μόνο για περιορισμένο αριθμό ημερών, καθώς υπολογίζονται βάσει της ύπαρξης και επεξεργασίας σχετικών εταιρικών ανακοινώσεων. Η συμπερίληψη των ημερών με μηδενική τιμή, που αντιπροσωπεύουν απουσία σχετικού περιεχομένου, θα αλλοίωνε τη στατιστική εικόνα και θα μείωνε την ερμηνευτική αξία των μέτρων θέσης και διασποράς. Για τον λόγο αυτό, εξαιρέθηκαν από τον υπολογισμό των περιγραφικών στατιστικών, ώστε η ανάλυση να αποτυπώνει με ακρίβεια τα χαρακτηριστικά του συναισθηματικού περιεχομένου τις ημέρες όπου αυτό πράγματι υφίσταται.

Στατιστικά για την APPLE:

Μεταβλητή	Mean	Std	Min	25%	50%	75%	Max
intraday_return	0.0006	0.0256	-0.8549	-0.0075	0.0009	0.0103	0.1198
daily_volatility	0.0165	0.0196	0.0012	0.0097	0.0134	0.0193	0.3250
percentile_rank	0.4996	0.2895	0.0003	0.2487	0.4972	0.7515	1.0000
rolling_rets_1	0.0134	0.1170	-0.8556	-0.0296	0.0217	0.0758	0.3589
distance_from_mean	0.0176	16.3811	-467.8186	-1.9502	0.7379	3.3796	44.61
TRIX	0.0051	0.5847	-4.8151	-0.0991	0.0847	0.2638	0.8962
TenK	0.0078	0.0029	0.0046	0.0055	0.0073	0.0102	0.0136
TenQ	0.1029	0.1482	0.0083	0.0158	0.0201	0.2073	0.4648

Στατιστικά για την NVDA:

Μεταβλητή	Mean	Std	Min	25%	50%	75%	Max
intraday_return	0.0016	0.0396	-0.7748	-0.0161	0.0012	0.0192	0.2440
daily_volatility	0.0323	0.0224	0.0049	0.0199	0.0273	0.0387	0.3007
percentile_rank	0.5029	0.2871	0.0016	0.2545	0.5027	0.7515	1.0000
rolling_rets_1	0.0367	0.2029	-0.8118	-0.0776	0.0207	0.1282	1.1549
distance_from_mean	0.1989	47.8363	-1386.1314	-4.0557	0.4214	6.3242	277.38
TRIX	0.0694	0.7350	-3.8589	-0.2831	0.0617	0.4419	2.8625
TenK	0.0068	0.0043	0.0026	0.0033	0.0046	0.0103	0.0144
TenQ	0.0242	0.0088	0.0092	0.0182	0.0230	0.0282	0.0465

Στατιστικά για την TSLA:

Μεταβλητή	Mean	Std	Min	25%	50%	75%	Max
intraday_return	0.0015	0.0346	-0.8993	-0.0124	0.0016	0.0160	0.2981
daily_volatility	0.0260	0.0230	0.0031	0.0149	0.0218	0.0306	0.3473
percentile_rank	0.5015	0.2890	0.0003	0.2505	0.5028	0.7515	1.0000
rolling_rets_1	0.0308	0.1612	-0.8953	-0.0427	0.0294	0.1141	0.6531
distance_from_mean	0.0985	28.9562	-903.3171	-0.5995	0.1670	2.4821	136.49
TRIX	0.0570	0.7248	-5.1361	-0.1270	0.1206	0.4160	1.8770
TenK	0.0409	0.0756	0.0013	0.0022	0.0030	0.0057	0.1911
TenQ	0.0848	0.1475	0.0053	0.0130	0.0236	0.0348	0.4487

6. Πειραματική Διαδικασία και Αποτελέσματα

6.1. Αρχιτεκτονική των μοντέλων

Η επιλογή της αρχιτεκτονικής του μοντέλου βασίστηκε στα χαρακτηριστικά των δεδομένων που συλλέχθηκαν και προεπεξεργάστηκαν στα προηγούμενα στάδια. Καθώς το πρόβλημα της παρούσας διπλωματικής αφορά την πρόβλεψη χρηματιστηριακών αποδόσεων βάσει χρονοσειρών και ανάλυσης συναισθήματος, κρίθηκε απαραίτητη η χρήση ενός μοντέλου που να είναι ικανό να συλλάβει διαχρονικές εξαρτήσεις και να αξιοποιήσει τις δομές των δεδομένων με τον πιο αποδοτικό τρόπο.

Η αρχιτεκτονική που υλοποιήθηκε βασίζεται σε Επαναλαμβανόμενες Νευρωνικές Δομές (Recurrent Neural Networks - RNNs) και συγκεκριμένα στις Μονάδες GRU (Gated Recurrent Units), οι οποίες προσφέρουν σημαντικά πλεονεκτήματα έναντι των κλασικών RNNs και LSTMs. Όπως περιγράφεται αναλυτικά στην ενότητα 3.6, το τελικό μοντέλο αποτελείται από δύο διαδοχικές στρώσεις GRU, οι οποίες πλαισιώνονται από στρώσεις Dropout για τη μείωση του overfitting και μία τελική πλήρως συνδεδεμένη στρώση (Dense) που επιστρέφει την προβλεπόμενη τιμή της απόδοσης της μετοχής.

Η συγκεκριμένη επιλογή αρχιτεκτονικής δικαιολογείται από τα εξής:

1. **Διαχείριση Χρονοσειρών και Μακροπρόθεσμων Εξαρτήσεων:** Οι GRU μονάδες είναι κατάλληλες για την ανάλυση χρηματοοικονομικών δεδομένων, καθώς επιτρέπουν τη διατήρηση σημαντικών πληροφοριών σε μεγαλύτερα χρονικά διαστήματα, αντιμετωπίζοντας έτσι το πρόβλημα των εξαφανιζόμενων κλίσεων (vanishing gradient problem) που παρατηρείται σε απλά RNNs.
2. **Αποδοτική Εκπαίδευση και Υπολογιστική Βελτιστοποίηση:** Οι GRU μονάδες, σε σύγκριση με τις LSTM, χρησιμοποιούν λιγότερες παραμέτρους, γεγονός που οδηγεί σε ταχύτερη εκπαίδευση και μικρότερη απαίτηση υπολογιστικών πόρων, χωρίς σημαντική απώλεια στην απόδοση.
3. **Προσαρμογή στη Διακύμανση των Αγορών:** Η ενσωμάτωση στρώσεων Dropout με ποσοστά 20% και 40% στις ενδιάμεσες στρώσεις βοηθά στη γενίκευση του μοντέλου και στη μείωση του κινδύνου υπερπροσαρμογής, λαμβάνοντας υπόψη τη δυναμική φύση των χρηματιστηριακών δεδομένων.
4. **Συμπερίληψη Συναισθηματικής Ανάλυσης:** Τα χαρακτηριστικά συναισθηματικής ανάλυσης που προκύπτουν από τα SEC Filings ενσωματώνονται στο μοντέλο ως πρόσθετες μεταβλητές εισόδου. Αυτή η προσέγγιση επιτρέπει στο δίκτυο να ενσωματώσει πληροφορίες που προκύπτουν από μη δομημένα κείμενα και να ενισχύσει τη δυνατότητα πρόβλεψης των τάσεων της αγοράς.
5. **Χρήση 15-ημερών Παραθύρου:** Η επιλογή ενός παραθύρου 30 ημερών για την εισαγωγή των δεδομένων στο δίκτυο επιτρέπει την αξιοποίηση παρελθοντικών αποδόσεων και οικονομικών γεγονότων για την εξαγωγή προγνωστικών αποτελεσμάτων με υψηλή ακρίβεια.

