Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Ηλεκτρικων Βιομηχανικων Διαταξεων και Συστηματων Αποφασεων
Μοναδα Προβλεψεων και Στρατηγικης

Πρόβλεψη χρονοσειρών με χρήση περιγραφικών συμμεταβλητών και τεχνικών επεξεργασίας φυσικής γλώσσας

Διπλωματικη Εργασια

του

ΑΝΑΣΤΑΣΙΟΥ ΣΤΕΡΓΙΟΠΟΥΛΟΥ

Επιβλέπων: Βασίλειος Ασημακόπουλος
Ομότιμος Καθηγητής Ε.Μ.Π.

Υπεύθυνος: Ευάγγελος Σπηλιώτης
Διδάκτωρ Ε.Μ.Π.

Αθήνα, Ιούλιος 2024

Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Ηλεκτρικων Βιομηχανικων Διαταξεων και Συστηματων Αποφασεων
Μοναδα Προβλεψεων και Στρατηγικης

Πρόβλεψη χρονοσειρών με χρήση περιγραφικών συμμεταβλητών και τεχνικών επεξεργασίας φυσικής γλώσσας

Διπλωματικη Εργασια

του

ΑΝΑΣΤΑΣΙΟΥ ΣΤΕΡΓΙΟΠΟΥΛΟΥ

Επιβλέπων: Βασίλειος Ασημακόπουλος
Ομότιμος Καθηγητής Ε.Μ.Π.

Υπεύθυνος: Ευάγγελος Σπηλιώτης
Διδάκτωρ Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 16η Ιουλίου 2024.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Βασίλειος Ασημακόπουλος	Δημήτριος Ασκούνης	Χρυσόστομος Δούκας
Ομότιμος Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2024

Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Ηλεκτρικων Βιομηχανικων Διαταξεων και Συστηματων Αποφασεων
Μοναδα Προβλεψεων και Στρατηγικης

ΔΗΛΩΣΗ ΜΗ ΛΟΓΟΚΛΟΠΗΣ ΚΑΙ ΑΝΑΛΗΨΗΣ ΠΡΟΣΩΠΙΚΗΣ ΕΥΘΥΝΗΣ

Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, δηλώνω ενυπογράφως ότι είμαι αποκλειστικός συγγραφέας της παρούσας Πτυχιακής Εργασίας, για την ολοκλήρωση της οποίας κάθε βοήθεια είναι πλήρως αναγνωρισμένη και αναφέρεται λεπτομερώς στην εργασία αυτή. Έχω αναφέρει πλήρως και με σαφείς αναφορές, όλες τις πηγές χρήσης δεδομένων, απόψεων, θέσεων και προτάσεων, ιδεών και λεκτικών αναφορών, είτε κατά κυριολεξία είτε βάσει επιστημονικής παράφρασης. Αναλαμβάνω την προσωπική και ατομική ευθύνη ότι σε περίπτωση αποτυχίας στην υλοποίηση των ανωτέρω δηλωθέντων στοιχείων, είμαι υπόλογος έναντι λογοκλοπής, γεγονός που σημαίνει αποτυχία στην Πτυχιακή μου Εργασία και κατά συνέπεια αποτυχία απόκτησης του Τίτλου Σπουδών, πέραν των λοιπών συνεπειών του νόμου περί πνευματικών δικαιωμάτων. Δηλώνω, συνεπώς, ότι αυτή η Πτυχιακή Εργασία προετοιμάστηκε και ολοκληρώθηκε από εμένα προσωπικά και αποκλειστικά και ότι, αναλαμβάνω πλήρως όλες τις συνέπειες του νόμου στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δεν μου ανήκει διότι είναι προϊόν λογοκλοπής άλλης πνευματικής ιδιοκτησίας.

(Υπογραφή)

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

16η Ιουλίου 2024

Περίληψη

Η τρίτη βιομηχανική επανάσταση σηματοδότησε τη μετάβαση της ανθρωπότητας στην εποχή της πληροφορίας που χαρακτηρίζεται από την αφθονία των δεδομένων. Τα τελευταία χρόνια, η πρόοδος των υπολογιστικών συστημάτων οδηγεί στην ευδοκίμηση της τεχνητής νοημοσύνης και μία νέα εποχή η οποία χαρακτηρίζεται ως τέταρτη βιομηχανική επανάσταση. Στον τομέα της πρόβλεψης χρονοσειρών, η μηχανική μάθηση κέρδισε έδαφος έναντι των παραδοσιακών κλασικών στατιστικών μεθόδων. Στους διαγωνισμούς πρόβλεψης M4 και M5, ηγούνται μοντέλα που χρησιμοποιούν τεχνικές μηχανικής μάθησης. Στον πρώτο, πρόκειται για συνδυασμούς μοντέλων μηχανικής μάθησης με κλασικές στατιστικές μεθόδους, ενώ στο δεύτερο, οι προσεγγίσεις με κορυφαίες επιδόσεις αποτελούνταν αποκλειστικά από τεχνικές μηχανικής μάθησης, ξεχωρίζοντας κάποια μοντέλα βασισμένα στα δέντρα αποφάσεων. Στη συνέχεια, ο διαγωνισμός πρόβλεψης M6 που ολοκληρώθηκε το 2023, ανέδειξε επίσης την αποτελεσματικότητα των γενικευμένων νευρωνικών δικτύων στην πρόβλεψη χρονοσειρών. Ένα από τα βέλτιστα μοντέλα του διαγωνισμού, εκπαιδεύτηκε σε περισσότερες από τις ζητούμενες χρονοσειρές, χρησιμοποιώντας κάποιες επιπλέον με παρόμοιες συμπεριφορές, που συνεισέφεραν στη βελτίωση της πρόβλεψης. Παρατηρώντας την τάση αυτή, και δεδομένου του αυξανόμενου όγκου με τον οποίο εκπαιδεύονται τα μοντέλα πρόβλεψης, δημιουργείται μία ανάγκη μετάβασης σε γενικευμένα θεμελιώδη μοντέλα (foundation models), τα οποία προεκπαιδεύονται με μεγάλο όγκο δεδομένων και στη συνέχεια μπορούν να μετεκπαιδευτούν με σκοπό τη γρήγορη εξειδίκευσή τους σε κάποιο ειδικό τομέα.

Στην παρούσα διπλωματική εργασία εξετάζεται η συνεισφορά των συμμεταβλητών όπως οι λεκτικές περιγραφές και τα ποιοτικά χαρακτηριστικά των χρονοσειρών, στην αποτελεσματικότητα ενός γενικευμένου μοντέλου πρόβλεψης. Αναπτύσσεται μία πειραματική διαδικασία η οποία περιλαμβάνει την ερμηνεία των λεκτικών περιγραφών που συνοδεύουν ένα μεγάλο όγκο χρονοσειρών με τεχνικές επεξεργασίας φυσικής γλώσσας. Συγκεκριμένα, υλοποιούνται μοντέλα παραγωγής διανυσμάτων προτάσεων με χρήση προεκπαιδευμένων θεμελιωδών γλωσσικών μοντέλων όπως τα μοντέλα BERT και GPT-2, καθώς και προγενέστερων μοντέλων ερμηνείας της φυσικής γλώσσας όπως τα Word2Vec και Doc2Vec. Στη συνέχεια, τα διανύσματα αυτά χρησιμοποιούνται για την ταξινόμηση των χρονοσειρών με τους αλγόριθμους k-NN και Random Forest. Επιπλέον εξετάζονται τεχνικές συσταδοποίησης όπως οι αλγόριθμοι k-Means, Agglomerative Hierarchical και DBSCAN, για την αντιστοίχιση κάθε χρονοσειράς σε με μία συστάδα, με βάση την περιγραφή της. Για την πρόβλεψη, σχεδιάζονται κάποια βαθιά νευρωνικά δίκτυα παράλληλης εισόδου. Τα αποτελέσματα δείχνουν πως παράγονται αντιπροσωπευτικά διανύσματα και συστάδες προτάσεων για τις περιγραφές. Επιπλέον, χρησιμοποιώντας πολυεπίπεδα νευρωνικά δίκτυα παράλληλης εισόδου των χρονοσειρών και των περιγραφών, το γενικευμένο μοντέλο παράγει πιο ακριβείς προβλέψεις από το μοντέλο αναφοράς.

Λέξεις Κλειδιά

Χρονοσειρές, Πρόβλεψη, Επεξεργασία Φυσικής Γλώσσας, Multimodal Learning, Foundation Models, LLMs, Ταξινόμηση, Συσταδοποίηση, Βαθιά Νευρωνικά Δίκτυα

Abstract

The third industrial revolution marked humanity's transition to the information age, characterized by the abundance of data. In recent years, advances in computing systems led to the flourishing of artificial intelligence and the emergence of a new era, the fourth industrial revolution. In the field of time series forecasting, machine learning has gained ground over traditional classical statistical methods. In the M4 and M5 forecasting competitions, models that utilize machine learning techniques have led the way. In the first competition, these models involved combinations of machine learning and classical statistical methods, whereas in the second, the top-performing approaches consisted solely of machine learning techniques, with some decision tree-based models standing out. Subsequently, the M6 forecasting competition, which concluded in 2023, also highlighted the effectiveness of generalized neural networks in time series forecasting. One of the optimal models from the competition was trained on more than the required time series, using some additional ones with similar behaviors, which contributed to improving the forecasts. Observing this trend and given the increasing volume with which forecasting models are being trained, there arises a need to transition to generalized foundation models, which are pre-trained with a large amount of data and can then be fine-tuned to quickly adapt to a specific domain.

This thesis examines the contribution of covariates such as verbal descriptions and qualitative characteristics of time series to the performance of a generalized forecasting model. An experimental process is developed, which includes interpreting the verbal descriptions accompanying a large volume of time series using natural language processing techniques. Models are implemented to generate sentence vectors using pre-trained foundational language models such as BERT and GPT-2, as well as earlier models like Word2Vec and Doc2Vec. These vectors are then used for classifying time series with algorithms like k-NN and Random Forest. Additionally, clustering techniques such as the k-Means, Agglomerative Hierarchical, and DBSCAN algorithms using the sentence vectors produced, are explored to group time series into clusters. For time series forecasting, some parallel input deep neural networks are designed to evaluate the impact of covariates. The results indicate that representative vectors and clusters of sentences for the descriptions are produced. Moreover, by using multi-level parallel input neural networks for both the time series and the descriptions, the generalized model generates more accurate forecasts than the baseline model, as evidenced by a reduction in error.

Keywords

Time Series, Forecasting, Natural Language Processing, Multimodal Learning, Foundation Models, LLMs, Classification, Clusterning, Deep Neural Networks

Ευχαριστίες

Θα ήθελα αρχικά να ευχαριστήσω τον Καθηγητή κ. Βασίλειο Ασημακόπουλο που μου έδωσε την ευκαιρία να εκπονήσω αυτήν τη διπλωματική εργασία στη Μονάδα Προβλέψεων και Στρατηγικής και να ασχοληθώ με την επιστήμη των δεδομένων και την πρόβλεψη των χρονοσειρών. Επίσης, θα ήθελα να ευχαριστήσω τον Καθηγητή κ. Χρυσόστομο Δούκα και τον Καθηγητή κ. Δημήτριο Ασκούνη για τη συμμετοχή τους στην εξεταστική επιτροπή της εργασίας.

Ευχαριστώ ιδιαίτερα τον Διδάκτορα κ. Ευάγγελο Σπηλιώτη, ο οποίος ήταν και η αιτία της ενασχόλησης μου με τον τομέα των προβλέψεων. Μέσω του μαθήματος, μου καλλιέργησε την περιέργεια να πειραματιστώ με τις προβλέψεις και το διαγωνισμό προβλέψεων M6 μέχρι εν τέλει την επίβλεψη αυτής της διπλωματικής, όπου η καθοδήγησή του ήταν καταλυτική.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένεια μου, τη Μυρσίνη και τους φίλους μου.

Αθήνα, Ιούλιος 2024

Αναστάσιος Στεργιόπουλος

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

1.1 Αντικείμενο Εργασίας

Η τρίτη βιομηχανική επανάσταση, αποτέλεσε τη μετάβαση της ανθρωπότητας στην εποχή της πληροφορίας (Castells, 1996). Τα δεδομένα πέρασαν από τη αναλογική μορφή στην ψηφιακή. Η τεχνολογία πλέον επιτρέπει τη συλλογή πληροφοριών από διάφορες πηγές, δημιουργώντας έτσι ένα εκτεταμένο απόθεμα που παραμένει όμως σε μεγάλο βαθμό αναξιοποίητο.

Στις αρχές του $21^{ου}$ αιώνα, με την πρόοδο των υπολογιστικών συστημάτων, αναπτύσσεται περαιτέρω η ιδέα της «τεχνητής» νοημοσύνης μου είχε πρωτοεμφανιστεί στα μέσα του προηγούμενου αιώνα. Πρόκειται για μία νέα εποχή η οποία χαρακτηρίζεται ως τέταρτη βιομηχανική επανάσταση (Schwab, 2017). Σε αυτό το πλαίσιο, η μηχανική μάθηση παίζει καθοριστικό ρόλο, επιτρέποντας τον εντοπισμό περίπλοκων σχέσεων στα δεδομένα με τα οποία εκπαιδεύεται, με σκοπό την πρόβλεψη ή τη λήψη κάποιας απόφασης. Ωστόσο, οι τεχνικές της μηχανικής μάθησης, απαιτούν τόσο η τεχνογνωσία για τον σχεδιασμό και την εκπαίδευσή τους, όσο και σημαντικούς υπολογιστικούς πόρους για να εκπαιδευτούν αποτελώντας μία σημαντική πρόκληση.

Στον τομέα της ανάλυσης και των προβλέψεων των χρονοσειρών, παραδοσιακά οι προβλέψεις στηρίζονται σε κλασικές στατιστικές μεθόδους, ενώ η μηχανική μάθηση εφαρμόζεται σταδιακά. Για πρώτη φορά στο διαγωνισμό προβλέψεων M4 (Makridakis, Spiliotis, & Assimakopoulos, 2020), ένας συνδυασμός μοντέλων μηχανικής μάθησης και στατιστικής επιτυγχάνει τη βέλτιστη επίδοση. Στο διαγωνισμό M5 (Makridakis, Spiliotis, & Assimakopoulos, 2022), τα βέλτιστα αποτελέσματα επιτυγχάνουν μοντέλα μηχανικής μάθησης. Ειδικότερα, πρόκειται για μοντέλα εκμάθησης συνόλου (ensemble learning), τα οποία συνδυάζουν πολλά ανεξάρτητα μοντέλα μηχανικής μάθησης για την παραγωγή της τελικής πρόβλεψης. Από το Φεβρουάριο του 2022 και για τους επόμενους 12 μήνες, διεξήχθη ο διαγωνισμός M6 (Makridakis et al., 2023). Οι διαγωνιζόμενοι παρήγαγαν εβδομαδιαίες προβλέψεις σχετικά με την τιμή 100 περιουσιακών στοιχείων (50 μετοχές εισηγμένες στο χρηματιστήριο που ανήκουν στο δείκτη S&P 500 και 50 διαπραγματεύσιμα αμοιβαία κεφάλαια). Στη διάθεση των συμμετεχόντων ήταν οποιαδήποτε υποστηρικτική πληροφορία μπορούσε να ενισχύσει τις προβλέψεις τους, όπως ειδήσεις, οικονομικές καταστάσεις ή και τιμές άλλων παρόμοιων μετοχών. Κορυφαία επίδοση παρουσίασε ένα νευρωνικό δίκτυο το οποίο εκπαιδεύεται με περισσότερα από τα αρχικά 100 περιουσιακά στοιχεία, με σκοπό να «μάθει» παρεμφερείς

συμπεριφορές μετοχών και αμοιβαίων κεφαλαίων. Ύστερα μέσω μίας παράλληλης εισόδου, παράγει προβλέψεις συνδυάζοντας τις πληροφορίες της εκάστοτε χρονοσειράς, με την ταυτότητά της.

Ο συνδυασμός της πληροφορίας για την παραγωγή μιας πρόβλεψης, μπορεί να επεκταθεί και στο συνδυασμό των μορφών της πληροφορίας που δέχεται ένα μοντέλο. Προσπάθειες για την αξιοποίηση των διαφορετικών μορφών δεδομένων από ένα μοναδικό μοντέλο έχουν γίνει και σε άλλους τομείς όπως αυτός της υγείας. Για παράδειγμα το μοντέλο HAIM (Soenksen et al., 2022) έχει ως σκοπό την αξιοποίηση εικόνων (ακτίνες X, μαγνητικές εξετάσεις), κειμένου (ιατρικές σημειώσεις από γιατρούς) και χρονοσειρών (οξυγονομέτρηση, ηλεκτροκαρδιογραφήματα) από ένα ποικιλότροπο (multimodal) μοντέλο δεδομένων, για την πρόβλεψη μιας επικείμενης νόσησης.

Την τελευταία δεκαετία, εμφανίζονται στο χώρο της τεχνητής νοημοσύνης τα «θεμελιώδη μοντέλα» (foundation models) (Bommasani et al., 2021). Πρόκειται για μοντέλα μηχανικής μάθησης που προ-εκπαιδεύονται με έναν μεγάλο όγκο δεδομένων από παράγοντες που κατέχουν υπολογιστικούς πόρους και τεχνογνωσία. Στη συνέχεια, τα μοντέλα αυτά μπορούν να προσαρμοστούν και να εξειδικευτούν σε συγκεκριμένες εφαρμογές παρόλο που αρχικά δε σχεδιάστηκαν για αυτές, διατηρώντας υψηλές επιδόσεις. Η μετεκπαίδευση αυτών των μοντέλων μπορεί να γίνει με σημαντικά λιγότερους πόρους, καθώς και με μικρότερο όγκο δεδομένων. Η ικανότητα του μοντέλου να λάβει αποφάσεις ή να παράγει προβλέψεις με μειωμένο δείγμα εκπαίδευσης (few-shot generalization) είναι αξιοσημείωτη, καθώς υπάρχουν περιπτώσεις όπου δεν υπάρχει διαθεσιμότητα αρκετών δεδομένων και αξιοποιούνται παρεμφερή στοιχεία με τα οποία έχει εκπαιδευτεί το θεμελιώδες μοντέλο.

Σημαντικό παράδειγμα των foundation models είναι τα μεγάλα γλωσσικά μοντέλα (Large Language Models), τα οποία είναι μοντέλα νευρωνικών δικτύων (δισεκατομμυρίων παραμέτρων) που έχουν προεκπαιδευτεί με μεγάλο όγκο κειμένων φυσικής γλώσσας, και χρησιμοποιούνται μεταξύ άλλων για την παραγωγή κειμένου, τη μετάφραση, και τη διαδραστική συνομιλία με χρήστες. Οι δυνατότητες αυτές, φαίνεται πως αυξάνονται με το μέγεθος του προεκπαιδευμένου μοντέλου καθώς δυνατότητες γλωσσικών μοντέλων για τις οποίες παρατηρήθηκε σημαντική επίδοση, επιτυγχάνονται μόνο σε συγκεκριμένες υπολογιστικές κλίμακες (Wei et al., 2022).

Στο χώρο των χρονοσειρών, γίνονται προσπάθειες για το σχεδιασμό και την εκπαίδευση θεμελιωδών γενικευμένων μοντέλων, με σκοπό την προεκπαίδευση τους σε μεγάλα σύνολα δεδομένων χρονοσειρών, και τη χρήση τους για ειδικότερους σκοπούς. Ένα τέτοιο μοντέλο, σε περίπτωση που μπορούσε να παράγει συστηματικά ακριβείς προβλέψεις, θα μπορούσε να οδηγήσει στην εξοικονόμηση των υπολογιστικών πόρων, καθώς και τον εκδημοκρατισμό των μοντέλων μηχανικής μάθησης, και την είσοδο νέων παραγόντων στο χώρο, όπως παρατηρήθηκε με την άνοδο των γλωσσικών μοντέλων τα τελευταία χρόνια. Παράδειγμα αυτής της προσπάθειας, είναι το μοντέλο Lag-Llama (Rasul et al., 2023), το οποίο παράγει προβλέψεις με βάση τις παρελθοντικές τιμές των χρονοσειρών, και λαμβάνει ως συμμεταβλητές κάποιες χρονικές υστερήσεις τους (lags). Η υλοποίησή του, αποτελείται από νευρωνικά δίκτυα, και πιο συγκεκριμένα μετασχηματιστές (όπως τα γλωσσικά μοντέλα) ενώ εκπαιδεύεται με έναν μεγάλο όγκο χρονοσειρών από διάφορες πηγές που αφορούν τομείς όπως την ενέργεια, τις μεταφορές, την οικονομία, τη φύση.

Ωστόσο, στα σύνολα δεδομένων χρονοσειρών που είναι διαθέσιμα στο διαδίκτυο, πέρα από τις ίδιες ακολουθίες τιμών που παρέχονται, συχνά αυτές συνοδεύονται από μία λεκτική περιγραφή που περιγράφουν το περιεχόμενό τους. Στη βιβλιογραφία οι περιγραφές αυτές δε χρησιμοποιούνται κατά την εκπαίδευση των μοντέλων. Φυσικά μία πρόκληση για ένα ενδεχόμενο μοντέλο που αξιοποιεί αυτήν την παράμετρο, είναι η κατάλληλη μετατροπή των διαφόρων μορφών δεδομένων, όπως χρονοσειρές και περιγραφές, σε ένα είδος πληροφορίας που μπορεί να επεξεργαστεί και να αφομοιωθεί από το υπολογιστικό σύστημα. Η διαδικασία της μετατροπής του κειμένου σε μία μορφή που μπορεί να ερμηνευθεί από τον υπολογιστή, ονομάζεται επεξεργασία φυσικής γλώσσας.

1.2 Σκοπός Εργασίας

Δεδομένης της προόδου των γενικευμένων μοντέλων πρόβλεψης, ο σκοπός της παρούσας διπλωματικής εργασίας είναι να εξετάσει την αποτελεσματικότητα του συνδυασμού χρονοσειρών και συμμεταβλητών όπως οι λεκτικές περιγραφές και τα ποιοτικά χαρακτηριστικά τους στην παραγωγή προβλέψεων. Πιο συγκεκριμένα, γίνεται η υπόθεση πως οι χρονοσειρές των οποίων οι λεκτικές περιγραφές είναι νοηματικά παρεμφερείς ή τα ποιοτικά χαρακτηριστικά τους μοιάζουν, ενδέχεται να μεταβάλλονται με παρεμφερή τρόπο και η συμπερίληψη των συμμεταβλητών αυτών να συμβάλλει στην καλύτερη εκπαίδευση του μοντέλου. Στην παρούσα εργασία, αναπτύσσεται μία πειραματική διαδικασία για την ερμηνεία περιγραφών των χρονοσειρών με τεχνικές επεξεργασίας φυσικής γλώσσας, τον υπολογισμό ποιοτικών χαρακτηριστικών και το σχεδιασμό βαθιών νευρωνικών δικτύων με παράλληλες εισόδους με σκοπό να μελετηθεί η συνεισφορά των συμμεταβλητών αυτών στην πρόβλεψη.

Σε έναν πρώτο χρόνο, θέλουμε να διερευνήσουμε ποια τεχνική ερμηνείας των περιγραφών παράγει τις πιο αντιπροσωπευτικές αναπαραστάσεις για τις περιγραφές. Για το σκοπό αυτό, οι αναπαραστάσεις που δημιουργούνται χρησιμοποιούνται για την ταξινόμηση των χρονοσειρών και ελέγχεται η επίδοσή τους σε σχέση με ένα γνωστό επιθυμητό αποτέλεσμα. Στη συνέχεια, αναλύεται η επίδοση των αλγορίθμων συσταδοποίησης προκειμένου να βρεθεί η βέλτιστη μέθοδος συσταδοποίησης των χρονοσειρών βάσει της αναπαράστασης των περιγραφών. Τέλος, σχεδιάζονται τρία βαθιά νευρωνικά δίκτυα για το σκοπό της πρόβλεψης. Συγκεκριμένα πρόκειται για ένα βαθύ νευρωνικό δίκτυο αναφοράς που λαμβάνει ως είσοδο μόνο τις χρονοσειρές (μοντέλο MLP-Baseline). Στη συνέχεια, σχεδιάζεται ένα βαθύ νευρωνικό δίκτυο παράλληλης εισόδου (μοντέλο MLP-Parallel) που επιτρέπει την παράλληλη είσοδο των αναπαραστάσεων των περιγραφών και των ποιοτικών χαρακτηριστικών παράλληλα με τις χρονοσειρές. Ακόμα, εξετάζεται ένα μοντέλο εκπαίδευσης σε δύο χρόνους, που βασίζεται στο MLP-Baseline το οποίο εκπαιδεύεται αρχικά χωρίς παράλληλες εισόδους, και στη συνέχεια μία δεύτερη φορά, με χρήση παράλληλης εισόδου για τις περιγραφές των χρονοσειρών (μοντέλο MtMs-NLP). Ο στόχος αυτών των τριών αρχιτεκτονικών είναι να αξιολογηθεί ποια παρέχει το βέλτιστο αποτέλεσμα όσον αφορά το σφάλμα πρόβλεψης και να διαπιστωθεί αν οι αρχιτεκτονικές που χρησιμοποιούν τις συμμμεταβλητές παράγουν καλύτερα αποτελέσματα.

1.3 Δομή Εργασίας

Στο Κεφάλαιο 2 μελετώνται οι βασικές αρχές της μηχανικής μάθησης και των νευρωνικών δικτύων. Εξετάζονται τα είδη της μηχανικής μάθησης και στη συνέχεια αναλύονται οι βασικές αρχές των νευρωνικών δικτύων, οι διαφορετικές αρχιτεκτονικές, η εκπαίδευση και η βελτιστοποίηση τους. Τέλος, αναφέρονται τεχνικές μεταεπεξεργασίας, όσον αφορά την καταπολέμηση της υπερπροσαρμογής και της υποπροσαρμογής, και τις τεχνικές συνδυασμού των μοντέλων.

Το Κεφάλαιο 3 επικεντρώνεται στην επεξεργασία φυσικής γλώσσας. Σε ένα πρώτο χρόνο παρουσιάζονται οι κύριες τεχνικές προεπεξεργασίας του κειμένου, όπως η τεχνική tokenization, η αφαίρεση των τετριμμένων λέξεων και οι μέθοδοι αποκατάληξης και λημματοποίησης. Στη συνέχεια, αναλύονται οι μέθοδοι αναπαράστασης κειμένου μέσω της παραγωγής διανυσμάτων λέξεων. Μελετώνται προσεγγίσεις όπως η απλή τεχνική Bag of Words, ο αλγόριθμος Word2Vec και η χρήση μετασχηματιστών όπως το BERT και GPT-2. Τέλος, αναφέρονται τρόποι μετασχηματισμού των διανυσμάτων λέξεων σε διανύσματα προτάσεων, για την κωδικοποίηση των περιγραφών των χρονοσειρών σε μία μορφή επεξεργάσιμη από τον υπολογιστή.

Στο Κεφάλαιο 4 εξετάζονται τεχνικές ταξινόμησης και συσταδοποίησης διανυσμάτων προτάσεων καθώς και μετρικές αξιολόγησης αυτών. Περιλαμβάνονται μέθοδοι ταξινόμησης όπως οι αλγόριθμοι k-Κοντινότερων Γειτόνων και τα Δέντρα Αποφάσεων, και τεχνικές συσταδοποίησης όπως η μέθοδος k-Means, η Ιεραρχική Συσταδοποίηση και η μέθοδος DBSCAN.

Στο Κεφάλαιο 5 γίνεται μια ανάλυση των ποιοτικών χαρακτηριστικών των χρονοσειρών και των μεθόδων πρόβλεψης. Αρχικά παρουσιάζεται η μέθοδος αποσύνθεσης, ενώ στη συνέχεια εξετάζονται ποιοτικά στατιστικά που μπορούν να χρησιμοποιηθούν για την πρόβλεψη και η παρουσίασή τους στον N-διάστατο χώρο. Ύστερα, αναφέρονται οι κλασικές στατιστικές μέθοδοι πρόβλεψης και οι μέθοδος με χρήση νευρωνικών δικτύων.

Στο Κεφάλαιο 6 αναλύεται η πειραματική διαδικασία που ακολουθήθηκε. Σε ένα πρώτο χρόνο, επεξηγείται η συνολική διαδικασία, και στη συνέχεια αναφέρονται τα δεδομένα που συλλέχθηκαν, η διαδικασία της παραγωγής διανυσμάτων προτάσεων, της ταξινόμησης και της συσταδοποίησης των χρονοσειρών καθώς και τα αποτελέσματα των πειραμάτων αυτών. Σε ένα δεύτερο χρόνο, παρουσιάζονται τα νευρωνικά μοντέλα που σχεδιάστηκαν για το σκοπό της πρόβλεψης, και αναλύονται οι επιδόσεις τους.

Τέλος, στο Κεφάλαιο 7 παρουσιάζονται τα συμπεράσματα της εργασίας όσον αφορά την παραγωγή διανυσμάτων προτάσεων και τη χρήση των συμμεταβλητών για την παραγωγή προβλέψεων, καθώς και προτάσεις για μελλοντικές προεκτάσεις και βελτιώσεις.

Κεφάλαιο 2

Μηχανική Μάθηση και Νευρωνικά Δίκτυα

Στην παρούσα διπλωματική εργασία, χρησιμοποιήθηκαν τεχνικές μηχανικής μάθησης τόσο για την επεξεργασία της φυσικής γλώσσας όσο και για τη δημιουργία προβλέψεων. Σε αυτή την ενότητα, θα γίνει μία εισαγωγή στη μηχανική μάθηση και τα νευρωνικά δίκτυα, πάνω στην οποία θα βασιστούν τα επόμενα κεφάλαια.

2.1 Μηχανική Μάθηση (Machine Learning)

Η μηχανική μάθηση, ορίζεται ως ένα σύνολο μεθόδων ανίχνευσης προτύπων σε δεδομένα με σκοπό την πρόβλεψη ή τη λήψη απόφασης (Murphy, 2012). Ενώ πρόκειται για έναν τομέα που εφευρέθηκε τον περασμένο αιώνα, η έλλειψη υπολογιστικών πόρων δεν επέτρεψε μεγάλες πρακτικές εφαρμογές και περαιτέρω έρευνα. Τα τελευταία χρόνια ωστόσο, η πρόοδος στον τομέα των ηλεκτρονικών υπολογιστών, τις επεξεργαστικές μονάδες και τις κάρτες γραφικών (GPU) συνέβαλλαν στην άνθιση της Μηχανικής Μάθησης, με εφαρμογές σε πολλές επιστήμες. Οι τεχνικές Μηχανικής Μάθησης μπορούν να διαχωριστούν σε τρεις κύριες κατηγορίες οι οποίες μελετώνται αναλυτικά παρακάτω καθώς αναφέρονται και κάποια παραδείγματα.

2.1.1 Είδη Μηχανικής Μάθησης

Επιβλεπόμενη Μάθηση (Supervised Learning)

Ο στόχος της επιβλεπόμενης μάθησης είναι η εκμάθηση μίας αντιστοίχισης από εισόδους x σε εξόδους y. Στη διάθεση ενός μοντέλου επιβλεπόμενης μάθησης, είναι ένα σύνολο ζευγαριών εισόδου-εξόδου $D = \{(x_i, y_i)\}_{i=1}^{N}$. Το D ονομάζεται σύνολο εκπαίδευσης και περιέχει N δείγματα. Η είσοδος $\vec{x}$ είναι ένα διάνυσμα όπου σε κάθε θέση περιέχει την τιμή για κάθε χαρακτηριστικό (feature) το οποίο παρέχει πληροφορίες για τον χαρακτηρισμό του δείγματος. Η έξοδος $\vec{y}$ ή ετικέτα (label) είναι γνωστή και υποδεικνύει τι πραγματικά είναι το δείγμα. Η έξοδος μπορεί να έχει τη μορφή μίας πραγματικής τιμής $y_i \in \mathbb{R}$ η οποία μας παραπέμπει σε ένα πρόβλημα παλινδρόμησης (regression) αλλά μπορεί να έχει τη μορφή μίας κατηγορίας, για παράδειγμα $y_i \in \{\text{Άντρας, Γυναίκα}\}$ δηλώνοντας ένα πρόβλημα ταξινόμησης (classification). Το χαρακτηριστικό της επιβλεπόμενης μάθησης είναι πως τα μοντέλα εκπαιδεύονται με γνωστές εξόδους. Παραδείγματα προβλημάτων στα οποία εφαρμόζεται η Επιβλεπόμενη Μηχανική Μάθηση είναι η πρόβλεψη μίας αρρώστιας με τη βοήθεια ιατρικών δεικτών (πα-

ράδειγμα ταξινόμησης) ή η πρόβλεψη της τιμής ενός ακινήτου με βάση την τοποθεσία του, το μέγεθος του και την κατάσταση στην οποία βρίσκεται.

Μη Επιβλεπόμενη Μάθηση (Unsupervised Learning)

Στη μη επιβλεπόμενη μάθηση, το σύνολο εκπαίδευσης αποτελείται μόνο από εισόδους $D = \{x_j\}_{j=1}^{N}$, και ο στόχος είναι να βρεθούν πρότυπα, «ανακαλύπτοντας» κάποια δομή και εξάγοντας γνώση από τα δεδομένα. Το πρόβλημα λοιπόν δεν είναι ρητά ορισμένο, αφού δεν υπάρχει κάποια ετικέτα. Επιπλέον, σε αντίθεση με την επιβλεπόμενη μάθηση κατά την οποία βρίσκεται στη διάθεση του μοντέλου η πραγματική έξοδος που παρατηρήθηκε, η μέτρηση του σφάλματος στα μοντέλα μη επιβλεπόμενης μάθησης δεν είναι τετριμμένη. Προβλήματα μη επιβλεπόμενης μάθησης είναι η συσταδοποίηση (clustering), η μείωση διαστάσεων (dimensionality reduction) και η ανίχνευση ανωμαλιών (anomaly detection). Παραδείγματα προβλημάτων στα οποία εφαρμόζονται τεχνικές μη επιβλεπόμενης μάθησης είναι η συσταδοποίηση πελατών βάσει των αγοραστικών τους συνηθειών για τη δημιουργία εξατομικευμένων προωθητικών διαφημίσεων, η ανίχνευση ανωμαλιών σε τραπεζικές συναλλαγές για τον εντοπισμό πιθανής απάτης και η μείωση των διαστάσεων των δεδομένων εικόνας για τη βελτίωση της απόδοσης των αλγορίθμων αναγνώρισης προσώπου.

Ενισχυτική Μάθηση (Reinforcement Learning)

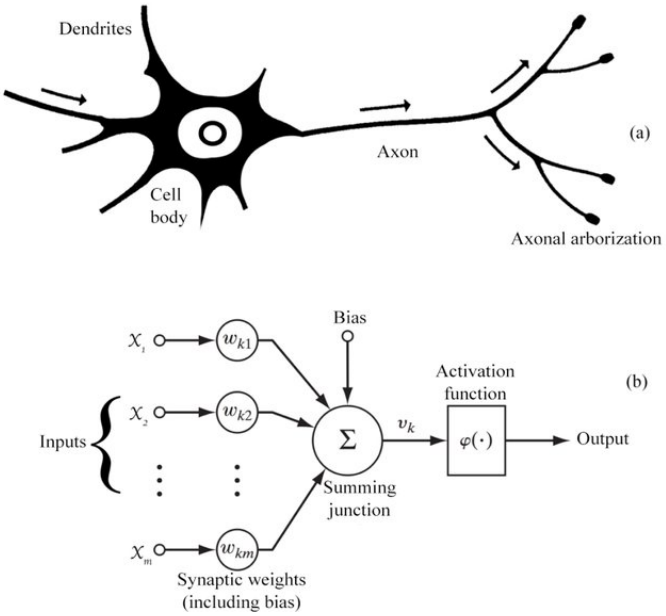
Στην ενισχυτική μάθηση, υπάρχει ένας αυτόνομος πράκτορας (agent) ο οποίος μαθαίνει να λαμβάνει αποφάσεις αλληλεπιδρώντας με ένα περιβάλλον. Ο στόχος του πράκτορα είναι να μεγιστοποιήσει τη συνολική ανταμοιβή του (reward) κατά τη διάρκεια μίας ακολουθίας ενεργειών και συνεπώς ταυτόχρονα να ελαχιστοποιήσει τις ποινές (punishments) που δέχεται όταν κάνει κάτι λάθος. Ο πράκτορας παρατηρεί την κατάσταση του περιβάλλοντος (state), λαμβάνει μία ενέργεια (action), και λαμβάνει μία ανταμοιβή βασισμένη στην ενέργεια που πραγματοποίησε. Στη συνέχεια, το περιβάλλον μεταβαίνει σε μία νέα κατάσταση και ο πράκτορας καλείται να ανταποκριθεί. Η ενισχυτική μάθηση χρησιμοποιείται, μεταξύ άλλων για την πλοήγηση των αυτόνομων οχημάτων τα οποία καλούνται να προχωρούν αυτόνομα στο χώρο και παράλληλα να μαθαίνουν από τις αλληλεπιδράσεις τους με το περιβάλλον.

2.2 Νευρωνικά Δίκτυα

Η σχεδίαση του πρώτου τεχνητού νευρωνικού δικτύου ονόματι Perceptron έγινε από τους McCulloch και Pitts (McCulloch & Pitts, 1943) ενώ η πρώτη υλοποίηση του (Mark I Perceptron) και περαιτέρω έρευνα έγινε από τον Frank Rosenblatt (Rosenblatt, 1958).

Η ιδέα βασίζεται στους νευρώνες του εγκεφάλου, οι οποίοι αποτελούνται από δενδρίτες που λαμβάνουν σήματα μέσω των συνάψεων, έναν πυρήνα που επεξεργάζεται τα σήματα, και έναν νευράξονα που μεταδίδει τα σήματα σε άλλους νευρώνες. Στο perceptron αντίστοιχα, υπάρχουν οι είσοδοι που έχουν συντελεστές (πίνακες βαρών), ένας κόμβος επεξεργασίας και η έξοδος όπως φαίνεται στην εικόνα 2.1.

Πλέον τα Νευρωνικά δίκτυα αποτελούν το πιο σημαντικό εργαλείο της μηχανικής μάθησης και παρακάτω θα μελετηθούν κάποιες κατηγορίες τους.



Σχήμα 2.1: *Έμπνευση των τεχνητών νευρώνων από τους βιολογικούς νευρώνες (Akgün & Demir, 2018).*

2.2.1 Perceptron

Στην αρχική μορφή των NN, η κατεύθυνση της πληροφορίας, όπως και στον ανθρώπινο εγκέφαλο, είναι από την είσοδο προς την έξοδο, εξού και ο χαρακτηρισμός εμπροσθιοτροφοδοτούμενα νευρωνικά δίκτυα. Αυτή η κατηγορία δικτύων, έχει εφαρμογές σε προβλήματα ταξινόμησης, παλινδρόμησης και συστημάτων ανίχνευσης ανωμαλιών καθώς σε σχέση με άλλα πιο περίπλοκα μοντέλα, παρέχουν μία λύση χαμηλότερου υπολογιστικού κόστους και υψηλής ταχύτητας. Ιδιαίτερο ενδιαφέρον παρουσιάζει η μελέτη του Perceptron (McCulloch & Pitts, 1943) καθώς αποτελεί τη βάση πάνω στην οποία αναπτύχθηκαν τα μεταγενέστερα πιο σύνθετα μοντέλα.

Υπολογισμός Εξόδου Perceptron

Όπως μπορεί να παρατηρηθεί στην εικόνα 2.1 β), για τον υπολογισμό της εξόδου ενός απλού τεχνητού νευρώνα, πραγματοποιούνται τα παρακάτω βήματα. Το διάνυσμα της εισόδου $\vec{x}$, πολλαπλασιάζεται με τον πίνακα συναπτικών βαρών $\vec{w}$ (εσωτερικό γινόμενο) και στη συνέχεια αθροίζονται. Στο άθροισμα αυτό, προστίθεται ο συντελεστής προκατάληψης (bias). Στη συνέχεια, το αποτέλεσμα μετασχηματίζεται από τη συνάρτηση ενεργοποίησης (activation function) και υπολογίζεται η έξοδος του κόμβου. Ο σκοπός του μετασχηματισμού με τη συνάρτηση ενεργοποίησης είναι η κανονικοποίηση της τιμής του αποτελέσματος σε ένα συγκεκριμένο εύρος.

Όσον αφορά τις συναρτήσεις ενεργοποίησης, για παράδειγμα οι McCulloch και Pitts (McCulloch & Pitts, 1943) χρησιμοποιούν τη βηματική συνάρτηση Heaviside που ορίζεται ως:

$$\varphi(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geq 0 \end{cases} \tag{2.1}$$

Κοινές συναρτήσεις ενεργοποίησης είναι η σιγμοειδής και η υπερβολική εφαπτομένη :

$$\text{Sigmoid} : \phi(x) = \frac{1}{1+e^{-x}}, \quad \tanh : \phi(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}, \tag{2.2}$$

Καθώς και η Identity και ReLU (Glorot, Bordes, & Bengio, 2011):

$$\text{Identity} : \phi(x) = x, \quad \text{ReLU}: \phi(x) = \begin{cases} 0, & x \leq 0 \\ x, & x > 0 \end{cases} \tag{2.3}$$

Η έξοδος ενός νευρώνα, τελικά υπολογίζεται ως:

$$y = \phi\left(\sum_{i=0}^{N} x_i w_i + b\right) \tag{2.4}$$

όπου:

- N είναι ο συνολικός αριθμός των εισόδων,
- x οι τιμές των εισόδων,
- w ο πίνακας των συνοπτικών βαρών,
- με b συμβολίζεται η προκατάληψη.

Εκπαίδευση Perceptron

Η εκπαίδευση ενός απλού εμπροσθιοτροφοδοτούμενου νευρωνικού δικτύου γίνεται υπολογίζοντας και ελαχιστοποιώντας τη συνάρτηση απώλειας (loss function). Η συνάρτηση αυτή είναι μια μετρική για την αξιολόγηση του σφάλματος μεταξύ μιας πρόβλεψης και της πραγματικής τιμής που παρέχεται κατά τη διάρκεια της εκπαίδευσης.