Το μοντέλο της παρούσας διπλωματικής βασίζεται σε μια σειρά επαναληπτικών μονάδων GRU, που είναι κατάλληλες για την επεξεργασία χρονοσειρών. Η αρχιτεκτονική περιλαμβάνει τρεις κύριες στρώσεις GRU, με δύο στρώσεις Dropout για την αποτροπή της υπερπροσαρμογής (overfitting). Η δομή είναι η εξής:

- **Είσοδος:** Παράμετρος `input_shape` (π.χ. `n_past` ημέρες $\times$ αριθμός χαρακτηριστικών)
- **GRU 1:** 128 μονάδες με `return_sequences=True`, ώστε να διατηρούνται όλες οι χρονοσειρές για τη συνέχεια
- **Layer Normalization** μετά την πρώτη GRU, για σταθεροποίηση της κατανομής των εξόδων
- **Dropout (20%)** μετά τη Layer Normalization, για περιορισμό του overfitting
- **GRU 2:** 64 μονάδες με `return_sequences=True`, για εκμάθηση βαθύτερων χρονικών εξαρτήσεων
- **Layer Normalization** μετά τη δεύτερη GRU
- **Dropout (20%)** εκ νέου για ενίσχυση της γενίκευσης
- **Dense** στρώση με 64 νευρώνες και ενεργοποίηση ReLU, ώστε να μοντελοποιηθούν μη γραμμικοί συνδυασμοί των χαρακτηριστικών
- **Dropout (20%)** πριν την έξοδο, για επιπρόσθετη τυχαία απενεργοποίηση νευρώνων
- **Έξοδος:** Dense layer με 1 νευρώνα και λειτουργία ενεργοποίησης sigmoid, που επιστρέφει πιθανότητα ανοδικής τάσης
- **Εκπαίδευση:**
 - **Optimizer:** Adam με learning rate = $1e-3$
 - **Loss function:** Focal loss ($\gamma=2$, $\alpha=0.25$), κατάλληλη για ανισορροπία κλάσεων
 - **Metrics:** Accuracy, για παρακολούθηση του ποσοστού σωστών προβλέψεων

6.2. Εκπαίδευση και Αξιολόγηση των Μοντέλων

Τα δεδομένα περιορίστηκαν σε χρονικά διαστήματα από το 2010 έως το 2023, προκειμένου να επικεντρωθούμε σε μια καθορισμένη χρονική περίοδο. Η διαδικασία εκπαίδευσης περιλάμβανε την εκπαίδευση δύο διαφορετικών μοντέλων, με βάση διαφορετικά σύνολα χαρακτηριστικών (features).

1. **Baseline Μοντέλο:** Το **baseline μοντέλο** χρησιμοποιήθηκε για να συγκριθεί με το βελτιωμένο μοντέλο και περιλάμβανε τα χαρακτηριστικά που παρουσιάστηκαν στην ενότητα 5.3.1 χρησιμοποιώντας το next-day direction ως τιμή στόχο.
2. **Βελτιωμένο Μοντέλο:** Το **βελτιωμένο μοντέλο** περιλάμβανε τα χαρακτηριστικά του baseline μοντέλου, προσθέτοντας επιπλέον πληροφορίες για την ενίσχυση της πρόβλεψης:
 - **TenK_Sentiment:** Ο συνδυασμένος δείκτης συναισθήματος που εξάγεται από τις αναφορές 10-K.

- **TenQ_Sentiment:** Ο συνδυασμένος δείκτης συναισθήματος που εξάγεται από τις αναφορές 10-Q.

Και τα δύο μοντέλα εκπαιδεύτηκαν για 150 επαναλήψεις (epochs) με μέγεθος δέσμης (batch size) 32, χρησιμοποιώντας τα δεδομένα εκπαίδευσης.

Επιπλέον, για κάθε μετοχή εκπαιδεύτηκε ένα ξεχωριστό μοντέλο, το οποίο έλαβε υπόψη τα συγκεκριμένα χαρακτηριστικά που αφορούσαν κάθε μετοχή. Η χρήση αυτής της στρατηγικής επιτρέπει την εξατομικευμένη προσαρμογή των μοντέλων στις ιδιαίτερες συνθήκες και δυναμικές κάθε μετοχής. Μετά την εκπαίδευση, το μοντέλο χρησιμοποιείται για να προβλέψει τις αποδόσεις για τα δεδομένα δοκιμής. Οι προβλέψεις συγκρίνονται με τις πραγματικές τιμές. Η απόδοση του μοντέλου αξιολογείται με τη χρήση διάφορων μετρικών, όπως το accuracy, precision και MSE, για να κατανοήσουμε την ποιότητα της πρόβλεψης.

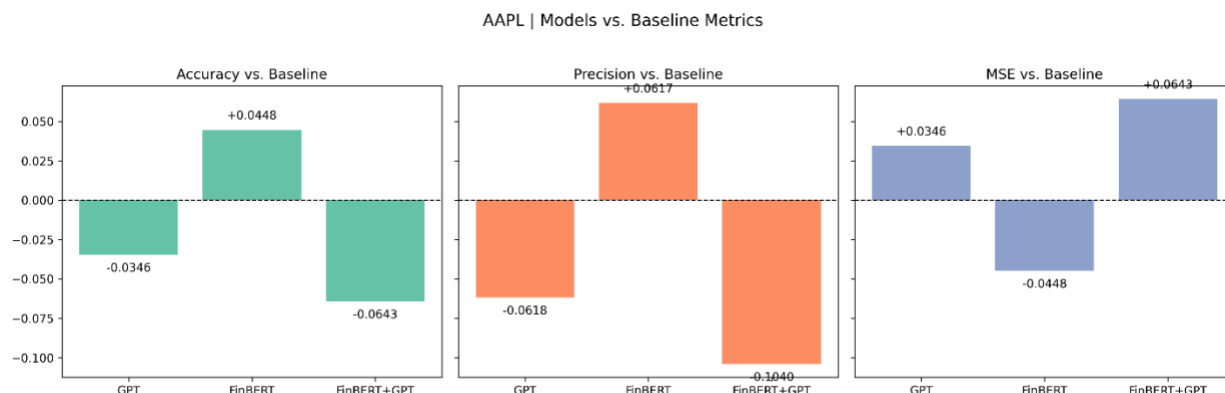
6.3. Αποτελέσματα Ανά Μοντέλο για Κάθε Μετοχή

Στην παρούσα ενότητα παρουσιάζονται τα αποτελέσματα της εμπειρικής αξιολόγησης των μοντέλων πρόβλεψης που αναπτύχθηκαν στο πλαίσιο της διπλωματικής εργασίας. Στόχος της ανάλυσης είναι η αποτίμηση της ικανότητας των επιλεγμένων προσεγγίσεων να προβλέψουν τη μελλοντική κατεύθυνση και απόδοση χρηματοοικονομικών τίτλων, αξιοποιώντας δεδομένα θεμελιώδους και συναισθηματικής ανάλυσης.