Η πιο κοινή συνάρτηση απώλειας είναι αυτή του μέσου τετραγωνικού σφάλματος (Mean Squared Error - MSE), η οποία υπολογίζει το μέσο όρο του τετραγώνου των διαφορών μεταξύ της εξόδου και της πραγματικής τιμής. Έστω λοιπόν η συνάρτηση απώλειας *MSE*:

$$\mathcal{E} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2. \tag{2.5}$$

Για να ελαχιστοποιηθεί η τιμή της συνάρτησης απώλειας, πρέπει να προσαρμοστούν τα συναπτικά βάρη των νευρώνων, δίνοντας λιγότερο ή περισσότερο σημασία στην είσοδο ενός νευρώνα με σκοπό να προσαρμοστεί η προβλεπόμενη έξοδος όσο πιο κοντά γίνεται στην πραγματική.

Η διαδικασία ελαχιστοποίησης της συνάρτησης απώλειας γίνεται με βάση την κλίσης της, με τη χρήση της μεθόδου Gradient Descent. Ο στόχος είναι να υπολογίζεται η κατεύθυνση της μέγιστης κλίσης της συνάρτησης απώλειας και στη συνέχεια να προσαρμόζονται τα βάρη προς την αντίθετη κατεύθυνση αυτής της κλίσης (προς το ελάχιστο).

Για να βρεθεί το ελάχιστο της συνάρτησης απώλειας μεταβάλλοντας τα βάρη του νευρωνικού δικτύου, υπολογίζεται η μερική παράγωγος της συνάρτησης απώλειας $\mathcal{E}$ ως προς τα βάρη w_{ij} ως:

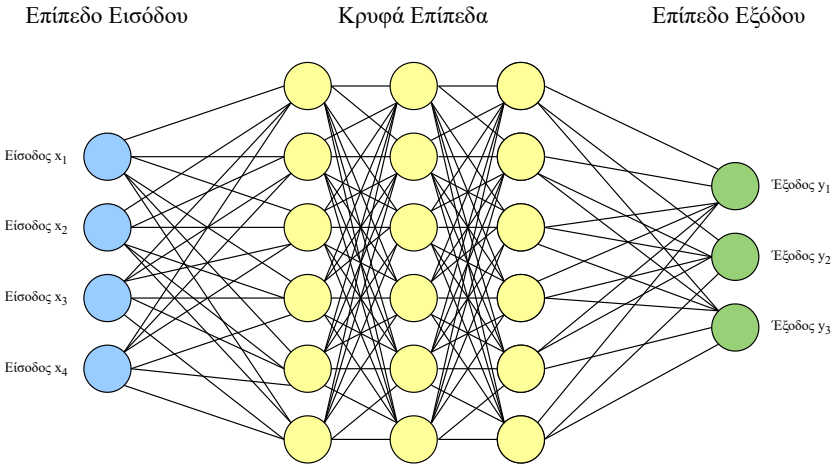
$$\Delta w_{ij} = -\eta \frac{\partial \mathcal{E}}{\partial w_{ij}} \Rightarrow w_{ij}^{t+1} = w_{ij}^{t} - \eta \frac{\partial \mathcal{E}}{\partial w_{ij}} \tag{2.6}$$

όπου η είναι ο συντελεστής μάθησης (learning rate), μια σταθερά που καθορίζει το μέγεθος του βήματος (step) κατά την ενημέρωση των βαρών.

Τα βάρη προσαρμόζονται συνεχώς μέχρι η συνάρτηση απώλειας να φτάσει σε ένα τοπικό ή ολικό ελάχιστο, επιτυγχάνοντας έτσι την εκπαίδευση του νευρωνικού δικτύου.

2.2.2 Βαθιά Μηχανική Μάθηση (Deep Learning)

Η ειδοποιός διαφορά μεταξύ ενός απλού νευρωνικού δικτύου και ενός Deep Learning πολυεπίπεδου μοντέλου, είναι ο αριθμός των κρυφών επιπέδων. Η είσοδος ενός νευρώνα ενδέχεται να είναι η είσοδος του επιπέδου εισόδου, δύναται να είναι όμως η έξοδος ενός άλλου νευρώνα ο οποίος με τη σειρά του μπορεί να επικοινωνεί με ακόμα περισσότερους.



Σχήμα 2.2: *Μοντελοποίηση ενός πλήρως συνδεδεμένου Νευρωνικού Δικτύου με τέσσερεις εισόδους, τρεις εξόδους και τρία κρυφά επίπεδα έξι νευρώνων.*

Τα πολλαπλά κρυφά επίπεδα, επιτρέπουν την εξαγωγή σύνθετων, μη γραμμικών εξαρτήσεων από τα δεδομένα (Brownlee, 2018) επιφέροντας όμως τον κίνδυνο της υπερπροσαρμογής σε αυτά ο οποίος θα μελετηθεί αναλυτικότερα στη συνέχεια.

Σε σχέση με τα απλά NN, χρησιμοποιούνται παρόμοιες συναρτήσεις ενεργοποίησης και απώλειας, αλλά η διαδικασία εκπαίδευσης διαφέρει εφόσον πρέπει να ενημερώνει πολλαπλούς πίνακες βαρών.

Οι αρχιτεκτονικές των νευρωνικών δικτύων ποικίλλουν ανάλογα με την εφαρμογή τους. Για τους σκοπούς αυτής της διπλωματικής εργασίας, θα μελετηθούν αναλυτικά τα Εμπροσθιοτροφοδοτούμενα Βαθιά Νευρωνικά Δίκτυα, οι Αυτοκωδικοποιητές και οι Μετασχηματιστές, ενώ θα γίνει αναφορά και στα Συνελικτικά καθώς και τα Αναδρομικά Νευρωνικά Δίκτυα.

Εμπροσθιοτροφοδοτούμενα Βαθιά Νευρωνικά Δίκτυα

Για την εκπαίδευση ενός εμπροσθιοτροφοδοτούμενου πολυεπίπεδου νευρωνικού δικτύου, όπως αυτό της εικόνας 2.2, δεδομένου πως το σφάλμα και τα βάρη εξαρτώνται από πολλά επίπεδα νευρώνων, η τεχνική εκπαίδευσης του μοντέλου διαφέρει από την σχέση 2.6 των απλών νευρωνικών δικτύων.

Επιπλέον, επειδή συχνά πρόκειται για δίκτυα με πολλές μονάδες εξόδου, η συνάρτηση απώλειας επαναπροσδιορίζεται για N συνολικά δείγματα εκπαίδευσης και Z αριθμούς εξόδου ως εξής:

$$E(\vec{w}) = \frac{1}{N \cdot Z} \sum_{i=1}^{N} \sum_{k=1}^{Z} (y_{ki} - \hat{y_{ki}})^2 \tag{2.7}$$

Ο αλγόριθμος Backpropagation (Linnainmaa, 1970), χρησιμοποιεί και αυτός τον gradient descent για να ελαχιστοποιήσει το τετραγωνικό σφάλμα μεταξύ των τιμών εξόδου του δικτύου και των τιμών-στόχων για αυτές τις εξόδους. Το πρόβλημα μάθησης που αντιμετωπίζει ο αλγόριθμος Backpropagation είναι να βρεθεί ανάμεσα σε όλες τις πιθανές τιμές βαρών για όλους τους κόμβους του δικτύου, η βέλτιστη λύση που θα ελαχιστοποιήσει τη συνάρτηση απώλειας.

Για την εκπαίδευση ενός τέτοιου μοντέλου, αρχικά υπολογίζεται μία πρόβλεψη *haty* για ένα δείγμα από το σύνολο δεδομένων από την είσοδο προς την έξοδο. Αυτό περιλαμβάνει τον υπολογισμό της εξόδου κάθε επιπέδου και την εισαγωγή της ως είσοδο στο επόμενο επίπεδο, μέχρι το τελικό επίπεδο της εξόδου που παράγει την πρόβλεψη όπως βλέπουμε στο σχήμα 2.2.

Στη συνέχεια, χρησιμοποιείται η συνάρτηση απώλειας για να αξιολογήσουμε τα σφάλματα της πρόβλεψης σε σύγκριση με την ετικέτα για κάθε έξοδο όπως εξηγήθηκε στη σχέση 2.7.

Για την ενημέρωση των βαρών, υπολογίζεται την κλίση της εξόδου σε σχέση με την ετικέτα και χρησιμοποιώντας τον κανόνα της αλυσίδας, υπολογίζουμε την κλίση του ακριβώς προηγούμενου επιπέδου πριν από την έξοδο. Έστω η συνάρτηση απώλειας $\mathcal{E}$, η έξοδος του δικτύου $\hat{y}$, και η έξοδος του i-οστού κρυμμένου επιπέδου η h_i, τότε ο κανόνας της αλυσίδας χρησιμοποιείται για να βρούμε την κλίση της απώλειας ως προς τα βάρη W_i του i-οστού στρώματος ως εξής:

$$\frac{\partial \mathcal{E}}{\partial W_i} = \frac{\partial \mathcal{E}}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial h_i} \cdot \frac{\partial h_i}{\partial W_i} \tag{2.8}$$

Τέλος, χρησιμοποιούμε την κλίση $\frac{\partial \mathcal{E}}{\partial W_i}$ για να ενημερώσουμε τα βάρη W_i του i-οστού κρυμμένου επιπέδου με την μέθοδο gradient descent:

$$W_i \leftarrow W_i - \eta \frac{\partial \mathcal{E}}{\partial W_i} \tag{2.9}$$

όπου η είναι ο ρυθμός μάθησης.

Η διαδικασία αυτή επαναλαμβάνεται για όλα τα κρυμμένα επίπεδα του δικτύου, προωθώντας τις κλίσεις προς το επίπεδο εισόδου, εξ ου και το όνομα backpropagation.

Η εκπαίδευση επαναλαμβάνεται σε όλα τα δείγματα του συνόλου δεδομένων μέχρι να επιτευχθεί η σύγκλιση του μοντέλου.

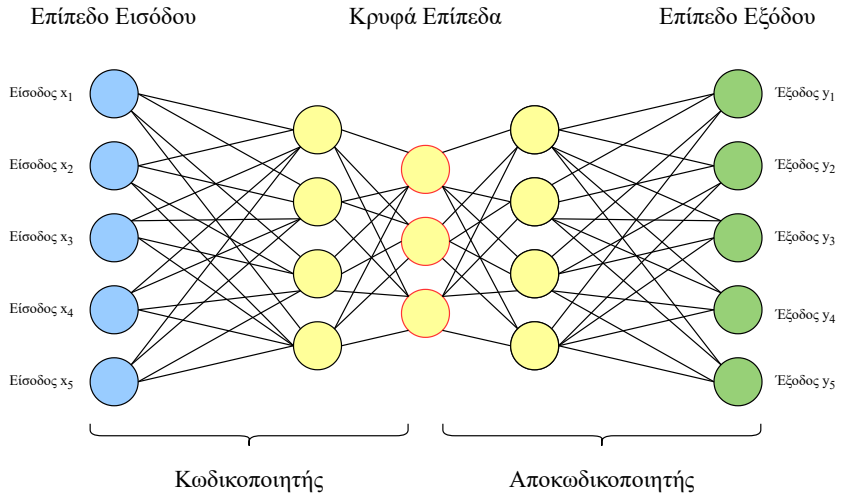
Βελτιστοποίηση Εκπαίδευσης - Στοχαστικό Gradient Descent

Το μειονέκτημα του απλού Gradient Descent που χρησιμοποιείται κατά την εκπαίδευση ενός νευρωνικού δικτύου γίνεται ολοένα και πιο εμφανές για μεγάλα σύνολα δεδομένων εκπαίδευσης. Η ανάγκη υπολογισμού της επιρροής κάθε δείγματος για κάθε βήμα της εκπαίδευσης, έχει ως αποτέλεσμα οι υπολογιστικοί πόροι που απαιτούνται για την εκπαίδευση να είναι ιδιαίτερα σημαντικοί. Η βελτιστοποίηση που προτείνεται από τον Στοχαστικού Gradient Descent ο οποίος χρησιμοποιήθηκε για πρώτη φορά στην εκπαίδευση νευρωνικών δικτύων από τον Amari (Amari, 1967), είναι πως τα δείγματα του συνόλου εκπαίδευσης ανακατεύονται, και ύστερα ομαδοποιούνται τυχαία σε παρτίδες (batches). Στη συνέχεια οι παρτίδες χρησιμοποιούνται για κάθε βήμα υπολογισμού, παρουσιάζοντας ουσιαστικά στο μοντέλο ως είσοδο μόνο τα δείγματα που εμπεριέχονται σε αυτές, αποκρύπτοντας τα υπόλοιπα δείγματα και εξοικονομώντας χρόνο και πόρους. Σε αντίθεση με τον απλό Gradient Descent ο οποίος υπολογίζει τις κλίσεις για όλα τα δείγματα και στη συνέχεια λαμβάνει το μέσο όρο των κλίσεων, ο Στοχαστικός GD αγνοεί πολλά δείγματα και δημιουργεί ένα μικρότερο υποσύνολο δεδομένων εκπαίδευσης με τα οποία υπολογίζονται οι κλίσεις. Έτσι, αντιμετωπίζει με πιο αποδοτικό τρόπο μεγάλα σύνολα δεδομένων, ενώ ταυτόχρονα βοηθά το μοντέλο να γενικεύει καλύτερα, καθώς η στοχαστική φύση του συχνά τον βοηθάει να ξεφεύγει από τοπικά ελάχιστα.

Αυτοκωδικοποιητές (Autoencoders)

Ο αυτοκωδικοποιητής είναι ένα νευρωνικό δίκτυο που εκπαιδεύεται με σκοπό να αναπαράγει την είσοδο που λαμβάνει στην έξοδο του. Στο εσωτερικό του περιέχει ένα ή περισσότερα κρυφά επίπεδα μικρότερης διάστασης από την είσοδο, (και την έξοδο αφού $\text{Διάσταση}_{\text{εισόδου}} = \text{Διάσταση}_{\text{εξόδου}}$) των οποίων ο στόχος είναι να συμπυκνώσουν την πληροφορία του επίπεδου εισόδου με τέτοιο τρόπο, ώστε να καθίσταται εφικτή η ανάκτηση της στο επίπεδο εξόδου. Συγκεκριμένα, για μία είσοδο $\bar{X} \in \mathbb{R}^n$, ο στόχος είναι ένας μετασχηματισμός κωδικοποίησης $K : \bar{X} \to \bar{Z}$ και ύστερα η ανακατασκευή της αρχικής πληροφορίας με έναν αποκωδικοποιητή $A : \bar{Z} \to \bar{X}$, όπου $dim(\bar{Z}) < dim(\bar{X})$. Η υλοποίηση τους, πρόκειται ουσιαστικά για μία παραλλαγή των εμπροσθιοτροφοδοτούμενων νευρωνικών δικτύων που αναλύθηκαν προηγουμένως, με την ιδιαιτερότητα πως πρόκειται για ένα συμμετρικό δίκτυο με τουλάχιστον ένα κρυφό επίπεδο, όπως βλέπουμε στην εικόνα 2.3. Η εκπαίδευσή τους πραγματοποιείται όπως μελετήθηκε στην υποενότητα 2.2.2.

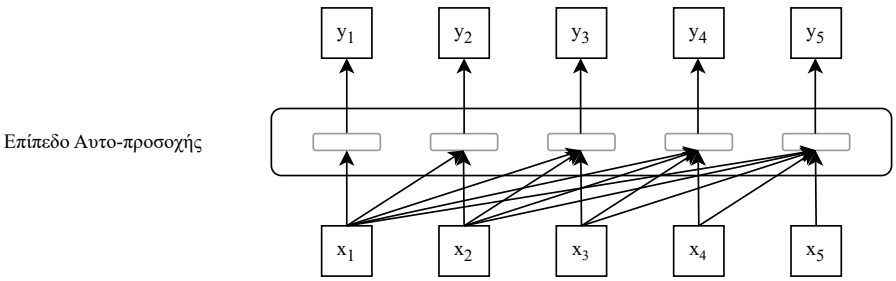
Η αρχιτεκτονική των αυτοκωδικοποιητών χρησιμοποιείται ευρέως για την μείωση διαστάσεων μεγάλων συνόλων δεδομένων την εξαγωγή χρήσιμων χαρακτηριστικών αλλά και για την ανίχνευση ανωμαλιών.



Σχήμα 2.3: *Μουτελοποίηση ενός αυτοκωδικοποιητή που σκοπό έχει να εκφράσει πληροφορία πέντε διαστάσεων σε τρεις (μέσω του κεντρικού κρυφού επιπέδου).*

Μετασχηματιστές (Transformers)

Οι μετασχηματιστές (transformers), δημοσιεύτηκαν από την ομάδα της Google, (Vaswani et al., 2017), και είναι βαθιά νευρωνικά δίκτυα που γρήγορα γίνονται τα νέα state-of-the-art μοντέλα για την επεξεργασία φυσικής γλώσσας τομέας που θα αναλυθεί σε βάθος στο κεφάλαιο 3 της εργασίας. Οι μετασχηματιστές αντιστοιχούν ακολουθίες από διανύσματα εισόδου $(x_1, \ldots, x_n)$ σε ακολουθίες εξόδου $(y_1, \ldots, y_n)$ του ίδιου μήκους. Η αρχιτεκτονική τους απαρτίζεται από πολλά «μπλοκ» μετασχηματιστών, καθένα από τα οποία είναι ένα πολυεπίπεδο νευρωνικό δίκτυο που αποτελείται είτε από απλά γραμμικά επίπεδα (linear layers), είτε από δίκτυα με προώθηση εισόδου (feedforward layers) είτε από επίπεδα αυτο-προσοχής (self-attention layers).



Σχήμα 2.4: *Το επίπεδο αυτο-προσοχής (Self Attention) των μετασχηματιστών.*

Στο σχήμα 2.4 απεικονίζεται η ροή πληροφοριών σε ένα επίπεδο αυτο-προσοχής. Όπως και με τον συνολικό μετασχηματιστή, ένα επίπεδο αυτο-προσοχής αντιστοιχεί ακολουθίες εισόδου $(x_1, \ldots, x_n)$ σε ακολουθίες εξόδου του ίδιου μήκους $(y_1, \ldots, y_n)$. Κατά την επεξεργασία κάθε στοιχείου στην είσοδο, το μοντέλο έχει πρόσβαση σε όλες τις εισόδους έως την i-οστή είσοδο που εξετάζεται χωρίς όμως να έχει πρόσβαση σε πληροφορίες για τις εισόδους πέρα από την τρέχουσα. Παρατηρείται λοιπόν πως η ανάλυση γίνεται «από αριστερά προς τα

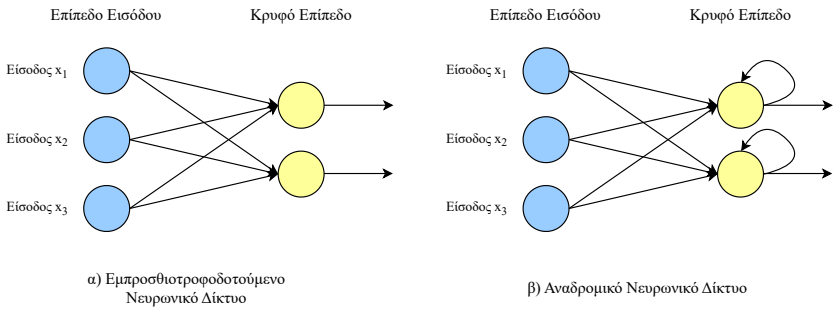
δεξιά» και όχι από τις δύο κατευθύνσεις.

Ο υπολογισμός που εκτελείται για κάθε στοιχείο είναι ανεξάρτητος από όλους τους άλλους υπολογισμούς συνεπώς μπορεί εύκολα να παραλληλοποιηθεί τόσο η προώθηση προς τα εμπρός όσο και η εκπαίδευση τέτοιων μοντέλων.

Με τα επίπεδα αυτο-προσοχής, οι μετασχηματιστές καταφέρνουν να ερμηνεύσουν το κείμενο φυσικής γλώσσας με βάση τα συμφραζόμενα των λέξεων, δημιουργώντας ένα νέο εργαλείο για την επεξεργασία και την παραγωγή της φυσικής γλώσσας, που χρησιμοποιούν η πλειοψηφία των Μεγάλων Γλωσσικών Μοντέλων (Large Language Models - LLMs). Περισσότερα για την εφαρμογή των μετασχηματιστών στην επεξεργασία φυσικής γλώσσας βρίσκονται στην υποενότητα 3.3.2.

Άλλες αρχιτεκτονικές Βαθιών Νευρωνικών Δικτύων

Αναδρομικά Νευρωνικά Δίκτυα: Τα Αναδρομικά Νευρωνικά Δίκτυα (Recurrent Neural Networks - RNN) είναι μια κατηγορία τεχνητών νευρωνικών δικτύων στην οποία οι συνδέσεις μεταξύ των κόμβων σχηματίζουν έναν κατευθυνόμενο γράφο. Σε αντίθεση με τα εμπροσθιοτροφοδοτούμενα δίκτυα όπου η κατεύθυνση της πληροφορίας είναι πάντα από την είσοδο προς την έξοδο, στα αναδρομικά δίκτυα παρουσιάζονται κύκλοι, όπως στην εικόνα 2.5. Η



Σχήμα 2.5: *Σύγκριση Εμπροσθιοτροφοδοτούμενων Νευρωνικών Δικτύων με τα Αναδρομικά Νευρωνικά Δίκτυα.*

αρχιτεκτονική αυτή, επιτρέπει να ληφθεί η έξοδος ενός κόμβου ως είσοδος την επόμενη χρονική στιγμή, λαμβάνοντας υπόψη τη χρονική διάσταση. Με τον τρόπο αυτό, τα δίκτυα «θυμούνται» προηγούμενες πληροφορίες τις οποίες χρησιμοποιούν για να επηρεάσουν τις τρέχουσες και μελλοντικές εισόδους, κάτι που τα καθιστά ιδιαίτερα χρήσιμα για εφαρμογές όπου η σειρά και η σχέση των δεδομένων είναι σημαντική. Ένα παράδειγμα αυτών των δικτύων είναι τα Long Short-Term Memory (LSTMs) νευρωνικά δίκτυα. Τα LSTMs (Hochreiter & Schmidhuber, 1997) δημοσιεύτηκαν για την αντιμετώπιση του υπολογισμού της κλίσης της συνάρτησης απώλειας για μεγάλες ακολουθίες δεδομένων εισόδου (όπως για παράδειγμα μεγάλες χρονοσειρές). Με τα «κύτταρα μνήμης» (memory cells), επιτυγχάνουν τη διατήρηση της πληροφορίας για παρατεταμένο χρονικό διάστημα, μαθαίνοντας μακροχρόνιες εξαρτήσεις για διαδοχικά δεδομένα όπου ο χρόνος είναι σημαντικός. Για το σκοπό

αυτό, τα «κύτταρα μνήμης» απαρτίζονται από την πύλη εισόδου, την πύλη εξόδου και την πύλη διαγραφής, οι οποίες ελέγχουν τη ροή των πληροφοριών επιτρέποντας στο μοντέλο να διατηρεί μία πληροφορία για να χρησιμοποιείται στο μέλλον ή να την απορρίπτει όταν κρίνει ότι δεν είναι πλέον χρήσιμη.

Συνελικτικά Νευρωνικά Δίκτυα: Τα Συνελικτικά Νευρωνικά Δίκτυα (Convolutional Neural Networks -CNNs) είναι ένας τύπος εμπροσθιοτροφοδοτούμενων νευρωνικών δικτύων που έχουν σχεδιαστεί με αρχικό στόχο την επεξεργασία δεδομένων όπως εικόνες και χρησιμοποιούνται ιδιαίτερα στον τομέα της όρασης των υπολογιστών. Χρησιμοποιούνται ευρέως στην επεξεργασία εικόνας και βίντεο, καθώς και σε άλλες εφαρμογές όπου τα δεδομένα μπορούν να αναπαρασταθούν ως πολυδιάστατα πλέγματα και δημοσιεύτηκαν για πρώτη φορά από την ομάδα του Lecun (LeCun et al., 1989).

Η αρχιτεκτονική τους αποτελείται από τρία είδη επιπέδων:

- Συνελικτικά Επίπεδα (Convolutional Layers):
  Τα συνελικτικά επίπεδα χρησιμοποιούν φίλτρα (kernels) για την εξαγωγή χαρακτηριστικών από τα δεδομένα εισόδου. Κάθε φίλτρο μετακινείται πάνω από την είσοδο και υπολογίζει το συνελικτικό γινόμενο $s(t) = \int x(a)w(t-a)da$ το οποίο συμβολίζεται ως $s(t) = (x * w)(t)$.

- Επίπεδα Υποδειγματοληψίας (Pooling Layers):
  Τα επίπεδα υποδειγματοληψίας μειώνουν τις διαστάσεις, διατηρώντας σημαντικές πληροφορίες και μειώνοντας την πολυπλοκότητα του μοντέλου.

- Πλήρως Συνδεδεμένα Επίπεδα (Fully Connected Layers): Τα πλήρως συνδεδεμένα επίπεδα λειτουργούν όπως στα παραδοσιακά νευρωνικά δίκτυα και χρησιμοποιούνται συνήθως στο τέλος του δικτύου για την ταξινόμηση ή την παραγωγή της πρόβλεψης, συνδυάζοντας τα χαρακτηριστικά που εξάχθηκαν από τα δύο προηγούμενα επίπεδα.

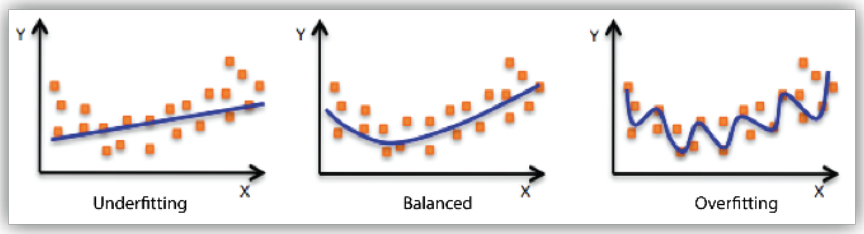
2.3 Τεχνικές Μεταεπεξεργασίας - Βελτιώσεις

2.3.1 Αντιμετώπιση Υπερπροσαρμογής/Υποπροσαρμογής

Ο σκοπός της εκπαίδευσης ενός νευρωνικού δικτύου είναι αυτό να «μάθει» από τα δεδομένα εκπαίδευσης και στη συνέχεια να γενικεύσει (generalization) τις γνώσεις του σε άγνωστα δεδομένα με τα οποία δεν έχει έρθει αντιμέτωπο στο παρελθόν. Ωστόσο, κάτι τέτοιο δεν επιτυγχάνεται πάντα, λόγω της υποπροσαρμογής ή της υπερπροσαρμογής του μοντέλου στα δεδομένα εκπαίδευσης.

Το φαινόμενο της υποπροσαρμογής (underfitting), παρατηρείται όταν το επίπεδο της συνάρτησης απώλειας που επιτυγχάνεται παραμένει υψηλό κατά την εκπαίδευση, κάτι που στη συνέχεια επιβεβαιώνεται από το σφάλμα κατά τη διάρκεια της αξιολόγησης με τα άγνωστα δεδομένα. Το μοντέλο αποτυγχάνει στην εκμάθηση των υποκείμενων σχέσεων στα δεδομένα εκπαίδευσης ενώ τρόποι αντιμετώπισης αποτελούν η αύξηση της πολυπλοκότητας του μοντέλου (προσθέτοντας περισσότερα επίπεδα και νευρώνες στο δίκτυο) και χρήση περισσότερων χαρακτηριστικών (features).

Αντιθέτως, η υπερπροσαρμογή overfitting συμβαίνει όταν το μοντέλο εκπαιδεύεται υπερβολικά καλά με δεδομένα εκπαίδευσης, συμπεριλαμβανομένου του θορύβου και των ανωμαλιών τους με αποτέλεσμα να μην επιτυγχάνει καλές προβλέψεις για τα άγνωστα δεδομένα. Οι τρόποι αντιμετώπισης της υπερπροσαρμογής αποτελούνται από το Dropout, τεχνική κατά την οποία διαγράφονται τυχαία ορισμένοι νευρώνες (τους επιβάλλεται μηδενικό βάρος), την πρόωρη διακοπή (Early Stopping) όπου η εκπαίδευση διακόπτεται σε περίπτωση που ύστερα από κάποιες επαναλήψεις εκπαίδευσης το σφάλμα ανεβαίνει αντί να πέφτει, και η αύξηση του μεγέθους των δεδομένων εκπαίδευσης είτε συμπεριλαμβάνοντας περισσότερα δεδομένα είτε παράγοντας συνθετικά δεδομένα.



Σχήμα 2.6: *Παράδειγμα Overfitting και Underfitting (URL, n.d.-a)*

2.3.2 Τεχνικές Εκμάθησης Συνόλου (Ensemble Learning)

Η τεχνική bagging (Breiman, 1996), είναι μία μέθοδος εκμάθησης συνόλου η οποία έχει ως στόχο την ελάττωση του σφάλματος λόγω υπερπροσαρμογής. Η ιδέα πίσω από τη μέθοδο αυτή είναι πως πολλά μοντέλα μπορούν να εκπαιδευτούν ξεχωριστά (συχνά με διαφορετικά υποσύνολα του συνόλου δεδομένων εκπαίδευσης) για τον ίδιο σκοπό, και στο τέλος να ψηφίσουν για την έξοδο των δεδομένων αξιολόγησης. Ο στόχος αυτής της διαδικασίας, είναι να αξιοποιηθεί το γεγονός πως εφόσον τα μοντέλα έχουν εκπαιδευτεί ξεχωριστά, αναμένεται να μην κάνουν όλα τα ίδια λάθη και έτσι κατόπιν ψηφοφορίας, να υπερισχύουν οι σωστές προβλέψεις.

Η τεχνική boosting, καταπολεμά το σφάλμα εκπαίδευσης μέσω διαδοχικών αντισταθμίσεων των αδυναμιών ενός μοντέλου κατά την εκπαίδευση. Παράδειγμα μίας τεχνικής boosting, αποτελεί η τεχνική Adaptative Boosting (Freund, Schapire, et al., 1996) που λειτουργεί σταδιακά, εστιάζοντας στις λάθος προβλέψεις ύστερα από κάθε εκπαίδευση του μοντέλου, και επαναλαμβάνοντας την εκπαίδευση με σκοπό την αντιστάθμιση των λάθος προβλέψεων διαδοχικά.

2.3.3 Μεταφορά Μάθησης (Transfer Learning)

Η μεταφορά μάθησης είναι η διαδικασία κατά την οποία μεταφέρεται η γνώση από ένα μοντέλο που έχει εκπαιδευτεί με κάποια δεδομένα εκπαίδευσης ενός περιβάλλοντος προέλευσης (source domain), έχοντας όμως ως στόχο η εφαρμογή της γνώσης να γίνει σε κάποιο άλλο περιβάλλον-στόχο (target domain) (Weiss, Khoshgoftaar, & Wang, 2016). Με αυτόν τον τρόπο, μπορεί να χρησιμοποιηθεί γνώση από κάποιο μοντέλο που εκπαιδεύτηκε με ένα

μεγάλο όγκο δεδομένων και του οποίου η εκπαίδευση χρειάστηκε πολλούς υπολογιστικούς πόρους, για την εκπαίδευση ενός δεύτερου, με σημαντικά λιγότερους πόρους και καλές επιδόσεις. Ως fine-tuning, ορίζεται η διαδικασία κατά την οποία η γνώση από το περιβάλλον προέλευσης προσαρμόζεται στο νέο περιβάλλον, ώστε να ανταποκρίνεται στο νέο στόχο του. Φυσικά πέρα από τα πλεονεκτήματα της εξοικονόμησης πόρων και της βελτίωσης της απόδοσης σε σχέσει με ένα μοντέλο που εκπαιδεύεται από την αρχή, υπάρχουν και προκλήσεις, αφού είναι πιθανό να υπάρχει δυσκολία στην προσαρμογή και εν τέλει να υπάρχει το φαινόμενο του negative learning (αρνητική μάθηση) που παρατηρείται όταν οι διαφορές μεταξύ των δύο περιβαλλόντων είναι τόσο σημαντική, ώστε οι γνώσεις που αποκτήθηκαν στο περιβάλλον προέλευσης είναι παραπλανητικές για την επίτευξη του στόχου στο νέο περιβάλλον. Στην ενότητα της επεξεργασίας φυσικής γλώσσας και ιδιαίτερα στην υποενότητα 3.3.2 υπογραμμίζεται η χρησιμότητα της μεταφοράς μάθησης για μοντέλα μετασχηματιστών τα οποία προ-εκπαιδεύονται για δισεκατομμύρια παραμέτρους με μεγάλο όγκο δεδομένων.

Κεφάλαιο 3

Επεξεργασία φυσικής γλώσσας (Natural Language Processing – NLP)

Η Επεξεργασία Φυσικής Γλώσσας (Natural Language Processing – NLP) έχει ως σκοπό την αποκωδικοποίηση της πληροφορίας σε μορφή κειμένου, με βάση τη διερμηνεία της και περιλαμβάνει γενικότερα την ανάπτυξη αλγορίθμων και μοντέλων για την κατανόηση, ερμηνεία και παραγωγή ανθρώπινης γλώσσας (Georgouli, 2015). Στην παρούσα εργασία, θέλοντας να διερευνήσουμε την επίπτωση των μεταδεδομένων των χρονοσειρών στην ακρίβεια μίας πρόβλεψης, η επεξεργασία της φυσικής γλώσσας είναι ιδιαίτερα σημαντική, αφού τα μεταδεδομένα που έχουμε, είναι μεταξύ άλλων και λεκτικές περιγραφές που αφορούν την εκάστοτε χρονοσειρά. Σε αυτή την ενότητα, θα δούμε αναλυτικά κάποιες μεθόδους διερμηνείας κειμένου από την προεπεξεργασία, ως την παραγωγή διανυσμάτων που μας βοηθούν να ερμηνεύσουμε προτάσεις όπως αυτές των μεταδεδομένων που έχουμε στη διάθεσή μας.

3.1 Προεπεξεργασία Κειμένου

Το πρώτο βήμα στη διαδικασία της επεξεργασίας της φυσικής γλώσσας, είναι η προεπεξεργασία του κειμένου, η οποία περιλαμβάνει μια σειρά από ενέργειες που αναλύονται παρακάτω.

3.1.1 Λεξικολογική Ανάλυση (Lexical Analysis)

Σε έναν πρώτο χρόνο, το κείμενο χωρίζεται σε κέρματα (tokens) τα οποία είναι λέξεις, αλλά και σύμβολα όπως τα σημεία στίξης. Ωστόσο, αγνοούνται διαχωριστικά όπως κενά, tabs και χαρακτήρες νέας γραμμής.

Αυτή	η	πρόταση	,	είναι	ένα	παράδειγμα	.

Πίνακας 3.1: *Διάσπαση της πρότασης σε tokens.*

Όπως βλέπουμε στο παράδειγμα 3.1, τα σημεία στίξης λαμβάνουν την ίδια σπουδαιότητα με τις κανονικές λέξεις, κάτι που δεν είναι πάντα επιθυμητό, για αυτό το λόγο, συχνά αφαιρούνται κοινά σημεία στίξης όπως «,», «.» και «'». Επίσης, οι χαρακτήρες συχνά μετατρέπονται από κεφαλαία σε πεζά γράμματα για να υπάρχει ομοιομορφία καθώς η ίδια λέξη με ή χωρίς κεφαλαία γράμματα αντιμετωπίζεται διαφορετικά από τους αλγόριθμους επεξεργασίας

κειμένου.

3.1.2 Αφαίρεση Τετριμμένων Λέξεων (Stopwords Removal)

Παρόμοια με τα σημεία στίξης, είναι εμφανές, πως λέξεις που εμφανίζονται συχνά σε ένα κείμενο όπως τα άρθρα, οι σύνδεσμοι και οι αντωνυμίες. Για παράδειγμα οι συνήθεις αγγλικές λέξεις “a”, “the”, “and”, δεν προσδίδουν επιπλέον σημασιολογική πληροφορία, χρήσιμη για την ανάλυση του κειμένου. Αντί αυτού, προσθέτουν θόρυβο και μεγαλώνουν τον αριθμό των tokens. Για το λόγο αυτό, είναι κοινή πρακτική να αφαιρούνται.

3.1.3 Αποκατάληξη και Λημματοποίηση (Stemming και Lemmatization)

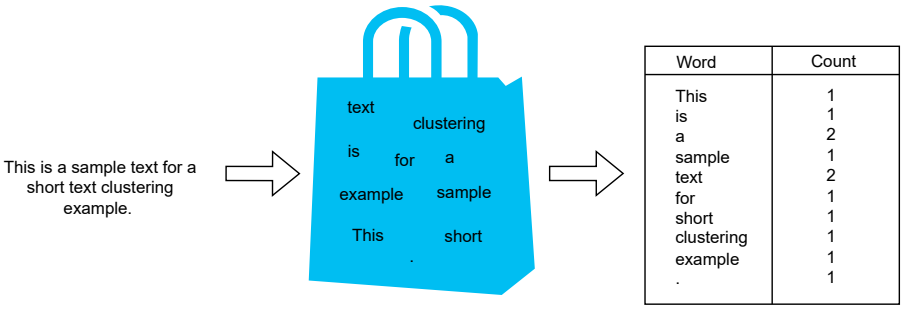
Παρατηρούμε επίσης, πως οι λέξεις “organizations”, “organized” και “organizing”, είναι τρεις λέξεις που ενώ νοηματικά μοιάζουν, προσμετρούνται σαν εντελώς ξεχωριστές έννοιες, αυξάνοντας το σύνολο των tokens.
Η τεχνική της αποκατάληξης Stemming, και ιδιαίτερα ο αλγόριθμος Porter Stemmer (Porter, 1980), αποκόπτει τις λέξεις και τις παράγωγές τους κρατώντας μόνο την κοινή τους ρίζα. Στο ανώτερο παράδειγμα, και οι τρεις λέξεις γίνονται “organ”, συνεπώς η λέξη που προκύπτει από τη μέθοδο του Stemming δεν είναι αναγκαστικά ένα υπαρκτό λήμμα καταχωρημένο σε λεξικό, αλλά η τεχνική ομαδοποιεί τις παραλλαγές μικραίνοντας το συνολικό αριθμό λέξεων. Παρόμοια, η τεχνική της λημματοποίησης Lemmatization κανονικοποιεί τις λέξεις σε μία εκδοχή κοντά σε κάποιο λήμμα, υπάρχον στο λεξικό της εκάστοτε γλώσσας, για να αποφύγει τις μορφολογικές παραλλαγές τους όπως ο πληθυντικός. Για παράδειγμα οι προαναφερθείσες λέξεις αντιστοιχίζονται στις λέξεις “organization”, “organized”, “organizing”.

3.1.4 Αλγόριθμος WordPiece

Πέρα από τις τεχνικές του Stemming και του Lemmatization, για την προεπεξεργασία του κειμένου, χρησιμοποιείται ο αλγόριθμος WordPiece που δημοσιεύτηκε από τη Google (Wu et al., 2016). Χωρίζοντας λέξεις σε «υπο-λέξεις» (sub-words), η τεχνική του WordPiece βοηθάει το μοντέλο που στη συνέχεια αναλύει το κείμενο να χειριστεί λέξεις που δεν έχει ξαναδεί αυτούσιες. Παραδείγματος χάρη, η λέξη «παράδειγμα» μπορεί να αναλυθεί στα συνθετικά «παρά» και «δείγμα». Τα αποδομημένα αυτά tokens συμβολίζονται με δύο σύμβολα δίεσης, [παρά] και [##δείγμα]. Η χρήση του αλγορίθμου γίνεται δημοφιλής με τη δημοσίευση του μοντέλου BERT που θα δούμε αναλυτικότερα στην υποενότητα των Μετασχηματιστών 3.3.2.

3.2 Μέθοδος Bag of Words

Η μέθοδος Bag of Words, είναι ένας τρόπος αναπαράστασης κειμένου σε μορφή διανυσμάτων με βάση τη συχνότητα εμφάνισης των λέξεων. Στην ουσία, αγνοείται η σειρά των λέξεων, και λαμβάνεται υπόψη μόνο η παρουσία ή η απουσία μιας λέξης στο κείμενο. Το αποτέλεσμα είναι ένα διάνυσμα που περιέχει τη συχνότητα εμφάνισης κάθε λέξης στο κείμενο.



Σχήμα 3.1: *Παράδειγμα Bag of Words.*

3.2.1 Term Occurence και Term Frequency

Για την αναπαράσταση των εγγράφων (μικρών κειμένων περιγραφής) που περιέχονται μέσα στο σύνολο των λεκτικών περιγραφών, αρχικά δημιουργείται ένα λεξικό $\mathcal{V}$ που περιέχει όλα τα μοναδικά tokens που εμφανίζονται στο σύνολο των εγγράφων. Η αναπαράσταση των εγγράφων γίνεται ως προς το λεξικό, σε ένα διάνυσμα μεγέθους ίσου με το συνολικό αριθμό μοναδικών όρων στο σύνολο των εγγράφων το οποίο συμβολίζουμε με $|\mathcal{V}|$. Για κάθε έγγραφο d στο σύνολο του κειμένου, δημιουργείται ένα διάνυσμα $\vec{X}_d$ διάστασης $|\mathcal{V}|$, ως εξής:

- Εμφάνιση Όρων (Term Occurence): Το διάνυσμα $\vec{X}_d$ παίρνει την τιμή 1 στη θέση του λεξικού που αντιστοιχεί στον όρο που περιέχεται στο έγγραφο που εξετάζεται και την τιμή 0 για κάθε άλλο όρο του λεξικού. Έστω λοιπόν d το έγγραφο (document) που εξετάζεται και $\mathcal{V}$ το λεξικό των μοναδικών όρων στο σύνολο των εγγράφων. Η μαθηματική σχέση του είναι:

$$x_i = \begin{cases} 1 & \text{αν ο όρος } t_i \text{ εμφανίζεται στο έγγραφο } d \\ 0 & \text{παντού αλλού} \end{cases} \tag{3.1}$$

- Συχνότητα Όρων (Term Frequency): Το διάνυσμα $\vec{X}_d$ παίρνει τη φυσική τιμή που αντιστοιχεί στη συχνότητα των παρουσιών του όρου μέσα στο έγγραφο, για όλους τους όρους του λεξικού (Luhn, 1957). Για κάθε όρο t_i του λεξικού $\mathcal{V}$, έστω f_{t_i} η συχνότητα του όρου t_i στο έγγραφο d. Το διάνυσμα δίνεται από τη σχέση $x_i = f_{t_i}$.

Για παράδειγμα, εάν το σύνολο απαρτίζεται από τα δύο έγγραφα: d_1 = «ο καιρός είναι καλός» και d_2 = «ο καιρός είναι βροχερός», τότε το λεξικό που προκύπτει είναι V = {ο, καιρός, είναι, καλός, βροχερός} και συνεπώς τα διανύσματα των εγγράφων είναι $\vec{X}_1|TO$ = $\vec{X}_1|TF$ = $[1,1,1,1,0]$ και $\vec{X}_2|TO$ = $\vec{X}_2|TF$ = $[1,1,1,0,1]$. Επιπλέον, αν ένα νέο έγγραφο εμφανιστεί d_3 = «ο καλός καιρός είναι ο βροχερός καιρός», τότε το λεξικό παραμένει ίδιο αλλά $\vec{X}_3|TO = [1,1,1,1,1]$ ενώ $\vec{X}_3|TF = [2,2,2,1,1]$.
Φυσικά τα διανύσματα που προκύπτουν είναι ιδιαίτερα αραιά (sparse vectors), ειδικά για σύνολα που περιέχουν ένα μεγάλο λεξικό και αυτό είναι ένα πρόβλημα της τεχνικής Bag of Words.

3.2.2 TF-IDF

Το μεγαλύτερο εμπόδιο στην επεξεργασία μικρών κειμένων είναι η αντιμετώπιση της αραιότητας (sparsity) των λέξεων που περιέχουν, αφού τα μικρά κείμενα συνήθως περιλαμβάνουν μικρό αριθμό λέξεων, και επομένως οι λέξεις εμφανίζονται σπάνια ή σε μικρές συχνότητες. Για την αντιμετώπιση αυτού του προβλήματος, χρησιμοποιείται ο αλγόριθμος TF-IDF που υπολογίζει ένα σκορ για κάθε λέξη, λαμβάνοντας υπόψη τόσο τη συχνότητα εμφάνισής μίας λέξης σε ένα από τα έγγραφα (Term Frequency - TF), όσο και τη συχνότητα εμφάνισής της σε όλα τα κείμενα του συνόλου των δεδομένων (Inverse Document Frequency IDF) (Sparck Jones, 1972). Έτσι, ευνοεί τις λέξεις με μεγάλη συχνότητα σε ένα έγγραφο, αλλά ταυτόχρονα σπανίζουν σε άλλα έγγραφα του συνόλου δεδομένων.
Πιο αναλυτικά, το TF-IDF σκορ υπολογίζεται ως:

$$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \times \text{IDF}(t, D) \tag{3.2}$$

$$\text{όπου TF}(t, d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \tag{3.3}$$

$$\text{και IDF}(t, D) = \log\left(\frac{N}{|\{d \in D : t \in d\}|}\right) \tag{3.4}$$

- $f_{t,d}$: συχνότητα όρου t στο έγγραφο,
- $\sum_{t' \in d} f_{t',d}$: συνολικός αριθμός συχνοτήτων όρων στο έγγραφο,
- N: αριθμός εγγράφων στο σύνολο των μεταδεδομένων,
- $|\{d \in D : t \in d\}|$: αριθμός εγγράφων όπου εμφανίζεται ο όρος t.

Το διάνυσμα $\vec{X}_d$ για κάθε έγγραφο d λαμβάνει τις τιμές των TF-IDF σκορ του κάθε όρου μέσα στο λεξικό V όλων των μοναδικών όρων. Έτσι βλέπουμε πως το διάνυσμα μπορεί και πάλι να είναι πολύ μεγάλων διαστάσεων και για αυτό το σκοπό συχνά αποκόπτεται, κρατώντας μόνο τις 1000 πιο συχνές λέξεις του λεξικού.

3.2.3 Επέκταση με N-grams - Περιορισμοί

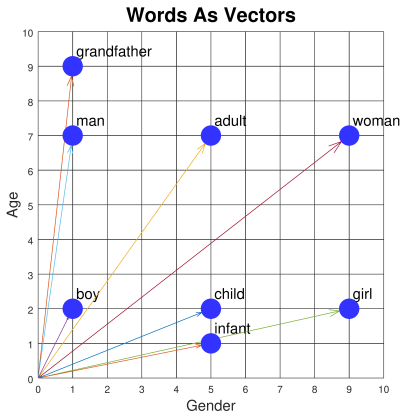
Το μοντέλο Bag of Words αγνοεί τη σειρά των tokens με αποτέλεσμα προτάσεις που έχουν διαφορετικό νόημα, να ερμηνεύονται με τον ίδιο τρόπο, για παράδειγμα, η πρόταση «το λιοντάρι κυνηγάει το βουβάλι» είναι διαφορετική νοηματικά από την πρόταση «το βουβάλι κυνηγάει το λιοντάρι» ωστόσο, οι προτάσεις έχουν το ίδιο διάνυσμα bag-of-words, αφού οι λέξεις που περιέχουν είναι οι ίδιες.

Μία επέκταση που μπορεί να γίνει για τη μερική αντιμετώπιση του προβλήματος αυτού, είναι η χρήση των N-grams, ακολουθίες από n λέξεις που προστίθενται στο λεξικό των κειμένων, συμπεριλαμβάνοντας περισσότερη πληροφορία από απλά tokens.

Μία πρόταση θα μπορούσε να διασπαστεί σε uni-grams (όπως είδαμε στο παράδειγμα) αλλά και σε bi-grams (ακολουθίες δύο λέξεων) ή tri-grams (ακολουθίες τριών λέξεων). Για το προηγούμενο παράδειγμα «Αυτή η πρόταση, είναι ένα παράδειγμα.», παρατηρούμε πως εάν συμπεριληφθούν τα bi-grams και τα tri-grams πέρα από τα uni-grams, το λεξικό V επεκτείνεται, και είναι V = {«Αυτή», «η», «πρόταση», «,», «είναι», «ένα», «παράδειγμα», «.», «Αυτή η», «η πρόταση», «πρόταση,», «,είναι», «είναι ένα», «ένα παράδειγμα», «παράδειγμα.», «Αυτή η πρόταση», «ή πρόταση,», «πρόταση, είναι», «, είναι ένα», «είναι ένα παράδειγμα», «ένα παράδειγμα.»}.
Η χρήση των N-grams συμβάλει στη συμπερίληψη των συμφραζομένων, αλλά ταυτοχρόνως επεκτείνει σημαντικά το λεξικό, αραιώνοντας ακόμα περισσότερο τα διανύσματα αναπαράστασης.

3.3 Word Vectors - Word Embeddings

Μια διαφορετική προσέγγιση στο πρόβλημα της διερμηνείας των προτάσεων είναι η αναπαράσταση κάθε token ξεχωριστά στο χώρο, με ένα διάνυσμα μικρότερου μεγέθους και πυκνότητας (dense vector). Ο σκοπός είναι tokens με παρόμοιο νόημα, να βρίσκονται κοντά στο χώρο, ενώ άλλα που δεν έχουν μεγάλη σχέση μεταξύ τους να βρίσκονται σε μεγαλύτερη απόσταση.



Σχήμα 3.2: *Παράδειγμα Word Vectors με διάνυσμα δύο συνιστωσών (URL, n.d.-g).*

Στο παράδειγμα της εικόνας 3.2, παρατηρούμε έναν απλό διανυσματικό χώρο δύο διαστάσεων, που ορίζεται από την ηλικία και το φύλλο, με την απεικόνιση κάποιων λέξεων με διανύσματα λέξεων. Η λέξη «παππούς» για παράδειγμα έχει μεγάλη ηλικία και αρσενικό φύλλο, ενώ αντίθετα η λέξη «κορίτσι» έχει θηλυκό γένος και μικρή ηλικία. Όπως είναι αναμενόμενο, αυτές οι λέξεις βρίσκονται αντιδιαμετρικά στο διανυσματικό χώρο. Επιπλέον, παρατηρούμε πως λέξεις όπως «ενήλικας», «παιδί» και «βρέφος» βρίσκονται στη μεσοκάθετη του ευθύγραμμου τμήματος που ορίζουν τα δύο γένη, αφού έχουν ουδέτερο γένος και διαφορετικές ηλικίες. Τέλος, μία ιδιαίτερα σημαντική διαπίστωση είναι πως καθίσταται δυνατό

να υπολογιστεί η παρακάτω αλγεβρική πράξη στα διανύσματα:

$$\upsilon_{\vec{\text{γυναίκα}}} - (\upsilon_{\vec{\text{άντρας}}} - \upsilon_{\vec{\text{αγόρι}}}) \approx \upsilon_{\vec{\text{κορίτσι}}} \tag{3.5}$$

Αυτή η παρατήρηση έγινε και από την ομάδα του Mikolov (Mikolov, Chen, Corrado, & Dean, 2013) παρουσιάζοντας το αποτέλεσμα της πράξης $\upsilon_{\vec{\text{Βασιλιάς}}} - \upsilon_{\vec{\text{Αντρας}}} + \upsilon_{\vec{\text{Γυναίκα}}} \approx \upsilon_{\vec{\text{Βασίλισσα}}}$, αποδεικνύοντας πως με τη σωστή εκπαίδευση των διανυσμάτων λέξεων, η αναπαράσταση ξεπερνάει τον υπολογισμό των απλών συντακτικών ομοιοτήτων, αποκωδικοποιώντας σχέσεις και σημασιολογικές ιδιότητες της φυσικής γλώσσας. Επιπλέον παρουσίασαν και άλλες σχέσεις που εντοπίστηκαν, ένα τμήμα των οποίων φαίνονται στον πίνακα 3.2, ζητώντας για παράδειγμα από το μοντέλο να αναφέρει την πιο κοντινή λέξη που αντιστοιχεί στη λέξη Ιταλία όταν του παρουσιάζεται η σχέση Γαλλία και Παρίσι, υπολογίζοντας δηλαδή το αποτέλεσμα της αλγεβρικής πράξης $\upsilon_{\vec{\text{Γαλλία}}} - \upsilon_{\vec{\text{Παρίσι}}} + \upsilon_{\vec{\text{Ιταλία}}}$.

Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium

Πίνακας 3.2: *Παραδείγματα Σχέσεων μεταξύ λέξεων, εντοπισμένα με αλγεβρικές πράξεις πάνω σε διανύσματα λέξεων.*

Η αναπαράσταση των λέξεων με διανύσματα, επιτρέπει την αναπαράσταση των tokens στο χώρο, και για την επίτευξη αυτού του σκοπού, χρησιμοποιούνται ευρέως, οι παρακάτω αλγόριθμοι.

3.3.1 Αλγόριθμος Word2Vec

Ο αλγόριθμος Word2Vec είναι μια τεχνική για την αναπαράσταση λέξεων σε ένα διανυσματικό χώρο χαμηλής διάστασης, που διατηρεί τη σημασιολογική τους πληροφορία. Δημοσιεύτηκε το 2013 από την ομάδα του Tomáš Mikolov (Mikolov et al., 2013) στην εταιρεία Google και υλοποιείται με δύο εκδοχές, με το μοντέλο Continuous Bag of Words - CBOW και με το μοντέλο Skip-Gram.

Εκδοχή Continuous Bag of Words - CBOW

Στο μοντέλο CBOW, για να υπολογιστούν τα διανύσματα λέξεων μιας πρότασης, λύνεται αρχικά το πρόβλημα πρόβλεψης μία λέξης με βάση τις $2m$ λέξεις γύρω από αυτή όπου m είναι το παράθυρο (window) των συμφραζομένων.

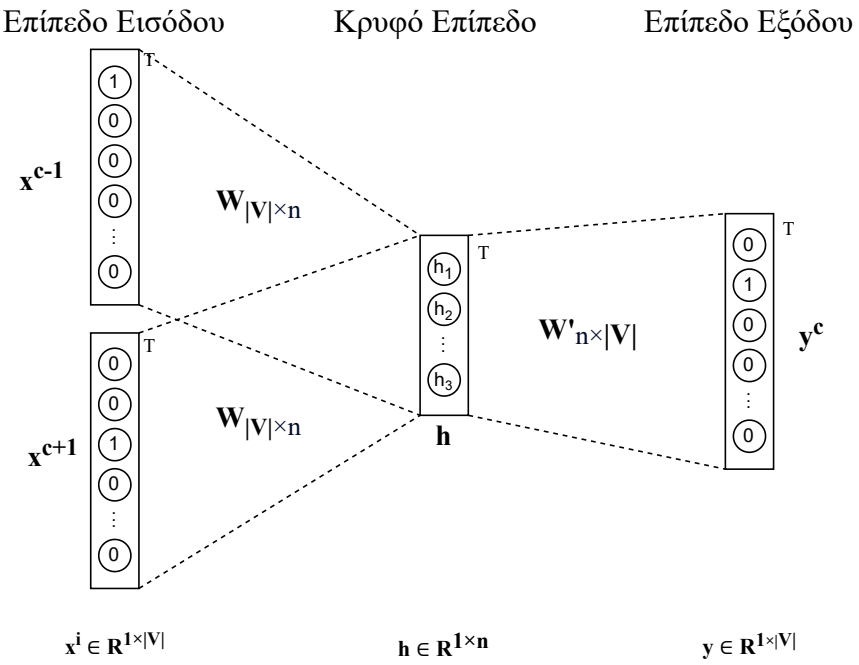
Το μοντέλο παίρνει ως είσοδο τα one-hot encodings των γύρω λέξεων που έχουν τη μορφή $\mathbf{x} \in \mathbb{R}^{1\times|V|}$ όπου με $|V|$ συμβολίζεται ο πληθικός αριθμός (cardinality) του συνόλου του λεξικού

όλων των λέξεων του corpus, και έχει μία και μόνο έξοδο, τη one-hot encoding μορφή της λέξης που θέλουμε να προβλέψουμε και ονομάζεται «κεντρική» λέξη.
Περισσότερα για την τεχνική κωδικοποίησης one-hot encoding και άλλες μεθόδους κωδικοποίησης για κατηγορικά δεδομένα βρίσκονται στο παράρτημα Α΄.

Συγκεκριμένα η είσοδος αποτελείται από τα εξής διανύσματα: $x^{(c-m)}, \ldots, x^{(c-1)}, x^{(c+1)}, .., x^{(c+m)}$ όπου c είναι ο δείκτης της κεντρικής λέξης που εξετάζεται με βάση τις γειτονικές.

Η έξοδος y^c του μοντέλου αποτελείται από το one-hot encoding της «κεντρικής» λέξης που εξετάζεται και θέλουμε το μοντέλο να προβλέψει.

Οι πίνακες $W_{|V|\times n}$ και $W'_{n\times|V|}$ είναι οι πίνακες που περιέχουν τα βάρη των νευρώνων και ανανεώνονται κατά την εκπαίδευση του μοντέλου. Στο τέλος τις εκπαίδευσης, αυτοί οι πίνακες, περιέχουν στην i-οστή γραμμή του πίνακα W και i-οστή στήλη του πίνακα W' αντίστοιχα, τα νέα διανύσματα που αντιστοιχούν στην i-οστή λέξη του λεξικού V και έχουν διάσταση $1\times n$.



Σχήμα 3.3: *Το νευρωνικό δίκτυο του μοντέλου CBOW.*

Όπως βλέπουμε στο σχήμα 3.3, το μοντέλο πρόκειται για ένα απλό νευρωνικό δίκτυο ενός κρυμμένου επιπέδου n κόμβων όπου n είναι η επιθυμητή διάσταση των διανυσμάτων λέξεων στο νέο διανυσματικό πυκνό χώρο. Οι ανάστροφοι των πινάκων έχουν χρησιμοποιηθεί για την καλύτερη οπτικοποίηση.

Οι είσοδοι πολλαπλασιάζονται με τον πίνακα βαρών W (εσωτερικό γινόμενο). Στη συνέχεια λαμβάνεται ο μέσος όρος αυτών και πολλαπλασιάζοντας με τον πίνακα βαρών W', υπολο-

γίζεται το αποτέλεσμα της συνάρτησης softmax που επιλέγεται ως συνάρτηση ενεργοποίησης (activation function) για να βρεθεί η πιθανότητα της κεντρικής λέξης.

Συγκεκριμένα ο υπολογισμός της συνάρτησης softmax για ένα διάνυσμα εισόδου $\mathbf{z} = [z_1, z_2, ..., z_{|V|}]$ είναι:

$$\text{softmax}(\mathbf{z}) = \left[\frac{\exp(z_1)}{\sum_{i=1}^{|V|} \exp(z_i)}, \frac{\exp(z_2)}{\sum_{i=1}^{|V|} \exp(z_i)}, ..., \frac{\exp(z_K)}{\sum_{i=1}^{|V|} \exp(z_i)} \right] \tag{3.6}$$

Η εκθετική συνάρτηση χρησιμοποιείται για να έχουμε θετικά αποτελέσματα, ενώ η διαίρεση με το άθροισμα για να κανονικοποιηθεί το διάνυσμα για να βρεθεί η πιθανότητα.

Στη συνέχεια, υπολογίζεται η συνάρτηση λάθους (loss function) με τη συνάρτηση cross entropy, η οποία υπολογίζεται ως:

$$H(y, \hat{y}) = -\sum_{j=1}^{|V|} y_j \log(\hat{y}_j) \tag{3.7}$$

όπου y είναι το one-hot encoding διάνυσμα της «κεντρικής» λέξης, και $\hat{y}$ είναι η έξοδος του μοντέλου. Για παράδειγμα, σε περίπτωση που το μοντέλο έχει επιτύχει πλήρως στην πρόβλεψη της κεντρικής λέξης, τότε $y = \hat{y}$ και αφού πρόκειται για δύο ίδια one-hot encodings, περιέχουν παντού την τιμή 0 εκτός από το δείκτη της κεντρικής λέξης όπου περιέχουν την τιμή 1. Συνεπώς για μία τέλεια πρόβλεψη έχουμε $H(1, 1) = -1 \log(1) = 0$. Αντιθέτως για μία λανθασμένη πρόβλεψη $\hat{y} = 0.1$, το αποτέλεσμα θα ήταν $H(1, 0.1) = -1 \log(0.1) \approx 2.30$.

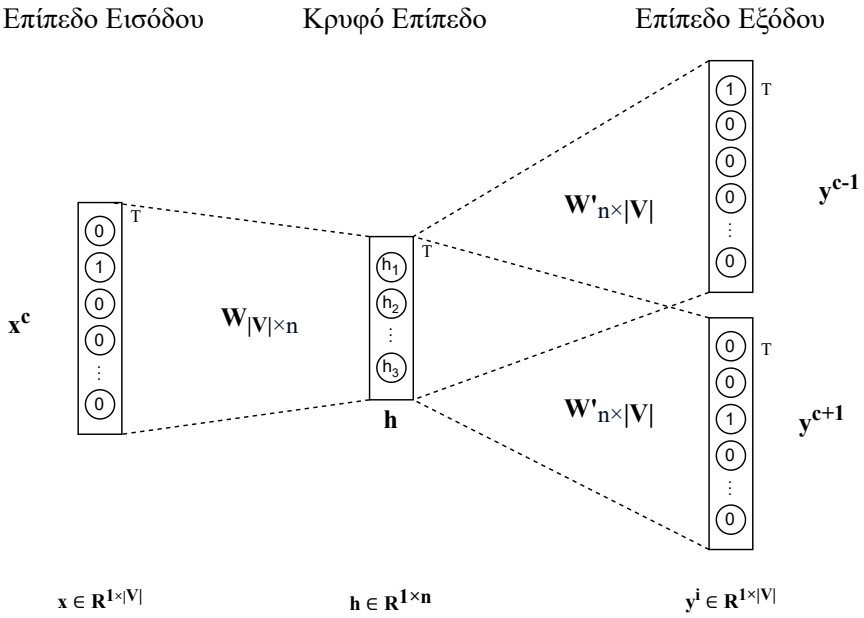
Για την ανεύρεση του ελαχίστου της συνάρτησης λάθους, ο αλγόριθμος CBOW χρησιμοποιεί την τεχνική της Στοχαστικής Κατάβασης Κλίσης, (Stochastic Gradient Descent - SGD) η οποία μελετήθηκε στο κεφάλαιο εκπαίδευσης των βαθιών νευρωνικόν δικτύων 2.2.2.

Εκδοχή Skip-Gram

Το μοντέλο Skip-Gram λειτουργεί αντίστροφα από το CBOW. Αντί να προβλέπει την «κεντρική» λέξη με βάση τις γύρω λέξεις, έχει στόχο να προβλέψει τις γύρω λέξεις με βάση την «κεντρική» λέξη. Δηλαδή, για μια πρόταση με λέξεις $w_1, w_2, ..., w_n$, το Skip-Gram προβλέπει τις λέξεις $w_{c-m}, w_{c-m+1}, ..., w_{c-1}, w_{c+1}, ..., w_{c+m}$ με βάση την τρέχουσα «κεντρική» λέξη w_c και το παράθυρο m.
Ο τρόπος λειτουργίας του Skip-Gram είναι ανάλογος με αυτόν του CBOW, χρησιμοποιώντας τη συνάρτηση softmax για τον υπολογισμό των πιθανοτήτων, τη συνάρτηση κόστους της εντροπίας και την τεχνική SGD για την ενημέρωση των πινάκων.

Επιλογή Μεταξύ CBOW και Skip-Gram

Η επιλογή μεταξύ των δύο εκδοχών του αλγορίθμου Word2Vec, πραγματοποιείται με βάση τις ανάγκες του προβλήματος.
Το Skip-Gram επιτυγχάνει καλή απόδοση με μικρές ποσότητες δεδομένων επειδή εστιάζει στην πρόβλεψη των γειτονικών λέξεων για μια κεντρική λέξη. Το γεγονός αυτό καθιστά



Σχήμα 3.4: *Το νευρωνικό δίκτυο του μοντέλου Skip-Gram.*

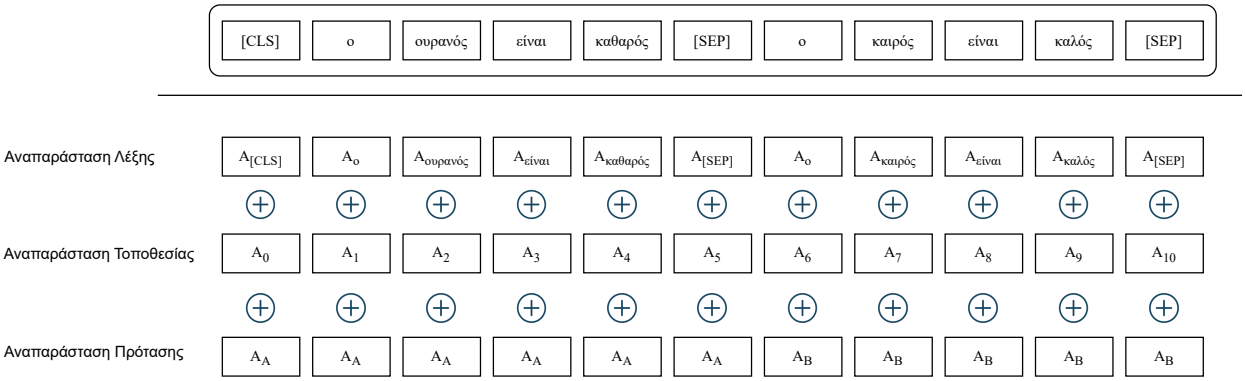
το Skip-Gram καλή επιλογή όταν ο όγκος των δεδομένων είναι μικρός, καθώς μπορεί να αντλήσει αρκετή πληροφορία από λίγα παραδείγματα. Επίσης, τα διανύσματα λέξεων που παράγονται από το Skip-Gram συχνά αποτυπώνουν καλύτερα τις σημασιολογικές σχέσεις μεταξύ των λέξεων, όπως συνώνυμα ή συγγενικές έννοιες.
Αντιθέτως, το CBOW είναι ταχύτερο στην εκπαίδευση επειδή προβλέπει την κεντρική λέξη δεδομένων των γειτονικών λέξεων. Αυτό το καθιστά καλή επιλογή για μεγάλα σύνολα δεδομένων ή για προβλήματα όπου η ταχύτητα είναι κρίσιμη. Ωστόσο, η απόδοση του CBOW μπορεί να μην είναι τόσο υψηλή για λέξεις με χαμηλή συχνότητα ή σε περιπτώσεις όπου η σημασιολογική ομοιότητα των λέξεων είναι πιο σημαντική.

3.3.2 Ανάλυση Συμφραζομένων και Μετασχηματιστές

Bidirectional Encoder Representations from Transformers - BERT

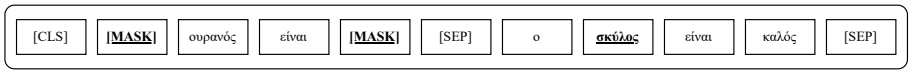
Βασισμένο στην τεχνολογία των μετασχηματιστών, το μοντέλο BERT (Devlin, Chang, Lee, & Toutanova, 2018) είναι ένα ιδιαίτερα χρήσιμο μοντέλο για την παραγωγή διανυσμάτων λέξεων. Σε αντίθεση με την τεχνική Word2Vec που είδαμε στην προηγούμενη υποενότητα, χρησιμοποιεί τα συμφραζόμενα μιας λέξης με πολύ διαφορετικό τρόπο. Για παράδειγμα η λέξη «τόξο» στις δύο προτάσεις Π_1: «το ουράνιο τόξο είναι όμορφο» και Π_2: «ο πολεμιστής άρπαξε το τόξο, τέντωσε αμέσως τη χορδή κι έριξε ένα βέλος» θα είχε το ίδιο διάνυσμα αν αυτό παραγόταν από τον αλγόριθμο Word2Vec ή από το μοντέλο Bag of Words. Η τεχνική του BERT όμως, έχει σκοπό να παράγει διαφορετικά διανύσματα λέξεων, ανάλογα με τις κοντινές τους λέξεις χρησιμοποιώντας δηλαδή δυναμικά το context και όχι στατικά όπως οι μέθοδοι που έχουν μελετηθεί ως αυτό το σημείο.
Η αρχιτεκτονική του μοντέλου BERT είναι ένας «αμφίδρομος» κωδικοποιητής (biderectional encoder) επειδή σε αντιδιαστολή με τους αρχικούς μετασχηματιστές, οι Devlin et al. (2018),

υποστηρίζουν πως η μέθοδος της μονόδρομης αυτο-προσοχής όπου μόνο οι προηγούμενες λέξεις στο κείμενο λαμβάνονται υπόψη, περιορίζει ιδιαίτερα τις ικανότητες του μοντέλου. Για να επιτευχθεί αυτό, προτείνουν να υπάρχει ως είσοδος στο μοντέλο πέρα από την αναπαράσταση της λέξης, η τοποθεσία αυτής (ο δείκτης της θέσης της λέξης στο κείμενο) καθώς και ο δείκτης της πρότασης στην οποία ανήκει.



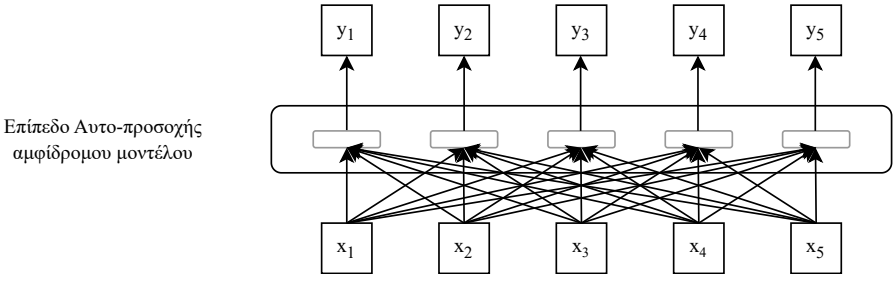
Σχήμα 3.5: *Η είσοδος του BERT μοντέλου για ένα κείμενο που αποτελείται από δύο προτάσεις* Π_1*: «ο ουρανός είναι καθαρός»,* Π_2*: «ο καιρός είναι καλός» ως το άθροισμα των αναπαραστάσεων των λέξεων, της θέσης και της πρότασης στην οποία ανήκουν.*

Επιπλέον, προτείνουν έναν νέο τρόπο εκπαίδευσης, το Masked Language Model, το οποίο «κρύβει» τυχαία μία λέξη του κειμένου και έχει ως σκοπό την πρόβλεψή της χρησιμοποιώντας μόνο τα συμφραζόμενα της και συνεπώς κωδικοποιεί μία αναπαράσταση αυτής, λαμβάνοντας υπόψη τόσο τα εκ του αριστερού συμφραζόμενα, όσο τα εκ του δεξιού.



Σχήμα 3.6: *15% των λέξεων του κειμένου επιλέγονται ως στόχοι της πρόβλεψης. Από αυτές, 80% αντικαθίσταται με τη λέξη [MASK], 10% αντικαθίσταται με μία τυχαία λέξη, και 10% μένει ως έχει.*

Το μοντέλο BERT χρησιμοποιεί ένα αμφίδρομο επίπεδο αυτο-προσοχής όπου σε αντίθεση με το 2.4 των αρχικών μετασχηματιστών που δε λαμβάνει υπόψη τις εκ του δεξιού λέξης ενός κειμένου, χρησιμοποιεί το σύνολο της εισόδου όπως είναι ορατό στο σχήμα 3.7.

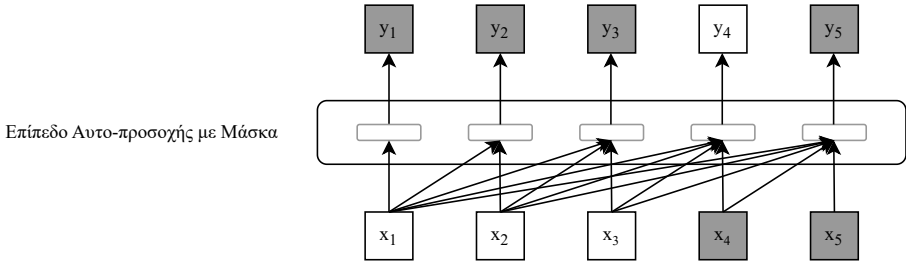


Σχήμα 3.7: *Αμφίδρομη αρχιτεκτονική στο επίπεδο αυτο-προσοχής.*

Generative Pre-trained Transformer 2 - GPT-2

Το 2019, αναπτύχθηκε και δημοσιεύτηκε το μοντέλο Generative Pre-trained Transformer 2 (GPT-2) (Radford et al., 2019) από την εταιρεία OpenAI που επίσης βασίζεται στην αρχιτεκτονική των μετασχηματιστών.
Σε αντίθεση με το BERT, η αρχιτεκτονική του GPT-2 δεν βασίζεται σε κωδικοποιητή αλλά σε αποκωδικοποιητή, με σκοπό να αποκωδικοποιεί το κείμενο και να λύνει το πρόβλημα της πρόβλεψης της επόμενης λέξης μίας πρότασης, χρησιμοποιώντας μόνο όσες λέξεις έχουν παρέλθει.

Η αρχιτεκτονική του GPT-2 δεν του επιτρέπει να γνωρίζει τα συμφραζόμενα που ακολουθούν τη λέξη την οποία μελετάει αλλά μόνο αυτά που έχουν παρέλθει και σε αντίθεση με το μοντέλο BERT, ο σκοπός της εκπαίδευσης του είναι η πρόβλεψη της ακριβώς επόμενης λέξης, και όχι μίας τυχαίας λέξης του κειμένου.



Σχήμα 3.8: *Επίπεδο αυτο-προσοχής με Μάσκα.*

Για να επιτευχθεί αυτό, στο μοντέλο χρησιμοποιείται η masked αυτο-προσοχή, όπου από το σχήμα 2.4 περνάμε στο σχήμα 3.8, και παρατηρούμε πως οι είσοδοι x_4 και x_5 δεν λαμβάνονται υπόψη, και πως η έξοδος του μοντέλου είναι η τιμή του y_4. Όπως και για τα προηγούμενα μοντέλα, μπορούμε να αντλήσουμε τα βάρη των διανυσμάτων μετά την εκπαίδευση, με σκοπό να λάβουμε τα διανύσματα λέξεων.

Οι ερευνητές της εταιρείας OpenAI εκπαιδεύουν το μοντέλο GPT-2 χρησιμοποιώντας το σύνολο δεδομένων WebText, το οποίο δημιούργησαν συλλέγοντας όλα τα κείμενα που εμπεριέχονται σε ιστοσελίδες που αναφέρονται στο μέσο κοινωνικής δικτύωσης Reddit (URL, n.d.-f) , και έχουν ψηφιστεί από τουλάχιστον τρεις χρήστες ως ενδιαφέρουσες. Το αποτέλεσμα είναι να συλλεχθούν 45 εκατομμύρια ιστοσελίδες που έχουν προταθεί από ανθρώπους.

Χρησιμοποιώντας το WebText, εκπαιδεύουν το μοντέλο GPT-2 σε τέσσερις διαφορετικές εκδοχές. Από την εκδοχή GPT-2 small με 117 εκατομμύρια παραμέτρους, 12 μπλοκ αποκωδικοποιητή-μετασχηματιστή και μέγεθος διανύσματος λέξης κάθε token τις 768 διαστάσεις, μέχρι το GPT-2 extra-large με 1,542 δισεκατομμύρια παραμέτρους, 48 μπλοκ αποκωδικοποιητή-μετασχηματιστή και μέγεθος διανύσματος λέξης κάθε token τις 1600 διαστάσεις. Τα προ-εκπαιδευμένα μοντέλα διατίθενται από την OpenAI στην ιστοσελίδα (URL,

n.d.-c).

3.4 Από Διανύσματα Λέξεων σε Διανύσματα Προτάσεων

Με τις μεθόδους Word2Vec, BERT και GPT-2, μπορούν να σχηματιστούν διανύσματα λέξεων. Ωστόσο, ο σκοπός της παρούσας εργασίας είναι η ερμηνεία των προτάσεων - εγγράφων του συνόλου των μεταδεδομένων των χρονοσειρών και όχι μεμονωμένων λέξεων. Για το λόγο αυτό, σε αυτή την υποενότητα, ερευνώνται τρόποι μετατροπής των διανυσμάτων λέξεων σε διανύσματα προτάσεων, καθώς και κάποιες τροποποιήσεις των μοντέλων που αναλύθηκαν, με σκοπό την απευθείας παραγωγή διανυσμάτων προτάσεων.

3.4.1 Τελεστές Μετατροπής

Η μέθοδος της πρόσθεσης είναι μια από τις απλούστερες προσεγγίσεις για τη μετατροπή διανυσμάτων λέξεων σε διανύσματα προτάσεων. Συγκεκριμένα, για μια πρόταση που αποτελείται από λέξεις με διανύσματα w_i, το διάνυσμα πρότασης $\vec{x_d}$ υπολογίζεται ως εξής:

$$\vec{x_d} = \sum_{i=1}^{N} w_i \tag{3.8}$$

κρατώντας τον ίδιο αριθμό διαστάσεων με την αρχική. Εύκολα μπορεί να παρατηρηθεί όμως πως μία πρόταση με περισσότερες λέξεις από μία άλλη, είναι πιθανό να έχει υψηλότερες τιμές και να απομακρυνθεί (στο χώρο) από μία άλλη χωρίς όμως να είναι απαραίτητα νοηματικά διαφορετικές. Για το σκοπό αυτό, η ομάδα του Arora (Arora, Liang, & Ma, 2017) από το Πανεπιστήμιο του Princeton υποστηρίζει πως μπορεί να ληφθεί η μέση τιμή των διανυσμάτων με τον υπολογισμό ενός απλού μέσου όρου

$$\vec{x_d} = \frac{1}{N} \sum_{i=1}^{N} w_i \tag{3.9}$$

Ακόμα, μπορεί να υπολογιστεί ο πολλαπλασιασμός μεταξύ των διανυσμάτων, τεχνική η οποία χρησιμοποιείται και από τη δημοσίευση της Google (Cer et al., 2018) και αναμένεται να δουλεύει καλύτερα.

$$\vec{x_d} = \frac{1}{N} \left(\prod_{i=1}^{N} w_i \right) \tag{3.10}$$

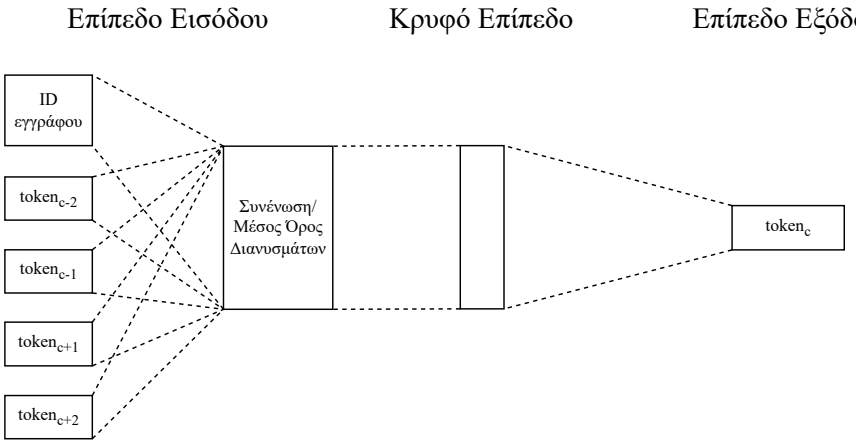
3.4.2 Μοντέλα Παραγωγής Διανυσμάτων Προτάσεων

Όπως ο αλγόριθμος TF-IDF ο οποίος παράγει απευθείας διανύσματα προτάσεων και όχι λέξεων, παρουσιάζονται ο αλγόριθμος Doc2Vec βασισμένος στον Word2Vec καθώς και ο Sentence-BERT που βασίζεται στο μοντέλο BERT.

Doc2Vec

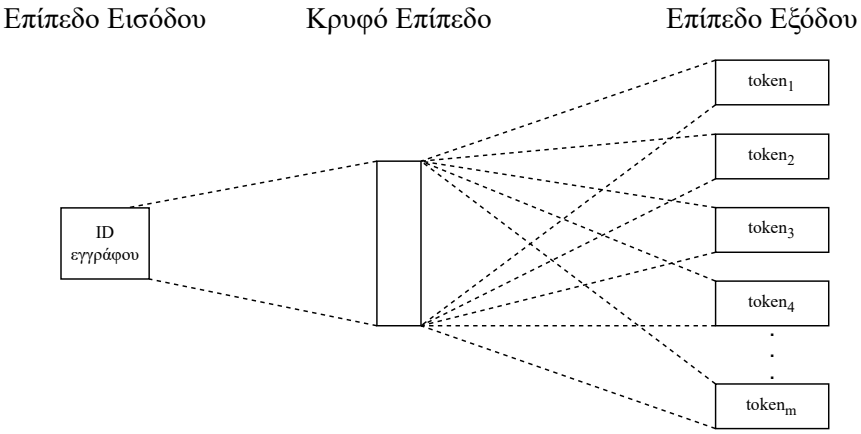
Ένα χρόνο μετά τη δημοσίευση του Word2Vec, η ομάδα της Google του Mikolov παρουσιάζει το 2014 το μοντέλο Paragraph Vector με την κοινότητα να χρησιμοποιεί το ψευδώνυμο

Doc2Vec (Le & Mikolov, 2014). Το μοντέλο Doc2Vec με ανάλογο τρόπο με το μοντέλο Word2Vec αποτελείται από δύο εκδοχές, τον αλγόριθμο Distributed Bag of Words - DBOW παραλλαγή του CBOW και τον αλγόριθμο Distributed Memory - DM ανάλογο του Skip-Gram, μέθοδοι που αναλύθηκαν στο 3.3.1. Συγκεκριμένα, για ένα παράθυρο m γύρω λέξεων, στην εκδοχή DM, ο κωδικός του εγγράφου λαμβάνεται ως είσοδος μαζί με τις m λέξεις του παραθύρου, και ο στόχος είναι η πρόβλεψη της κεντρικής λέξεις (χρησιμοποιώντας δηλαδή τις γύρω λέξεις και ταυτόχρονα την πληροφορία σε ποιο έγγραφο/παράγραφο βρίσκεται η λέξη που πρέπει να προβλεφθεί όπως φαίνεται στην αρχιτεκτονική του 3.9).



Σχήμα 3.9: *Αρχιτεκτονική του μοντέλου DM, όπου τα tokens και οι παράγραφοι αναπαρίστανται με διανύσματα.*

Εν αντιθέσει, στην εκδοχή DBOW 3.10 χρησιμοποιείται μόνο ο κωδικός του εγγράφου με σκοπό την πρόβλεψη των μ λέξεων του παραθύρου.



Σχήμα 3.10: *Αρχιτεκτονική του μοντέλου DBOW, όπου τα tokens και οι παράγραφοι αναπαρίστανται με διανύσματα.*

Sentence-BERT

Το Sentence-BERT είναι μια επέκταση του βασικού BERT (Bidirectional Encoder Representations from Transformers) μοντέλου, για τη δημιουργία διανυσμάτων προτάσεων που μπορούν να χρησιμοποιηθούν σε προβλήματα σύγκρισης και κατάταξης προτάσεων. Δημοσιεύτηκε από τους Reimers και Gurevych (2019) για να αντιμετωπίσει το πρόβλημα της υπολογιστικής αποδοτικότητας και της ποιότητας στις αναπαραστάσεις προτάσεων, το οποίο δεν μπορούσε να επιλυθεί αποτελεσματικά με το αρχικό BERT μοντέλο (Reimers & Gurevych, 2019).

Για την εκπαίδευση του Sentence-BERT, υπολογίζονται αρχικά τα διανύσματα των λέξεων που περιέχει η πρόταση με τη μέθοδο BERT και στη συνέχεια με τη μέθοδο pooling που ομαδοποιεί τα διανύσματα αυτά. Στη συνέχεια εξετάζονται ζεύγη προτάσεων που είναι είτε «θετικά» (σημασιολογικά παρόμοια) είτε αρνητικά (σημασιολογικά διαφορετικά). Το μοντέλο εκπαιδεύεται ώστε να ελαχιστοποιηθεί η συνάρτηση απώλειας για τα θετικά ζεύγη, δηλαδή να μειώνει την απόσταση μεταξύ των διανυσμάτων προτάσεων που είναι σημασιολογικά παρόμοιες, και να μεγιστοποιεί την απόσταση μεταξύ των διανυσμάτων προτάσεων που είναι σημασιολογικά διαφορετικές.

Κεφάλαιο 4

Ταξινόμηση και Συσταδοποίηση Προτάσεων σε μορφή διανύσματος

Έχοντας αναπαραστήσει τα μεταδεδομένα από μορφή κειμένου σε μορφή διανυσμάτων με διαφορετικές μεθόδους, ο σκοπός είναι αφ' ενός να επιτευχθεί η καλύτερη δυνατή ομαδοποίηση αυτών, με απώτερο στόχο τη βέλτιστη πρόβλεψη των δεδομένων, αφ' ετέρου να επαληθευτεί με τις κατάλληλες μεθόδους αξιολόγησης η ομαδοποίηση και η ταξινόμηση τους.