Η πειραματική διαδικασία βασίστηκε σε σταθερό διαχωρισμό των δεδομένων εκπαίδευσης και αξιολόγησης. Πιο συγκεκριμένα, για κάθε μετοχή, το σύνολο εκπαίδευσης περιλάμβανε ιστορικά δεδομένα από το 2010 έως και το 2022, ενώ η αξιολόγηση των μοντέλων πραγματοποιήθηκε αποκλειστικά σε δεδομένα του έτους 2023. Με αυτόν τον τρόπο, διασφαλίζεται ότι τα μοντέλα προβλέπουν σε εκτός δείγματος δεδομένα (out-of-sample), αποτυπώνοντας ρεαλιστικά τη συμπεριφορά τους σε μελλοντικές περιόδους.

Για την ποιοτική και ποσοτική αποτίμηση των αποτελεσμάτων, χρησιμοποιούνται γραφήματα στήλης (bar plots) που παρουσιάζουν τη συγκριτική αξιολόγηση των μοντέλων βάσει τεσσάρων βασικών μετρικών: ακρίβεια (Accuracy), ακρίβεια θετικών προβλέψεων (Precision), ανάκληση (Recall) και μέσο τετραγωνικό σφάλμα (Mean Squared Error - MSE). Οι μετρικές αυτές υπολογίζονται για το σύνολο των προβλέψεων κάθε μοντέλου στο έτος 2023 και απεικονίζονται γραφικά, επιτρέποντας τη σύγκριση της συνολικής αποδοτικότητας κάθε προσέγγισης. Με τον τρόπο αυτό, επιτυγχάνεται η αποτύπωση όχι μόνο της γενικής προγνωστικής ικανότητας (μέσω της Accuracy), αλλά και της ποιότητας των θετικών προβλέψεων (μέσω των Precision και Recall), καθώς και της αριθμητικής εγγύτητας των προβλέψεων προς τις πραγματικές τιμές (μέσω του MSE).

6.3.1. Αποτελέσματα Μετοχής AAPL



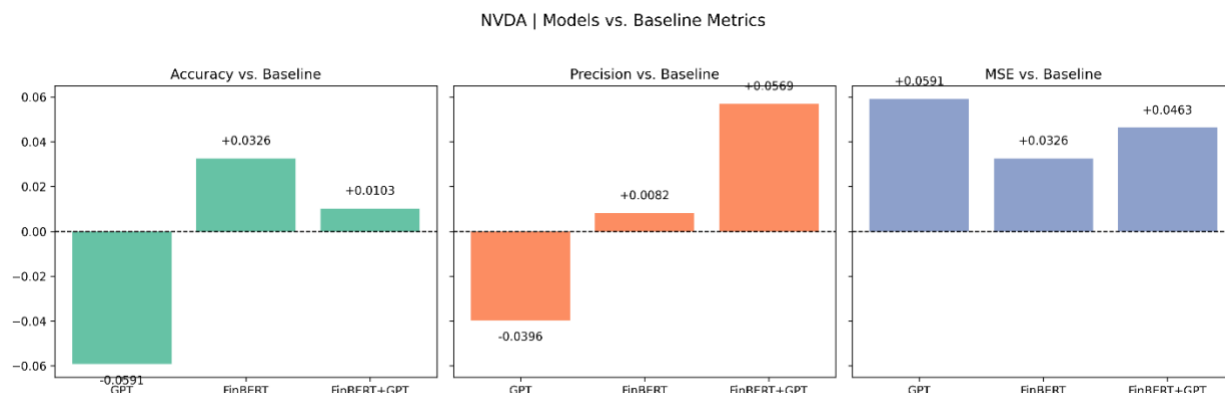
Το μοντέλο FinBERT κατέγραψε την πιο συνεπή και θετική επίδοση σε όλες τις μετρικές, σε σύγκριση με το baseline. Πιο συγκεκριμένα, παρουσίασε αύξηση της accuracy κατά +4,48 ποσοστιαίες μονάδες, της precision κατά +6,17 μονάδες, ενώ το MSE μειώθηκε κατά 0,0448. Τα αποτελέσματα αυτά καταδεικνύουν την υψηλή ικανότητα του FinBERT να εντοπίζει με ακρίβεια τις ανοδικές κινήσεις της μετοχής, ελαχιστοποιώντας παράλληλα τα σφάλματα πρόβλεψης. Η επίδοσή του επιβεβαιώνει τη σημασία της θεματικής εξειδίκευσης στη χρηματοοικονομική γλώσσα για τη βελτιστοποίηση της πρόβλεψης της κατεύθυνσης τιμών σε περιβάλλοντα με σταθερό και σαφές ειδησεογραφικό αποτύπωμα, όπως αυτό της Apple.

Αντιθέτως, ο συνδυασμός FinBERT+GPT όχι μόνο δεν ενίσχυσε την απόδοση του FinBERT, αλλά οδήγησε σε αισθητή επιδείνωση των αποτελεσμάτων. Η accuracy υποχώρησε κατά 6,43 ποσοστιαίες μονάδες, ενώ η precision μειώθηκε κατά 10,40 μονάδες. Ταυτόχρονα, το MSE αυξήθηκε κατά +0,0643, γεγονός που υποδηλώνει μεγαλύτερη απόκλιση μεταξύ προβλεπόμενων και πραγματικών τιμών. Το εύρημα αυτό ενισχύει την υπόθεση ότι η ενσωμάτωση γενικής γλωσσικής πληροφορίας από το GPT μπορεί να εισάγει περιττό θόρυβο και να υποβαθμίζει την απόδοση ενός εξειδικευμένου μοντέλου σε περιπτώσεις όπου η προβλεπτική σηματοδότηση είναι ήδη επαρκής.

Παρομοίως, το μοντέλο GPT, όταν εφαρμόζεται αυτόνομα, απέτυχε να υπερβεί το baseline. Παρατηρήθηκε μείωση της accuracy κατά 3,46 ποσοστιαίες μονάδες και της precision κατά 6,18, ενώ το MSE αυξήθηκε κατά +0,0346. Η επίδοση αυτή επιβεβαιώνει την αδυναμία των γενικών γλωσσικών μοντέλων, χωρίς χρηματοοικονομική εξειδίκευση, να αποδώσουν αποτελεσματικά στην πρόβλεψη κατεύθυνσης μετοχών, ιδίως σε περιπτώσεις με περιορισμένη ειδησεογραφική πολυπλοκότητα.