Σε αυτό το κεφάλαιο θα μελετηθούν μέθοδοι ταξινόμησης και συσταδοποίησης των διανυσμάτων προτάσεων που χρησιμοποιήθηκαν, καθώς και μετρικές αξιολόγησης αυτών των μεθόδων.

4.1 Ταξινόμηση Προτάσεων (Sentence Classification)

Η ταξινόμηση είναι μία επιβλεπόμενη μέθοδος μηχανικής μάθησης (supervised learning), κατά την οποία ταξινομούνται δεδομένα σε κλάσεις με βάση τα χαρακτηριστικά τους. Χρησιμοποιούνται δεδομένα εκπαίδευσης για τα οποία ξέρουμε την ετικέτα τους (label), με σκοπό το μοντέλο να μάθει να ταξινομεί δεδομένα που δεν έχει ξαναδεί. Σε αυτή την ενότητα αναλύονται οι κύριες μέθοδοι ταξινόμησης.

4.1.1 k-Κοντινότεροι Γείτονες (k-Nearest Neighbors)

Ο αλγόριθμος k-Nearest Neighbors (k-NN) (Cover & Hart, 1967) βασίζει την ταξινόμηση ενός νέου σημείο «δείγματος», στην ταξινόμηση των k κοντινότερων γειτόνων του, από το σύνολο εκπαίδευσης. Για τον υπολογισμό της απόστασης αυτής, συχνά χρησιμοποιείται η μετρική της Ευκλείδειας απόστασης, όπως φαίνεται στον παρακάτω ψευδοκώδικα.
Ο αλγόριθμος χρησιμοποιεί τα δεδομένα εκπαίδευσης για να προβλέψει την κατηγορία στην οποία ανήκει ένα νέο σημείο που δεν έχει ξαναδεί, βασίζοντας την απόφαση αυτή στο πόσο κοντά είναι στον n-διάστατο χώρο με άλλα σημεία.

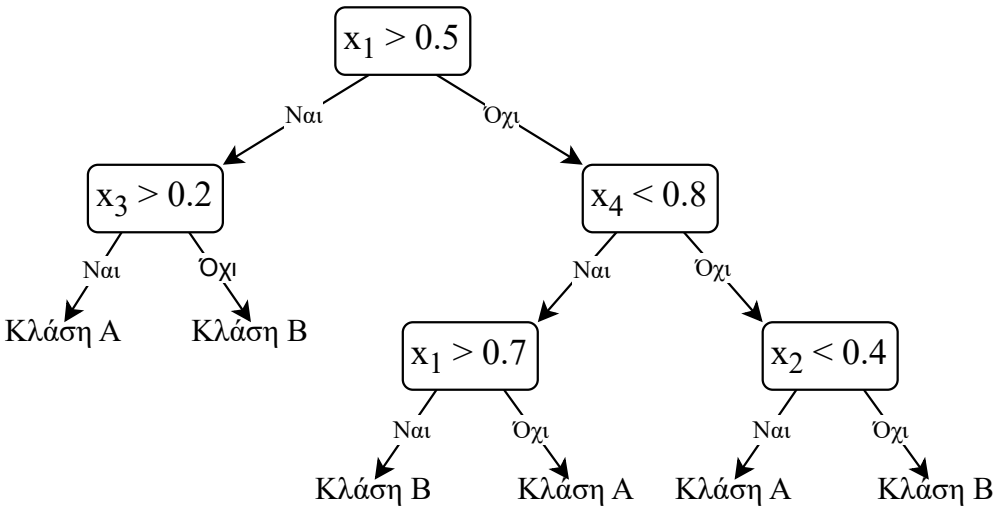
4.1.2 Δέντρα Αποφάσεων και Τυχαίο Δάσος (Random Forest)

Τα δέντρα αποφάσεων είναι μία δημοφιλής μέθοδος ταξινόμησης στη μηχανική μάθηση (Quinlan, 1986). Λειτουργούν δημιουργώντας διακριτά σημεία αποφάσεων, χρησιμοποι-

Αλγοριθμος 4.1: *k-NN*

1: **procedure** κ-NN(σύνολο δεδομένων εκπαίδευσης X_{train}, ετικέτες κλάσεων εκπαίδευσης Y_{train}, αριθμός κοντινότερων γειτόνων k, σύνολο δειγμάτων προς ταξινόμηση X_{test})
2: **for all** $x \in X_{test}$ **do**
3: Υπολογισμός και αποθήκευση των αποστάσεων $dist(x, X_{train}) = \sqrt{\sum_{n=1}^{|X_{train}|}(x - x_n)^2}$
4: **end for**
5: **for all** $x \in X_{test}$ **do**
6: Ταξινόμηση του x στην κλάση στην οποία ανήκει η πλειοψηφία των k
7: κοντινότερων γειτόνων του.
8: **end for**
9: **return** ετικέτες κλάσεων δεδομένων δοκιμής Y_{test}
10: **end procedure**

ώντας τα χαρακτηριστικά (features) που παρέχονται για κάθε δείγμα, οδηγώντας σε διακλαδώσεις. Η επιλογή του σημείου διαίρεσης γίνεται με βάση την πλειοψηφία των στοιχείων που ανήκουν σε μια κλάση και τα κριτήρια προσαρμόζονται έως ότου επιτευχθεί η βέλτιστη ταξινόμηση για τα δεδομένα εκπαίδευσης.

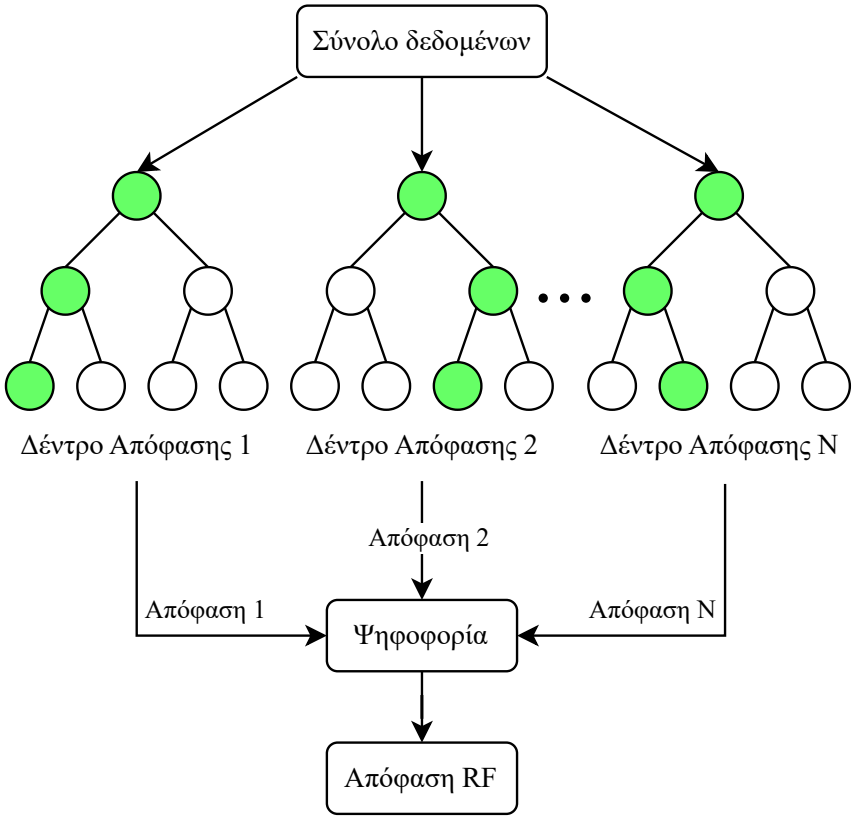


Σχήμα 4.1: *Παράδειγμα δέντρου απόφασης για ταξινόμηση σε δύο κλάσεις Α και Β, με διάνυσμα εισόδου* $\vec{x} = (x_1, x_2, x_3, x_4)$ *τεσσάρων διαστάσεων.*

Ο αλγόριθμος Random Forest (Breiman, 2001) είναι μια μέθοδος συνδυασμού πολλαπλών μοντέλων (ensemble learning), στην προκειμένη περίπτωση πολλαπλών δέντρων αποφάσεων, με σκοπό τη βελτίωση της ταξινόμησης. Αποτελείται από έναν μεγάλο αριθμό ανεξάρτητων δέντρων αποφάσεων που λειτουργούν ως ένα «δάσος» (forest).

Για να δημιουργηθεί το δάσος, εξάγονται πολλά διαφορετικά δείγματα εκπαίδευσης από το αρχικό σύνολο δεδομένων χρησιμοποιώντας την τεχνική bootstrap sampling όπου κάθε δείγμα δημιουργείται επιλέγοντας τυχαία δεδομένα από το αρχικό σύνολο (κάθε παρατήρηση μπορεί να επιλεγεί περισσότερες από μία φορές σε ένα δείγμα αλλά και καμία).

Για κάθε δείγμα εκπαίδευσης, κατασκευάζεται ένα δέντρο απόφασης. Σε κάθε διακλάδωση,



Σχήμα 4.2: *Λήψη απόφασης με τον αλγόριθμο Random Forest.*

επιλέγεται ένα τυχαίο υποσύνολο των συνολικών features για να βρεθεί η βέλτιστη διακλάδωση (split). Αυτό εισάγει την τυχαιότητα στον αλγόριθμο και διασφαλίζει ότι τα δέντρα δεν είναι πανομοιότυπα μεταξύ τους.

Ο υψηλός αριθμός των δέντρων στο δάσος είναι μια σημαντική υπερπαράμετρος που συνεισφέρει στη μείωση του overfitting.

Για την τελική πρόβλεψη, οι προβλέψεις όλων των δέντρων συνδυάζονται. Στην περίπτωση της ταξινόμησης, η πρόβλεψη γίνεται με ψηφοφορία πλειοψηφίας.

Αλγοριθμος 4.2: *Random Forest*

procedure Random Forest(σύνολο δεδομένων εκπαίδευσης S, χαρακτηριστικά F, αριθμός επιθυμητών δέντρων T)
    $H \leftarrow \{\}$ *αρχικοποίηση του δάσους* H
    for i ∈ 1,2, ..., T **do**
        Δημιουργία τυχαίου υποσυνόλου του S με τη μέθοδο bootstrap sampling
        Υπολογισμός του δέντρου αποφάσεων h_i με το υποσύνολο εκπαίδευσης S_i,
        χρησιμοποιώντας σε κάθε διακλάδωση τυχαίο υποσύνολο των χαρακτηριστικών F_i
        Προσθήκη του νέου δέντρου h_i στο δάσος H
        return δάσος H
    end for
end procedure

4.1.3 Αξιολόγηση Ταξινομητών

Πίνακας Σύγχυσης

Ο πίνακας σύγχυσης (confusion matrix) είναι ένας πίνακας που περιγράφει την απόδοση ενός αλγορίθμου ταξινόμησης. Εμφανίζει τις πραγματικές κλάσεις των δεδομένων, σε σχέση με τις κλάσεις στις οποίες ταξινομήθηκαν.
Για ένα μοντέλο ταξινόμησης δύο κλάσεων, ο πίνακας σύγχυσης είναι ο 4.1.

	Προβλεπόμενο Θετικό	Προβλεπόμενο Αρνητικό
Πραγματικό Θετικό	Αληθές Θετικό (TP)	Ψευδές Αρνητικό (FN)
Πραγματικό Αρνητικό	Ψευδές Θετικό (FP)	Αληθές Αρνητικό (TN)

Πίνακας 4.1: *Πίνακας Σύγχυσης Ταξινόμησης δύο κλάσεων (Θετικό - Αρνητικό).*

Ενώ για ένα μοντέλο πολλών κλάσεων, ο πίνακας σύγχυσης επεκτείνεται στον πίνακα 4.2.

	Προβλεπόμενη Κλάση 1	Προβλεπόμενη Κλάση 2	Προβλεπόμενη Κλάση 3
Πραγματική Κλάση 1	TP_1	$FP_{1,2}$	$FP_{1,3}$
Πραγματική Κλάση 2	$FN_{2,1}$	TP_2	$FP_{2,3}$
Πραγματική Κλάση 3	$FN_{3,1}$	$FN_{3,2}$	TP_3

Πίνακας 4.2: *Πίνακας Σύγχυσης για ταξινόμηση πολλών κλάσεων.*

Μετρικές Αξιολόγησης

Οι πιο κοινές μετρικές για την αξιολόγηση ενός μοντέλου ταξινόμησης είναι οι παρακάτω:

$$\text{Accuracy - Ορθότητα} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.1}$$

$$\text{Sensitivity - Ευαισθησία} = \frac{TP}{TP + FN} \tag{4.2}$$

$$\text{Precision - Ακρίβεια} = \frac{TP}{TP + FP} \tag{4.3}$$

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}} \tag{4.4}$$

Οι οποίες εύκολα επεκτείνονται για ένα μοντέλο πολλών κλάσεων, λαμβάνοντας τον μέσο όρο της κάθε μετρικής που μελετήθηκε για τι δύο κλάσεις. Για παράδειγμα, η μετρική της ακρίβειας C κλάσεων, υπολογίζεται ως:

$$\text{Μέσος Όρος Ακρίβειας} = \frac{1}{C}\sum_{i=1}^{C}\frac{TP_i}{TP_i + FP_i}$$

4.2 Συσταδοποίηση Προτάσεων - Sentence Clustering

Συσταδοποίηση ονομάζεται η διαδικασία κατά την οποία καταμερίζονται ετερογενή δεδομένα σε συστάδες (clusters). Πρόκειται για μία μέθοδο μή επιβλεπομένης μάθησης (unsupervised learning), όπου οι κατηγορίες στις οποίες κατηγοριοποιούνται τα δεδομένα δεν είναι προκαθορισμένες όπως στην ταξινόμηση, αλλά προκύπτουν με βάση της ομοιότητες των δεδομένων. Παρακάτω, μελετούνται οι πιο χρησιμοποιημένοι αλγόριθμοι συσταδοποίησης που κάνουν χρήση του κριτηρίου των κέντρων, της ιεραρχίας και της πυκνότητας των δεδομένων. Οι αλγόριθμοι αυτοί, μπορούν να λάβουν ως είσοδο τα διανύσματα προτάσεων που έχουν παραχθεί, με σκοπό την ομαδοποίηση συγγενικών προτάσεων.

4.2.1 Κριτήριο Κέντρων - Παράδειγμα k-Means

Η μέθοδος k-Means είναι μια από τις πιο διαδεδομένες μεθόδους συσταδοποίησης (clustering). Στην αρχή του αλγορίθμου, ορίζεται ένας αριθμός k συστάδων (clusters) στις οποίες επιθυμούμε να κατηγοριοποιηθούν τα στοιχεία του συνόλου των δεδομένων. Ο στόχος του αλγορίθμου είναι η ελαχιστοποίηση της απόστασης κάθε στοιχείου από το κέντρο της συστάδας στην οποία ανήκει. Η διαδικασία επαναλαμβάνεται έως ότου τα κέντρα των συστάδων σταθεροποιηθούν.

Ο αλγόριθμος k-Means προέρχεται από τον Stuart Lloyd της εταιρείας Bell Labs και δημοσιεύτηκε στο άρθρο του Least squares quantization in PCM το 1982 (Lloyd, 1982), ενώ πολλές βελτιώσεις και επεκτάσεις έχουν προταθεί έκτοτε. Παρακάτω, παρατίθεται ο ψευδοκώδικας του απλού k-Means ο οποίος λαμβάνει σαν είσοδο τον αριθμό των συστάδων που επιθυμούμε να δημιουργηθούν και επιστρέφει τις συστάδες.

Αλγοριθμος 4.3: *k-Means*

procedure K-Means(σύνολο δεδομένων D, αριθμός επιθυμητών συστάδων - k)
    Αρχικοποίηση των k κέντρων επιλέγοντας k τυχαία σημεία του συνόλου δεδομένων.
    repeat
        Υπολογισμός $dist(x, \mu_i)$, $\forall x \in D$ και κάθε κέντρο μ_i των clusters C_i.
        Ταξινόμηση του σημείο στο πιο κοντινό κέντρο μ_i.
        Υπολογισμός των νέων κέντρων για κάθε cluster C_i (Υπολογίζοντας $\frac{1}{|C_i|}\sum_{x \in C_i} x$)
    until Σύγλιση (νέα κέντρα ισούνται με τα προηγούμενα)
    return Συσταδοποιημένα δεδομένα και κέντρα συστάδων
end procedure

Συνεπώς το πρόβλημα ανάγεται στη συσταδοποίηση που ελαχιστοποιεί την εξής συνάρτηση:

$$\arg\min_{\mathbf{C}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in C_i} \left\| \mathbf{x} - \mu_i \right\|^2$$

όπου

k: αριθμός συστάδων.

$\mathbf{C}$: σύνολο αναθέσεων σε συστάδες.

C_i: υποσύνολο στοιχείων που έχουν ανατεθεί στη συστάδα i.

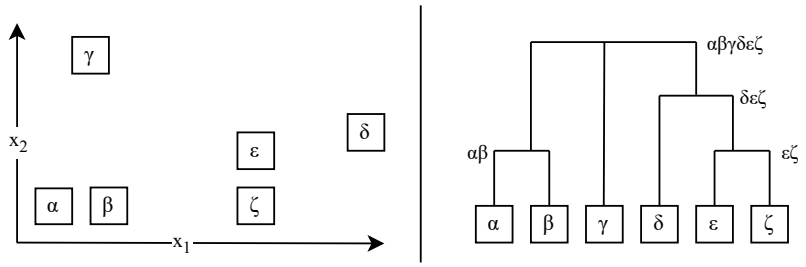
$\mathbf{x}$: κάποιο σημείο που ανήκει στη συστάδα i.

μ_i: συντεταγμένες του κέντρου της συστάδας i.

$\left\| \mathbf{x} - \mu_i \right\|^2$: Ευκλείδεια απόσταση μεταξύ του στοιχείου $\mathbf{x}$ και του κέντρου μ_i της συστάδας i.

4.2.2 Κριτήριο Ιεραρχίας - Ιεραρχική Συσταδοποίηση

Η ιεραρχική συσταδοποίηση (Ward Jr, 1963) είναι μια μέθοδος που παράγει μια ιεραρχία συστάδων σε μορφή δένδρου. Έχει ως πλεονέκτημα πως μπορεί να επιλεχθεί οποιοδήποτε βάθος του δένδρου αναπαράστασης των συστάδων, με σκοπό τη γενικότερη ή την ειδικότερη επιλογή συστάδων μεγαλύτερου ή μικρότερου μεγέθους.



Σχήμα 4.3: *Παράδειγμα Ιεραρχικής Συσταδοποίησης με βάση την ευκλείδεια απόσταση μεταξύ των στοιχείων α,β,γ,δ,ε,ζ στο χώρο.*

Για παράδειγμα, στο σχήμα 4.3, μπορούμε να διαλέξουμε να διαχωρίσουμε τα δεδομένα σε {α,β,γ,δ,ε,ζ} ή στις συστάδες {αβ, γ, δ, εζ} ή ένα επίπεδο πιο πάνω {αβ, γ, δεζ}.
Υπάρχουν δύο βασικές προσεγγίσεις για την ιεραρχική συσταδοποίηση:

- η bottom-up - agglomerative προσέγγιση όπου εκκινείται από χαμηλά στην ιεραρχία με n συστάδες για n δεδομένα και ανεβαίνοντας επίπεδα γίνονται συγχωνεύσεις.
- και η top-down - divisive όπου όλες οι παρατηρήσεις αρχικά ανήκουν σε μια μεγάλη συστάδα και στη συνέχεια διαιρούνται σε μικρότερες.

4.2.3 Κριτήριο Πυκνότητας - Μέθοδος DBSCAN

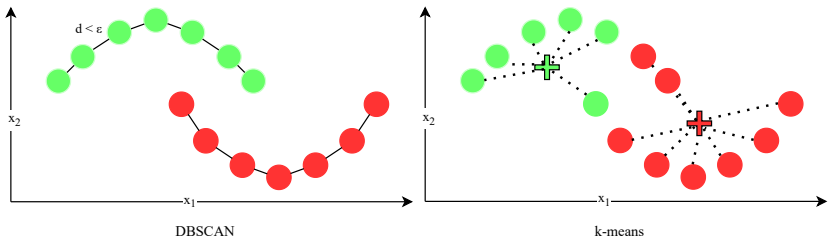
Η μέθοδος DBSCAN (Density-Based Spatial Clustering of Applications with Noise) προτάθηκε από τους Ester, M., Kriegel, et al. (1996) στο άρθρο A density-based algorithm for discovering clusters in large spatial databases with noise (Ester, Kriegel, Sander, Xu,

et al., 1996). Είναι μια μέθοδος συσταδοποίησης που βασίζεται στην πυκνότητα των δεδομένων. Στην DBSCAN, ορίζονται δύο παράμετροι: η παράμετρος ε η οποία καθορίζει την εμβέλεια γειτονίας για ένα δεδομένο στοιχείο, και τον ελάχιστο αριθμό γειτόνων (minPoints) που πρέπει να έχει ένα σημείο για να μη θεωρηθεί απομονωμένο. Κάθε παρατήρηση του συνόλου των δεδομένων εξετάζεται για να διαπιστωθεί εάν ανήκει σε μια συστάδα με βάση την απόστασή της από τις άλλες, και με τον αριθμό γειτονικών παρατηρήσεων ενώ τα σημεία που δεν ανήκουν σε καμία συστάδα θεωρούνται θόρυβος και αγνοούνται.

Αλγοριθμος 4.4: *DBSCAN*

```
1:  procedure DBSCAN(σύνολο δεδομένων D, ε, minPoints)
2:      V ← {} αρχικοποίηση του πίνακα επισκέψεων ως κενό σύνολο
3:      for all (x ∈ D) ∩ (x ∉ V) do
4:          εύρεση των γειτόνων N του x με βάση την απόσταση-εμβέλεια ε
5:          if |N| < minPoints then
6:              το σημείο x είναι θόρυβος
7:          else
8:              C ← {x} αρχικοποίηση ενός νέου cluster
9:              for all x′ in N do - για όλους τους γείτονες
10:                 N ← N \ {x′}
11:                 if x′ ∉ V then - εάν δεν έχει επισκεφθεί
12:                     V ← V ∪ {x′}
13:                     έλεγχος των γειτόνων N′, του x′
14:                     if |N′| ≥ minPoints then
15:                         N ← N ∪ N′
16:                     end if
17:                 end if
18:                 if x′ ∉ C then
19:                     C ← C ∪ {x′} προσθήκη στο cluster
20:                 end if
21:             end for
22:         end if
23:     end for
24: end procedure
```

Με αυτόν τον τρόπο, εντοπίζονται clusters με σχήματα πέρα από απλούς κύκλους όπως εντοπίζει ο αλγόριθμος k-Means, όπως μπορούμε να δούμε στο σχήμα 4.4.



Σχήμα 4.4: *Παράδειγμα Συσταδοποίησης δύο κλάσεων με τις μεθόδους DBSCAN και k-Means. Με το σύμβολο «+» απεικονίζονται τα κέντρα του αλγορίθμου k-Means.*

4.2.4 Αξιολόγηση Συσταδοποίησης

Το ότι οι κατηγορίες που προκύπτουν από τη συσταδοποίηση δεν είναι προκαθορισμένες αλλά προκύπτουν από τα ίδια τα δεδομένα, καθιστά την αξιολόγηση των διαφορετικών μεθόδων δυσκολότερη από αυτήν της ταξινόμησης. Παρόλα αυτά, η ακαδημαϊκή κοινότητα βασίζεται σε κάποιες μετρικές οι οποίες αποτελούν έναν σημαντικό δείκτη ως προς την ποιότητα του αποτελέσματος.

Silhouette Score

Το Silhouette Score (Rousseeuw, 1987) είναι ένα μέτρο ομοιότητας ενός στοιχείου με τη συστάδα στην οποία έχει ανατεθεί (συνοχή) σε σύγκριση με άλλες συστάδες (διαχωρισμός). Το σκορ αυτό κυμαίνεται από -1 έως +1, όπου μια υψηλή τιμή δείχνει ότι το αντικείμενο ταιριάζει καλά με τη δική του συστάδα και ελάχιστα με τις γειτονικές συστάδες.
Ο υπολογισμός του προκύπτει από τα αποτελέσματα των συναρτήσεων a και b όπου:

$$a(i) = \frac{1}{|C_x| - 1} \sum_{j \in C_x, i \neq j} distance(i, j)$$

Πρόκειται για τον υπολογισμό της μέσης απόστασης μεταξύ του σημείου i και όλων των άλλων σημείων που έχουν ανατεθεί στην ίδια συστάδα. Η συστάδα στην οποία έχει ανατεθεί το σημείο i ονομάζεται C_x και περιέχει αριθμό στοιχείων $|C_x|$. Η πράξη $|C_x| - 1$ οφείλεται στο ότι δεν προσμετράται το ζευγάρι $distance(i, i)$ για τον υπολογισμό. Το σκορ $a(i)$ μας δείχνει λοιπόν πόσο καλή ανάθεση του στοιχείου i στο cluster C_x έχει γίνει.
Στη συνέχεια υπολογίζεται η «ανομοιότητα» (dissimilarity) του στοιχείου i με στοιχεία άλλων clusters.

$$b(i) = \min_{y \neq x} \frac{1}{|C_y|} \sum_{j \in C_y} distance(i, j)$$

όπου υπολογίζεται η ελάχιστη μέση απόσταση του i με όλα τα άλλα στοιχεία σε οποιαδήποτε άλλο cluster C_y, όπου $y \neq x$. Το cluster που επιτυγχάνει την ελαχιστοποίηση αυτή, θεωρείται γειτονικό (αμέσως επόμενη καλύτερη λύση από το cluster στο οποίο έχει ήδη ανατεθεί).
Τέλος, το Silhouette Score υπολογίζεται ως:

$$s(i) = \begin{cases} \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, & \text{if } |C_x| > 1 \\ 0, & \text{if } |C_x| = 1 \end{cases}$$

Adjusted Random Index

Ο δείκτης Adjusted Random Index (ARI) (Rand, 1971) είναι μία μετρική της ομοιότητας μεταξύ δύο διαφορετικών αναθέσεων σε συστάδες. Έχοντας δηλαδή τα αποτελέσματα από μία συσταδοποίηση Σ_1 και αυτά από μία άλλη Σ_2, υπολογίζει την ομοιότητα μεταξύ των δύο αποτελεσμάτων, χρησιμοποιώντας τις κοινές κατηγοριοποιήσεις μεταξύ των δύο, αλλά διορθώνοντας παράλληλα για τυχούσες ομοιότητες που προέκυψαν τυχαία. Πρόκειται για μία επέκταση του απλού Random Index (RI), ο οποίος είναι ένας δείκτης που δεν λαμβάνει

υπόψη την τυχαιότητα και υπολογίζεται όπως φαίνεται παρακάτω.

Αλγοριθμος 4.5: *Adjusted Random Index*

procedure ARI(Αποτελέσματα Συσταδοποίησης: Σ_1, Σ_2)
    N αριθμός δειγμάτων στο σύνολο δεδομένων
    Σ_1, Σ_2 οι δύο διαφορετικές συσταδοποιήσεις για το σύνολο δεδομένων.
    a αριθμός ζευγαριών από σημεία που βρίσκονται στο ίδιο cluster στη Σ_1 **και** στη Σ_2
    β αριθμός ζευγαριών από σημεία που βρίσκονται σε διαφορετικό cluster στη Σ_1 **και** στη Σ_2
    Υπολογισμός του αριθμού όλων των πιθανών ζευγαριών για N δείγματα: $\binom{N}{2} = \frac{N(N-1)}{2}$
    Υπολογισμός του απλού Rand Index ως: **RI** = $\frac{a+\beta}{\binom{N}{2}}$
    Υπολογισμός της αναμενόμενης τιμής του Rand Index για τυχαίες συσταδοποιήσεις:
    $E = \frac{\sum_i \binom{n_i}{2} \sum_j \binom{n_j}{2}}{\binom{N}{2}}$, όπου n_i είναι ο αριθμός των δειγμάτων στο cluster i και n_j είναι ο
    αριθμός των δειγμάτων στο cluster j
    Υπολογισμός του Adjusted Rand Index ως: **ARI** = $\frac{RI-E}{1-E}$
end procedure

Η μετρική ARI μας επιτρέπει ως εκ τούτου, να συγκρίνουμε δύο συσταδοποιήσεις που έχουν προκύψει από δύο διαφορετικές μεθόδους. Κυμαίνεται από -1 έως 1, όπου το 1 υποδηλώνει τέλεια συμφωνία μεταξύ των δύο συσταδοποιήσεων, το 0 παραπέμπει σε τυχαία συμφωνία και το -1 σε εντελώς διαφορετικές συσταδοποιήσεις.

Κεφάλαιο 5

Χρονοσειρές - Ποιοτικά Χαρακτηριστικά - Μέθοδοι Πρόβλεψης

5.1 Χρονοσειρές

Μία χρονοσειρά (time series) είναι μία ακολουθία $\{x_t : t = 0, 1, 2, ...\}$ όπου κάθε x_t εκφράζει την τιμή ενός συστήματος κατά την χρονική στιγμή t. Συνήθως, οι τιμές αυτές καταγράφονται με σταθερό βήμα και δεν περιέχουν απουσιάζουσες τιμές (ομογενείς χρονοσειρές), αλλά κάτι τέτοιο δεν είναι απαραίτητο (ετερογενείς χρονοσειρές). Οι τιμές που καταγράφονται ενδέχεται να συμβολίζουν από χιλιοστά βροχόπτωσης για κάποια χρονική στιγμή, μέχρι καρδιακούς χτύπους ενός ασθενή και τιμές χρηματιστηριακών αγαθών.

Οι χρονοσειρές δύναται να χαρακτηρίζονται από το θόρυβο ή την τυχαιότητα (στοχαστικές χρονοσειρές), ενώ άλλες κρύβουν μέσα τους κάποιες σχέσεις εξάρτησης των μελλοντικών παρατηρήσεων από τις προηγούμενες (ντετερμινιστικές χρονοσειρές). Και για τα δύο είδη χρονοσειρών, η διαδικασία της ανάλυσης τους, αποκαλύπτοντας τα ποιοτικά χαρακτηριστικά τους, μπορεί να φανεί ιδιαίτερα χρήσιμη για την κατανόηση της συμπεριφοράς τους και την πρόβλεψη των μελλοντικών τιμών τους.

5.1.1 Ποιοτικά Χαρακτηριστικά Χρονοσειρών

Τα ποιοτικά χαρακτηριστικά των χρονοσειρών, είναι η τάση (trend), η κυκλικότητα (cycles), η εποχιακότητα (seasonality) και η τυχαιότητα (randomness) (Makridakis, Wheelwright, & Hyndman, 2008).

Τάση (Trend)

Η τάση αναφέρεται στη γενική κατεύθυνση που ακολουθεί μια χρονοσειρά σε μακροπρόθεσμο ορίζοντα. Αντιπροσωπεύει τη μακροπρόθεσμη αύξηση ή μείωση των τιμών της χρονοσειράς, η οποία μπορεί να είναι γραμμική παρουσιάζοντας μία σταθερή μεταβολή ή μη γραμμική, έχοντας για παράδειγμα εκθετική ή λογαριθμική μορφή.

Κυκλικότητα (Cycles)

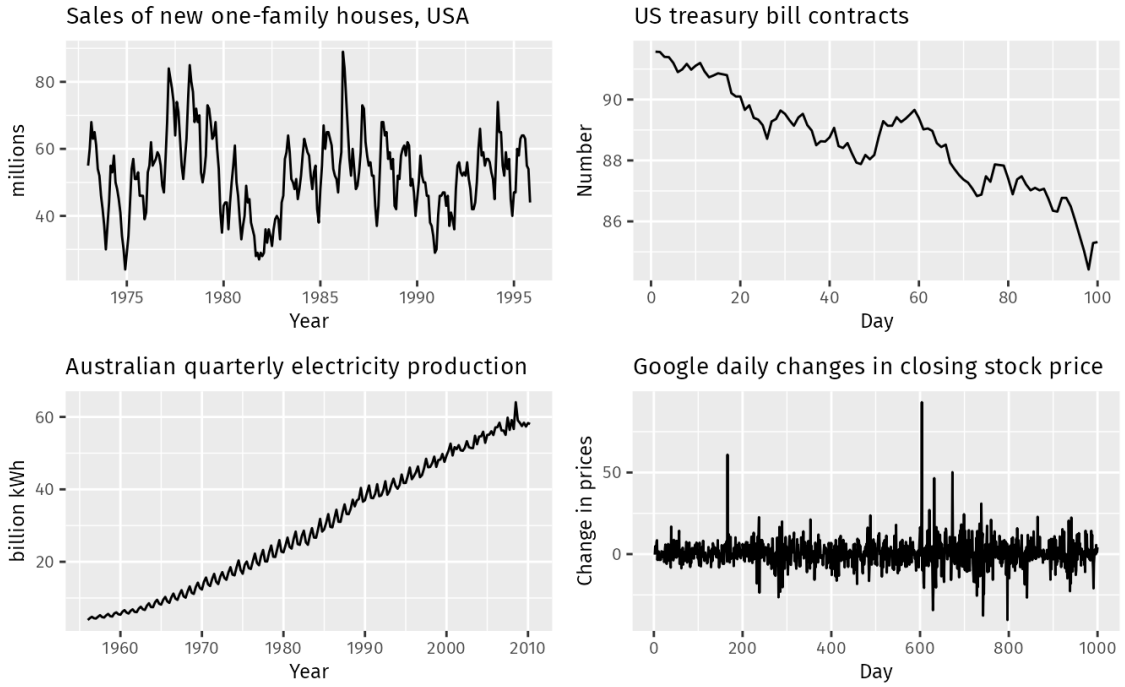
Η κυκλικότητα αναφέρεται στις διακυμάνσεις της χρονοσειράς που εμφανίζονται σε μακροπρόθεσμο ορίζοντα λόγω εξωγενών συνθηκών. Αυτές οι διακυμάνσεις έχουν συνήθως μεγαλύτερη διάρκεια από τις εποχιακές και δεν έχουν απαραίτητα σταθερή περίοδο. Για παράδειγμα, οι οικονομικοί κύκλοι περιλαμβάνουν φάσεις ανάπτυξης και ύφεσης.

Εποχιακότητα (Seasonality)

Η εποχιακότητα αναφέρεται στις περιοδικές διακυμάνσεις της χρονοσειράς που επαναλαμβάνονται σε σταθερά διαστήματα, όπως μήνες, τρίμηνα και εξάμηνα. Αυτές οι διακυμάνσεις οφείλονται σε εποχιακούς παράγοντες, όπως οι καιρικές συνθήκες, οι αργίες και οι κοινωνικές συνήθειες. Η εποχιακότητα μπορεί να επηρεάσει σημαντικά τη ζήτηση προϊόντων και υπηρεσιών.

Τυχαιότητα (Randomness)

Η τυχαιότητα αναφέρεται στις απρόβλεπτες διακυμάνσεις της χρονοσειράς που δεν μπορούν να εξηγηθούν από την τάση, την κυκλικότητα ή την εποχιακότητα. Αυτές οι διακυμάνσεις είναι τυχαίες και μπορεί να προκύπτουν από ακανόνιστες ή απροσδόκητες επιρροές.



Σχήμα 5.1: *Παραδείγματα χρονοσειρών (Hyndman & Athanasopoulos, 2018).*

Στην εικόνα 5.1, παρατηρούνται τέσσερεις χρονοσειρές που παρουσιάζουν κάποια από τα ποιοτικά χαρακτηριστικά που αναφέρθηκαν. Η επάνω αριστερά χρονοσειρά παρουσιάζει έντονα χαρακτηριστικά εποχιακότητας μέσα στο χρόνο αλλά και μία κυκλική συμπεριφορά κάθε 6 με 10 έτη όσον αφορά τις αγορές κατοικιών. Επάνω δεξιά, η χρονοσειρά χαρακτηρίζεται από πτωτική τάση που αναπαριστά τον αριθμό συμβολαίων κρατικών γραμματίων των ΗΠΑ. Κάτω αριστερά μπορεί να παρατηρηθεί μια χρονοσειρά με έντονη ανοδική τάση και

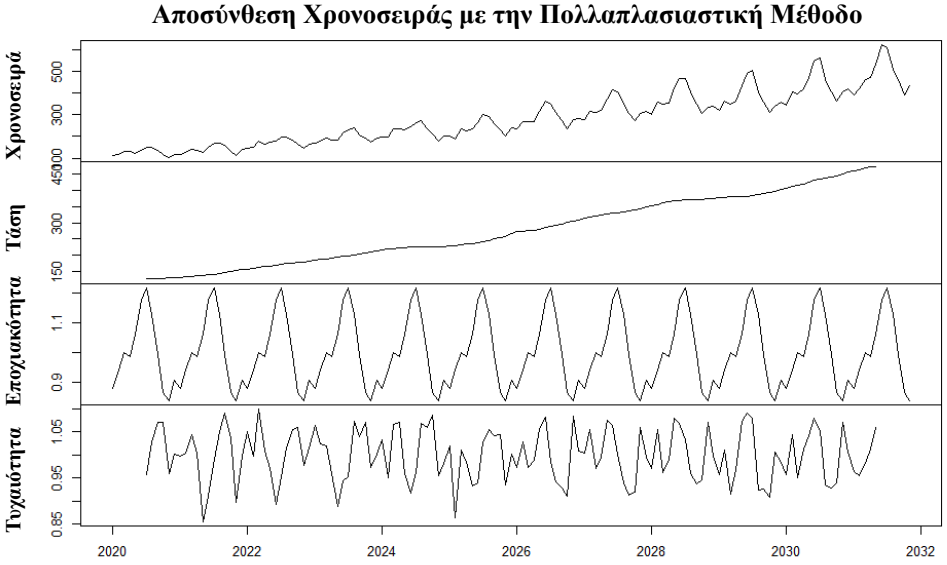
σημαντική εποχιακότητα που αναπαριστά την παραγωγή ηλεκτρισμού στην Αυστραλία ενώ από την τυχαιότητα χαρακτηρίζεται η κάτω δεξιά χρονοσειρά που αναπαριστά την ημερήσια τιμή της μετοχής της Google.

5.1.2 Αποσύνθεση Χρονοσειρών (Time Series Decomposition)

Ο στόχος της αποσύνθεσης μίας χρονοσειράς, είναι η έκφραση της με τις τέσσερεις βασικές συνιστώσες που αναφέρθηκαν παραπάνω και διατυπώνεται ως :

$$Y_t = f(S_t, T_t, C_t, R_t) \tag{5.1}$$

Όπου Y_t είναι η παρατηρούμενη τιμή της χρονοσειράς τη χρονική στιγμή t, S_t η εποχιακότητα, T_t η τάση, C_t η κυκλική συνιστώσα και R_t η τυχαιότητα.



Σχήμα 5.2: *Παράδειγμα αποσύνθεσης χρονοσειράς με πολλαπλασιαστικό μοντέλο χρησιμοποιώντας το πακέτο stats της προγραμματιστικής γλώσσας R σε συνιστώσες τάσης, εποχιακότητας και τυχαιότητας.*

Ενώ δύο μέθοδοι μαθηματικής διατύπωσης της αποσύνθεσης είναι το πολλαπλασιαστικό μοντέλο:

$$Y_t = S_t \times T_t \times C_t \times R_t \tag{5.2}$$

και το προσθετικό μοντέλο:

$$Y_t = S_t + T_t + C_t + R_t \tag{5.3}$$

5.1.3 Ποσοτικοποίηση Ποιοτικών Χαρακτηριστικών Χρονοσειρών

Πέρα από τα τέσσερα κύρια ποιοτικά χαρακτηριστικά μίας χρονοσειράς, έχει παρατηρηθεί πως είναι χρήσιμο να εξάγονται επιπλέον χαρακτηριστικά τα οποία βοηθούν στο χαρακτηρισμό μιας χρονοσειράς. Ο στόχος, είναι η εξαγωγή κάποιων χαρακτηριστικών που μπορούν να «περιγράψουν» μία χρονοσειρά με τον καλύτερο δυνατό τρόπο, ώστε να μπορούν να χρησιμοποιηθούν για την πρόβλεψη των μελλοντικών τιμών της.

Οι (Kang, Hyndman, & Smith-Miles, 2017) επιλέγουν έξι χαρακτηριστικά μεταξύ των οποίων είναι η τάση και η εποχιακότητα αλλά όχι μόνο. Συγκεκριμένα χρησιμοποιούνται:

- Η φασματική εντροπία (Spectral Entropy)

$$F_1 = \int_{-\pi}^{+\pi} f_x(\lambda) \log(f_x(\lambda))\, d\lambda \tag{5.4}$$

  Πρόκειται για μία εκτίμηση της εντροπίας Shannon (Shannon, 1948) της συνάρτησης της φασματικής πυκνότητας $f_x(\lambda)$ που αναπαριστά την προβλεψιμότητα (forecastability) μιας χρονοσειράς.

- Η ένταση της τάσης

$$F_2 = 1 - \frac{\text{var}(R_t)}{\text{var}(x_t - S_t)} \tag{5.5}$$

  όπου var είναι η διακύμανση, ενώ R_t και S_t, είναι η τυχαιότητα και η εποχιακότητα όπως αυτές υπολογίζονται από την αποσύνθεση STL (Cleveland, Cleveland, McRae, Terpenning, et al., 1990), $x_t = S_t + T_t + R_t$.

- Η ένταση της εποχιακότητας

$$F_3 = 1 - \frac{\text{var}(R_t)}{\text{var}(x_t - T_t)} \tag{5.6}$$

- Η συχνότητα της εποχιακότητας ης χρονοσειράς ($F_4 = 4$ για τριμηνιαία δεδομένα, $F_4 = 12$ για μηνιαία κ.ο.κ).

- Η αυτο-συσχέτιση πρώτου βαθμού (First Order Autocorrelation) που υπολογίζεται ως η συσχέτιση μεταξύ της χρονοσειράς και της χρονοσειράς με υστέρηση μία χρονική στιγμή:

$$F_5 = \text{Corr}(R_t, R_{t-1}) \tag{5.7}$$

  Όπου η συσχέτιση υπολογίζεται ως

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{(n\sum x^2 - (\sum x)^2)(n\sum y^2 - (\sum y)^2)}} \tag{5.8}$$

- Η βέλτιστη παράμετρος του μετασχηματισμού Box-Cox (Box & Cox, 1964) που προκύπτει από τον υπολογισμό:

$$w_t = \begin{cases} \frac{x_t^{\lambda}-1}{\lambda}, & \text{if } \lambda \neq 0 \\ \log(x_t), & \text{if } \lambda = 0 \end{cases} \tag{5.9}$$

  Ως F_6 επιλέγεται η τιμή $\lambda \in (0, 1)$ που σταθεροποιεί τη διακύμανση της χρονοσειράς μεγιστοποιώντας την πιθανοφάνεια. Ο δείκτης λ, μετρά τη στασιμότητα της χρονοσειράς.