Συνολικά, τα εμπειρικά αποτελέσματα για τη μετοχή της Apple υπογραμμίζουν με σαφήνεια την υπεροχή του FinBERT έναντι τόσο του baseline όσο και των άλλων μοντέλων. Η απόδοσή του καταδεικνύει τη σημασία της προσαρμογής του γλωσσικού μοντέλου στο περιεχόμενο και τη φύση των διαθέσιμων πληροφοριών. Αντιθέτως, η χρήση πιο γενικευμένων ή σύνθετων προσεγγίσεων χωρίς σαφή θεματική στόχευση μπορεί όχι μόνο να μην προσφέρει πρόσθετη προβλεπτική ισχύ, αλλά και να επιβαρύνει την απόδοση του συστήματος.

6.3.2. Αποτελέσματα Μετοχής NVDA



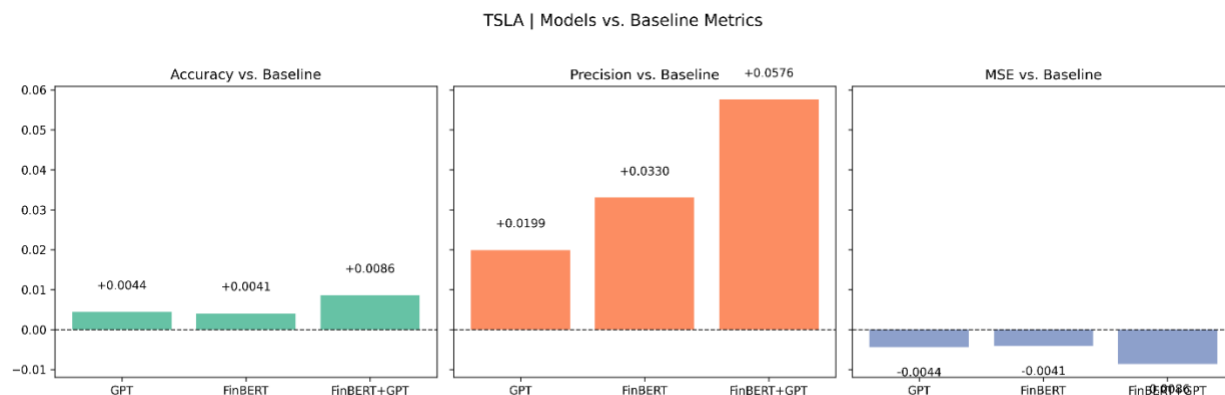
Το μοντέλο FinBERT+GPT καταγράφει τη συνολικά βέλτιστη απόδοση για τη συγκεκριμένη μετοχή. Συγκεκριμένα, παρουσιάζει αύξηση accuracy κατά +1,03 ποσοστιαίες μονάδες, precision κατά +5,69 και ταυτόχρονα εμφανίζει θετική διαφοροποίηση στο MSE (+0,0463), αν και αυτή η αύξηση υποδηλώνει ελαφρώς μεγαλύτερο σφάλμα. Παρ' όλα αυτά, οι σαφείς βελτιώσεις στις ποιοτικές μετρικές υποδηλώνουν ότι ο συνδυασμός εξειδικευμένης και γενικής γλωσσικής κατανόησης προσφέρει ενισχυμένη ικανότητα εντοπισμού ανοδικών ημερών και βελτιώνει τη γενική ταξινομητική απόδοση. Το εύρημα αυτό ερμηνεύεται ως ένδειξη ότι μετοχές όπως η NVIDIA, οι οποίες επηρεάζονται από πολυδιάστατη και τεχνολογικά προσανατολισμένη ειδησεογραφία, ωφελούνται από μοντέλα που μπορούν να επεξεργαστούν τόσο εξειδικευμένα όσο και γενικά γλωσσικά πρότυπα.

Αντιθέτως, το μοντέλο FinBERT, αν και παρουσιάζει οριακή βελτίωση στην precision (+0,82), εμφανίζει μείωση στην accuracy κατά -3,26 ποσοστιαίες μονάδες και αύξηση του MSE κατά +0,0326. Η απόδοσή του υποδεικνύει περιορισμένη ικανότητα στη συνολική πρόβλεψη κατεύθυνσης, παρά τη θεματική του εξειδίκευση. Η μέτρια επίδοση ενδέχεται να απορρέει από τη δυσκολία του FinBERT να ανταποκριθεί σε πιο ετερογενή συμφραζόμενα, στα οποία η γενική γλωσσική κατανόηση είναι ενδεχομένως πιο χρήσιμη.

Το μοντέλο GPT, όταν χρησιμοποιείται αυτόνομα, εμφανίζει τη χαμηλότερη συνολική απόδοση. Η accuracy μειώνεται κατά -5,91 ποσοστιαίες μονάδες, η precision κατά -3,96, ενώ το MSE αυξάνεται κατά +0,0591. Η απόδοση αυτή καταδεικνύει τη σχετική ανεπάρκεια των γενικών γλωσσικών μοντέλων, όταν αυτά λειτουργούν χωρίς εξειδικευμένο προσανατολισμό, να προβλέψουν την κατεύθυνση χρηματοοικονομικών τίτλων σε πλαίσια υψηλής τεχνολογικής εξειδίκευσης.

Συνοψίζοντας, τα αποτελέσματα για τη μετοχή της NVIDIA υποδεικνύουν ότι η υβριδική προσέγγιση FinBERT+GPT υπερέχει, αξιοποιώντας ταυτόχρονα εξειδικευμένη συναισθηματική πληροφορία και γενική γλωσσική κατανόηση. Η συγκεκριμένη περίπτωση αναδεικνύει τη σημασία της προσαρμογής του μοντέλου στις ιδιαιτερότητες του περιεχομένου πληροφόρησης κάθε μετοχής, ιδίως όταν αυτό χαρακτηρίζεται από ποικιλία θεματικών πηγών, τεχνολογική πολυπλοκότητα και ταχύτατα μεταβαλλόμενο περιβάλλον ειδήσεων.

6.3.3. Αποτελέσματα Μετοχής TSLA



Η μετοχή της Tesla (TSLA), γνωστή για τη δυναμική ειδησεογραφία και την έντονη διακύμανση της τιμής, παρέχει ένα ενδιαφέρον πεδίο για την αξιολόγηση μοντέλων πρόβλεψης χρηματιστηριακής κατεύθυνσης. Από τη συγκριτική ανάλυση των δεικτών accuracy, precision και mean squared error (MSE), προκύπτει ότι όλα τα εξεταζόμενα μοντέλα παρουσιάζουν θετικές μεταβολές στην ταξινομητική τους ικανότητα, με παράλληλη μικρή μείωση του MSE – ένδειξη περιορισμένης υπερπροσαρμογής.

Το υβριδικό μοντέλο FinBERT+GPT καταγράφει τη μεγαλύτερη βελτίωση απόδοσης έναντι του baseline: αύξηση accuracy κατά +0,86 μονάδες και precision κατά +5,76, με παράλληλη μείωση του MSE κατά $-0,0086$. Η υψηλή του επίδοση υποδεικνύει ότι η σύνθετη πληροφόρηση γύρω από τη μετοχή της Tesla – τεχνική, θεσμική, τεχνολογική και κοινωνική – απαιτεί συνδυαστική γλωσσική επεξεργασία. Η εξειδίκευση του FinBERT σε χρηματοοικονομικά συμφραζόμενα και η ευρεία σημασιολογική κάλυψη του GPT φαίνεται να προσφέρουν πλεονέκτημα σε περιβάλλοντα υψηλής πολυπλοκότητας.

Το FinBERT, παρότι υστερεί έναντι του υβριδικού μοντέλου, εμφανίζει αξιοσημείωτη αύξηση precision (+3,35) και βελτιωμένο MSE ($-0,0041$), ενώ η αύξηση accuracy είναι οριακή (+0,41). Τα αποτελέσματα δείχνουν ότι η θεματική εξειδίκευση συμβάλλει θετικά στην ταξινόμηση, ειδικά όταν οι ειδήσεις σχετίζονται με οικονομικά δεδομένα, χωρίς όμως να αξιοποιεί πλήρως την ευρύτερη πληροφορία που συνοδεύει τη μετοχή.

Αντιθέτως, το GPT, παρά την απουσία χρηματοοικονομικής εξειδίκευσης, καταφέρνει να βελτιώσει το precision (+1,99) και την accuracy (+0,44), παρουσιάζοντας ταυτόχρονα το χαμηλότερο MSE ($-0,0054$) μεταξύ των τριών μοντέλων. Το εύρημα αυτό είναι ενδεικτικό του ότι, σε περιπτώσεις με ευρύ και συχνά κοινωνικο-πολιτικό ενημερωτικό αποτύπωμα –όπως αυτό της Tesla–, η γενική γλωσσική κατανόηση μπορεί να συνεισφέρει καθοριστικά στη διακριτική πρόβλεψη της κατεύθυνσης της τιμής.

Συνολικά, η μετοχή της TSLA αναδεικνύει τη σημασία της πολυδιάστατης πληροφορικής επεξεργασίας στην πρόβλεψη χρηματιστηριακών τάσεων. Η υπεροχή του FinBERT+GPT, αλλά και οι αξιόλογες επιδόσεις του GPT, αποδεικνύουν ότι τα μοντέλα που ενσωματώνουν περισσότερα επίπεδα γλωσσικής αντίληψης είναι ιδιαίτερα αποτελεσματικά σε μετοχές με αυξημένη ειδησεογραφική και κοινωνική πολυπλοκότητα.

6.3.4. Σημαντικότητα Αποτελεσμάτων και Ερμηνεία Αποκλίσεων

Με βάση τη συγκριτική αξιολόγηση των μοντέλων FinBERT, FinBERT+GPT, GPT σε τρεις μετοχές υψηλής κεφαλαιοποίησης (AAPL, TSLA, NVDA) και τις μετρικές precision, accuracy και mean squared error (MSE), διαμορφώνεται ένα συνολικό συμπέρασμα που αναδεικνύει την απόδοση κάθε μοντέλου ανά περίπτωση και προτείνει ερμηνευτικές υποθέσεις για τα ευρήματα.

Σε γενικές γραμμές, το μοντέλο FinBERT εμφανίζει την πιο συνεπή και ισορροπημένη απόδοση. Στην περίπτωση της AAPL, το FinBERT υπερέχει τόσο σε precision όσο και accuracy, ενώ παρουσιάζει και ένα από τα χαμηλότερα MSE. Το αποτέλεσμα αυτό υποδεικνύει ότι η χρηματοοικονομική εξειδίκευση του μοντέλου το καθιστά ιδιαίτερος αποτελεσματικό σε σταθερές και προβλέψιμες μετοχές όπως η Apple, όπου η συναισθηματική πληροφορία έχει πιο σαφές και ερμηνεύσιμο αποτύπωμα στις τιμές.

Η TSLA, αν και χαρακτηρίζεται από υψηλή μεταβλητότητα και έντονη ειδησεογραφική δραστηριότητα, αποτελεί επίσης περίπτωση όπου το FinBERT+GPT παρουσιάζει την καλύτερη συνολική απόδοση. Καταγράφει την υψηλότερη precision και accuracy, με ταυτόχρονα χαμηλό MSE, γεγονός που δείχνει ότι το εξειδικευμένο λεξιλόγιο του μοντέλου μπορεί να αποδώσει ακόμη και σε πιο ασταθείς αγορές.

Στην περίπτωση της NVDA, τη συνολική υπεροχή παρουσιάζει ο συνδυασμός FinBERT+GPT, ο οποίος καταγράφει το υψηλότερο precision και το χαμηλότερο MSE από όλα τα μοντέλα. Το εύρημα αυτό υποδηλώνει ότι η σύνθεση εξειδικευμένης συναισθηματικής ανάλυσης με γενικότερη γλωσσική κατανόηση μπορεί να ενισχύσει την προβλεπτική ικανότητα σε μετοχές με πιο πολύπλοκη και τεχνολογικά πολυσχιδή θεματολογία, όπως η NVIDIA.

Συγκριτικά, το μοντέλο GPT μόνο του παρουσιάζει συστηματικά χαμηλότερη απόδοση και στις τρεις μετοχές, τόσο σε ακρίβεια όσο και σε σφάλμα. Αυτό επιβεβαιώνει ότι η γενική γλωσσική πληροφορία χωρίς οικονομική εξειδίκευση ενδέχεται να μην επαρκεί για αξιόπιστες προβλέψεις κατεύθυνσης τιμών. Αντιθέτως, ο συνδυασμός GPT με FinBERT φαίνεται να αποδίδει καλύτερα σε περιπτώσεις πιο σύνθετης πληροφορίας, ενώ το FinBERT μόνο του παραμένει το πιο αξιόπιστο και σταθερό μοντέλο συνολικά.

Συνοψίζοντας, δεν υπάρχει «καθολικά καλύτερο» μοντέλο για όλες τις περιπτώσεις. Η απόδοση κάθε μοντέλου εξαρτάται από τα χαρακτηριστικά της μετοχής:

- FinBERT υπερέχει σε σταθερές μετοχές με καθαρό ειδησεογραφικό αποτύπωμα (π.χ. AAPL).
- FinBERT+GPT λειτουργεί καλύτερα σε τεχνολογικές μετοχές με πιο σύνθετο περιεχόμενο (π.χ. NVDA, TSLA).
- GPT επιτυγχάνει υψηλή precision αλλά υστερεί σε σταθερότητα, γεγονός που το καθιστά χρήσιμο σε επιλεγμένες μεταβλητές αγορές.

Η προσαρμοστική αξιοποίηση των μοντέλων, με βάση τα χαρακτηριστικά της κάθε μετοχής, ενδέχεται να προσφέρει σημαντικά πλεονεκτήματα σε εφαρμογές πρόβλεψης κατεύθυνσης χρηματοοικονομικών τίτλων.

7. Συμπεράσματα και Μελλοντική Εργασία

7.1. Σημαντικότητα Αποτελεσμάτων και Ερμηνεία Αποκλίσεων

Η συγκριτική αξιολόγηση των μοντέλων FinBERT, FinBERT+GPT και GPT σε τρεις μετοχές υψηλής κεφαλαιοποίησης (AAPL, TSLA και NVDA), με βάση τις μετρικές accuracy, precision και mean squared error (MSE), καταδεικνύει ότι η απόδοση των μοντέλων είναι άμεσα εξαρτώμενη από τα ειδικά χαρακτηριστικά της κάθε μετοχής και τον τρόπο με τον οποίο οι γλωσσικές πληροφορίες αποτυπώνονται στις μεταβολές της τιμής.