Με αυτά τα έξι χαρακτηριστικά, παρατηρείται πως μπορεί να επιτευχθεί μία «συμπυκνωμένη» περιγραφή οποιασδήποτε χρονοσειράς, ανεξαρτήτως του μήκους της, με ένα διάνυσμα $\vec{F} = (F_1, F_2, F_3, F_4, F_5, F_6)$

Στη δημοσίευση Exploring the representativeness of the M5 competition data (Theodorou et al., 2022), μελετάται η αντιπροσωπευτικότητα των δεδομένων του διαγωνισμού πρόβλεψης «M5» (2020). Στον «M5» ζητήθηκε να προβλεφθούν οι ιεραρχικές πωλήσεις ανά μονάδα προϊόντος της μεγαλύτερης εταιρείας λιανικής πώλησης στον κόσμο, της Walmart για τρεις κατηγορίες αγαθών: Τρόφιμα, Οικιακά Προϊόντα και Χόμπι. Για την αντιπροσώπευση των χρονοσειρών, εξετάζονται αρχικά 2508 χαρακτηριστικά, από τα οποία επιλέγονται εν τέλει 42. Τα χαρακτηριστικά που επιλέχθηκαν, συμπεριλαμβάνουν, μεταξύ άλλων, τους συντελεστές του διακριτού μετασχηματισμού Fourier, τη διακύμανση, την κατανομή των δεδομένων, τα διαφορικά χαρακτηριστικά, την εντροπία, τη γραμμική τάση της σειράς, τις διαλείψεις που παρουσιάζονται και το μέγιστο αριθμό από διπλότυπες παρατηρήσεις κάθε χρονοσειράς.

5.1.4 Παρουσίαση Ποιοτικών Χαρακτηριστικών σε Ν-Διάστατο Χώρο

Η μείωση των διαστάσεων είναι ένα κρίσιμο βήμα στην ανάλυση δεδομένων υψηλής διάστασης, καθώς συμβάλλει στην απλοποίηση του μοντέλου και τη βελτίωση της απόδοσης των μοντέλων που χρησιμοποιούνται για την πρόβλεψη. Στην παρούσα υποενότητα θα εξεταστούν δύο δημοφιλείς τεχνικές μείωσης διαστάσεων: Η ανάλυση σε κύριες συνιστώσες (Principal Component Analysis - PCA), και οι αυτοκωδικοποιητές που μελετήθηκαν σύντομα στο κεφάλαιο 2 .

Ανάλυση σε Κύριες Συνιστώσες (Principal Component Analysis - PCA)

Η ανάλυση σε κύριες συνιστώσες (Pearson, 1901) είναι μια γραμμική μέθοδος μείωσης διαστάσεων που στοχεύει στον εντοπισμό των κύριων συνιστωσών που εξηγούν τη μεγαλύτερη διασπορά των δεδομένων. Με την προβολή των δεδομένων στις κύριες συνιστώσες, η μέθοδος αυτή μπορεί να μειώσει τις διαστάσεις του αρχικού συνόλου χαρακτηριστικών, διατηρώντας όσο το δυνατόν περισσότερη πληροφορία.

Για την αναπαράσταση των δεδομένων σε λιγότερες συνιστώσες, το πρώτο βήμα είναι η κανονικοποίηση των δεδομένων. Έχοντας υπολογίσει το μέσο όρο και την τυπική απόκλιση κάθε χαρακτηριστικού, αρχικά αφαιρείται από τις τιμές ο μέσος όρος και στη συνέχεια το αποτέλεσμα διαιρείται από την τυπική απόκλιση.

Στη συνέχεια, υπολογίζεται ο πίνακας συνδιακύμανσης που για παράδειγμα για τρία χαρακτηριστικά F_1, F_2, F_3 είναι :

$$\begin{bmatrix} Cov(F_1,F_1) & Cov(F_1,F_2) & Cov(F_1,F_3) \\ Cov(F_2,F_1) & Cov(F_2,F_2) & Cov(F_2,F_3) \\ Cov(F_3,F_1) & Cov(F_3,F_2) & Cov(F_3,F_3) \end{bmatrix} \tag{5.10}$$

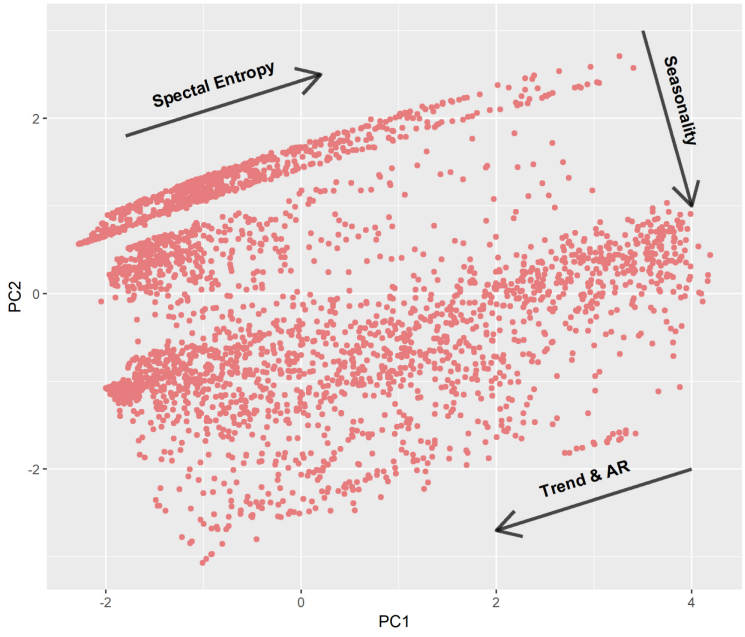
$$\Rightarrow \begin{bmatrix} Var(F_1) & Cov(F_1,F_2) & Cov(F_1,F_3) \\ Cov(F_2,F_1) & Var(F_2) & Cov(F_2,F_3) \\ Cov(F_3,F_1) & Cov(F_3,F_2) & Var(F_3) \end{bmatrix} \tag{5.11}$$

Όπου $Var(F_i) = \frac{1}{N}\sum_{n=1}^{N}(F_{i,n} - \bar{F}_i)^2$ είναι η διακύμανση του χαρακτηριστικού F_i, και $Cov(F_i,F_j) = \frac{1}{N}\sum_{n=1}^{N}(F_{i,n} - \bar{F}_i)(F_{j,n} - \bar{F}_j)$ είναι η συνδιακύμανση μεταξύ των χαρακτηριστικών F_i και F_j.

Έχοντας τον πίνακα συνδιακύμανσης, υπολογίζονται τα ιδιοδιανύσματα και οι ιδιοτιμές (για παράδειγμα για τα τρία χαρακτηριστικά θα υπάρχουν τρία ιδιοδιανύσματα και τρεις ιδιοτιμές).
Το ιδιοδιάνυσμα το οποίο αντιστοιχεί στη μεγαλύτερη ιδιοτιμή είναι η πιο σημαντική συνιστώσα καθώς εξηγεί το μεγαλύτερο ποσοστό της διακύμανσης των δεδομένων, και ονομάζεται $PC1$, αντίστοιχα και τα υπόλοιπα σε φθίνουσα σειρά. Για τη μείωση διαστάσεων, μπορούμε να επιλέξουμε να χρησιμοποιήσουμε μόνο τις x πιο σημαντικές κύριες συνιστώσες.
Για να μετασχηματιστούν τα αρχικά χαρακτηριστικά στη νέα μειωμένη μορφή, το διάνυσμα για παράδειγμα F_1, μετασχηματίζεται ως:

$$PC_1 = \vec{v}_1^T \vec{F_1} \tag{5.12}$$

όπου v_1 είναι το ιδιοδιάνυσμα που αντιστοιχεί στο F_1, και $\vec{F_1}$ το αρχικό διάνυσμα που περιέχει τις τιμές των δειγμάτων για το χαρακτηριστικό F_1.
Αναλύοντας τα χαρακτηριστικά σε κύριες συνιστώσες, είναι πλέον εφικτό να τα παρουσιάσουμε στο δισδιάστατο χώρο, χρησιμοποιώντας τις δύο κύριες συνιστώσες τους.
Στο σχήμα 5.3 το σύνολο των χρονοσειρών του διαγωνισμού M3, αναλύθηκε με βάση τα 6 χαρακτηριστικά που μελετήθηκαν στο 5.1.3, και στη συνέχεια αναπαρίσταται από τις δύο κύριες συνιστώσες του στο χώρο (Spiliotis et al., 2020). Με τον τρόπο αυτό, μπορούν να εντοπιστούν χρονοσειρές που έχουν παρόμοιες τιμές στα έξι χαρακτηριστικά που εξετάστηκαν, αφού βρίσκονται κοντά η μία στην άλλη σε αυτόν τον χώρο.



Σχήμα 5.3: *Παράδειγμα αντιπροσώπευσης του συνόλου των χρονοσειρών του διαγωνισμού προβλέψεων M3 στο χώρο, με τη μέθοδο ανάλυσης σε κύριες συνιστώσες (Spiliotis et al., 2020).*

Αυτοκωδικοποιητές

Η ανάλυση σε κύριες συνιστώσες, επιτυγχάνει πολλές φορές να μειώσει τις διαστάσεις το δεδομένων, ωστόσο αυτό είναι εφικτό μόνο για δεδομένα που παρουσιάζουν γραμμικές σχέσεις. Αντιθέτως, οι αυτοκωδικοποιητές μπορούν να συλλάβουν μη γραμμικές σχέσεις μεταξύ των δεδομένων, έχοντας όμως ιδιαίτερα υψηλότερες απαιτήσεις σε υπολογιστικούς πόρους.
Αποτελούνται από δύο μέρη: τον κωδικοποιητή (encoder) και τον αποκωδικοποιητή (decoder). Ο κωδικοποιητής συμπιέζει τα χαρακτηριστικά σε έναν χώρο χαμηλών διαστάσεων, ενώ ο αποκωδικοποιητή επιχειρεί να τα ανακατασκευάσει από τη συμπιεσμένη αναπαράσταση. Έστω λοιπόν η κωδικοποίηση $\mathbf{h} = f(\mathbf{F})$, όπου $\mathbf{h}$ είναι η συμπιεσμένη αναπαράσταση των δεδομένων και f είναι η συνάρτηση του κωδικοποιητή. Τα ανακατασκευασμένα δεδομένα $\mathbf{F}' = g(\mathbf{h})$, είναι επιθυμητό να βρίσκονται όσο πιο κοντά γίνεται στα αρχικά. Έχοντας πλέον έναν εκπαιδευμένο αυτοκωδικοποιητή, είναι εφικτό να μειωθούν οι διαστάσεις με τη συνάρτηση κωδικοποίησης.

5.2 Μέθοδοι Πρόβλεψης Χρονοσειρών

Ιστορικά, απλές στατιστικές μέθοδοι έχουν ιδιαίτερα καλή ακρίβεια στην πρόβλεψη και χρειάζονται πολύ λιγότερους υπολογιστικούς πόρους από μεθόδους που χρησιμοποιούν τεχνικές μηχανικής μάθησης. Ωστόσο, τα τελευταίοι χρόνια οι τεχνικές μηχανικής μάθησης χρησιμοποιούνται ευρέως σε πολλούς τομείς, καθώς φαίνεται να μπορούν να επιτύχουν καλές επιδόσεις και στις προβλέψεις (Διαγωνισμοί Πρόβλεψης M5 2020 και M6 2022). Στην ενότητα αυτή θα μελετηθούν αρχικά κάποιες απλές στατιστικές μέθοδοι, με σκοπό τη χρήση τους ως μέτρο σύγκρισης για την επίδοση των πιο πολύπλοκων μεθόδων μηχανικής μάθησης που παρουσιάζονται στη συνέχεια. Η ειδοποιός διαφορά των δύο προσεγγίσεων είναι πως οι μέθοδοι μηχανικής μάθησης είναι δυνατό να εντοπίσουν περίπλοκα πρότυπα και σχέσεις μεταξύ των παρατηρήσεων οι οποίες δεν είναι εμφανείς και εύκολες να εντοπιστούν από απλές στατιστικές μεθόδους.

5.2.1 Κλασικές Στατιστικές Μέθοδοι Πρόβλεψης

Με τον όρο «στατιστική μέθοδος», χαρακτηρίζεται η εφαρμογή ενός μαθηματικού μοντέλου πάνω σε μια σειρά δεδομένων προκειμένου να παραχθεί με συστηματικό τρόπο η πρόβλεψη της πορείας της χρονοσειράς. Στην παρούσα εργασία, θα μελετηθούν κάποιες απλές μέθοδοι όπως η μέθοδος Naïve, η απλή μέθοδος γραμμικής παλινδρόμησης και η απλή εκθετική εξομάλυνση.

Μέθοδος Naïve

Η πιο βασική μέθοδος πρόβλεψης χρονοσειρών είναι η αφελής μέθοδος, γνωστή ως Naïve. Στη μέθοδο αυτή, γίνεται η υπόθεση πως η αμέσως επόμενη τιμή μίας χρονοσειράς, θα είναι ίδια την παρούσα τελευταία παρατήρηση, χρησιμοποιώντας δηλαδή ως πρόβλεψη για την επόμενη χρονική περίοδο την αμέσως προηγούμενη παρατήρηση:

$$F_t = Y_{t-1} \tag{5.13}$$

όπου F_t είναι η προβλεπόμενη τιμή για την περίοδο t και Y_{t-1} είναι η πραγματική τιμή της χρονοσειράς για την περίοδο $t-1$.
Δεδομένου ότι αυτή η μέθοδος δεν λαμβάνει υπόψη τις πιθανές διακυμάνσεις στο επίπεδο της χρονοσειράς ή την τάση, οι προβλέψεις της είναι συνήθως χαμηλής ακρίβειας, ιδιαίτερα για μακροπρόθεσμες προβλέψεις. Ωστόσο, χρησιμοποιείται συχνά ως σημείο αναφοράς (benchmark) για την αξιολόγηση της απόδοσης άλλων, πιο περίπλοκων μεθόδων πρόβλεψης.

Απλή Γραμμική Παλινδρόμηση (Simple Linear Regression)

Μία ιδιαίτερα διαδεδομένη μέθοδος πρόβλεψης είναι αυτή της απλής γραμμικής παλινδρόμησης. Η παραδοχή της μεθόδου αυτής είναι πως οι τιμές της χρονοσειράς μπορούν να μοντελοποιηθούν σχετικά καλά από μία ευθεία, η οποία μπορεί να εκφρασθεί με τη βοήθεια του μαθηματικού τύπου:

$$F_t = a + bX + e \tag{5.14}$$

όπου F_t είναι η προβλεπόμενη τιμή για τη χρονική στιγμή t, a είναι το αρχικό σημείο για $(X = 0)$, b η κλίση της ευθείας και e το σφάλμα ($e_t = Y_t - F_t$, για πραγματική τιμή Y_t και F_t την τιμή πρόβλεψης).
Για την ανεύρεση των βέλτιστων a και b, χρησιμοποιείται η μέθοδος των ελαχίστων τετραγώνων ως εξής:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - a - bX_i)^2 \tag{5.15}$$

$$\Rightarrow b = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}, \tag{5.16}$$

$$\Rightarrow a = \bar{Y} - b\bar{X} \tag{5.17}$$

Μέθοδοι Εκθετικής Εξομάλυνσης (Exponential Smoothing)

- Απλή Εκθετική Εξομάλυνση (Simple Exponential Smoothing - SES) Τα μοντέλα εκθετικής εξομάλυνσης, είναι μοντέλα που χρησιμοποιούν σταθμισμένους κινητούς μέσους όρους με βάρη που φθίνουν εκθετικά. Η βασική αρχή της μεθόδου είναι ότι τα πιο πρόσφατα δεδομένα περιέχουν πιο σημαντική πληροφορία. Έτσι τα μοντέλα εκθετικής εξομάλυνσης δίνουν μεγάλη βαρύτητα στα πρόσφατα δεδομένα, η οποία φθίνει εκθετικά καθώς κινούμαστε προς το παρελθόν.

- Εκθετική Εξομάλυνση με Τάση (Holt's Linear Trend Model) Η μέθοδος αυτή επεκτείνει την απλή εκθετική εξομάλυνση λαμβάνοντας υπόψη την τάση στα δεδομένα. Χρησιμοποιεί δύο εξισώσεις εκθετικής εξομάλυνσης: μία για το επίπεδο (level) και μία για την τάση (trend).

- Εκθετική Εξομάλυνση με Τάση και Εποχιακότητα (Holt-Winters Seasonal Model) Αυτή η μέθοδος προσαρμόζει περαιτέρω την εκθετική εξομάλυνση για να συμπεριλάβει

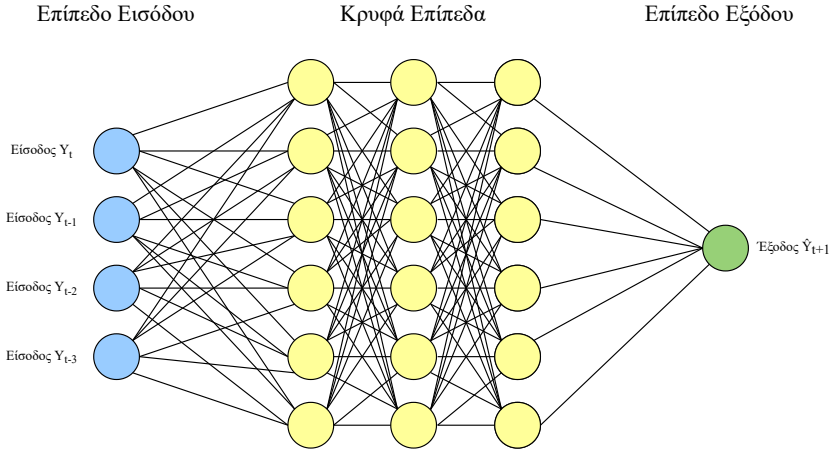
εποχιακές διακυμάνσεις. Υπάρχουν δύο εκδοχές του μοντέλου αυτού: το προσθετικό (additive) και το πολλαπλασιαστικό (multiplicative).

5.2.2 Μέθοδοι Πρόβλεψης με Χρήση Νευρωνικών Δικτύων

Σε αντίθεση με τις προηγούμενες απλές στατιστικές μεθόδους, οι μέθοδοι των νευρωνικών δικτύων δεν έχουν κάποιους ακριβείς μαθηματικούς κανόνες αλλά «μαθαίνουν» από τα δεδομένα με σκοπό την ελαχιστοποίηση του σφάλματος όπως μελετήθηκε σε βάθος στο κεφάλαιο 2.

Η είσοδος του νευρωνικού δικτύου αποτελείται από παρελθοντικές τιμές της χρονοσειράς (lags). Για μία χρονοσειρά Y, οι είσοδοι θα είναι $\Upsilon_t, Y_{t-1}, Y_{t-2}, \ldots, Y_{t-k}$ όπου k είναι ο αριθμός των παρελθοντικών τιμών που λαμβάνονται υπόψη (look-back window) και ο σκοπός της εκπαίδευση, είναι να παραχθούν οι προβλέψεις $Y_{t+1}, Y_{t+2}, \ldots, Y_{t+h}$, όπου το h ονομάζεται ορίζοντας πρόβλεψης με όσο λιγότερο σφάλμα γίνεται. Πέρα από τις παρελθοντικές τιμές, η είσοδος μπορεί να περιλαμβάνει και άλλα χαρακτηριστικά, όπως αυτά που μελετήθηκαν στην υποενότητα 5.1.3.

Σε περίπτωση που ο ορίζοντας πρόβλεψης h είναι μεγαλύτερος από μία μελλοντική χρονική στιγμή το μοντέλο πρέπει να παράξει πολλές εξόδους. Προκειμένου το δίκτυο να παράξει πολλαπλές εξόδους, είναι εφικτό είτε να εκπαιδευτεί στην παραγωγή όλων των προβλέψεων ταυτόχρονα (και συνεπώς να διαθέτει h κόμβους εξόδου), είτε να παραχθούν πολλαπλά νευρωνικά δίκτυα (ενός κόμβου εξόδου), το κάθε ένα από τα οποία έχει ως στόχο την πρόβλεψη μίας τιμής του ορίζοντα h.



Σχήμα 5.4: *Παράδειγμα ενός Πολυεπίπεδου Νευρωνικού Δικτύου (MLP) με τρία κρυφά επίπεδα για την πρόβλεψη της χρονοσειράς $Y(t)$ για ορίζοντα πρόβλεψης $h = 1$ και τέσσερις παρελθοντικές τιμές.*

5.3 Μετρικές Αξιολόγησης Προβλέψεων σε Χρονοσειρές

Έχοντας παράξει προβλέψεις, είναι σημαντικό να μπορεί να αξιολογηθεί η διαφορά μεταξύ μίας πραγματικής παρατήρησης της χρονοσειράς, με αυτή που παράχθηκε από το μοντέλο

πρόβλεψης. Αυτή η διαφορά ονομάζεται σφάλμα, και οι μέθοδοι αξιολόγησης του σφάλματος που χρησιμοποιούνται μπορούν να αναδείξουν την προκατάληψη, ή την ακρίβεια της πρόβλεψης.

Απλοί δείκτες σφάλματος που χρησιμοποιούνται είναι το απόλυτο μέσο σφάλμα (MAE = MAE = $\frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{T}_i|$) το οποίο εντοπίζει ένα μέσο όρο της αστοχίας της πρόβλεψης, χωρίς όμως να περιγράφει την κατεύθυνση της. Συχνά χρησιμοποιείται και ο δείκτης της ρίζας του τετραγωνικού σφάλματος ($RMSE = \sqrt{\frac{1}{k}\sum_{t=1}^{n}(Y_t - \hat{Y}_t)^2}$) καθώς δίνει μεγάλο βάρος και εκφράζεται στις μονάδες της αρχικής χρονοσειράς.

Θέλοντας συχνά να συγκριθεί μία μέθοδος πρόβλεψης για διαφορετικές χρονοσειρές που κυμαίνονται σε διαφορετικά επίπεδα, δείκτες σφάλματος όπως το MAE και $RMSE$ δεν είναι ιδιαίτερα χρήσιμοι. Αντιθέτως, δείκτες όπως το μέσο απόλυτο ποσοστιαίο σφάλμα:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_i - F_i}{Y_i}\right| \cdot 100 \tag{5.18}$$

που βοηθάει για τη σύγκριση του μοντέλου πρόβλεψης σε χρονοσειρές που έχουν διαφορετικά επίπεδα και μέση τιμή, αλλά και το συμμετρικό μέσο απόλυτο ποσοστιαίο σφάλμα (Goodwin & Lawton, 1999):

$$sMAPE = \frac{2}{h}\sum_{t=n+1}^{n+h}\frac{|Y_t - F_t|}{|Y_t| + |F_t|} \cdot 100 \tag{5.19}$$

το οποίο μπορεί να πάρει τιμές από 0% έως 200%.

Επιπλέον, χρησιμοποιείται το μέσο απόλυτο κανονικοποιημένο σφάλμα (Mean Absolute Scaled Error) (Hyndman & Koehler, 2006):

$$MASE = \frac{\frac{1}{n-m}\sum_{t=m+1}^{n}|Y_t - F_t|}{\frac{1}{n-m}\sum_{t=m+1}^{n}|Y_t - Y_{t-m}|} \tag{5.20}$$

όπου m είναι η συχνότητα των δεδομένων, για παράδειγμα $m = 12$ για μηνιαία χρονοσειρά, $m = 4$ για τριμηνιαία κ.ο.κ. Αυτή η μετρική σφάλματος ουσιαστικά συνδυάζει το μέσο απόλυτο σφάλμα (αριθμητής) με μία κοινωνικοποίηση ως προς τη μέθοδο Naïve (παρονομαστής) που μελετήθηκε στην υποενότητα 5.2.1. Δεδομένου πως η χρονοσειρά δεν είναι στάσιμη, το αποτέλεσμα του δείκτη $MASE$ είναι μία τιμή η οποία όταν μικρότερη της μονάδας, η μέθοδος πρόβλεψης έχει κατά μέσο όρο καλύτερη απόδοση από τη μέθοδο Naïve.

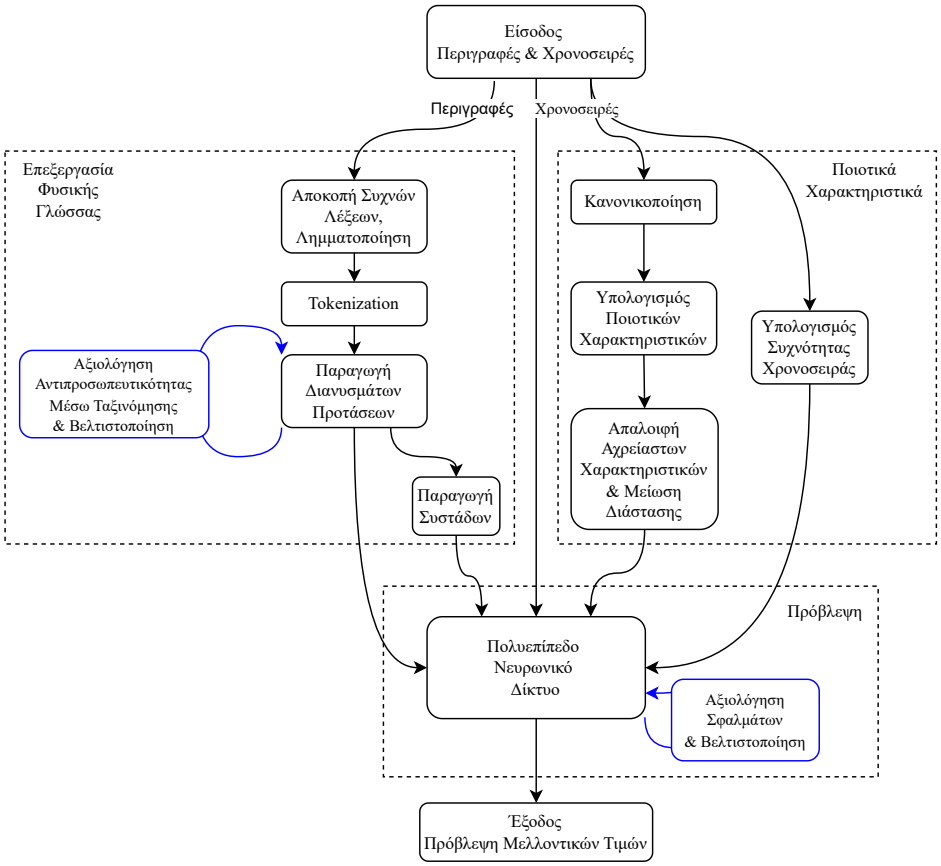
Οι μετρικές $MASE$ και $sMAPE$ είναι ευρέως χρησιμοποιημένες ωστόσο εμπεριέχουν προβλήματα μεροληψίας (η μετρική $sMAPE$ τιμωρεί περισσότερο θετικές αποκλίσεις από ότι αρνητικές), καθώς και υπερεκτίμησης της απόδοσης της μεθόδου Naïve (κατά τον υπολογισμό του $MASE$). Στη δημοσίευση (Semenoglou, Spiliotis, Makridakis, & Assimakopoulos, 2021), προτείνεται τη μετρική $AvgRelMAE$ που βασίζεται στο μέσο απόλυτο σφάλμα και τη μέθοδο Naïve_2, μία παραλλαγή της μεθόδου Naïve, προσαρμοσμένη για εποχιακότητα αν παρατηρείται στη χρονοσειρά, με τη μέθοδο αποσύνθεσης (η ύπαρξη εποχιακότητας κρίνεται από έναν έλεγχο αυτοσυσχέτισης 90%).

$$AvgRelMAE = \exp\left(\frac{1}{N}\sum_{i=1}^{N}\ln\left(\frac{MAE_{Model,i}}{MAE_{\text{Naïve}_{2},i}}\right)\right) \tag{5.21}$$

Κεφάλαιο 6

Πειραματική Διαδικασία

Για να διαπιστωθεί εάν οι λεκτικές περιγραφές των χρονοσειρών είναι δυνατό να συμβάλλουν στην βελτίωση των προβλέψεων, στο κεφάλαιο αυτό παρουσιάζεται η πειραματική διαδικασία που ακολουθήθηκε καθώς και τα αποτελέσματα. Στο σχήμα 6.1, παρουσιάζεται το μοντέλο που αναπτύχθηκε και θα επεξηγηθεί στις παρακάτω ενότητες.



Σχήμα 6.1: *Προτεινόμενη μεθοδολογία για την παραγωγή προβλέψεων (με μπλε χρώμα συμβολίζονται οι διαδικασίες που χρησιμοποιούνται κατά την εκπαίδευση και βελτιστοποίηση του μοντέλου).*

6.1 Συλλογή Δεδομένων

Η πειραματική διαδικασία που ακολουθήθηκε, περιλαμβάνει την ανάλυση χρονοσειρών από μία μεγάλη βάση δεδομένων. Για το σκοπό αυτό επιλέχθηκε η πλατφόρμα Nasdaq Data Link, (παλαιότερα γνωστό ως Quandl) η οποία είναι ένα αποθετήριο χρονοσειρών διαφόρων ειδών χρονοσειρών από πολλαπλές πηγές, όπως χρηματιστήρια, εταιρείες, οργανισμούς και άλλους παρόχους δεδομένων. Τα είδη χρονοσειρών που μπορούν να βρεθούν στο Nasdaq Data Link είναι από ιστορικές τιμές μετοχών, οικονομικούς δείκτες, και δεδομένα εξαγωγών πετρελαίου, μέχρι τιμές ακινήτων, κρυπτονομισμάτων, μετάλλων, και αγροτικών ειδών. Οι χρονοσειρές αυτές, συνοδεύονται από ένα αρχείο μεταδεδομένων, το οποίο επεξηγεί τη φυσική τιμή που αναπαρίσταται από την εκάστοτε χρονοσειρά. Τα δεδομένα από την πλατφόρμα μπορούν να αποκτηθούν με τη χρήση της υπηρεσίας της διεπαφής προγραμματισμού εφαρμογών (Application Programming Interface - API) η οποία διευκολύνει τη διαδικασία, όσον αφορά την «αυτόματη» δημιουργία ενός μεγάλου σε όγκο συνόλου δεδομένων.

6.1.1 Σύνολο Δεδομένων FRED

Για το σκοπό της ανάλυσης των περιγραφών χρονοσειρών για τη βελτίωση των προβλέψεων, από την πλατφόρμα Nasdaq Data Link επιλέχθηκε το σύνολο δεδομένων Federal Reserve Economic Data - FRED. Η βάση δεδομένων FRED, έχει δημιουργηθεί και συντηρείται από την ομοσπονδιακή τράπεζα St. Louis, μία εκ των δώδεκα Ομοσπονδιακών Τραπεζών των Ηνωμένων Εθνών της Αμερικής. Το σύνολο δεδομένων FRED περιέχει μια μεγάλη ποικιλία οικονομικών δεδομένων και στατιστικών στοιχείων. Αυτό το σύνολο δεδομένων είναι ένα από τα πιο δημοφιλή και εκτενή αποθετήρια οικονομικών δεδομένων παγκοσμίως και περιλαμβάνει δεδομένα που καλύπτουν μια ευρεία γκάμα θεμάτων, όπως οι μακροοικονομικοί δείκτες (Ακαθάριστο Εγχώριο Προϊόν - ΑΕΠ, Ανεργία, Πληθωρισμός, Επιτόκια), τα δεδομένα για την αγορά εργασίας (Απασχόληση και ανεργία, μέσος όρος Ωριαίων Αποδοχών, ο Αριθμός ανοικτών θέσεων εργασίας), και άλλα δημοσιονομικά και κυβερνητικά δεδομένα (Δημόσιο χρέος, Δημοσιονομικό έλλειμμα και πλεόνασμα, Φορολογικά έσοδα).

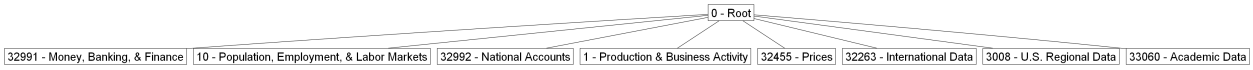
Το πλεονέκτημα του συνόλου δεδομένων FRED, είναι πως πρόκειται για 830,000 αμερικάνικες και διεθνείς χρονοσειρές, οι οποίες πέρα από μία σύντομη περιγραφή, έναν κωδικό, και την ημερομηνία τελευταίας τροποποίησης, περιέχουν έναν αριθμό κατηγορίας στην οποία ανήκουν. Η κατηγορία αυτή μπορεί να χρησιμοποιηθεί μετά από τον κατάλληλο μετασχηματισμό, ως μέτρο σύγκρισης για την ταξινόμηση των χρονοσειρών με βάση τη λεκτική περιγραφή τους, υπάρχει δηλαδή ένα κατηγοριοποιημένο, ground truth σύνολο δεδομένων. Ωστόσο στην πλατφόρμα Nasdaq Data Link, από τις ανακοινωθείσες 830,000 χρονοσειρές του FRED συμπεριλαμβάνονται οι 339,645.

6.1.2 Κατηγορίες Χρονοσειρών στο FRED

Χρησιμοποιώντας την υπηρεσία του API, του FRED (URL, n.d.-b), είναι εφικτό δεδομένου ενός κωδικού μίας χρονοσειράς, να αποκτηθούν με την κλήση fred/series/categories οι κατηγορίες στις οποίες ανήκει η χρονοσειρά. Έτσι, για κάθε μία από τις 339,645 χρονοσειρές που παρουσιάζονται στην πλατφόρμα Nasdaq Data Link, μπορούμε να ανακτήσουμε την κατηγορία στην οποία ανήκουν.

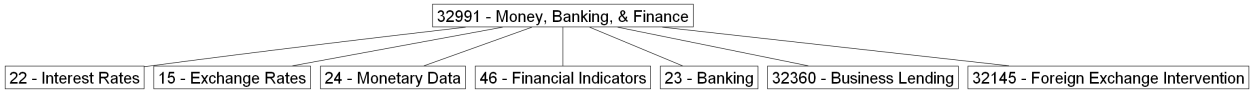
Ωστόσο, το αποτέλεσμα που επιστρέφει η κλήση fred/series/categories αποτελείται από τις «ειδικές» κατηγορίες στις οποίες ανήκει η χρονοσειρά, οι οποίες ξεπερνούν τις 5,500 στον αριθμό, εξού και η αναγκαιότητα του μετασχηματισμού αυτών στις κατηγορίες «γονείς» τους. Η δομή των χρονοσειρών στη βάση FRED, είναι δενδρική. Όλες οι κατηγορίες υπάγονται στην κατηγορία ρίζα με κωδικό 0, ενώ η κάθε κατηγορία περιέχει κάποιες κατηγορίες «παιδιά» οι οποίες με τη σειρά τους αντίστοιχα, περιέχουν και αυτές επιπλέον κατηγορίες.

Με αλλεπάλληλες κλήσεις fred/category/children στο API της βάσης, ξεκινώντας από την κατηγορία ρίζα νούμερο 0, κάνοντας ουσιαστικά μία αναζήτηση κατά πλάτος (Breadth First Search - BFS), ανακαλύπτουμε όλες τις κατηγορίες και τις σχέσεις μεταξύ τους, με τον τρόπο αυτό, μπορούμε να ανάγουμε κάθε ειδική κατηγορία σε μία πό τις 8 «υπερκατηγορίες» που βρίσκονται ακριβώς κάτω από την κατηγορία ρίζα 6.2.



Σχήμα 6.2: *Οι 8 κατηγορίες «παιδιά» της κατηγορίας ρίζα.*

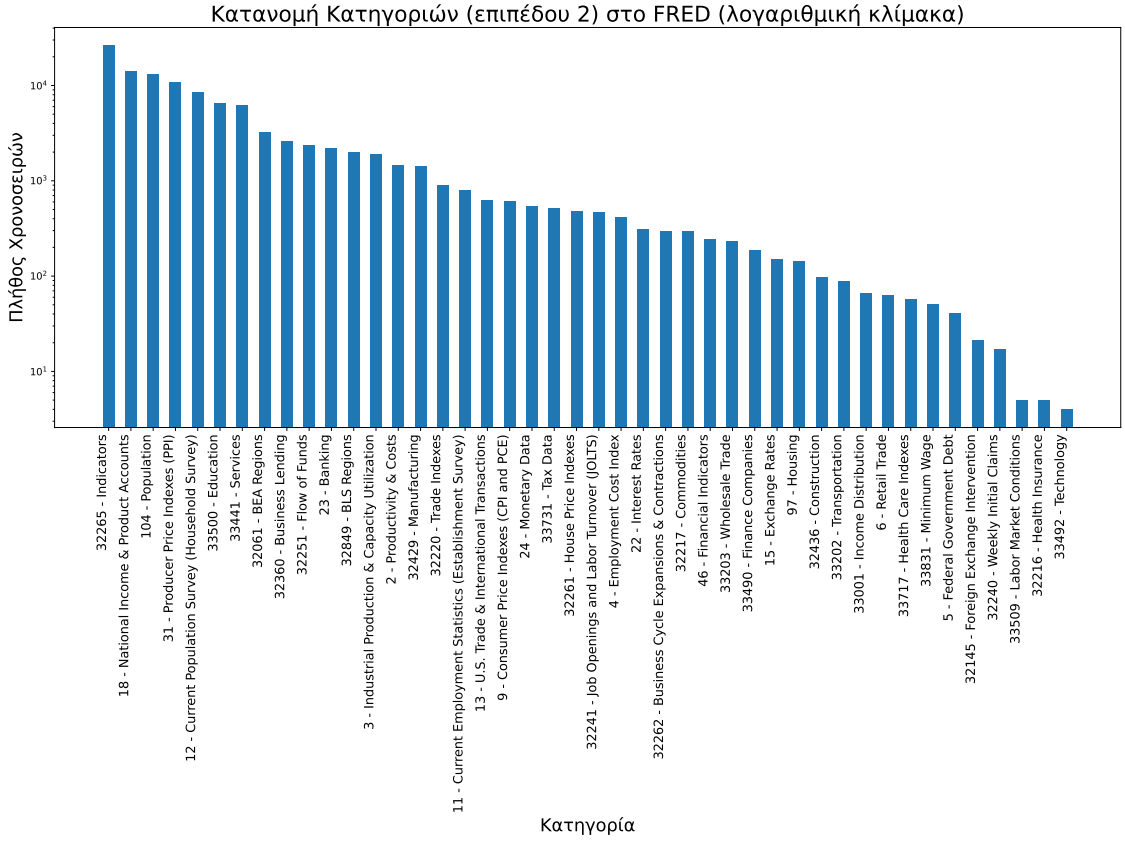
Κάθε μία από αυτές τις κατηγορίες, περιέχει αντίστοιχα τις δικές τις υποκατηγορίες, για παράδειγμα, για την κατηγορία 32991 - Money, Banking, & Finance, παρατηρούμε πως υπάρχουν οι εξής κατηγορίες «παιδιά»:



Σχήμα 6.3: *Οι κατηγορίες «παιδιά» της κατηγορίας 32991 - Money, Banking, & Finance.*

Στην πραγματικότητα, το δέντρο των κατηγοριών παρουσιάζει μέγιστο βάθος 9 το οποίο είναι ιδιαίτερα μεγάλο και μας ωθεί στο να διαλέξουμε ένα πιο γενικό επίπεδο αναπαράστασης (πιο κοντά στην κατηγορία ρίζα). Ένα λογικό επίπεδο κατηγοριών για τη δημιουργία ενός ground truth dataset για τους σκοπούς της ταξινόμησης, είναι το επίπεδο δύο, κρατώντας δηλαδή, τις κατηγορίες παιδιά των 8 κύριων κατηγοριών (δύο επίπεδα κάτω από τη ρίζα του δέντρου), έτσι, όλες οι χρονοσειρές του συνόλου δεδομένων ανήκουν σε μία από 70 κατηγορίες.
Επειδή παρατηρήθηκε πως πολλές χρονοσειρές ανήκουν, πέρα από την κατηγορία που έχει να κάνει με το περιεχόμενο της χρονοσειράς, σε μία κατηγορία τοποθεσίας (εκεί από όπου έχουν συλλεχθεί τα δεδομένα), κρίθηκε αναγκαίο να αφαιρεθούν οι κατηγορίες που σχετίζονται με την τοποθεσία των χρονοσειρών και συνεπώς οι κατηγορίες επιπέδου μειώθηκαν σε 44.

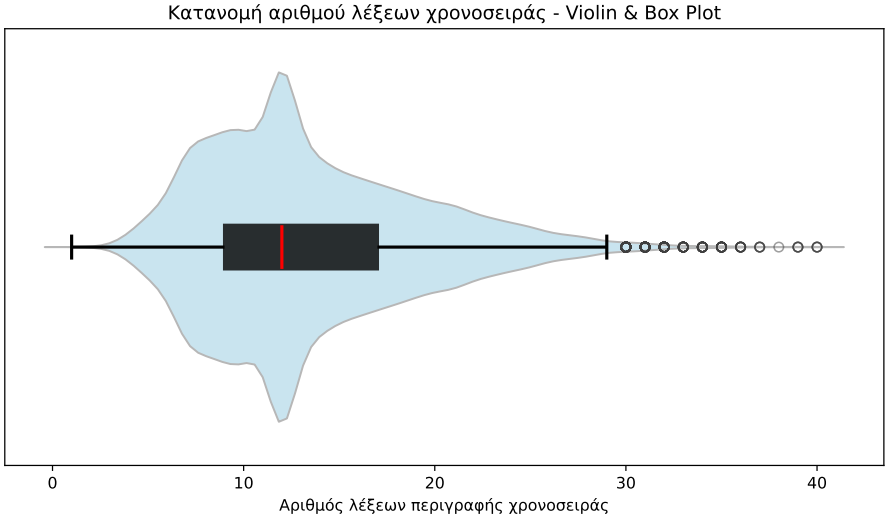
Με τον τρόπο αυτό, γίνεται η ανάθεση των χρονοσειρών στις κατηγορίες επιπέδου 2, και το αποτέλεσμα της κατανομής τους φαίνεται στο γράφημα 6.4.



Σχήμα 6.4: *Κατανομή των κατηγοριών δευτέρου επιπέδου, εξαιρουμένων των κατηγοριών τοποθεσίας.*

6.1.3 Περιγραφές Χρονοσειρών στο FRED

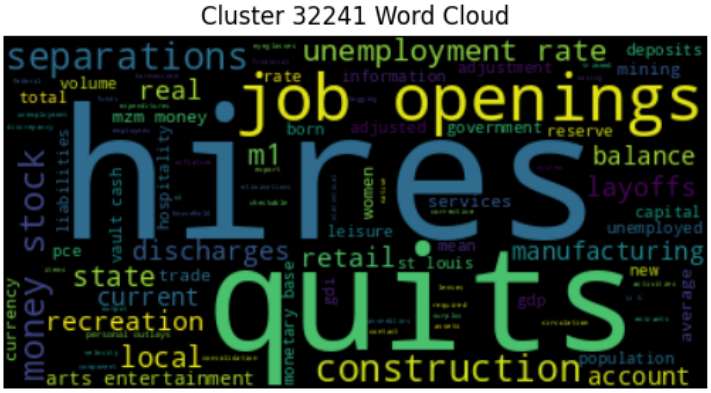
Οι λεκτικές περιγραφές που περιέχει το σύνολο δεδομένων, είναι μικρές περιγραφές και περιέχουν κατά μέσο όρο 13.45 λέξεις. Συγκεκριμένα, η κατανομή του αριθμού των λέξεων των περιγραφών φαίνεται στο γράφημα 6.5, το οποίο επιδεικνύει με τη βοήθεια των box και violin γραφημάτων πως λίγες χρονοσειρές περιέχουν περιγραφές με περισσότερες από 30 λέξεις, ενώ το 50% των χρονοσειρών διαθέτουν 9 με 17 λέξεις.



Σχήμα 6.5: *Περιγραφές χρονοσειρών - Κατανομή αριθμού λέξεων.*

Ένα παράδειγμα μίας λεκτικής περιγραφής στο FRED είναι «Average Hourly Earnings of Production Employees: Durable Goods: Motor Vehicle Parts Manufacturing in Michigan.» όπου παρατηρείται πως δίνεται πληροφορία για το περιεχόμενο της μέτρησης, καθώς και τους τομείς στους οποίους υπάγεται. Επιπλέον, διευκρινίζεται η τοποθεσία της μέτρησης.
Για να αφαιρεθούν λέξεις που είναι κοινές μεταξύ όλων των περιγραφών, όπως αναφέρθηκε στην υποενότητα 3.1.2, από τις περιγραφές αφαιρούνται οι λέξεις που εμφανίζονται συχνά σε ένα κείμενο όπως τα άρθρα, οι σύνδεσμοι και οι αντωνυμίες καθώς και οι χώρες του κόσμου, τα έθνη τις Αμερικής και κάποιες επιπλέον λέξεις σχετιζόμενες με τη συχνότητα των χρονοσειρών, τις μονάδες μέτρησης και τα διαστήματα εμπιστοσύνης. Τέλος, όλες οι λέξεις μετατρέπονται σε πεζά γράμματα για να μην αντιμετωπίζεται κάποια λέξη ως διαφορετική σε περίπτωση που υπάρχει γραμμένη στα κεφαλαία γράμματα και αφαιρούνται τα σημεία στίξης.
Διαβάζοντας όλες τις περιγραφές ανά κατηγορία, είναι εφικτό να δημιουργηθούν κάποια Word Clouds, μια γραφική αναπαράσταση του συνολικού κειμένου που εμφανίζεται σε όλες τις περιγραφές των χρονοσειρών που ανήκουν σε μία κατηγορία, όπου οι λέξεις που εμφανίζονται συχνότερα στο κείμενο αναπαρίστανται με μεγαλύτερη γραμματοσειρά.

Ένα παράδειγμα των λέξεων που περιέχονται σε περιγραφές των χρονοσειρών της κατηγορίας 32241 - Job Openings and Labour Turnover, είναι αυτό που φαίνεται στην εικόνα 6.6. Όπως είναι αναμενόμενο, οι πιο συνήθεις λέξεις είναι hires (προσλήψεις) & quits (παραιτήσεις) αφού πρόκειται για χρονοσειρές για ανοίγματα θέσεις εργασίας.



Σχήμα 6.6: *Word Cloud των λέξεων που εμπεριέχονται στις περιγραφές των χρονοσειρών που ανήκουν στην κατηγορία 32241 - Job Openings and Labour Turnover.*

Τέλος, οι περιγραφές διαχωρίζονται σε tokens με τη βοήθεια του Natural Language Toolkit - NLTK (URL, n.d.-e) για τους απλούς αλγόριθμους, ενώ ο διαχωρισμός για τους μετασχηματιστές είναι σχετικά διαφορετικώς, δεδομένου πως πρέπει να χρησιμοποιηθούν τα κατάλληλα επιπλέον tokens [CLS], [SEP], [PAD] για τον αλγόριθμο BERT/SBERT και και τα [BOS] (Beginning of Sentence) και [EOS] (End of Sentence) για τον αλγόριθμο GPT-2.

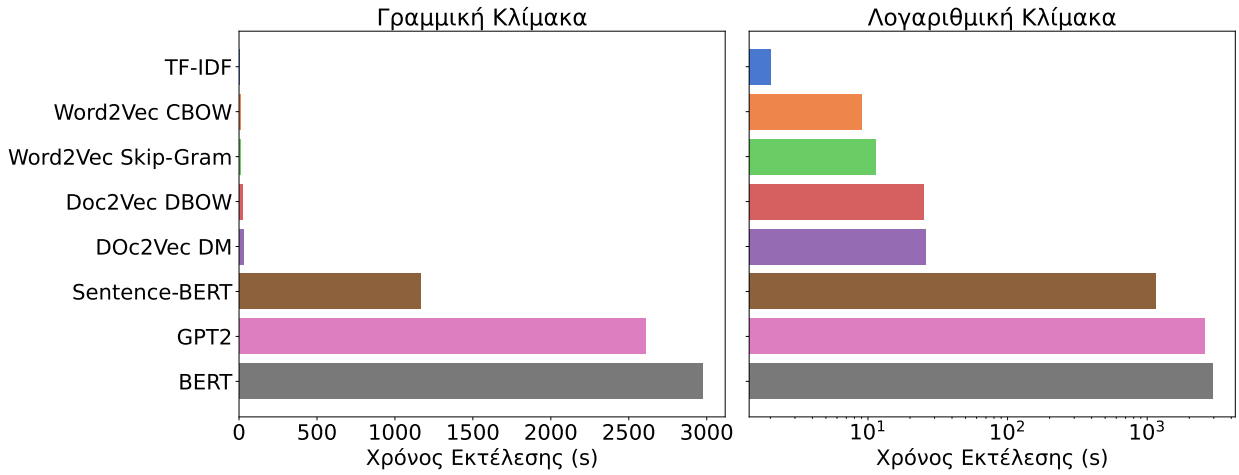
6.2 Παραγωγή Διανυσμάτων Προτάσεων

Με τους αλγόριθμους παραγωγής διανυσμάτων προτάσεων που μελετήθηκαν σε βάθος στο κεφάλαιο 3, παράγονται τα διανύσματα προτάσεων που έχουν ως σκοπό να χρησιμοποιηθούν για την αναπαράσταση των χρονοσειρών. Οι αλγόριθμοι TF-IDF, Sentence-Bert και Doc2Vec (DBOW & DM), παράγουν άμεσα διανύσματα προτάσεων, ενώ έμμεσα, μέσω μετατροπής από διανύσματα λέξεων σε διανύσματα προτάσεων, παράγονται με τους αλγόριθμους Word2Vec (Skip-Gram & CBOW), BERT και GPT-2.

Για τους μετασχηματιστές, χρησιμοποιήθηκαν προ-εκπαιδευμένα μοντέλα σε ξένα σύνολα δεδομένων, τα οποία στη συνέχεια προσαρμόστηκαν στα δεδομένα των λεκτικών περιγραφών των χρονοσειρών. Για το GPT-2, χρησιμοποιήθηκε ένα προ-εκπαιδευμένο GPT-2 στο σύνολο δεδομένων WebText στη μορφή GPT-2 small (117 εκατομμύρια παραμέτρους, 12 μπλοκ αποκωδικοποιητή-μετασχηματιστή και μέγεθος διανύσματος λέξης κάθε token τις 768 διαστάσεις). Το μοντέλο Sentence-BERT εκπαιδεύτηκε σε ένα σύνολο δεδομένων ενός δισεκατομμυρίου ζευγαριών από προτάσεις, που εξάχθηκαν από 32 σύνολα δεδομένων όπως τα περιεχόμενα του Wikihow, Stack Exchange, Yahoo Answers και παράγει διανύσματα προτάσεων 768 διαστάσεων. Τέλος, για την προ-εκπαίδευση του μοντέλου BERT, χρησιμοποιήθηκε ένα σύνολο δεδομένων που αποτελείται από τη Βικιπαίδεια και το Toronto Book Corpus (7,000 βιβλία) και τα διανύσματα λέξεων που υπολογίζει στην έξοδο είναι διάστασης 768. Τα μοντέλα αυτά, είναι διαθέσιμα στο αποθετήριο προ-εκπαιδευμένων μοντέλων και datasets Hugging Face (URL, n.d.-d).

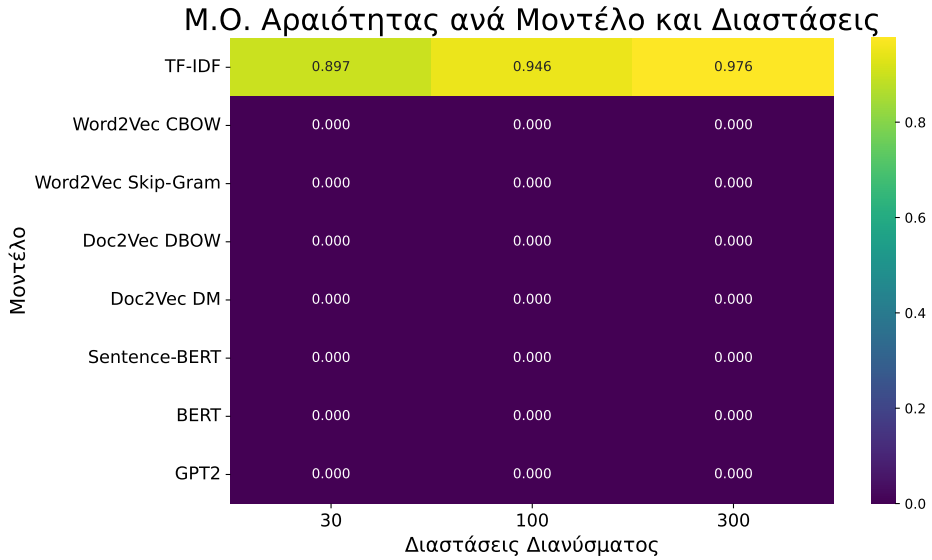
Τα διανύσματα παράχθηκαν σε τρεις εκδοχές διαστάσεων (30, 100, 300), με σκοπό να εξεταστεί στη συνέχεια η επίδραση της διάστασης στην επίλυση του προβλήματος. Για τα μοντέλα των μετασχηματιστών που δεν μπορούν να δεχθούν ως είσοδο την επιθυμητή διάσταση, πραγματοποιείται η μείωση των διαστάσεων με τη μέθοδο της ανάλυσης σε κύριες συνιστώσες (5.1.4).

Ενώ ο χρόνος εκτέλεσης των μοντέλων παραγωγής διανυσμάτων προτάσεων είναι ιδιαίτερα χαμηλός για τα απλά μοντέλα, δεν ισχύει το ίδιο για τα μοντέλα των μετασχηματιστών. Συγκεκριμένα ενώ η απλή τεχνική των συχνοτήτων εκτελείται κατά μέσο όρο για 2 δευτερόλεπτα, το μοντέλο BERT, χρειάζεται πάνω από 43 λεπτά για τον ίδιο ακριβώς όγκο δεδομένων. Επιπλέον, παρατηρούμε πως πράγματι η εκδοχή CBOW του αλγορίθμου Word2Vec είναι πιο γρήγορη από την εκδοχή Skip-Gram όπως εξηγήθηκε στο θεωρητικό κομμάτι. Στο γράφημα 6.7 βρίσκονται τα αποτελέσματα ως προς το χρόνο εκτέλεσης των διαφορετικών μοντέλων.



Σχήμα 6.7: *Χρόνος εκτέλεσης για την παραγωγή διανυσμάτων προτάσεων διάστασης 100.*

Επιπλέον, μια σημαντική παρατήρηση είναι πως η τεχνική TF-IDF, λόγω της απλοϊκής της φύσης όπου μετρά και κανονικοποιεί τις συχνότητες εμφάνισης των λέξεων, παράγει ιδιαίτερα αραιά διανύσματα. Στο σχήμα 6.8, φαίνεται ένας χάρτης θερμότητας για την οπτικοποίηση της μέσης αραιότητας ανά μοντέλο παραγωγής και διάσταση διανύσματος. Το μοντέλο TF-IDF, όπως είναι αναμενόμενο παράγει ιδιαίτερα αραιά διανύσματα, τα οποία είναι ολοένα και πιο αραιά όσο αυξάνονται οι διαστάσεις των διανυσμάτων. Εν αντιθέσει, όλα τα υπόλοιπα μοντέλα παράγουν ιδιαίτερα πυκνά διανύσματα.



Σχήμα 6.8: *Μέση αραιότητα ανά μοντέλο και διάσταση διανύσματος - αραιότητα του TF-IDF.*

6.3 Ταξινόμηση Χρονοσειρών Βάσει Διανυσμάτων Προτάσεων

Για να διαπιστωθεί αν τα διανύσματα που παράχθηκαν, αναπαριστούν με σωστό τρόπο τις περιγραφές των χρονοσειρών, η πειραματική διαδικασία συνεχίζει, προσπαθώντας να λύσει το πρόβλημα της ταξινόμησης. Συγκεκριμένα, από το σύνολο δεδομένων δημιουργείται ένα σύνολο δεδομένων που περιέχει 11 κατηγορίες που περιέχουν περισσότερες από 2,000 χρονοσειρές. Για κάθε μία από τις επιλεγμένες κατηγορίες, στο σύνολο δεδομένων εμπεριέχονται κατά μέγιστο 3,000 κατηγορίες (και κατ' ελάχιστο 2,000). Αναπτύσσεται συνεπώς ένα σχετικά ισορροπημένο dataset με κατά μέσο όρο 2,828 χρονοσειρές ανά κατηγορία. Ως ετικέτες (labels) λαμβάνονται οι κατηγορίες όπως αυτές έχουν ανατεθεί στο FRED ενώ τα χαρακτηριστικά features είναι οι περιγραφές των χρονοσειρών σε μορφή διανύσματος. Η διάσπαση σε σύνολο εκπαίδευσης και σύνολο αξιολόγησης (train-test split γίνεται επιλέγοντας τυχαία 80% των αρχικών δεδομένων για την εκπαίδευση, και 20% δεδομένα (6,200 δείγματα) τα οποία το μοντέλο δεν «βλέπει» ποτέ κατά τη διάρκεια της εκπαίδευσης και χρησιμοποιούνται μόνο για την αξιολόγησή του. Μέσω της ταξινόμησης και της αξιολόγησης των αποτελεσμάτων, μπορεί να διαπιστωθεί κατά πόσο οι λεκτικές περιγραφές των χρονοσειρών αναπαρίστανται σωστά η όχι από τα διανύσματα που παράχθηκαν.
Για την ταξινόμηση, χρησιμοποιούνται οι αλγόριθμοι k-NN και Random Forest ενώ για την αξιολόγηση, υπολογίζονται οι μετρικές της ορθότητας (accuracy), της ευαισθησίας (sensitivity/recall), της ακρίβειας (precision) και του F1-Score.

6.3.1 Ταξινόμηση - Αλγόριθμος k-NN

Ο αλγόριθμος k-NN είναι ένας απλός αλγόριθμος, για τον οποίο μελετήθηκε μόνο η επίδραση του αριθμού των k κοντινότερων γειτόνων. Σε συνδυασμό με την εναλλαγή της διάστασης των διανυσμάτων προτάσεων και το μοντέλο από το οποίο παράχθηκαν, έχουμε τον εξής χώρο αναζήτησης:

Ταξινόμηση με k-NN	
Υπερπαράμετροι	**Παραλλαγές**
Μοντέλο Παραγωγής Διανύσματος	Doc2Vec DBOW, Doc2Vec DM, Word2Vec Skip-Gram, Word2Vec CBOW, TF-IDF, Sentence-BERT, GPT-2, BERT
Διάσταση Διανύσματος	30, 100, 300
Αριθμός Γειτόνων	1, 3, 5, ..., 20

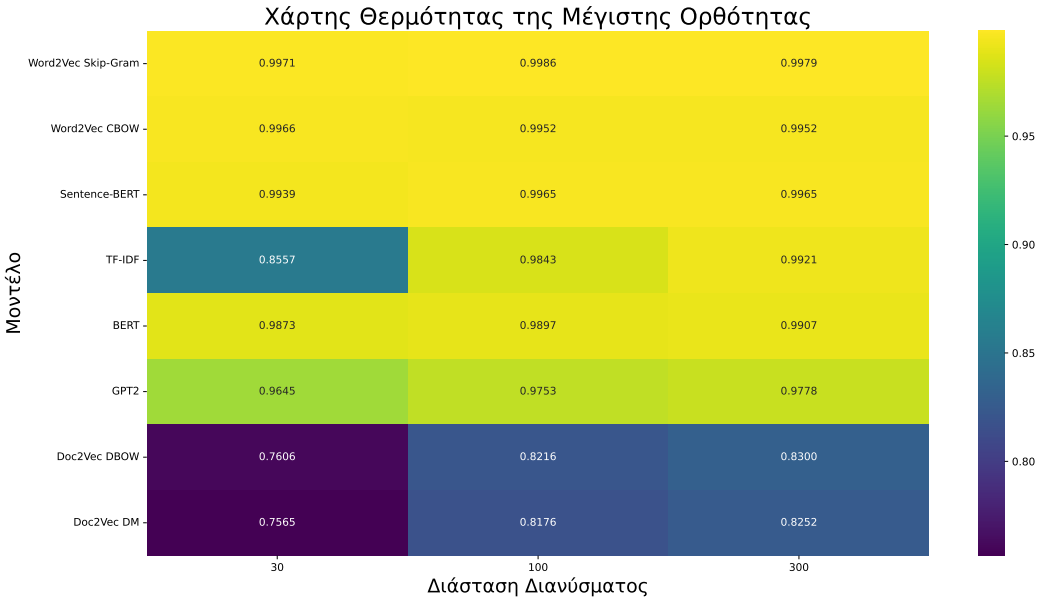
Πίνακας 6.1: *Υπερπαράμετροι Ταξινόμησης με k-NN.*

Αποτελέσματα

Σε γενικές γραμμές τα περισσότερα μοντέλα παραγωγής διανυσμάτων προτάσεων έχουν καλές επιδόσεις, με το μοντέλο Word2Vec Skip-Gram, να υπερτερεί σε όλες τις μετρικές, ιδιαίτερα κοντά στην παραλλαγή CBOW καθώς και στο Sentence-BERT.

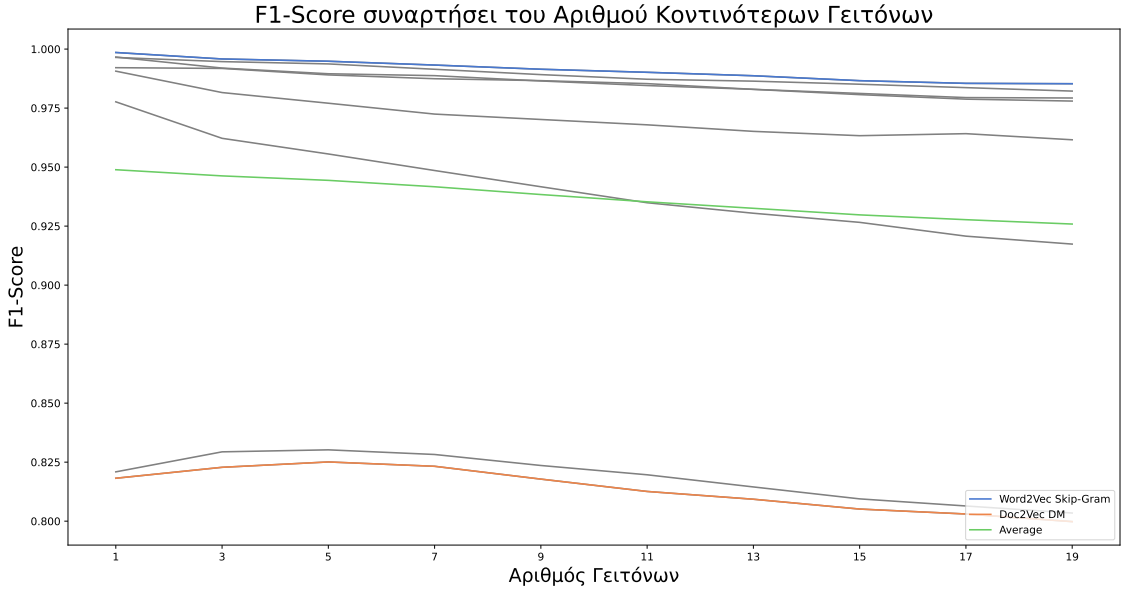
Συγκεκριμένα η ταξινόμηση των διανυσμάτων που παράχθηκαν από το Word2Vec Skip-Gram, επιτυγχάνει για διάνυσμα διάστασης 100 και $k = 3$ κοντινότερους γείτονες, Accuracy 0.9985, Precision 0.9985 και Recall 0.9984, αποδεικνύοντας πως με τις περιγραφές των χρονοσειρών σε μορφή διανυσμάτων μπορεί να επιτευχθεί εξαιρετική ταξινόμηση.

Παρακάτω παρατίθενται κάποια γραφήματα που επιδεικνύουν το ρόλο της διάστασης των διανυσμάτων, και τον αριθμό κοντινότερων γειτόνων, για το αποτέλεσμα της ορθότητας και του F1-Score.
Επιπλέον, παρατηρούμε πως τα μοντέλα Word2Vec (και οι δύο εκδοχές του) καθώς και το Sentence-BERT επιτυγχάνουν ιδιαίτερα καλά αποτελέσματα για οποιοδήποτε μέγεθος διανύσματος. Το μοντέλο TF-IDF επιφέρει χαμηλά αποτελέσματα στην εκδοχή των λίγων διαστάσεων ενώ οι δύο εκδοχές Doc2Vec δεν έχουν καλή απόδοση και ξεπερνούν το 80% στην ορθότητα.



Σχήμα 6.9: *Χάρτης Θερμότητας της Μέγιστης Ορθότητας ανά Μοντέλο και Διάσταση Διανύσματος.*

Στο σχήμα 6.9 διακρίνονται οι μέγιστες τιμές τις ορθότητας ανάλογα με το μοντέλο παραγωγής διανυσμάτων και το μέγεθος αυτών. Εκεί επιβεβαιώνεται πως η εκδοχή CBOW του μοντέλου Word2Vec είναι καλύτερη 3.3.1 από την εκδοχή CBOW για την παραγωγή διανυσμάτων μικρών προτάσεων όπως οι περιγραφές των χρονοσειρών (με κατά μέσο όρο 13 λέξεις πριν την αποκοπή των περιττών λέξεων).



Σχήμα 6.10: *Επίδραση του αριθμού κοντινότερων γειτόνων στο αποτέλεσμα του F1-Score. Με χρώμα παρουσιάζονται η καλύτερη, η χειρότερη και η μέση επίδοση μεταξύ των μοντέλων παραγωγής διανυσμάτων ενώ έχουν σχεδιαστεί για σύγκριση και οι επιδόσεις των άλλων μοντέλων σε γκρι χρώμα.*

Τέλος, η επίδραση του αριθμού των γειτόνων οι οποίοι χρησιμοποιούνται για την ταξινόμηση

φαίνεται να έχει μεγάλη επίπτωση στα αποτελέσματα της ορθότητας της ταξινόμησης, καθώς εξαιρουμένων των διανυσμάτων του TF-IDF μικρής διάστασης παρατηρείται πτώση όσο αυτός αυξάνεται.

6.3.2 Ταξινόμηση - Αλγόριθμος Random Forest

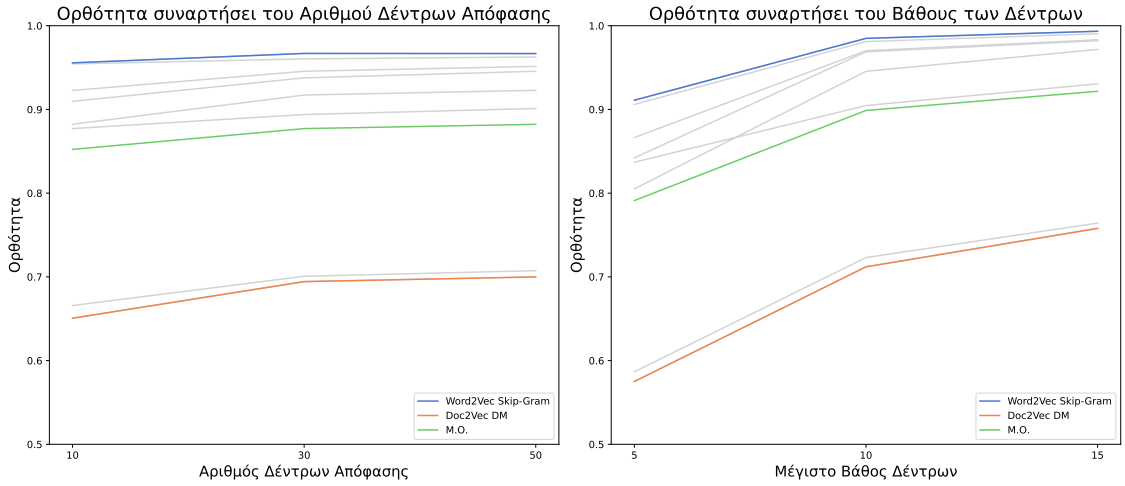
Για τον αλγόριθμο Random Forest, οι υπερπαράμετροι που χρησιμοποιήθηκαν είναι το μέγιστο βάθος και ο αριθμός των δέντρων που αποτελούν το τυχαίο δάσος. Ο χάρτης αναζήτησης είναι ο εξής:

Ταξινόμηση με Random Forest	
Υπερπαράμετροι	**Παραλλαγές**
Μοντέλο Παραγωγής Διανύσματος	Doc2Vec DBOW, Doc2Vec DM, Word2Vec Skip-Gram, Word2Vec CBOW, TF-IDF, Sentence-BERT, GPT-2, BERT
Διάσταση Διανύσματος	30, 100, 300
Αριθμός Δέντρων	10, 30, 50
Μέγιστο Βάθος Δέντρων	5, 10, 15

Πίνακας 6.2: *Υπερπαράμετροι Ταξινόμησης με Random Forest.*

Αποτελέσματα

Τα αποτελέσματα είναι παρόμοια με αυτά του αλγορίθμου k-NN, οι δύο εκδοχές του Word2Vec επιτυγχάνουν τις καλύτερες επιδόσεις με βάσει όλες τις μετρικές, με τη σημαντική άνοδο του GPT-2 που είναι στην τρίτη θέση κατά μέσο όρο. Η επίδραση του αριθμού των δέντρων καθώς και αυτή του βάθους των δέντρων απόφαση, είναι θετική, παρατηρώντας πως όσο αυξάνονται επιφέρουν καλύτερα αποτελέσματα, ωστόσο παρατηρείται πως για τον αριθμό των δέντρων, η αύξηση από τα 30 στα 50 δεν επιφέρει σημαντικά καλύτερα αποτελέσματα.

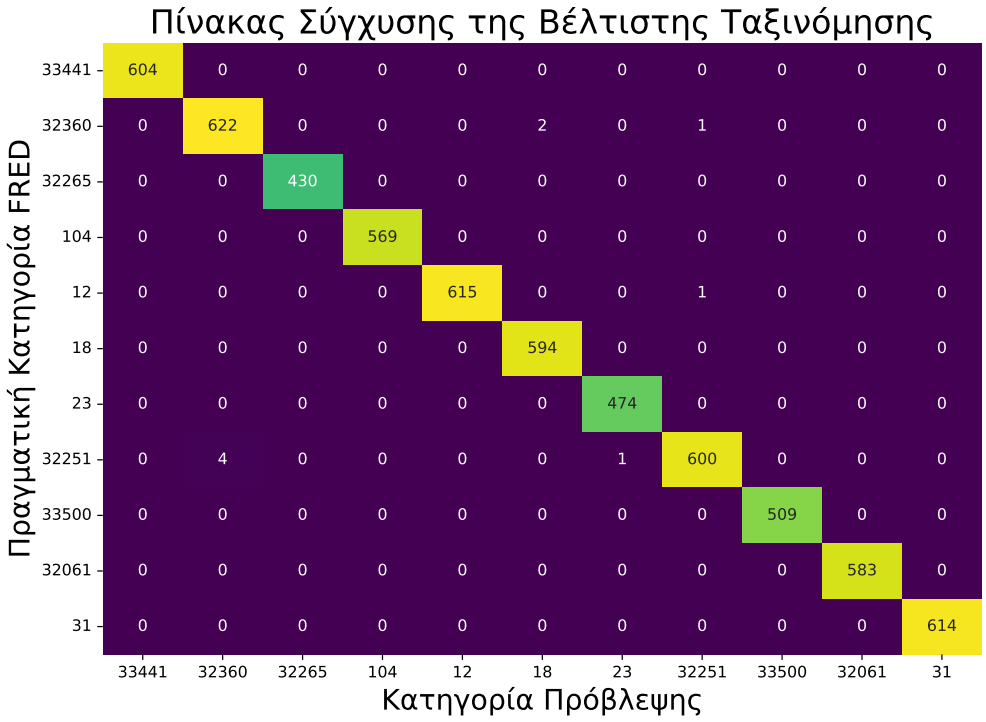


Σχήμα 6.11: *Επίδραση του αριθμού καθώς και του μέγιστου βάθους των δέντρων που αποτελούν το τυχαίο δάσος του αλγορίθμου Random Forest. Στο διάγραμμα παρουσιάζεται η καλύτερη επίδοση του Word2Vec Skip-Gram, η χειρότερη του Doc2Vec DM και ο μέσος όρος, ενώ για σύγκριση έχουν σχεδιαστεί και οι υπόλοιπες γραμμές σε γκρι χρώμα.*

6.3.3 Συνολικά Αποτελέσματα Ταξινόμησης - Παρατηρήσεις

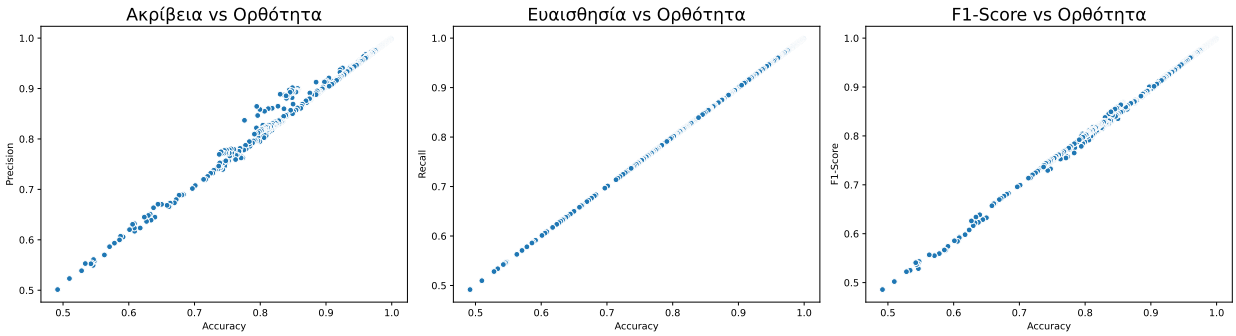
Τα αποτελέσματα της ταξινόμησης των χρονοσειρών βάσει των διανυσμάτων που παράχθηκαν είναι ιδιαίτερα καλά, οι τιμές τόσο της Ορθότητας, όσο και της ευαισθησίας και της ακρίβειας τείνουν ασυμπτωτικά στο 1, πράγμα που σημαίνει ότι τα διανύσματα που έχουν παραχθεί από τα μοντέλα επεξεργασίας της φυσικής γλώσσας παρέχουν αρκετή πληροφορία, για να επιτευχθεί η σωστή κατηγοριοποίηση.
Σε γενικές γραμμές, και με τις δύο μεθόδους ταξινόμησης, τα βέλτιστα αποτελέσματα προέρχονται από τον αλγόριθμο Word2Vec Skip-Gram, με την εκδοχή CBOW να έχει ελάχιστα χειρότερη επίδοση. Οι τεχνικές Sentence-BERT και GPT-2 καταλαμβάνουν την τρίτη θέση, ενώ στον αντίποδα, οι τεχνικές Doc2Vec και TF-IDF φαίνεται πως δεν έχουν καλά αποτελέσματα.
Ένα άλλο σημαντικό ζήτημα του πειράματος αυτού είναι ο προσδιορισμός της σημαντικότητας του μεγέθους των διαστάσεων των διανυσμάτων, που ενώ φαίνεται πως επηρεάζει πολύ το TF-IDF και Doc2Vec, δείχνει να μην έχει σημαντικό ρόλο για τα υπόλοιπα μοντέλα δοκιμάζοντας για 30, 100 και 300 διαστάσεις.

Παρακάτω παρατίθεται ο πίνακας σύγχυσης για την ταξινόμηση από το συνδυασμό με την καλύτερη ορθότητα, ακρίβεια και ευαισθησία καθώς μόνο 8 δείγματα ταξινομήθηκαν λανθασμένα. Αυτό το αποτέλεσμα προκύπτει από τα διανύσματα του μοντέλου Word2Vec Skip-Gram, για και ως παράμετρο $k = 3$ κοντινότερο γείτονα και διάνυσμα $\vec{x}$ διάστασης $dim(\vec{x}) = 100$.



Σχήμα 6.12: *Πίνακας σύγχυσης για τη βέλτιστη ταξινόμηση, τα IDs των κατηγοριών είναι αυτά που δίνονται από το FRED και χρησιμοποιήθηκαν ως ετικέτες.*

Μία ακόμα παρατήρηση είναι πως η ακρίβεια, η ευαισθησία και το F1-Score είναι ιδιαίτερα συσχετισμένα (με τη συσχέτιση τους να πλησιάζει την τιμή 1), το οποίο σημαίνει πως το σύνολο δεδομένων είναι ισορροπημένο και πως τα καλά μοντέλα επιτυγχάνουν υψηλά σκορ και στις τέσσερις μετρικές παράλληλα. Παρακάτω φαίνονται τα διαγράμματα διασποράς μεταξύ της ορθότητας και των τριών μετρικών.



Σχήμα 6.13: *Πίνακας σύγχυσης για τη βέλτιστη ταξινόμηση, τα IDs των κατηγοριών είναι αυτά που δίνονται από το FRED και χρησιμοποιήθηκαν ως ετικέτες.*

Ταξινόμητης	Μοντέλο Διαν.	Μέγεθος Διαν.	Παράμετροι	Accuracy
k-NN	Wd2Vc Skip-Gram	100	Γείτονες:3	0.9985
k-NN	Wd2Vc CBOW	30	Γείτονες:1	0.9966
k-NN	BERT	100	Γείτονες:3	0.9964
Random F.	Wd2Vc Skip-Gram	100	#Δέντρων:50, Μέγιστο Βάθος:15	0.9958
Random F.	Wd2Vc CBOW	100	#Δέντρων:30, Μέγιστο Βάθος:15	0.9934
Random F.	GPT-2	100	#Δέντρων:50, Μέγιστο Βάθος:15	0.9905

Πίνακας 6.3: *Βέλτιστα Αποτελέσματα Ταξινόμησης.*

6.4 Συσταδοποίηση Χρονοσειρών Βάσει Διανυσμάτων Προτάσεων

Όπως φαίνεται στο σχήμα 6.1, έπειτα από την παραγωγή των διανυσμάτων, δημιουργείται ακόμα ένα χαρακτηριστικό, αυτό της συστάδας στην οποία βρίσκεται η περιγραφή, που ουσιαστικά υπολογίζει την κατηγορία στην οποία ανήκει η χρονοσειρά με βάση την περιγραφή της. Για τη συσταδοποίηση, χρησιμοποιούμε το ίδιο dataset, που χρησιμοποιήθηκε για την ταξινόμηση, με τη διαφορά πως επειδή πρόκειται για μία μέθοδο μη επιβλεπόμενης μάθησης, στους αλγορίθμους συσταδοποίησης k-Means, Agglomerative Hierarchical Clustering και DBSCAN, παρέχονται ως είσοδοι μόνο τα διανύσματα προτάσεων (και όχι οι κατηγορίες στις οποίες ανήκουν).

6.4.1 Συσταδοποίηση - Αλγόριθμοι k-Means, Agglomerative Hierarchical και DBSCAN

Για το σκοπό της συσταδοποίησης, δοκιμάστηκαν οι τεχνικές k-Means, Agglomerative Hierarchical και DBSCAN, για τις οποίες μεταβάλλονται οι υπερπαράμετροι που βλέπουμε στον παρακάτω πίνακα 6.4. Η είσοδος αποτελείται από το dataset των 11 κατηγοριών, και για την αξιολόγηση των συστάδων χρησιμοποιείται ο δείκτης ARI για τον οποίο ως συσταδοποίηση σύγκρισης έχει τεθεί η συσταδοποίηση που έχει γίνει από τους ειδικούς του FRED, και το σκορ Silhouette το οποίο υπολογίζει την ομοιότητα και την ανομοιότητα ανάμεσα στα clusters. Για τα μοντέλα k-Means και Agglomerative Hierarchical Clustering, ζητήθηκε η παραγωγή 11 συστάδων (για να μπορέσει να γίνει η σύγκριση με το FRED), ενώ για το μοντέλο DBSCAN που δεν επιτρέπει την άμεση εισαγωγή κάποιου αριθμού συστάδων, αγνοήθηκαν όσα αποτελέσματα παράχθηκαν που περιείχαν κάτω από 11 συστάδες.

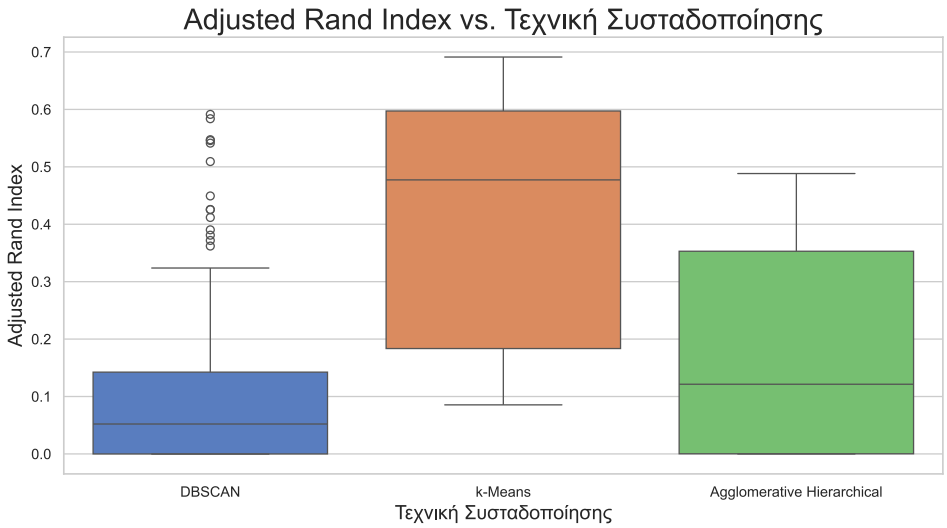
Γενικά για τα τρία μοντέλα	
Υπερπαράμετροι	**Παραλλαγές**
Μοντέλο Παραγωγής Διανύσματος	Doc2Vec DBOW, Doc2Vec DM, Word2Vec Skip-Gram, Word2Vec CBOW, TF-IDF, Sentence-BERT, GPT-2, BERT
Διάσταση Διανύσματος	30, 100, 300
Συσταδοποίηση με k-Means	
Υπερπαράμετροι	**Παραλλαγές**
Ανοχή για Σύγκλιση	0.0001, 0.001, 0.01
Συσταδοποίηση με Agglomerative Hierarchical Clustering	
Υπερπαράμετροι	**Παραλλαγές**
Μετρική Απόστασης	euclidean, manhattan
Συσταδοποίηση με DBSCAN	
Υπερπαράμετροι	**Παραλλαγές**
Ελάχιστος αριθμός γειτόνων	3, 5, 7
EPS (εμβέλεια γειτονίας)	0.3, 0.4, 0.5, 0.6

Πίνακας 6.4: *Υπερπαράμετροι Συσταδοποίησης.*

6.4.2 Αποτελέσματα

ARI

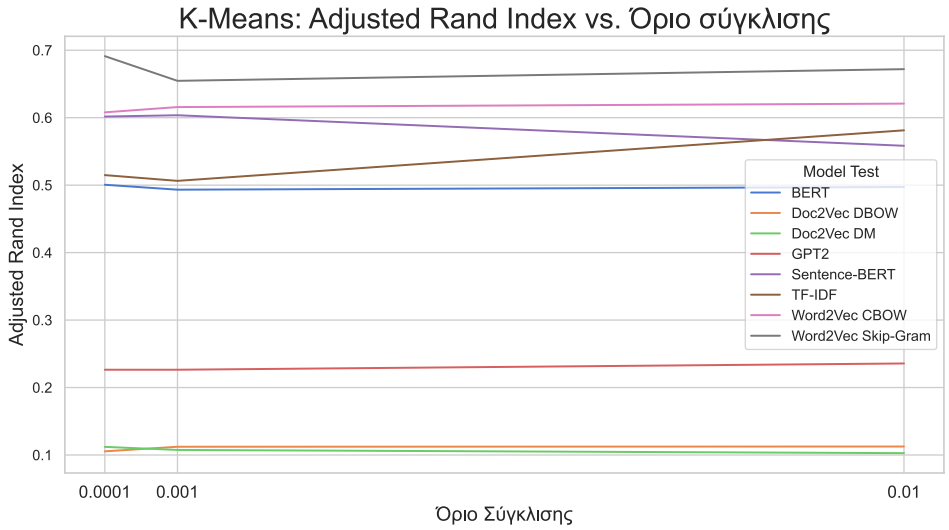
Τα αποτελέσματα της συσταδοποίησης δείχνουν πως ο πιο απλός k-Means επιτυγχάνει ιδιαίτερα καλά αποτελέσματα με τη μετρική ARI.