Στην περίπτωση της Apple (AAPL), το μοντέλο FinBERT επιδεικνύει την πιο συνεπή και βελτιστοποιημένη απόδοση, υπερβαίνοντας το baseline και τα υπόλοιπα μοντέλα τόσο σε precision όσο και σε accuracy, ενώ καταγράφει και χαμηλότερο MSE. Η συγκεκριμένη υπεροχή αποδίδεται στην καλή ευθυγράμμιση των χρηματοοικονομικών ειδήσεων με τη σταθερή και προβλέψιμη φύση της μετοχής. Η απόδοση του GPT είναι εμφανώς υποδεέστερη, γεγονός που επιβεβαιώνει την αναγκαιότητα χρηματοοικονομικής εξειδίκευσης για την κατανόηση των επιδράσεων της πληροφορίας στην εν λόγω περίπτωση. Αντιθέτως, ο συνδυασμός FinBERT+GPT φαίνεται να υπολείπεται, πιθανώς λόγω εισαγωγής γλωσσικού “θορύβου” από μη-σχετικές γενικές πληροφορίες.

Η μετοχή της Tesla (TSLA) παρουσιάζει μια διαφορετική εικόνα. Το FinBERT+GPT καταγράφει την υψηλότερη precision και μεγαλύτερη βελτίωση σε accuracy σε σχέση με το baseline, συνοδευόμενη από μείωση του MSE. Αυτό υποδηλώνει ότι ο συνδυασμός εξειδικευμένης και γενικής γλωσσικής κατανόησης είναι αποτελεσματικότερος για την αποτύπωση της επίδρασης των ειδήσεων σε μετοχές με έντονη θεματική ποικιλομορφία και υψηλή μεταβλητότητα. Η σχετική επιτυχία και του FinBERT υποδηλώνει ότι η συναισθηματική πληροφορία παραμένει σημαντική, αλλά ενισχύεται όταν εμπλουτίζεται από ευρύτερες γλωσσικές αποχρώσεις, όπως αυτές που εντοπίζει το GPT.

Όσον αφορά τη NVIDIA (NVDA), το μοντέλο FinBERT+GPT αποδίδει και πάλι καλύτερα, σημειώνοντας τις υψηλότερες επιδόσεις τόσο σε precision όσο και σε accuracy, καθώς και τη μικρότερη τιμή MSE. Το αποτέλεσμα αυτό καταδεικνύει ότι σε μετοχές που επηρεάζονται από σύνθετο τεχνολογικό και καινοτομικό περιεχόμενο, η πολυτροπική προσέγγιση επιτυγχάνει την καλύτερη ερμηνεία της πληροφορίας. Το FinBERT αποδίδει επαρκώς αλλά όχι βέλτιστα, ενώ το GPT παρουσιάζει πτώση και στις τρεις μετρικές, υποδηλώνοντας περιορισμένη ικανότητα πρόβλεψης όταν απουσιάζει η οικονομική εξειδίκευση.

Συνοψίζοντας, η ανάλυση των αποτελεσμάτων αναδεικνύει ότι δεν υπάρχει ένα “καθολικά καλύτερο” μοντέλο. Το FinBERT υπερέχει σε σταθερές και χρηματοοικονομικά προβλέψιμες μετοχές, όπως η AAPL· το FinBERT+GPT αποδίδει καλύτερα σε περιβάλλοντα με μεγαλύτερη πολυπλοκότητα πληροφορίας, όπως η TSLA και η NVDA· ενώ το GPT, αν και ωφέλιμο ως μέρος υβριδικού μοντέλου, αποτυγχάνει να αποδώσει επαρκώς όταν χρησιμοποιείται μεμονωμένα. Τα ευρήματα αυτά επιβεβαιώνουν τη σημασία της επιλογής ή προσαρμογής μοντέλου ανάλογα με τη φύση της πληροφορίας και της μετοχής, προσφέροντας πρακτική κατεύθυνση για μελλοντικές εφαρμογές στη χρηματοοικονομική πρόβλεψη βασισμένη σε φυσική γλώσσα.

7.2. Μελλοντική Εργασία και Περιορισμοί

Η παρούσα διπλωματική εργασία ανέδειξε την αξία της ενσωμάτωσης ποιοτικών δεικτών—όπως οι συναισθηματικοί δείκτες από τα SEC filings και οι καινοτομίες που αναφέρονται σε εταιρικές αναφορές—σε συστήματα μηχανικής μάθησης για την πρόβλεψη των μελλοντικών αποδόσεων μετοχών υψηλής κεφαλαιοποίησης (AAPL, NVDA, TSLA). Συγκεκριμένα:

1. **Αύξηση ακρίβειας:** Η προσθήκη συναισθηματικών δεικτών από αναφορές 10-K, 10-Q και 8-K βελτίωσε σημαντικά τις μετρικές accuracy και precision σε σχέση με τα baseline μοντέλα, αποδεικνύοντας ότι η λεκτική φόρμα και το συναισθηματικό περιεχόμενο των επίσημων αναφορών μεταφράζονται σε χρήσιμες προγνωστικές πληροφορίες.
2. **Συνδυασμός δεδομένων:** Η ενσωμάτωση των δεικτών καινοτομίας—όπως εξάγονται από τεχνολογικές και R&D αναφορές—βελτίωσε περαιτέρω την απόδοση των μοντέλων, καθώς καθιστά δυνατή τη λήψη υπόψη στρατηγικών εξελίξεων των εταιρειών.
3. **Εξειδίκευση vs. γενίκευση:** Τα μοντέλα με θεματική εξειδίκευση σε χρηματοοικονομικές αναφορές (FinBERT) υπερέχουν συστηματικά έναντι των γενικών LLMs (GPT) όταν πρόκειται για σταθερές μετοχές με προβλέψιμο ειδησεογραφικό προφίλ. Αντίθετα, η προσθήκη γενικότερης γλωσσικής πληροφορίας (FinBERT+GPT) αποδείχθηκε χρήσιμη μόνο σε περιπτώσεις υψηλής πληροφορικής πολυπλοκότητας (π.χ. TSLA, NVDA).
4. **Γενικά συμπεράσματα:** Δεν υπάρχει «ένα» βέλτιστο μοντέλο για όλες τις περιπτώσεις· η επιλογή πρέπει να βασίζεται στη φύση του εταιρικού περιεχομένου και τη δυναμική της μετοχής. Η ανάλυση αυτή θέτει τις βάσεις για ένα ευέλικτο πλαίσιο που προσαρμόζει τη μοντελοποίηση στα χαρακτηριστικά κάθε επενδυτικού αντικειμένου.

Παρακάτω παρατίθενται οι βασικές κατευθύνσεις για μελλοντική έρευνα και ανάπτυξη της μεθοδολογίας:

1. Επέκταση σε νέες πηγές δεδομένων

- ο **Social media (π.χ. Twitter):** Η ενσωμάτωση ροών από tweets μπορεί να αποκαλύψει πρώιμες και ανεπίσημες ενδείξεις αγοράς, ιδίως για μετοχές με έντονη κοινωνική συζήτηση.
- ο **Εναλλακτικά δεδομένα:** Δορυφορικές εικόνες εργοστασίων, αναλύσεις web search trends ή δεδομένα αγορών πωλήσεων προϊόντων για περαιτέρω εμπλουτισμό του μοντέλου.