Σχήμα 6.14: *Αξιολόγηση ARI για τις τρεις τεχνικές συσταδοποίησης.*

Στο γράφημα 6.14 παρατηρείται το εύρος των τιμών ARI, όταν μεταβάλλονται οι παράμετροι που παρουσιάζονται στον πίνακα 6.4.
Η τεχνική k-Means είναι αυτή που επιτυγχάνει τη μέγιστη τιμή 0.691 ARI, η οποία κυμαίνεται από το -1 στο 1 και υποδηλώνει τη συμφωνία της συσταδοποίησης που παράγει ο k-Means με τη συσταδοποίηση από τους ειδικούς του FRED. Αναλύοντας την επίδοση

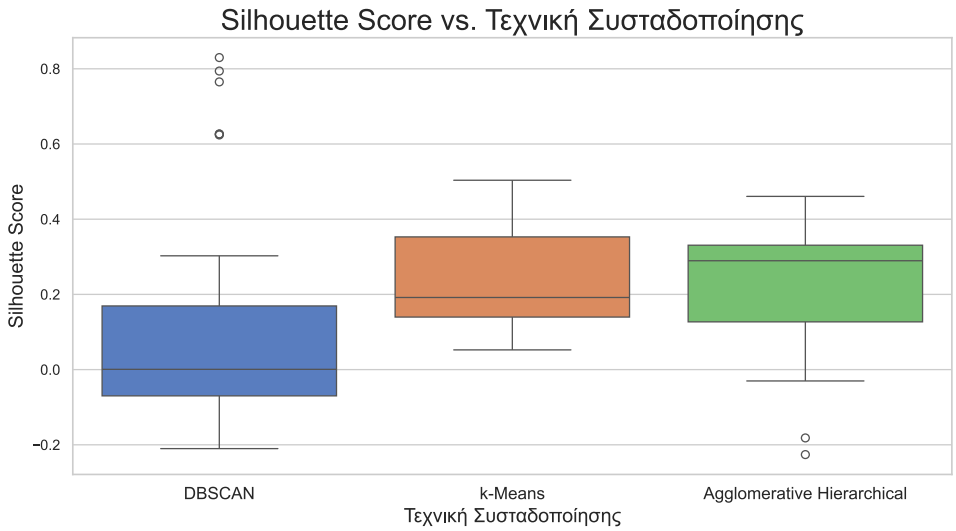
του k-Means, φαίνεται πως το μέγεθος των διανυσμάτων δεν παίζει ιδιαίτερο ρόλο καθώς η αύξηση ή η μείωση του δεν μπορεί να συσχετιστεί με καλύτερα ή χειρότερα αποτελέσματα. Παρατηρώντας τη συμπεριφορά του ARI για διαφορετικές τιμές του ορίου σύγκλισης και του μεγέθους των διανυσμάτων, δεν μπορεί να εξαχθεί κάποιο συμπέρασμα. Ωστόσο στο γράφημα 6.15, φαίνεται πως την πραγματική επίδραση στα αποτελέσματα την έχει το μοντέλο παραγωγής διανυσμάτων και όχι η υπερπαράμετρος του ορίου σύγκλισης ή του μεγέθους των διανυσμάτων. Οι τιμές ARI ανά μοντέλο παραμένουν σχεδόν στάσιμες όταν μεταβάλλεται το όριο σύγκλισης.



Σχήμα 6.15: *Αξιολόγηση ARI για τον k-Means: Επίδραση του μουτέλου παραγωγής διανυσμάτων.*

Silhouette Score

Αρκετά διαφορετικά αποτελέσματα παρουσιάζονται για τη μετρική Silhouette Score, η οποία μετρά την ομοιότητα ενός στοιχείου με άλλα στοιχεία στη συστάδα στην οποία έχει ανατεθεί, σε σύγκριση με στοιχεία άλλων συστάδων.
Και οι τρεις τεχνικές συσταδοποίησης, επιτυγχάνουν μέγιστα Silhouette Score, με τα διανύσματα TF-IDF, το οποίο είναι αναμενόμενο αφού είναι διανύσματα ιδιαίτερα αραιά, και συνεπώς ευνοούν την απόσταση μεταξύ των σημείων και την καλύτερη αξιολόγηση της συσταδοποίησης από το Silhouette Score.



Σχήμα 6.16: *Αξιολόγηση ARI για τις τρεις τεχνικές συσταδοποίησης.*

Η τεχνική DBSCAN ξεχωρίζει στη μετρική Silhouette, επειδή είναι η μόνη που έχει την ευχέρεια να επιλέξει έναν μεγάλο αριθμό συστάδων, και συνεπώς μπορεί να δημιουργήσει πολλές μικρές συστάδες όπου τα στοιχεία μεταξύ τους είναι όμοια και αρκετά απομονωμένα από τα άλλα των άλλων συστάδων.

6.4.3 Συνολικά Αποτελέσματα Συσταδοποίησης - Παρατηρήσεις

Όσον αφορά τα μοντέλα που παράγουν τα διανύσματα, όπως είναι αναμενόμενο, υψηλό Silhouette σκορ, επιτυγχάνουν τα διανύσματα TF-IDF, καθώς πρόκειται για αραιά (sparse) διανύσματα τα οποία ευνοούν την απομόνωση των συστάδων όπως παρατηρήθηκε στο σχήμα 6.8. Βέλτιστο δείκτη ARI, επιτυγχάνει το μοντέλο Word2Vec Skip-Gram, το οποίο επίσης αναδείχθηκε ιδιαίτερα καλό στην ταξινόμηση των περιγραφών.
Μεταξύ των αλγορίθμων συσταδοποίησης, ο k-Means ξεχωρίζει και αναδείχθηκε ιδιαίτερα ικανός για την παραγωγή συστάδων για το συγκεκριμένο σκοπό.
Τέλος, σημειώνεται πως η χαμηλότερη διάσταση διανυσμάτων δε φαίνεται να επηρεάζει σημαντικά την αναπαράσταση των προτάσεων.
Στον πίνακα 6.5, φαίνονται οι βέλτιστες τιμές ARI και Silhouette, καθώς και οι παράμετροι που οδήγησαν σε αυτό.

Μοντέλο Συσταδ.	Μοντέλο Διαν.	Μέγεθος Διαν.	Παράμετροι	ARI	Silhouette
k-Means	Wd2Vc Skip-Gram	300	Σύγκλιση:0.001	**0.691**	0.389
k-Means	Wd2Vc CBOW	30	Σύγκλιση:0.01	**0.620**	0.393
k-Means	BERT	300	Σύγκλιση:0.001	**0.603**	0.17
DBSCAN	TF-IDF	30	EPS:0.3, Min:7	0.32	**0.830**
k-Means	TF-IDF	30	Σύγκλιση:0.001	0.513	**0.5036**
Hierarchical	Doc2Vec DM	300	Αποστ: Manhattan	0.025	**0.447**

Πίνακας 6.5: *Βέλτιστα Αποτελέσματα συσταδοποίησης.*

6.5 Προεπεξεργασία Χρονοσειρών

Έχοντας ολοκληρώσει το πειραματικό κομμάτι της επεξεργασίας της φυσικής γλώσσας, παράγοντας αντιπροσωπευτικά διανύσματα για τις περιγραφές των χρονοσειρών, το επόμενο βήμα είναι ο υπολογισμός των ποιοτικών χαρακτηριστικών που θα χρησιμοποιηθούν για την πρόβλεψη.

Πλήθος Παρατηρήσεων Από τις χρονοσειρές του συνόλου FRED, επιλέγονται 53,248 χρονοσειρές. Το κριτήριο επιλογής είναι μόνο ο αριθμός των παρατηρήσεων, καθώς λαμβάνονται υπόψη μόνο όσες χρονοσειρές έχουν πάνω από 50 παρατηρήσεις. Έτσι προκύπτει το σύνολο δεδομένων των 53,248 χρονοσειρών, με μέσο πλήθος παρατηρήσεων τις 341.36.

Υπολογισμός Συχνότητας Η συχνότητα των χρονοσειρών υπολογίζεται από τις ημερομηνίες των παρατηρήσεων. Ανάλογα με τη διαφορά των ημερών ανάμεσα σε δύο παρατηρήσεις, οι χρονοσειρές διαχωρίζονται σε Ημερήσιες, Εβδομαδιαίες, Μηνιαίες, Τριμηνιαίες, Εξαμηνιαίες, Ετήσιες.

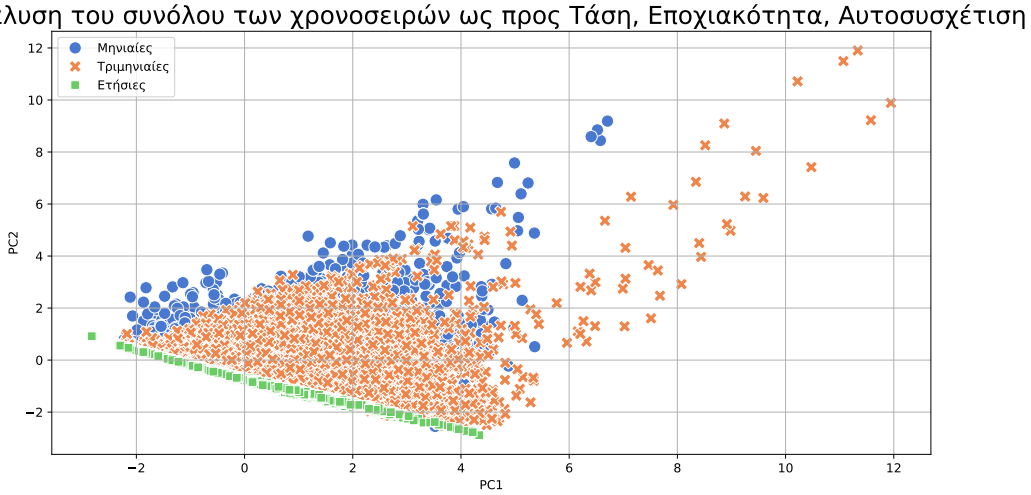
Κανονικοποίηση Οι χρονοσειρές, κανονικοποιούνται από το 0 στο 1 με τον εξής απλό υπολογισμό.

$$\text{Κονανικοποιημένη Χρονοσειρά} = \frac{\text{Χρονοσειρά} - \min(\text{Χρονοσειράς})}{\max(\text{Χρονοσειράς}) - \min(\text{Χρονοσειράς})} \tag{6.1}$$

Η κανονικοποίηση αυτή, βοηθάει τόσο στη μεταγενέστερη χρήση όσον αφορά την είσοδο των νευρωνικών δικτύων, όσο και στο να παραχθεί ένα κοινό σημείο αναφοράς, με συγκρίσιμα στατιστικά μεταξύ χρονοσειρών που δεν εξαρτώνται από το μέγεθος των παρατηρήσεων.

Υπολογισμός Ποιοτικών Χαρακτηριστικών Τα 52 ποιοτικά χαρακτηριστικά που υπολογίστηκαν βρίσκονται αναλυτικά στο παράρτημα B′.
Επειδή ο σκοπός των χαρακτηριστικών αυτών είναι να προσφέρουν πληροφορία που θα βελτιώσει τη δυνατότητα του πολυεπίπεδου νευρωνικού μοντέλου να προβλέπει τις μελλοντικές στιγμές της χρονοσειράς, είναι επιθυμητό το κάθε χαρακτηριστικό να λαμβάνεται υπόψη μόνο εάν πραγματικά είναι ανεξάρτητο και χρήσιμο, και ως εκ τούτου παρέχει μία ασυσχέτιστη με τις άλλες πληροφορία.
Για το λόγο αυτό, υπολογίζονται οι τιμές των στατιστικών χαρακτηριστικών για το σύνολο των χρονοσειρών, με σκοπό να εντοπισθούν τα χρήσιμα στατιστικά που θα αξιοποιηθούν στην επόμενη φάση για την πρόβλεψη.
Παρατηρείται πως η τιμή του χαρακτηριστικού variance_larger_than_standard_deviation (αν η διακύμανση είναι μεγαλύτερη από την τυπική απόκλιση) είναι διαρκώς 0 και αμετάβλητη συνεπώς διαγράφεται. Επιπλέον, εφόσον πρόκειται για κανονικοποιημένα στο εύρος δεδομένων 0-1, δεν υπάρχει λόγος να χρησιμοποιηθεί στην παρούσα φάση οι μετρικές count_above_0 και count_bellow_0.
Τέλος, για να απαλειφτούν τα υψηλά συσχετιζόμενα στατιστικά, υπολογίζεται ο συντελεστής συσχέτισης Pearson για κάθε ζευγάρι. Τα χαρακτηριστικά που συσχετίζονται κατά περισ-

σότερο από 90% ή λιγότερο από -90% με κάποια άλλα (και συνεπώς δεν επιφέρουν κάποια επιπλέον γνώση), απαλείφονται.
Με τον τρόπο αυτό από τα αρχικά 52 στατιστικά, μένουν τα 34, ενώ τα χαρακτηριστικά που δεν επιλέχθηκαν έχουν σημειωθεί με «*» στο παράρτημα Β΄.

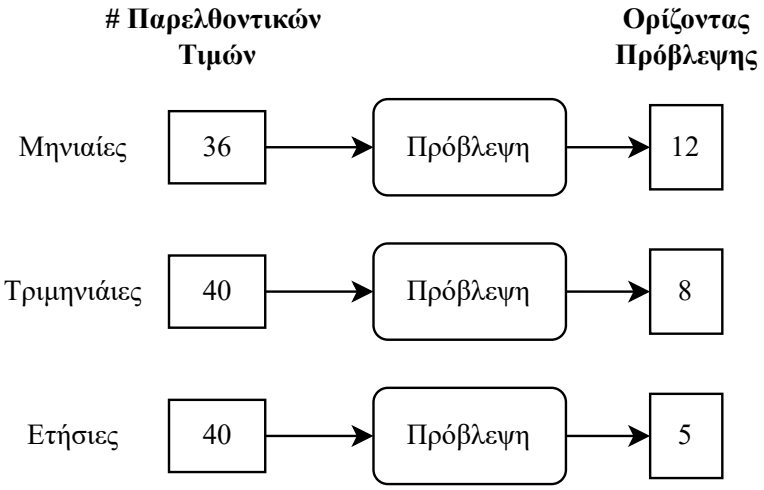


Σχήμα 6.17: *Αναπαράσταση των χρονοσειρών του πειράματος μέσω της τάσης, εποχιακότητας και αυτοσυσχέτισης πρώτου βαθμού.*

Στο σχήμα 6.17, φαίνεται μία αναπαράσταση στο χώρο του συνόλου των δεδομένων με τη μέθοδο ανάλυσης κυρίων συνιστωσών για τα ποιοτικά χαρακτηριστικά της τάσης, εποχιακότητας και αυτοσυσχέτισης πρώτου βαθμού. Η κύρια συνιστώσα PC1 εξηγεί το 61.07% της διακύμανσης, ενώ η PC2 το 25.74%, συνεπώς αθροιστικά εξηγούν το 86.81% της διακύμανσης των τριών χαρακτηριστικών. Στο διάγραμμα παρατηρούμε πως οι τρεις τύπου συχνοτήτων (Μηνιαίες, Τριμηνιαίες, Ετήσιες Χρονοσειρές) μοιράζονται διαφορετικά στο χώρο, δείχνοντας για παράδειγμα, πως οι ετήσιες χρονοσειρές, παρουσιάζουν συνήθως διαφορετικές τιμές στα τρία αυτά μεγέθη απ ότι οι μηνιαίες.

6.6 Προβλέψεις

Το τελικό στάδιο του πειράματος για να διευκρινιστεί εάν οι λεκτικές περιγραφές των χρονοσειρών μπορούν να συμβάλλουν στη βελτίωση των προβλέψεων, είναι το στάδιο της πρόβλεψης.
Αρχικά, είναι σημαντικό να ορισθεί ο ορίζοντας της πρόβλεψης, καθώς και ο αριθμός των παρελθοντικών τιμών που θα χρησιμοποιηθούν. Στο σχήμα 6.18, παρουσιάζεται η πρόβλεψη η οποία θα κληθεί να παράξουν τα μοντέλα που αναπτύσσονται παρακάτω.



Σχήμα 6.18: *Ορίζοντες Πρόβλεψης ανά συχνότητα χρονοσειράς.*

6.6.1 Μοντέλα Πρόβλεψης - Αρχιτεκτονικές Νευρωνικών Δικτύων

Παρακάτω αναλύονται οι αρχιτεκτονικές των μοντέλων που αναπτύχθηκαν για τους σκοπούς αυτής της διπλωματικής. Οι αρχιτεκτονικές αυτές υλοποιήθηκαν με την Python 3.12, και συγκεκριμένα με τη βιβλιοθήκη Keras. Επειδή στη συνέχεια δοκιμάζονται πολλές υπερπαράμετροι, στις περιγραφές των αρχιτεκτονικών οι αριθμοί των κόμβων, των εισόδων, των εξόδων κ.ο.κ. αναφέρονται με δυναμικό τρόπο σε μορφή μεταβλητών.

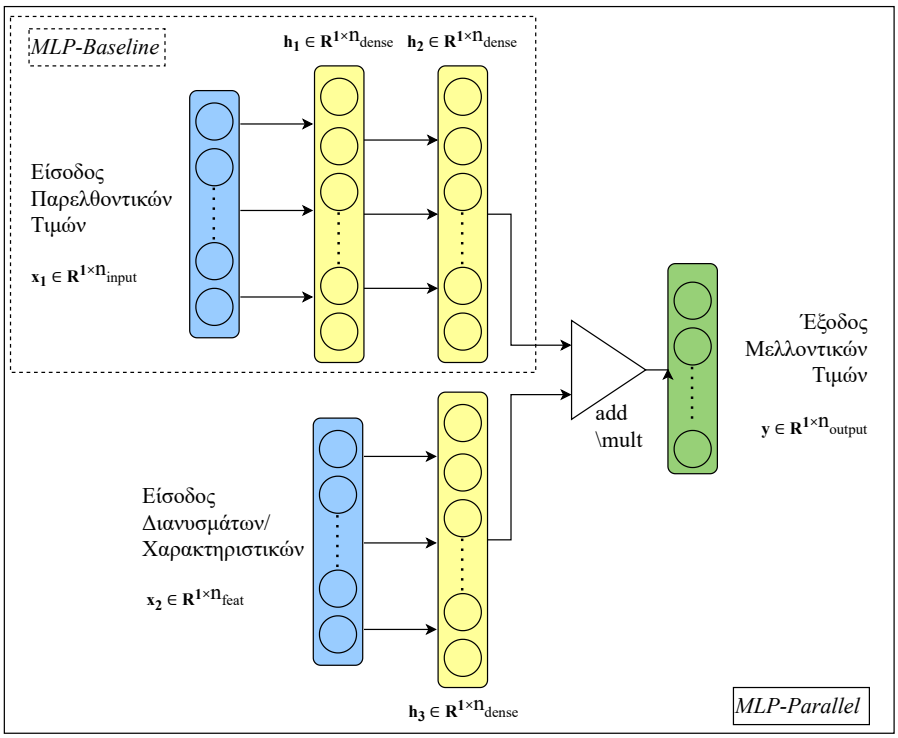
MLP-Baseline

Πρόκειται για την απλή εκδοχή ενός Πολυεπίπεδου Νευρωνικού Μοντέλου το οποίο περιέχει ένα επίπεδο εισόδου με n_{input} εισόδους παρελθοντικών παρατηρήσεων.
Ακολουθούν δύο κρυφά επίπεδα με n_{dense} κόμβους οι οποίοι συνδέονται με τους κόμβους του επόμενου επιπέδου με πιθανότητα 1-$p_{dropout}$ με σκοπό τυχαία, κάποιες συνδέσεις να διαγράφονται για την αποφυγή του overfitting.
Το επίπεδο εξόδου αποτελείται από n_{output} στοιχεία, ανάλογα με τον ορίζοντα πρόβλεψης.
Οι συναρτήσεις ενεργοποίησης είναι ReLU για τα κρυφά επίπεδα και γραμμική για το επίπεδο εξόδου.
Επιπλέον δίνεται η δυνατότητα να ενεργοποιηθεί ή όχι η τεχνική early stopping όπως εξηγήθηκε στο κεφάλαιο 2.

MLP-Parallel

Το μοντέλο MLP-Parallel πρόκειται για ένα μοντέλο που συνδυάζει το MLP-Baseline με μία ή δύο παράλληλες εισόδους μεγέθους n_{feat}, οι οποίες περνώντας από ένα κρυφό επίπεδο με n_{dense} κόμβους, συνδυάζονται με την έξοδο του δεύτερου κρυφού επιπέδου του MLP-Baseline με μία εκ των μεθόδων *combine_method*, για να παραχθούν n_{output} έξοδοι.
Οι παραλλαγές MLP-Parallel(ts, stat_features), MLP-Parallel(ts, nlp), είναι ανάλογα με το

ποια είναι η δεύτερη είσοδος που δίνεται (τα ποιοτικά χαρακτηριστικά ή τα διανύσματα προτάσεων). Επίσης, στην παραλλαγή MLP-Parallel(all), παρέχονται με παράλληλο τρόπο και τα διανύσματα προτάσεων και τα ποιοτικά χαρακτηριστικά.



Σχήμα 6.19: *Το μοντέλο MLP-Parallel που εμπεριέχει το MLP-Baseline.*

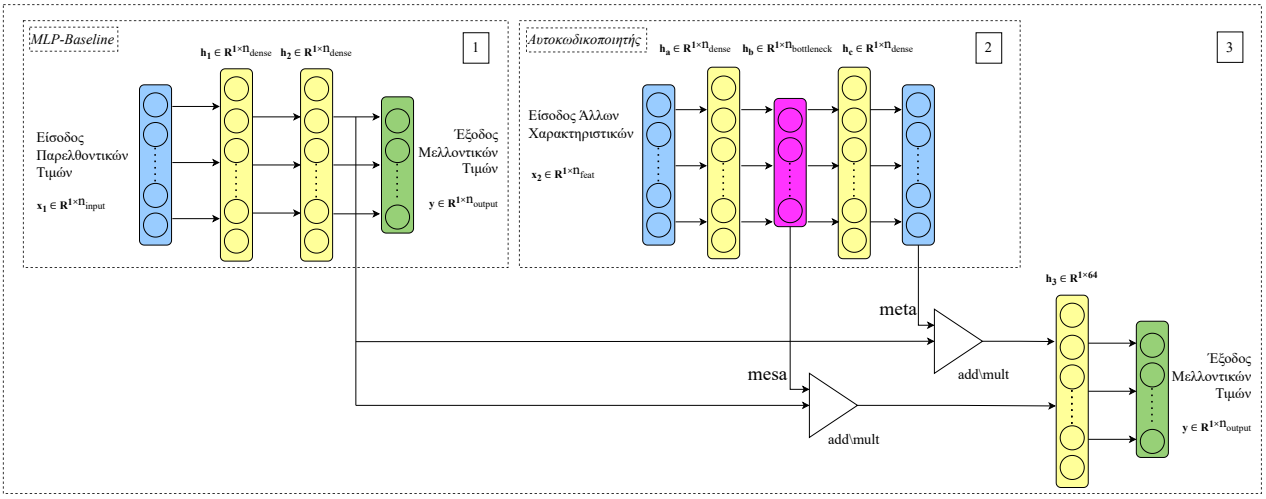
MLP-Concat

Η παραλλαγή MLP-Concat προκύπτει από τη συνένωση των εισόδων των χρονοσειρών, με τις εισόδους των χαρακτηριστικών/διανυσμάτων. Στη συνέχεια, η πρόβλεψη παράγεται όπως στο MLP-Baseline. Το μέγεθος της εισόδου είναι $n_{input} + n_{feat}$, όπου n_{feat} το μέγεθος του διανύσματος των ποιοτικών χαρακτηριστικών ή των διανυσμάτων που αναπαριστούν τις περιγραφές (ή και των δύο στην περίπτωση MLP-Concat(all)).

MtMs-NLP

Το πιο σύνθετο μοντέλο που σχεδιάστηκε, είναι το MtMs-NLP. Αυτό βασίζεται στην ιδέα που κατέκτησε την πρώτη θέση στο διαγωνισμό προβλέψεων M6, όπου το ζητούμενο ήταν η μηνιαία πρόβλεψη των τιμών 50 μετοχών του χρηματιστηρίου της Αμερικής και 50 ETFs (Διαπραγματεύσιμα Αμοιβαία Κεφάλαια). Οι πληροφορίες που μπορούσαν να χρησιμοποιηθούν ήταν απεριόριστες, ωστόσο η αντιμετώπιση του Filip Stanek που παρουσιάζεται στο (Staněk, 2023), είναι πως ένα πολυεπίπεδο νευρωνικό μοντέλο εκπαιδεύεται για την πρόβλεψη πολύ περισσότερων από 100 περιουσιακών αγαθών, αλλά στη συνέχεια, με τη χρήση ενός αυτοκωδικοποιητή, συνδυάζεται η γενική γνώση του μοντέλου για την ειδική πρόβλεψη ενός συγκεκριμένου περιουσιακού στοιχείου.

Η υλοποίηση του MtMs-NLP απαρτίζεται από ένα μοντέλο όπως το MLP-Baseline, το οποίο αρχικά εκπαιδεύεται στην παραγωγή προβλέψεων. Παράλληλα, εκπαιδεύεται ένας αυτοκωδικοποιητής στο να κωδικοποιεί και να αποκωδικοποιεί τα διανύσματα, με σκοπό να «μάθει» στο επίπεδο στένωσης του, να παράγει τη βέλτιστη κωδικοποίηση του διανύσματος, η οποία όμως επιτρέπει την ανακατασκευή του. Η διάσταση αυτή είναι η $n_{bottlencek}$.
Σε ένα πρώτο χρόνο εκπαιδεύεται το MLP-Baseline και ο αυτοκωδικοποιητής, ενώ στη συνέχεια, γίνεται μία συνολική εκπαίδευση με τα ήδη προ-εκπαιδευμένα μοντέλα συνδυάζοντας είτε την έξοδο του κωδικοποιητή (παραλλαγή mesa), είτε την έξοδο του αποκωδικοποιητή (παραλλαγή meta).

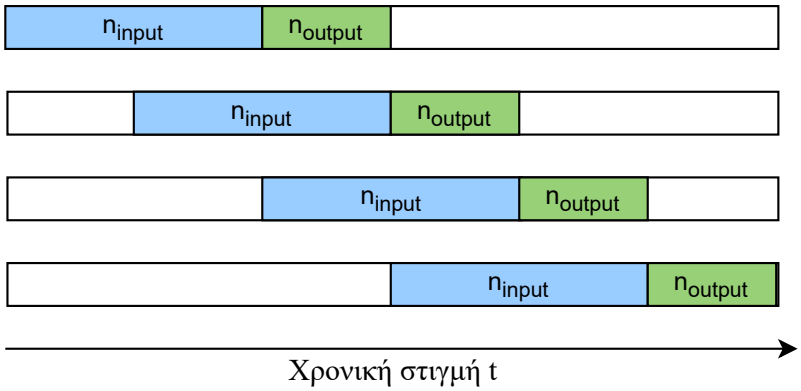


Σχήμα 6.20: *Η Αρχιτεκτονική του μοντέλου MtMs που εμπεριέχει το MLP-Baseline και έναν αυτοκωδικοποιητή.*

6.6.2 Εκπαίδευση Μοντέλων Πρόβλεψης

Προετοιμασία Χρονοσειρών Εκπαίδευσης

Κυλιόμενα Παράθυρα Με σκοπό τα μοντέλα που αναπτύχθηκαν να λάβουν υπόψη τις ιδιαιτερότητες των χρονοσειρών σε βάθος, και να αξιοποιήσουν περισσότερες παρατηρήσεις, για κάθε χρονοσειρά παράγονται n_{train} κυλιόμενα παράθυρα (rolling windows). Για να γίνει αυτό, δημιουργούνται παράθυρα εισόδου μεγέθους n_{input} καθώς και εξόδου n_{output} όπως φαίνεται στο σχήμα 6.21.



Σχήμα 6.21: *Δημιουργία τεσσάρων κυλιόμενων παραθύρων εκπαίδευσης από μία χρονοσειρά.*

Καθορισμός και Επιλογή Υπερπαραμέτρων

Η επιλογή των υπερπαραμέτρων στην εκπαίδευση των μοντέλων που αναπτύχθηκαν, είναι ιδιαίτερα σημαντική, καθώς έχουν μεγάλη επίδραση. Σχετίζονται με τον αριθμό των κόμβων στα κρυφά επίπεδα (n_{dense}), τη δυνατότητα της διακοπής της εκπαίδευσης σε περίπτωση το μοντέλο δε βελτιώνεται για πάνω από 5 συνεχόμενες εποχές (early_stopping), την πιθανότητα διαγραφής συνδέσεων ($p_{dropout}$), τον τρόπο συνδυασμού των εξόδων των κρυφών επιπέδων (combine_method), τον αριθμό των εποχών εκπαίδευσης (n_{epochs}), τον μέγιστο αριθμό κυλιόμενων παραθύρων (n_{train}) και τα μοντέλα παραγωγής διανυσμάτων (nlp_model) που αναδείχτηκαν ως ποιοτικά στο προηγούμενο βήμα καθώς και τις συστάδες που παράχθηκαν από τον k-Means για τα διανύσματα του Word2Vec Skip-Gram σε μορφή one-hot encoding. Όσον αφορά τη διάσταση των διανυσμάτων προτάσεων, επιλέχθηκε η διάσταση 30 καθώς στο προηγούμενο βήμα της πειραματικής παρατηρήθηκε πως οι πιο μεγάλες διαστάσεις δεν επιφέρουν κάποια επιπλέον πληροφορία, ωστόσο δυσκολεύουν την εκπαίδευση. Οι παράμετροι καταγράφονται στον πίνακα 6.6.

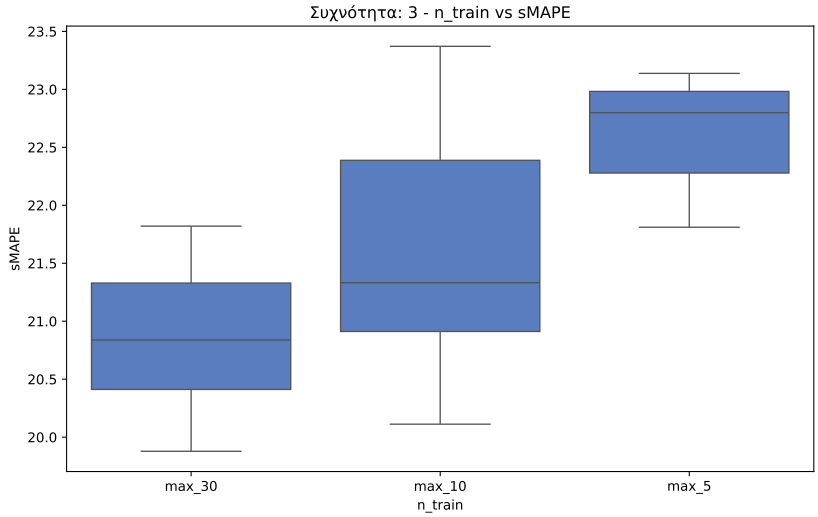
Γενικά	
Παράμετροι	**Παραλλαγές**
Συχνότητα Χρονοσειρών	Μηνιαίες, Τριμηνιαίες, Ετήσιες
Επιλογή Μοντέλου	
Παράμετροι	**Παραλλαγές**
model	MLP-Baseline, MLP-Parallel(ts, stat_features), MLP-Parallel(ts, nlp), MLP-Parallel(all), MLP-Concat(ts, stat_features), MLP-Concat(ts, nlp), MLP-Concat(all), MtMs-NLP(mesa),MtMs-NLP(meta)
Υπερπαράμετροι Αρχιτεκτονικής & Εκπαίδευσης	
Υπερπαράμετροι	**Παραλλαγές**
n_{dense}	100, 200, 300
$p_{dropout}$	0.1, 0.2, 0.3
early_stopping	True (Ανοχή 5 εποχές)
n_{epochs}	30, 60, 100
batch_size	512*, 1024, 2048 (*512 για MtMs-NLP)
combine_method	mult, add
n_{train}	max_5, max_10, max_30
Υπερπαράμετροι Χαρακτηριστικών	
Υπερπαράμετροι	**Παραλλαγές**
nlp_model	TF-IDF, Word2Vec Skip-Gram, GPT-2, Sentence-BERT, BERT, Συστάδες
stat_features	Υποσύνολο των 34 στατιστικών χαρακτηριστικών

Πίνακας 6.6: *Υπερπαράμετροι Μοντέλου Πρόβλεψης.*

6.6.3 Αποτελέσματα Μοντέλων Πρόβλεψης

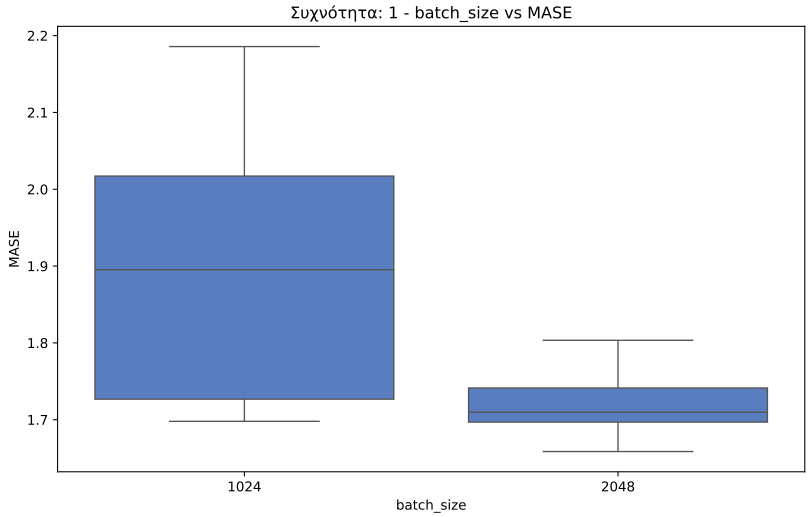
Αποτελέσματα Μοντέλου MLP-Baseline

Η επιρροή του αριθμού των κυλιόμενων παραθύρων που λαμβάνεται υπόψη για το σύνολο των δεδομένων εκπαίδευσης, είναι σαφής. Για όλες τις συχνότητες, το αυξημένο n_{train} οδηγεί σε μικρότερο σφάλμα. Όταν το n_{train} θέτεται ίσο με max_30, αυτό σημαίνει πως αν αυτό είναι εφικτό, λαμβάνονται υπόψη έως 30 δείγματα για κάθε χρονοσειρά. Αυτό έχει ως αποτέλεσμα για τις μηνιαίες χρονοσειρές, τα συνολικά δείγματα εκπαίδευσης που παράγονται να είναι 201,440 ενώ είναι 184,364 για τις τριμηνιαίες και 126,661 για τις ετήσιες.



Σχήμα 6.22: *Επιρροή του αριθμού των κυλιόμενων παραθύρων στην επίδοση του MLP-Baseline.*

Στο σχήμα 6.22 γίνεται μία σύγκριση της μετρικής sMAPE για ετήσιες χρονοσειρές, σε σχέση με τον αριθμό των κυλιόμενων παραθύρων που χρησιμοποιήθηκε.
Επιπλέον, σημαντική είναι και η επιρροή της μεταβλητής batch_size που καθορίζει τον αριθμό των δειγμάτων που επεξεργάζεται η αλγόριθμος σε κάθε εποχή της εκπαίδευσης του μοντέλου. Στο σχήμα 6.23, φαίνεται η επιρροή της τιμής του batch_size (1024 και 2048) στο σφάλμα MASE για τις προβλέψεις των μηνιαίων χρονοσειρών (με τις υπόλοιπες υπερπαραμέτρους να μεταβάλλονται). Για το συγκεκριμένο μοντέλο, το υψηλό βατση_σιζε παράγει καλύτερα αποτελέσματα.



Σχήμα 6.23: *Επιρροή του batch_size στην επίδοση του MLP-Baseline.*

Όσον αφορά τον αριθμό εποχών εκπαίδευσης που χρησιμοποιήθηκε, τα βέλτιστα αποτελέσματα χρησιμοποιούν συνήθως 100 εποχές.

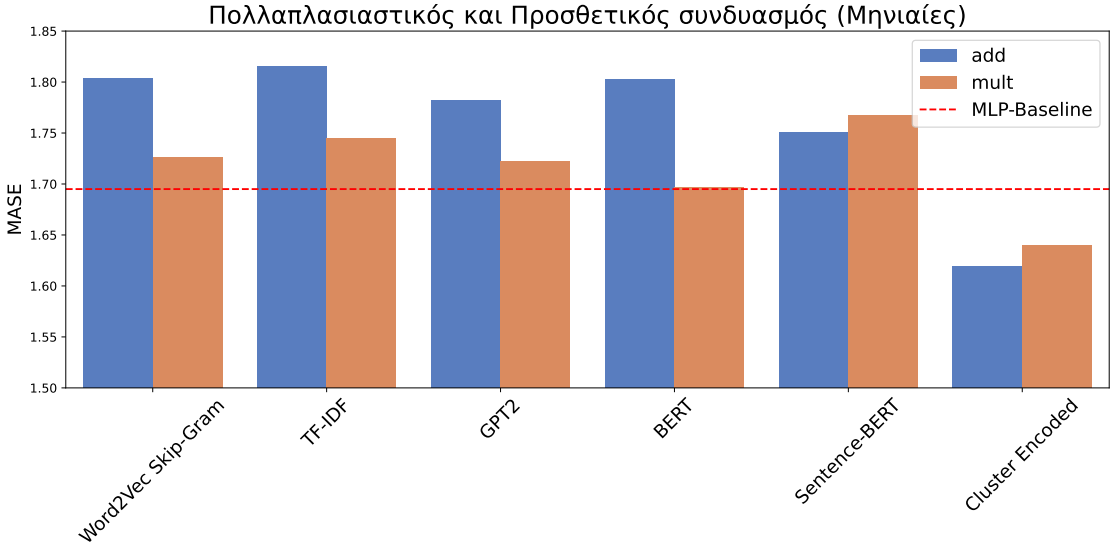
Τα καλύτερα αποτελέσματα του μοντέλου MLP-Baseline για τις τρεις συχνότητες βρίσκονται στον πίνακα 6.7 και θα χρησιμοποιηθούν ως μέτρο αναφοράς (benchmark) για τα υπόλοιπα μοντέλα.

Συχνότητα	n_{epochs}	batch_size	n_{train}	n_{dense}	sMAPE (%)	MASE
Μηνιαίες	100	2048	max_30	300	28.891	1.695
Τριμηνιαίες	100	2048	max_30	300	30.420	1.408
Ετήσιες	100	2048	max_30	200	19.878	5.154

Πίνακας 6.7: *Βέλτιστα Αποτελέσματα MLP-Baseline ($n_{dropout}$ = 0.2).*

Αποτελέσματα Μοντέλων MLP-Parallel και MLP-Concat

Ενσωμάτωση Περιγραφών Το μοντέλο MLP-Parallel(ts, nlp) με παράλληλη είσοδο στις χρονοσειρές τις περιγραφές, επιτυγχάνει βελτίωση της πρόβλεψης στις ετήσιες και μηνιαίες χρονοσειρές σε σχέση με το μοντέλο MLP-Baseline.
Βελτίωση επιτυγχάνεται μόνο με τη χρήση των συστάδων που δημιουργήθηκαν με τον αλγόριθμο συσταδοποίησης k-Means από τα διανύσματα προτάσεων Word2Vec Skip-Gram. Συγκεκριμένα, η ενσωμάτωση των διανυσμάτων προτάσεων κάθε χρονοσειράς αποδείχθηκε χειρότερη από την ενσωμάτωση πιο απλοϊκών one-hot encodings που αντιστοιχούν στη συστάδα στην οποία ανατέθηκε η χρονοσειρά κατά τη συσταδοποίηση της.



Σχήμα 6.24: *Σφάλμα MASE των προβλέψεων από το μοντέλο MLP-Parallel(ts, nlp), συναρτήσει της μεθόδου συνδυασμού των κρυφών επιπέδων - Μηνιαίες Χρονοσειρές.*

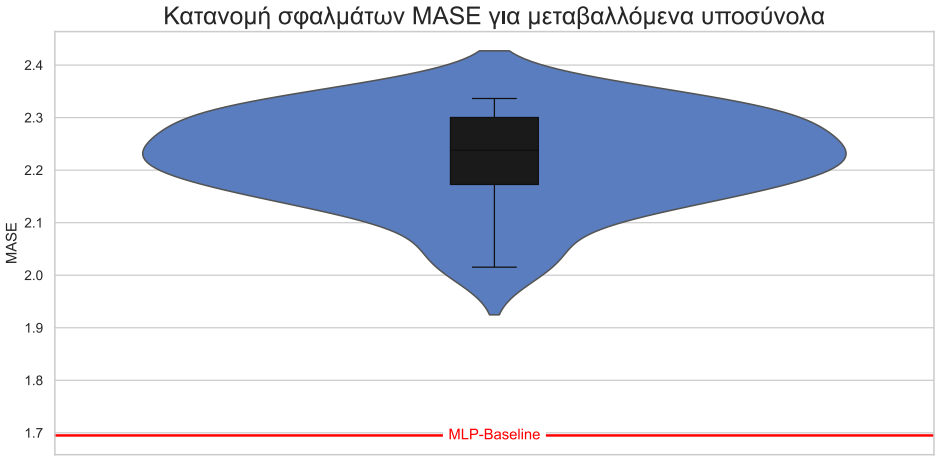
Στο μοντέλο με παράλληλες εισόδους MLP-Parallel(ts, nlp), για την συγχώνευση των δεδομένων των κρυφών επιπέδων, δε φαίνεται να υπάρχει κάποια ξεκάθαρη καλύτερη τεχνική. Για τις μηνιαίες χρονοσειρές, στα δύο μοντέλα που οδηγούν στο ελάχιστο σφάλμα MASE έχει χρησιμοποιηθεί η προσθετική μέθοδος, ενώ η πολλαπλασιαστική μέθοδος παρουσιάζει καλύτερα αποτελέσματα για τις υπόλοιπες.