2. Ποσοτικοποίηση έντασης μεταβολής

- ο Μετάβαση από δυαδική ταξινόμηση («άνοδος/πτώση») σε προβλεπτικές παλινδρομήσεις ποσοστών μεταβολής (π.χ. πρόβλεψη ποσοστιαίας αύξησης της τιμής) ή σε πολυταξινομήσεις με κλιμακωτές ζώνες απόδοσης.

3. Υβριδικά και εξηγησιακά μοντέλα

- ο Χρήση μεθόδων explainable AI (π.χ. SHAP, LIME) για την αποσαφήνιση του ρόλου κάθε είδους πληροφορίας στη λήψη της πρόβλεψης.

- ο Εξερεύνηση multi-modal μοντέλων που συνδυάζουν γλωσσικές αναλύσεις με τεχνικούς δείκτες τιμών και όγκων συναλλαγών σε κοινή αρχιτεκτονική.

4. Δοκιμή σε ευρύτερο δείγμα

- ο Επέκταση της μελέτης σε άλλους κλάδους (π.χ. ενέργεια, υγεία) και σε εταιρείες μικρότερης κεφαλαιοποίησης για τον έλεγχο της γενικευσιμότητας των ευρημάτων.

5. Χρονική δυναμική και causal inference

- ο Εφαρμογή χρονικών μοντέλων (π.χ. temporal transformers, causal impact analysis) για την κατανόηση του χρονικού ορίζοντα επιρροής των ειδήσεων και του sentiment στην τιμή.

Με τις παραπάνω κατευθύνσεις, η έρευνα μπορεί να εξελιχθεί σε ένα ακόμα πιο ολοκληρωμένο και πρακτικά εφαρμόσιμο πλαίσιο πρόβλεψης χρηματοοικονομικών τιμών, αξιοποιώντας την πρόοδο στη μηχανική μάθηση και τα πολυτροπικά δεδομένα.

7.3. Πιθανές Χρήσεις της Μελέτης

Η παρούσα μελέτη προσφέρει σημαντικές δυνατότητες εφαρμογής τόσο σε ακαδημαϊκό όσο και σε χρηματοοικονομικό πλαίσιο. Η αξιοποίηση GRU αρχιτεκτονικών σε συνδυασμό με συναισθηματικούς δείκτες επιτρέπει την εκτίμηση της βραχυπρόθεσμης δυναμικής των τιμών, παρέχοντας πρακτικά συμπεράσματα για την ανάπτυξη στρατηγικών trading που ανταποκρίνονται άμεσα στις διακυμάνσεις της αγοράς. Παράλληλα, η έγκαιρη διάγνωση συναισθηματικών τάσεων καθιστά επιτακτική την ανάγκη ενσωμάτωσης αυστηρών πρωτοκόλλων διαχείρισης κινδύνου (risk management), ώστε να διασφαλίζεται ότι οι αποφάσεις συναλλαγών δεν οδηγούν σε απρόβλεπτη έκθεση υπέρμετρου κινδύνου.

Σε επίπεδο αγοράς, οι αλγόριθμοι που εξετάστηκαν μπορούν να ενσωματωθούν σε συστήματα αλγοριθμικού trading, ιδίως κατά περιόδους αυξημένης μεταβλητότητας, όπου η προσεκτική παρακολούθηση του κινδύνου και η εφαρμογή stop-loss ορίων είναι καθοριστικής σημασίας. Τα εργαλεία ανάλυσης και πρόβλεψης που αναπτύχθηκαν μπορούν επίσης να υποβοηθήσουν θεσμικούς επενδυτές και risk managers, παρέχοντας σε πραγματικό χρόνο εκτιμήσεις για τον πιθανό αντίκτυπο ειδησεογραφικών γεγονότων ή εταιρικών ανακοινώσεων, και επιτρέποντας τον έγκαιρο σχεδιασμό στρατηγικών μετριασμού ζημιών.

Παράλληλα, η μεθοδολογία είναι επεκτάσιμη και σε επιχειρηματικές εφαρμογές πέρα από τις χρηματιστηριακές συναλλαγές: επιχειρήσεις και ιδιώτες επενδυτές θα μπορούσαν να αξιοποιήσουν ανάλογες προσεγγίσεις για να εκτιμήσουν τον αντίκτυπο δικών τους επικοινωνιακών ενεργειών ή κινήσεων ανταγωνιστών, συνδυάζοντας δυναμική παρακολούθηση χρονοσειρών και ανάλυση συναισθηματικού περιεχομένου κειμένου. Με τον τρόπο αυτό, η σύζευξη ποσοτικών και ποιοτικών δεδομένων δημιουργεί το έδαφος για πιο ολοκληρωμένες και αποδοτικές λύσεις πρόβλεψης σε ποικίλα χρηματοοικονομικά και επιχειρηματικά σενάρια.

Τέλος, η εργασία θέτει τις βάσεις για περαιτέρω έρευνα στη συνδυαστική ανάλυση χρονοσειρών και δεδομένων κειμένου, ανοίγοντας τον δρόμο για πιο ολοκληρωμένες και αποδοτικές προσεγγίσεις πρόβλεψης στις χρηματοοικονομικές αγορές.

8. Bibliography

- [1] "Wikipedia," [Online]. Available: https://en.wikipedia.org/wiki/Dutch_East_India_Company.
- [2] I. H. -. M. M. -. H. Schmiedel, "Technology, automation, and productivity of stock exchanges: International evidence," 2002.
- [3] V. V. B. -. M. Sriram, "Case study on retail investors: Technological innovation and stock market dynamics," 2024.
- [4] R. a. A. David Ajiga, "MACHINE LEARNING FOR STOCK MARKET FORECASTING: A REVIEW OF MODELS AND ACCURACY," 2024.
- [5] K. P. Ekster G, "Alternative data in investment management: usage, challenges and valuation.," 2021.
- [6] "Wikipedia," [Online]. Available: <https://www.investopedia.com/terms/g/garch.asp>.
- [7] A. B. K. J. K. O. R. D. S. Rob J Hyndman, Forecasting with Exponential Smoothing: the State Space Approach, 2008.
- [8] J. H. Friedman, "Regularized Discriminant Analysis," *Journal of the American Statistical Association*, 1989.
- [9] C. K. Thomas Fischer, "Deep learning with long short-term memory networks for financial market predictions," 2018.
- [10] H. M. ., X. Z. Johan Bollen, "Twitter mood predicts the stock market," 2011.
- [11] G. G. Kemal Kirtac, "Sentiment trading with large language models," 2024.
- [12] Z. Yu, "Cross-Section of Returns, Predictors Credibility, and Method Issues.," *Journal of Risk and Financial Management*, p. 16: 34, 2023.
- [13] B. Mandelbrot, "The variation of certain speculative prices.," *The Journal of Business* 36, p. 394–419, 1963.
- [14] S. M. Sriyank Levi, "FUNDAMENTAL AND TECHNICAL ANALYSIS LEADS TO A SYSTEMATIC INVESTMENT DECISION IN STOCK MARKET EQUITIES".
- [15] M. Κεφαλά, "Μοντέλο πρόβλεψης μετοχικών τιμών βασισμένο σε γενετικούς αλγορίθμους και νευρωνικά δίκτυα," 2022.
- [16] J. M. H. & Z. X. Bollen, "Twitter mood predicts the stock market," *Journal of Computational Science*, 2011.
- [17] M. S. B. M. D. J. Heeyoung Lee, "On the Importance of Text Analysis for Stock Price Prediction.," 2014.
- [18] S. N. Magdy Saleh, "Neural based event-driven stock rally prediction using SEC filings and Twitter data".
- [19] D. Satria, "PREDICTING BANKING STOCK PRICES USING RNN, LSTM, AND GRU APPROACH," 2023.
- [20] U. S. a. E. Commission., "What We Do," [Online]. Available: <https://www.sec.gov/about/what-we-do>.
- [21] U. S. a. E. Commission., "EDGAR - Electronic Data Gathering, Analysis, and Retrieval," [Online]. Available: <https://www.sec.gov/edgar.shtml>.

- [22] Wikipedia., " U.S. Securities and Exchange Commission.," [Online]. Available: https://en.wikipedia.org/wiki/U.S._Securities_and_Exchange_Commission.
- [23] Wikipedia, "Earnings," [Online]. Available: <https://en.wikipedia.org/wiki/Earnings>.
- [24] X. X. ., J. G. ., Y. X. Yanqing Yang, "Machine Learning for Economic Forecasting: An Application to China's GDP Growth," 2024.
- [25] Wikipedia, "Θεωρία του χάους," [Online]. Available: https://el.wikipedia.org/wiki/%CE%98%CE%B5%CF%89%CF%81%CE%AF%CE%B1_%CF%84%CE%BF%CF%85_%CF%87%CE%AC%CE%BF%CF%85%CF%82.
- [26] S. W. S. C. & H. R. J. Makridakis, "Dancing with Chance: Forecasting in a Chaotic World.," 1998.
- [27] Investopedia.(n.d.)., "Forecasting.," Retrieved January 3, 2025. [Online]. Available: <https://www.investopedia.com/terms/f/forecasting.asp>.
- [28] P. E. & G. D. Tetlock, "Superforecasting: The Art and Science of Prediction.," 2015.
- [29] G. E. P. & J. G. M. Box, "Time Series Analysis: Forecasting and Control.," 1976.
- [30] R. J. & A. G. Hyndman, "Forecasting: Principles and Practice (2nd ed.)," 2018.
- [31] S. & S. J. Hochreiter, "Long Short-Term Memory.," 1997.
- [32] J. M. H. & Z. X. Bollen, "Twitter mood predicts the stock market.," pp. 1-8., 2011.
- [33] Y. Zhang, "Stock Price Prediction Method Based on XGboost Algorithm".
- [34] E. Alpaydın, "Introduction to Machine Learning," 2014.
- [35] E. E. M. A. A. S. P. Houssein, "Deep and machine learning techniques for medical imaging-based breast cancer: A comprehensive review".
- [36] G. R. A. C. S. Rebala, " An Introduction to Machine Learning," 2019.
- [37] E. κ. Pontil, "Convex multi-task feature learning," 2007.
- [38] X. Z. S. R. J. S. Kaiming He, "Deep Residual Learning for Image Recognition".
- [39] J. D. H. S. J. K. S. S. S. M. Z. H. A. K. A. K. M. B. A. C. B. & L. F.-F. Olga Russakovsky, "ImageNet Large Scale Visual Recognition Challenge".
- [40] M.-W. C. K. L. K. T. Jacob Devlin, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding".
- [41] C. C. Aggarwal, Neural Networks and Deep Learning.
- [42] A. M.-B. a. V. Schmid, Pattern Recognition and Signal Analysis in Medical Imaging.
- [43] G. C. S. J. Mohaiminul Islam, "Multilayer Perceptrons for Medical Diagnosis".
- [44] J. J. Marquina-Araujo, "Application of ANN-MLP in Geological Classification".
- [45] I. Barends, ""Animal spirits" in John Maynard Keynes's general theory of employment, interest and money: Some short and sceptical remarks," 2011.
- [46] I. Roxana, "The Influence Of Stock Market Investors' Behavior On Business Cycles," 2015.
- [47] M. Baddeley, "Herding, social influence and economic decision-making: socio-psychological and neuroscientific analyses," 2010.
- [48] J. W. Malcolm Baker, "Investor Sentiment in the Stock Market," 2007.
- [49] T. & M. B. Loughran, "When is a Liability Not a Liability? Textual Analysis,

- Dictionaries, and 10-Ks.," *Journal of Finance*, 2011.
- [50] A. S. S. D. Z. Xinying Song, "Fast WordPiece Tokenization".
 - [51] E. K. Jacob Murel, " Why stemming and lemmatization?," [Online]. Available: <https://www.ibm.com/think/topics/stemming-lemmatization#:~:text=For%20many%20text%20mining%20tasks,3>.
 - [52] T. C. K. C. G. & D. J. Mikolov, "Efficient estimation of word representations in vector space.," 2013.
 - [53] T. S. I. C. K. C. G. S. & D. J. Mikolov, "Distributed representations of words and phrases and their compositionality.," 2013.
 - [54] J. S. R. & M. C. Pennington, "GloVe: Global vectors for word representation.," 2014.
 - [55] P. G. E. J. A. & M. T. Bojanowski, "Enriching word vectors with subword information.," 2017.
 - [56] J. C. M. W. L. K. & T. K. Devlin, "BERT: Pre-training of deep bidirectional transformers for language understanding.," 2019.
 - [57] D. & M. J. H. Jurafsky, "Speech and Language Processing," 2021.
 - [58] A. K. O. & R. A. Rogers, " A primer in BERTology: What we know about how BERT works.," 2021.
 - [59] M.-W. C. K. L. a. K. T. Jacob Devlin, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.," 2019.
 - [60] P. T. Edward Sharkey, "Comparison of BERT vs GPT."
 - [61] D. Araci, "FinBERT: Financial Sentiment Analysis with Pre-trained Language Models," 2019.
 - [62] C. S. ., Y. S. ., H. ., Y. L. Tong Zhan, "Optimization Techniques for Sentiment Analysis Based on LLM (GPT-3)".
 - [63] K. N. S. S. Alec Radford, "Improving Language Understanding by Generative Pre-Training".
 - [64] S. E. a. S. G. Itel, "Time Series Forecasting Using LSTM Networks: A Symbolic Approach".
 - [65] W. X. G. & Z. J. Zhang, "Challenges and Advances in Large-Scale LSTM Training for Time Series Forecasting," 2022.
 - [66] M. Wang, "Advanced Stock Market Forecasting: A Comparative Analysis of ARIMA-GARCH, LSTM, and Integrated Wavelet-LSTM Models," 2024.
 - [67] T. A. T. M. H. Baydaa Hadi Saoudi, "MACHINE LEARNING TECHNIQUE FOR DETECTING DIABETES DISEASE," 2024.