Όσον αφορά το μοντέλο MLP-Concat(ts, nlp), στο οποίο τα διανύσματα συγχωνεύονται στην είσοδο, αυτό έχει χειρότερα αποτελέσματα από το MLP-Parallel(ts, nlp) ωστόσο στις μηνιαίες χρονοσειρές επιτυγχάνει και αυτό καλύτερο αποτέλεσμα από το μοντέλο αναφοράς. Παρακάτω φαίνονται κάποια βέλτιστα αποτελέσματα τα οποία παρουσιάζονται με έντονα γράμματα όταν είναι καλύτερα από την αναφορά.

Μοντέλο	Συχνότητα	n_{dense}	Συνδυασμός	sMAPE (%)	MASE
MLP-Baseline	Μηνιαίες	300	-	28.891	1.695
MLP-Parallel(ts, συστάδες)	Μηνιαίες	300	mult	**28.109**	**1.640**
MLP-Parallel(ts, συστάδες)	Μηνιαίες	300	add	**28.222**	**1.619**
MLP-Concat(ts, συστάδες)	Μηνιαίες	300	-	**28.502**	**1.671**
MLP-Baseline	Τριμηνιαίες	300	-	30.420	1.408
MLP-Parallel(ts, TF-IDF)	Τριμηνιαίες	300	mult	30.722	1.44
MLP-Concat(ts, συστάδες)	Τριμηνιαίες	300	-	32.050	1.560
MLP-Baseline	Ετήσιες	200	-	19.878	5.154
MLP-Parallel(ts, συστάδες)	Ετήσιες	300	mult	**19.365**	**5.042**
MLP-Concat(ts, συστάδες)	Ετήσιες	300	-	21.770	6.058

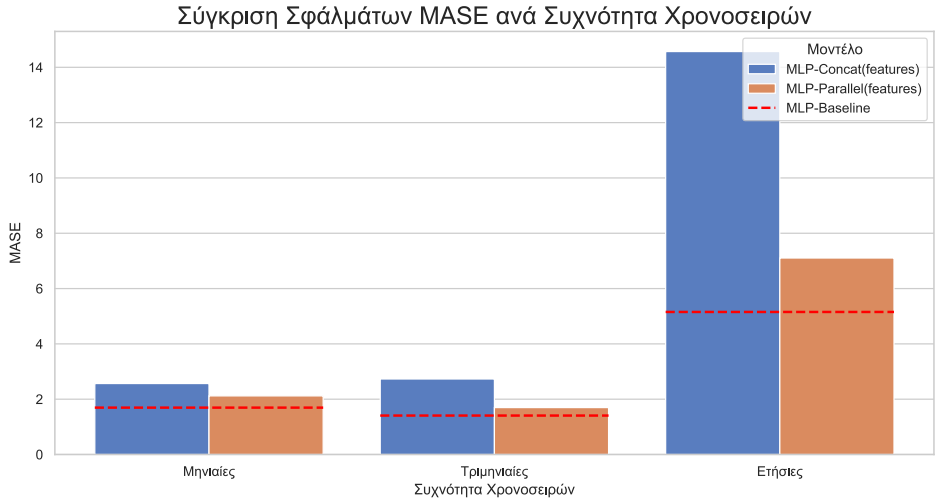
Πίνακας 6.8: *Βέλτιστα Αποτελέσματα MLP-Parallel(ts, nlp) και MLP-Concat(ts, nlp), Σταθερές:* n_{epochs}*=100, batch_size=2048,* n_{train}*=max_30,* $n_{dropout}$ = 0.2.

Ενσωμάτωση Ποιοτικών Χαρακτηριστικών Κατά τη διάρκεια δοκιμών υποσυνόλων των 34 χαρακτηριστικών με τη μέθοδο backward elimination, ξεχωρίζει το υποσύνολο των 22 ποιοτικών χαρακτηριστικών F' = ['trend', 'seasonality', 'abs_energy', 'absolute_maximum', 'absolute_sum_of_changes', 'first_location_of_maximum', 'first_location_of_minimum', 'has_duplicate', 'has_duplicate_max', 'has_duplicate_min', 'kurtosis', 'longest_strike_above_mean', 'longest_strike_below_mean', 'mean', 'mean_abs_change','mean_change', 'percentage_of_reoccurring_values_to_all_values', 'skewness', 'sum_of_reoccurring_values', 'binned_entropy', 'cid_ce', 'large_standard_deviation'].
Ωστόσο, τα αποτελέσματα τόσο του μοντέλου MLP-Parallel(ts, stat_features), όσο και του MLP-Concat(ts, stat_features) είναι κατώτερου επιπέδου σε σχέση με το μοντέλο MLP-Baseline. Μεταβάλλοντας το υποσύνολο των χαρακτηριστικών που τροφοδοτείται στο μοντέλο, στο σχήμα 6.26 παρατηρείται πως το σφάλμα για τις χρονοσειρές είναι σε όλες τις περιπτώσεις αρκετά υψηλότερο από αυτό του MLP-Baseline.



Σχήμα 6.25: *Κατανομή σφαλμάτων MASE των προβλέψεων από το μοντέλο MLP-Parallel(ts, stat_features), για μεταβαλλόμενα υποσύνολα των ποιοτικών χαρακτηριστικών - Μηνιαίες Χρονοσειρές.*

Χαρακτηριστική είναι επίσης η μειωμένη απόδοση της εκδοχής MLP-Concat(ts, stat_features) σε σχέση με αυτήν του MLP-Parallel(ts, stat_features) όπως φαίνεται στο σχήμα με την επίδοση ανά συχνότητα 6.26.



Σχήμα 6.26: *Σύγκριση Ελαχίστων σφαλμάτων MLP-Concat(ts, stat_features) και MLP-Parallel(ts, stat_features) ανά συχνότητα.*

Μοντέλο	Συχνότητα	n_{dense}	Συνδυασμός	sMAPE (%)	MASE
MLP-Baseline	Μηνιαίες	300	-	28.891	1.695
MLP-Parallel(ts, stat_features)	Μηνιαίες	300	mult	33.569	2.119
MLP-Concat(ts, stat_features)	Μηνιαίες	200	-	39.000	2.566
MLP-Baseline	Τριμηνιαίες	300	-	30.420	1.408
MLP-Parallel(ts, stat_features)	Τριμηνιαίες	200	mult	32.646	1.696
MLP-Concat(ts, stat_features)	Τριμηνιαίες	300	-	41.388	2.730
MLP-Baseline	Ετήσιες	200	-	19.878	5.154
MLP-Parallel(ts, stat_features)	Ετήσιες	300	mult	23.795	7.104
MLP-Concat(ts, stat_features)	Ετήσιες	300	-	38.862	14.569

Πίνακας 6.9: *Βέλτιστα Αποτελέσματα MLP-Parallel(ts, stat_features) και MLP-Concat(ts, stat_features), Σταθερές:* n_{epochs}*=100, batch_size=2048,* n_{train}*=max_30,* $n_{dropout}$ = 0.2.

Ενσωμάτωση και των περιγραφών και των ποιοτικών χαρακτηριστικών Στο μοντέλο MLP-Parallel(all) παρέχονται χρονοσειρές, περιγραφές σε μορφή συστάδων και όλα τα διαθέσιμα ποιοτικά χαρακτηριστικά με παράλληλο τρόπο, ενώ στο MLP-Concat(all) με συνένωση στην είσοδο. Παρατηρούμε και πάλι, πως η τεχνική της παράλληλης εισόδου παράγει καλύτερα αποτελέσματα συγκριτικά με αυτήν της συγχώνευσης. Τα βέλτιστα αποτελέσματα προέρχονται και πάλι από με τις παραμέτρους n_{dense} = 300 και την πολλαπλασιαστική μέθοδο συνδυασμού των κρυφών επιπέδων. Τα αποτελέσματα του παράλληλου μοντέλου είναι όπως αναμένεται καλύτερα από αυτά του MLP-Parallel(ts, stat_features), και χειρότερα από το MLP-Parallel(ts, συστάδες). Σταθερά κάτω από το MLP-Baseline, φαίνεται πως τα ποιοτικά χαρακτηριστικά δεν προσφέρουν κάποιο πλεονέκτημα στο μοντέλο πρόβλεψης.

Μοντέλο	Συχνότητα	n_{dense}	Συνδυασμός	sMAPE (%)	MASE
MLP-Baseline	Μηνιαίες	300	-	28.891	1.695
MLP-Parallel(all)	Μηνιαίες	300	mult	33.383	2.071
MLP-Concat(all)	Μηνιαίες	300	mult	41.329	2.789
MLP-Baseline	Τριμηνιαίες	300	-	30.420	1.408
MLP-Parallel(all)	Τριμηνιαίες	300	mult	32.720	1.592
MLP-Concat(all)	Τριμηνιαίες	300	mult	47.161	3.519
MLP-Baseline	Ετήσιες	200	-	19.878	5.154
MLP-Parallel(all)	Ετήσιες	300	mult	23.309	7.171
MLP-Concat(all)	Ετήσιες	300	mult	33.697	12.678

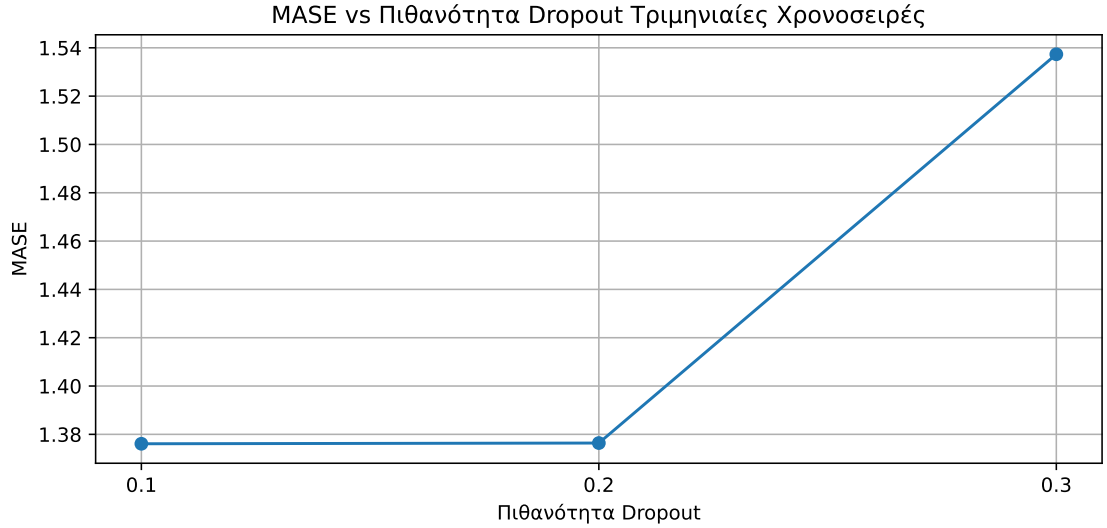
Πίνακας 6.10: *Βέλτιστα Αποτελέσματα MLP-Parallel(all), MLP-Concat(all), Σταθερές:* n_{epochs}*=100, batch_size=2048,* n_{train}*=max_30,* $n_{dropout}$ = 0.2.

Αποτελέσματα Μοντέλου MtMs-NLP

Τα αποτελέσματα του μοντέλου MtMs-NLP είναι στην περίπτωση των τριμηνιαίων χρονοσειρά καλύτερα από αυτά του MLP-Parallel και διαρκώς καλύτερα από το MLP-Baseline. Ως αριθμός στένωσης στον αποκωδικοποιητή δοκιμάστηκαν οι τιμές 5 και 2, ωστόσο η τιμή 5 είναι αυτή που παράγει όλα τα βέλτιστα αποτελέσματα.

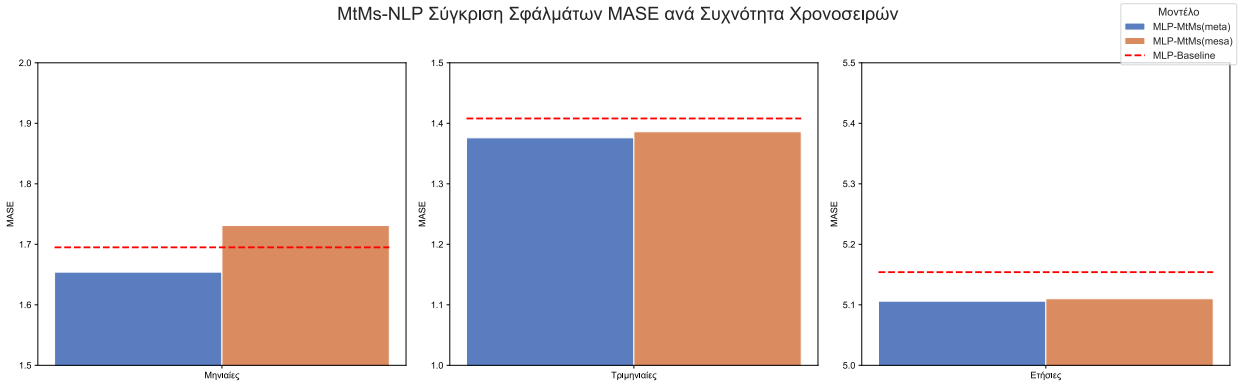
Η παραλλαγή meta, παρότι ουσιαστικά αποτελεί το επαναδημιουργημένο διάνυσμα των συστάδων (ύστερα από την κωδικοποίηση και την αποκωδικοποίηση του), καταφέρνει με τη μέθοδο της μετεκπαίδευσης να ξεπεράσει την απλοϊκή παράλληλη είσοδο. Η παραλλαγή mesa, με τη μειωμένη διάσταση του διανύσματος των χρονοσειρών, έχει χειρότερα αποτελέσματα.

Λόγω της αρχιτεκτονικής του μοντέλου, ένας κίνδυνος που πρέπει να ληφθεί υπόψη, είναι αυτός της υπερπροσαρμογής. Για το σκοπό αυτό, εξετάζονται διαφορετικές τιμές για τη πιθανότητα μίας ακμής να διαγραφεί ($p_{dropout}$). Στο σχήμα 6.27 παρατηρούμε πως μεταβάλλοντας την τιμή του $p_{dropout}$ από το 0.1 στο 0.2, η τιμή του σφάλματος είναι σχεδόν στάσιμη, ενώ η τιμή 0.3 (δηλαδή να διαγραφούν 30% των ακμών) ανεβάζει το σφάλμα. Για το λόγο αυτό επιλέγεται η τιμή 0.2.



Σχήμα 6.27: *Σφάλμα MASE των προβλέψεων συναρτήσει της πιθανότητας Dropout.*

Η οπτικοποίηση της βελτίωσης σε σχέση με το μοντέλο αναφοράς, παρουσιάζεται στο γράφημα 6.28 για τη μετρική MASE, όπου για την κάθε συχνότητα, έχει πραγματοποιηθεί μεγέθυνση σε ένα εύρος 0.5 MASE.



Σχήμα 6.28: *Αποτελέσματα MtMs-NLP σε σχέση με το MLP-Baseline.*

Μοντέλο	Συχνότητα	batch_size	$n_{bottleneck}$	sMAPE (%)	MASE
MLP-Baseline	Μηνιαίες	2048	-	28.891	1.695
MLP-MtMs(meta)	Μηνιαίες	1024	5	**28.25**	**1.635**
MLP-MtMs(mesa)	Μηνιαίες	1024	5	29.331	1.731
MLP-Baseline	Τριμηνιαίες	2048	-	30.420	1.408
MLP-MtMs(meta)	Τριμηνιαίες	512	5	**29.806**	**1.376**
MLP-MtMs(mesa)	Τριμηνιαίες	1024	5	**30.004**	**1.386**
MLP-Baseline	Ετήσιες	200	-	19.878	5.154
MLP-MtMs(meta)	Ετήσιες	512	5	**19.870**	**5.106**
MLP-MtMs(mesa)	Ετήσιες	512	5	20.040	**5.11**

Πίνακας 6.11: *Βέλτιστα αποτελέσματα MLP-MtMs με Περιγραφές*
Σταθερές: n_{epochs} = 100, n_{train} = *max_30*, n_{dense} = 300, *Συνδυασμός* = *mult*, $n_{dropout}$ = 0.2.

Συγκριτικά Αποτελέσματα Μοντέλων Πρόβλεψης - Παρατηρήσεις

Στον πίνακα 6.12, παρουσιάζονται τα συνολικά αποτελέσματα των μοντέλων. Παρατηρείται πως τα μοντέλα που χρησιμοποιούν τις συστάδες που δημιουργήθηκαν με τα διανύσματα που αναπαριστούν τις λεκτικές περιγραφές των χρονοσειρών, επιτυγχάνουν καλύτερη πρόβλεψη. Ωστόσο, η διαφορά με το μοντέλο αναφοράς χωρίς τις περιγραφές, δεν είναι μεγάλη. Το πιο σύνθετο MtMs-NLP, επιτυγχάνει καλύτερη πρόβλεψη για τις τριμηνιαίες χρονοσειρές, όπου το MLP-Parallel(ts, nlp) πετυχαίνει βέλτιστη επίδοση κάτω από το σημείο αναφοράς. Αντιθέτως, το μοντέλο MLP-Parallel(ts, nlp) παρότι πιο απλοϊκό, επιτυγχάνει βέλτιστα αποτελέσματα για τις μηνιαίες και ετήσιες χρονοσειρές. Τα ποιοτικά χαρακτηριστικά δεν ενσωματώνονται με επιτυχία, παράγοντας τα χειρότερα αποτελέσματα.

Μηνιαίες		
MLP-Parallel(ts, συστάδες)	28.222	1.619
MLP-MtMs(meta)	28.250	1.635
MLP-Parallel(ts, συστάδες)	28.109	1.640
MLP-Concat(ts, συστάδες)	28.502	1.671
MLP-Baseline	28.891	1.695
MLP-MtMs(mesa)	29.331	1.731
MLP-Parallel(all)	33.383	2.071
MLP-Parallel(ts, stat_features)	33.569	2.119
MLP-Concat(ts, stat_features)	39.000	2.566
MLP-Concat(all)	41.329	2.789
Τριμηνιαίες		
MLP-MtMs(meta)	29.806	1.376
MLP-MtMs(mesa)	30.004	1.386
MLP-Baseline	30.420	1.408
MLP-Parallel(ts, TF-IDF)	30.722	1.440
MLP-Concat(ts, συστάδες)	32.050	1.560
MLP-Parallel(all)	32.720	1.592
MLP-Parallel(ts, stat_features)	32.646	1.696
MLP-Concat(ts, stat_features)	41.388	2.730
MLP-Concat(all)	47.161	3.519
Ετήσιες		
MLP-Parallel(ts, συστάδες)	19.365	5.042
MLP-MtMs(meta)	19.870	5.106
MLP-MtMs(mesa)	20.040	5.110
MLP-Baseline	19.878	5.154
MLP-Concat(ts, συστάδες)	21.770	6.058
MLP-Parallel(ts, stat_features)	23.795	7.104
MLP-Parallel(all)	23.309	7.171
MLP-Concat(all)	33.697	12.678
MLP-Concat(ts, stat_features)	38.862	14.569
Μοντέλο	**sMAPE (%)**	**MASE**

Πίνακας 6.12: *Συνολική άποψη βέλτιστων αποτελεσμάτων (Ταξινόμηση MASE Αύξων).*

Κεφάλαιο 7

Συμπεράσματα και Προεκτάσεις

Η παρούσα διπλωματική εργασία, είχε ως σκοπό να εξετάσει την αποτελεσματικότητα του συνδυασμού χρονοσειρών και συμμεταβλητών όπως οι περιγραφές των χρονοσειρών και τα ποιοτικά χαρακτηριστικά τους, προκειμένου να διαπιστωθεί αν βελτιώνεται η πρόβλεψη.
Στα πρώτα κεφάλαια, επεξηγήθηκε το θεωρητικό υπόβαθρο στο οποίο βασίζεται η εργασία. Έγινε μία εισαγωγή στη μηχανική μάθηση και τα νευρωνικά δίκτυα ενώ μετέπειτα αναφέρθηκαν πρακτικές της επεξεργασίας της φυσικής γλώσσας και παραγωγής διανυσμάτων προτάσεων. Μελετήθηκαν αλγόριθμοι ταξινόμησης για την αξιολόγηση των διανυσμάτων αυτών σε πρακτικές εφαρμογές καθώς και τεχνικές συσταδοποίησης για την παραγωγή συστάδων που χρησιμοποιούνται κατά τη διάρκεια της πρόβλεψης. Τέλος, αναλύθηκαν προσεγγίσεις για την πρόβλεψη των χρονοσειρών.
Στο κεφάλαιο 6, παρουσιάστηκε η πειραματική διαδικασία που ακολουθήθηκε. Έγινε αναφορά στο σύνολο των δεδομένων FRED χρονοσειρών και περιγραφών που συλλέχθηκε, την προεπεξεργασία την οποία υπέστη, και τα μοντέλα παραγωγής διανυσμάτων προτάσεων που χρησιμοποιήθηκαν. Κάποια από αυτά, βασίζονται σε θεμελιώδη προεκπαιδευμένα LLMs όπως τα BERT και GPT-2 τα οποία προσαρμόστηκαν καταλλήλως, ενώ άλλα όπως το Word2Vec Skip-Gram εκπαιδεύτηκαν μόνο με το σύνολο των δεδομένων των περιγραφών. Στη συνέχεια της πειραματικής διαδικασίας, εξηγήθηκαν τα ποιοτικά χαρακτηριστικά τα οποία εξήχθησαν, και σχεδιάστηκαν τα βαθιά νευρωνικά μοντέλα για το σκοπό της πρόβλεψης.

7.1 Συμπεράσματα

Φτάνοντας στο τέλος της παρούσας διπλωματικής εργασίας, είναι σημαντικό να γίνει μία σύνοψη των κύριων ευρημάτων και να αναδειχτούν τα συμπεράσματα που προέκυψαν από τη μελέτη. Όσον αφορά την ερμηνεία της φυσικής γλώσσας:

- Μέσω της παραγωγής των διανυσμάτων προτάσεων με μοντέλα όπως το Word2Vec και προεκπαιδευμένα θεμελιώδη LLMs όπως τα BERT και GPT-2 που προσαρμόζονται για το συγκεκριμένο σκοπό, παρατηρήθηκε πως μπορεί να επιτευχθεί ιδιαίτερα καλή ερμηνεία και αναπαράσταση των περιγραφών.

- Η ποιότητα των διανυσμάτων αναπαράστασης, επιβεβαιώθηκε χρησιμοποιώντας τα διανύσματα αυτά για να ταξινομηθούν χρονοσειρές με βάση την περιγραφή τους και συγκρίνοντας στη συνέχεια την ταξινόμηση που έγινε από ειδικούς στη βάση δεδομένων

FRED. Επιπλέον, παρατηρήθηκε πως η μείωση των διαστάσεων από 300 σε 100 ή 30 για τα διανύσματα δεν επέφερε ιδιαίτερα χειρότερη αναπαράσταση των περιγραφών. Η εκδοχή Skip-Gram του μοντέλου Word2Vec έχει ιδιαίτερα καλά αποτελέσματα στην αντιπροσωπευτικότητα των διανυσμάτων που παράγει για τις μικρές περιγραφές των χρονοσειρών, ωστόσο οι αναπαραστάσεις που παράχθηκαν από τα μοντέλα GPT-2 και BERT είναι ελάχιστα κατώτερης ποιότητας.

- Την καλύτερη συσταδοποίηση των περιγραφών των χρονοσειρών επιτυγχάνει ο αλγόριθμος k-Means με τα διανύσματα προτάσεων του μοντέλου Word2Vec Skip-Gram όπου η μεταβολή του ορίου σύγκλισης από το 0.1 στο 0.001 είχε ασήμαντη επιρροή.

- Οι τεχνικές που περιλαμβάνουν μεγάλα γλωσσικά μοντέλα έχουν πολύ υψηλότερες απαιτήσεις σε υπολογιστικούς πόρους ακόμα και στην περίπτωση της μετεκπαίδευσής τους, με το χρόνο εκτέλεσης του προγράμματος παραγωγής διανυσμάτων να απαιτεί κατά προσέγγιση 500 φορές περισσότερο χρόνο από τις απλούστερες τεχνικές.

Σχετικά με τα μοντέλα προβλέψεων:

- Όλα τα μοντέλα επηρεάζονται θετικά από τα επιπλέον δεδομένα εκπαίδευσης που παράγουν τα κυλιόμενα παράθυρα. Ωστόσο, λόγω του μεγάλου όγκου των δεδομένων, είναι αναγκαίος ο υψηλός αριθμός εποχών εκπαίδευσης και κόμβων ανά κρυφό επίπεδο για να αφομοιώσουν την πληροφορία που τους παρέχεται. Ενδεικτικά, παρατηρούμε πως για πέντε κυλιόμενα παράθυρα ανά χρονοσειρά, τα μοντέλα συγκλίνουν σε λιγότερες από 50 εποχές, ενώ για τριάντα κυλιόμενα παράθυρα, η εκπαίδευση συγκλίνει στις 100 εποχές.

- Τα μοντέλα που χρησιμοποιούν τις περιγραφές των χρονοσειρών είτε παράλληλα, είτε συνενώνοντας τες στα δεδομένα των χρονοσειρών, επιτυγχάνουν βελτίωση των αποτελεσμάτων σε σχέση με το αρχικό Baseline μοντέλο που λαμβάνει υπόψη μόνο τα δεδομένα των χρονοσειρών. Όσον αφορά το βέλτιστο τρόπο του συνδυασμού, αυτός είναι η παράλληλη είσοδος του MLP-Parallel και MtMs-NLP και όχι η συνένωση του MLP-Concat, η οποία κατά μέσο όρο έχει 48.44% μεγαλύτερο σφάλμα. Επίσης, για το συνδυασμό των παράλληλων κρυφών επιπέδων, η πολλαπλασιαστική μέθοδος αποδεικνύεται καλύτερη.

- Γενικά για όλες τις αρχιτεκτονικές που χρησιμοποιούν τις περιγραφές, η απλουστευμένη μορφή της συστάδας στην οποία ανήκει μία περιγραφή σε κωδικοποίηση one-hot (που παράχθηκε με τον αλγόριθμο k-Means και τα διανύσματα του μοντέλου Word2Vec Skip-Gram), έχει καλύτερες επιδόσεις από τα διανύσματα προτάσεων που περιέχουν περισσότερη πληροφορία.

- Σχετικά με τα ποιοτικά χαρακτηριστικά, φαίνεται πως οι αρχιτεκτονικές που προτάθηκαν δεν επιτυγχάνουν να τα αξιοποιήσουν κατάλληλα, αφού τα μοντέλα που προσπαθούν να τα ενσωματώσουν δεν επιτυγχάνουν καλύτερες προβλέψεις από το μοντέλο αναφοράς.

- Το πιο σύνθετο μοντέλο MtMs-NLP, ξεπερνά σε επίδοση το απλούστερο MLP-Parallel στην περίπτωση των τριμηνιαίων χρονοσειρών. Ωστόσο, το μοντέλο αυτό, λόγω της αυξημένης πολυπλοκότητας του, πέρα από σημαντικούς υπολογιστικούς πόρους για την εκπαίδευσή του, είναι ευάλωτο στην υπερπροσαρμογή.

7.2 Μελλοντικές Προεκτάσεις

Η παρούσα εργασία, εξέτασε τη συνεισφορά συμμεταβλητών όπως περιγραφές και ποιοτικά χαρακτηριστικά των χρονοσειρών στην παραγωγή προβλέψεων με χρήση πολυεπίπεδων νευρωνικών δικτύων. Σχετικά με την παραγωγή των διανυσμάτων για την αναπαράσταση των περιγραφών, αυτά δείχνουν πως έχουν καλά αποτελέσματα και η βελτίωση τους δεν αποτελεί προτεραιότητα.

Όσον αφορά τα μοντέλα πρόβλεψης και την εκπαίδευση τους, είναι εφικτό να πραγματοποιηθούν επεκτάσεις ως προς την αύξηση του όγκου των χρονοσειρών εκπαίδευσης, μέσω της συλλογής δεδομένων από επιπλέον πηγές. Επιπρόσθετα, συμπεριλαμβάνοντας λεκτικές περιγραφές όπως άρθρα ειδήσεων, σχόλια από ειδικούς ή ακόμα και από πολίτες σε κοινωνικά δίκτυα, ενδεχομένως να καθίσταται δυνατό καθοριστούν καλύτερα οι εξωγενείς παράγοντες που επηρεάζουν τις χρονοσειρές σε τομείς όπως αυτοί των των μακροοικονομικών και των χρηματοοικονομικών δεδομένων. Πρέπει να σημειωθεί ότι με την αύξηση του όγκου των δεδομένων, είτε αυτά προέρχονται από χρονοσειρές είτε από λεκτικές περιγραφές, αυξάνεται και η ανάγκη για περισσότερους υπολογιστικούς πόρους.

Για την καλύτερη αφομοίωση της γνώσης, δύναται να χρησιμοποιηθούν πιο περίπλοκα νευρωνικά δίκτυα όπως για παράδειγμα τα Long Short-Term Memory (Hochreiter & Schmidhuber, 1997) που ενδείκνυνται για μεγάλες ακολουθίες δεδομένων εισόδου και έχουν τη δυνατότητα να «θυμούνται» μοτίβα στις χρονοσειρές από το πιο μακρινό παρελθόν. Για παράδειγμα, τα LSTM δίκτυα, έχουν χρησιμοποιηθεί για την πρόβλεψη χρηματιστηριακών αγαθών όπου φαίνεται να υπερτερούν σε σχέση με τα MLP (Cao, Li, & Li, 2019). Λόγω της ιδιότητας της μνήμης μπορούν να εντοπίσουν πιο εύκολα κάποια τυχούσα κυκλικότητα που μπορεί να παρουσιάζουν τα δεδομένα, την οποία στη συνέχεια μπορούν να εκμεταλλευτούν στην πρόβλεψη. Τα δίκτυα αυτά, επιφέρουν ωστόσο μεγαλύτερο υπολογιστικό κόστος όσον αφορά την εκπαίδευσή τους σε σχέση με αυτά που χρησιμοποιήθηκαν στην παρούσα εργασία.

Ένα μοντέλο με σημαντικό όγκο δεδομένων και ανάγκες σε υπολογιστικούς πόρους, δύσκολα δημιουργείται και συντηρείται από κάθε ερευνητή ξεχωριστά. Συνεπώς, μια πρόταση είναι η δημιουργία ένα θεμελιώδους μοντέλου πρόβλεψης χρονοσειρών, ύστερα από μία προεκπαίδευση γενικού σκοπού, να προσαρμόζεται από κάθε ενδιαφερόμενο για το σκοπό της πρόβλεψης που θέλει να επιτύχει. Μία πρόταση για την εκπαίδευση ενός τέτοιου μοντέλου, είναι να γίνει με τεχνικές ομοσπονδιακής μάθησης (federated learning) (Kairouz et al., 2021), με αποκεντροποιημένο τρόπο από πολλούς υπολογιστές ανά τον κόσμο με πολλά διαφορετικά σύνολα δεδομένων, συμψηφίζοντας τις τοπικές ενημερώσεις

στα συναπτικά βάρη. Επιπρόσθετα, με τη διάθεση της αρχιτεκτονικής και της υλοποίησης του μοντέλου στο ευρύ κοινό, θα υπήρχε η δυνατότητα συνεισφοράς με προτάσεις για αλλαγές όσον αφορά τον πηγαίο προγραμματιστικό κώδικα. Για την ενίσχυση της ασφάλειας και της αξιοπιστίας σε μία τέτοια περίπτωση ενδείκνυται να χρησιμοποιηθεί η τεχνολογία blockchain. Με την ενσωμάτωση της, κάθε ενημέρωση του μοντέλου μπορεί να καταγράφεται ως ένα block στην αλυσίδα, καθιστώντας τις ενημερώσεις αναλλοίωτες και ανιχνεύσιμες από όλους. Ανά πάσα στιγμή μπορεί να εξεταστεί το μητρώο των block για να διαπιστωθεί αν κάποιος κόμβος έχει ενεργήσει κακόβουλα, όπως για παράδειγμα να διαπιστωθεί εάν έχει αλλοιώσει την εκπαίδευση του μοντέλου. Επιπλέον, είναι εύκολο να υλοποιηθεί ένα σύστημα ανταμοιβής ή τιμωρίας, με σκοπό να παρέχει κίνητρο στους παράγοντες που συμμετέχουν στην εκπαίδευση να συνεχίσουν να διαθέτουν πόρους προς αυτόν το σκοπό.

Ακόμα, για να γίνει το μοντέλο αυτό πιο προσβάσιμο, προτείνεται να δημιουργηθεί μία διεπαφή διαλόγου (chat) που χρησιμοποιεί κάποιο μεγάλο γλωσσικό μοντέλο, με σκοπό ο χρήστης να «επεξηγεί» τις ιδιαιτερότητες της χρονοσειράς που κατέχει και στη συνέχεια να εισάγει τη χρονοσειρά για την οποία επιθυμεί να λάβει μία πρόβλεψη. Το γλωσσικό μοντέλο, ερμηνεύοντας τη φυσική γλώσσα με την οποία επικοινωνεί ο χρήστης, θα καλεί την υπηρεσία του θεμελιώδους μοντέλου πρόβλεψης και στη συνέχεια θα επιστρέφει κάποιες προβλέψεις σχετικά με τη χρονοσειρά, χρησιμοποιώντας τη γνώση που κατέχει από την εκπαίδευσή του και τα συμφραζόμενα που του δόθηκαν σχετικά με το είδος και την περιγραφή της χρονοσειράς. Σκοπός της προέκτασης αυτής, είναι ο εκδημοκρατισμός της τεχνητής νοημοσύνης για χρήστες οι οποίοι δεν έχουν οικειότητα με τον προγραμματισμό.

Τέλος, προτείνεται να δημιουργηθεί ένα ανοιχτό αποθετήριο δεδομένων όπου ο κάθε ενδιαφερόμενος θα είχε τη δυνατότητα να εισάγει δεδομένα που έχει συλλέξει αυξάνοντας τον όγκο των δεδομένων εκπαίδευσης, καθώς και τη διαφάνεια του μοντέλου αυτού. Η διαφάνεια ως προς τα δεδομένα εκπαίδευσης ενός μοντέλου είναι κύριο ζήτημα. Για το σκοπό αυτό δημιουργήθηκε ο δείκτης The Foundation Model Transparency Index (Bommasani et al., 2023) από ερευνητές των πανεπιστημίων Stanford, Princeton και MIT και έχει ως σκοπό την αξιολόγηση όσον αφορά τη διαφάνεια των γλωσσικών θεμελιωδών μοντέλων και μπορεί να επεκταθεί κατάλληλα και για άλλα είδη θεμελιωδών μοντέλων.

Παραρτήματα

Παράρτημα Α΄

Κωδικοποίηση Κατηγορικών Δεδομένων

Υπάρχουν διάφορες μέθοδοι κωδικοποίησης που χρησιμοποιούνται για τη μετατροπή κατηγορικών δεδομένων σε μορφή που μπορεί να χρησιμοποιηθεί από μοντέλα μηχανικής μάθησης. Σε αυτό το παράρτημα, αναφέρονται οι κύριες τεχνικές.

Α΄.1 Κωδικοποίηση One-hot

Η κωδικοποίηση one-hot αντιπροσωπεύει κάθε κατηγορία με ένα διάνυσμα που έχει τιμή 1 σε μία μόνο θέση και 0 σε όλες τις υπόλοιπες. Αυτό σημαίνει ότι η διάσταση του διανύσματος που προκύπτει είναι ίση με το συνολικό αριθμό κατηγοριών.

$$\vec{x} \in \{0,1\}^n \quad \text{και} \quad \sum_{i=1}^{n} x_i = 1, \text{όπου } n \text{ ο αριθμός των κατηγοριών} \tag{Α΄.1}$$

Για παράδειγμα, για τρεις κατηγορίες: πορτοκάλι, μήλο, ρόδι, το αποτέλεσμα θα ήταν:

$$\vec{x}_{\text{πορτοκάλι}} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad \vec{x}_{\text{μήλο}} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad \vec{x}_{\text{ρόδι}} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \tag{Α΄.2}$$

Η τεχνική one-hot, παρότι παρουσιάζει μεγάλες διαστάσεις, είναι ευρέως χρησιμοποιημένη στη μηχανική μάθηση επειδή με την κωδικοποίηση αυτή, τα διανύσματα που προκύπτουν είναι πλήρως ασυσχέτιστα (αφού κάθε ένα από αυτά είναι ανεξάρτητα) και συνεπώς αντιμετωπίζονται όλα με τον ίδιο τρόπο.

Α΄.2 Απλή Κωδικοποίηση (Dummy Encoding)

Η κωδικοποίηση dummy είναι παρόμοια με την κωδικοποίηση one-hot, χωρίς όμως τον περιορισμό του να υπάρχει κάπου η τιμή 1. Έτσι επιτυγχάνεται η κωδικοποίηση με μικρότερα διανύσματα, μιας και μία από τις κατηγορίες κωδικοποιείται μόνο με μηδενικά. Παρόλα αυτά, η τεχνική του dummy συχνά αποφεύγεται στη μηχανική μάθηση, επειδή μία μηδενική είσοδος διαχειρίζεται διαφορετικά από το μοντέλο.

$$\vec{x} \in \{0,1\}^{n-1}, \text{όπου } n \text{ ο αριθμός των κατηγοριών} \tag{Α΄.3}$$

Για παράδειγμα, για τρεις κατηγορίες: πορτοκάλι, μήλο, ρόδι, το αποτέλεσμα θα ήταν:

$$\vec{x}_{\text{πορτοκάλι}} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \vec{x}_{\text{μήλο}} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \vec{x}_{\text{ρόδι}} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \tag{Α΄.4}$$

Α΄.3 Κωδικοποίηση Ετικέτας (Label Encoding)

Με αυτή τη μέθοδο, κάθε κατηγορία παίρνει έναν ακέραιο αριθμό από το 0 μέχρι $n-1$ όπου n είναι ο αριθμός των κατηγοριών.

$$\vec{x} \in \mathbb{N} \tag{Α΄.5}$$

Για παράδειγμα, για τρεις κατηγορίες: πορτοκάλι, μήλο, ρόδι, το αποτέλεσμα θα ήταν:

$$\text{πορτοκάλι} = 0, \quad \text{μήλο} = 1, \quad \text{ρόδι} = 2 \tag{Α΄.6}$$

Το μειονέκτημα αυτής της μεθόδου είναι πως πολλές φορές ένας μεγάλος ακέραιος αριθμός αντιμετωπίζεται διαφορετικά από έναν μικρό στα μοντέλα μηχανικής μάθησης. Έτσι για κατηγορίες όπως τα φρούτα, θεωρείται λάθος να κωδικοποιηθούν με αυτή τη μέθοδο, μιας και μπορεί το ένα φρούτο με μεγάλη τιμή κωδικοποίησης να λάβει περισσότερη σημασία στο μοντέλο. Η κωδικοποίηση αυτή ενδεικνύεται για παράδειγμα για κατηγορίες που υποδηλώνουν κάποια σειρά, όπως τα μεγέθη ρούχων όπως το Small, Medium, Large.

Α΄.4 Δυαδική Κωδικοποίηση (Binary Encoding)

Η δυαδική κωδικοποίηση χρησιμοποιεί δυαδικούς αριθμούς για την αναπαράσταση των κατηγοριών και έχει το πλεονέκτημα πως μπορεί να εκφράσει πολλές κατηγορίες με μικρές σχετικά διαστάσεις (κατά μέγιστο $\lceil \log_2 n \rceil$ για n κατηγορίες).

$$\vec{x} \in \{0, 1\}^m \quad \text{όπου} \quad m = \lceil \log_2 n \rceil \tag{Α΄.7}$$

Για παράδειγμα, για τέσσερις κατηγορίες: πορτοκάλι, μήλο, ρόδι, καρπούζι, τα διανύσματα μπορούν να εκφραστούν με $\log_2(4) = 2$ διαστάσεις και το αποτέλεσμα είναι:

$$\text{πορτοκάλι} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \text{μήλο} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \text{ρόδι} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \text{καρπούζι} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \tag{Α΄.8}$$

Παράρτημα Β΄

Ποιοτικά Χαρακτηριστικά Χρονοσειρών

Πίνακας Β΄.1: *Πίνακας των στατιστικών χαρακτηριστικών (με **έντονα γράμματα** σημειώνονται τα χαρακτηριστικά που λαμβάνουν μόνο δυαδικές τιμές 0 ή 1, ενώ με * συμβολίζονται τα χαρακτηριστικά που δεν επιλέχθηκαν για τη συνέχεια του πειράματος).*

No	**Feature Name**	**Περιγραφή**
1	trend	Η τάση που παρουσιάζει η χρονοσειρά.
2	seasonality	Η εποχιακότητα της χρονοσειράς.
3	first_order_acf	Η αυτοσυσχέτιση πρώτης τάξης της χρονοσειράς.
4	st_dev*	Η τυπική απόκλιση της χρονοσειράς.
5	variation_coefficient	Η κανονικοποιημένη τυπική απόκλιση ως προς το μέσο όρο της χρονοσειράς.
6	abs_energy	Η απόλυτη ενέργεια της χρονοσειράς ($E = \sum_{i=1}^{n} x_i^2$).
7	absolute_maximum	Η απόλυτη μέγιστη τιμή της χρονοσειράς.
8	absolute_sum_of_-changes	Το απόλυτο άθροισμα των αλλαγών στη χρονοσειρά ($\sum_{i=1}^{n-1} \mid x_{i+1} - x_i \mid$).
9	skewness	Η ασυμμετρία της κατανομής της χρονοσειράς.
10	benford_correlation	Η συσχέτιση του Benford στη χρονοσειρά.
11	binned_entropy	Η εντροπία της χρονοσειράς όταν χωριστεί σε κάδους των 10.
12	**variance_larger_-than_standard_-deviation***	Αν η διακύμανση είναι μεγαλύτερη από την τυπική απόκλιση.
13	sum_of_-reoccurring_data_-points*	Το άθροισμα των επαναλαμβανόμενων σημείων δεδομένων.
14	cid_ce	Η πολυπλοκότητα της χρονοσειράς μέσω του cid_ce ($\sqrt{\sum_{i=1}^{n-1}(x_i - x_{i-1})^2}$).
15	count_above_0*	Ο αριθμός των τιμών πάνω από το 0.
16	count_above_mean*	Ο αριθμός των τιμών πάνω από τον μέσο όρο.
17	count_below_0*	Ο αριθμός των τιμών κάτω από το 0.
18	count_below_mean*	Ο αριθμός των τιμών κάτω από τον μέσο όρο.

No	**Feature Name**	**Περιγραφή**
19	first_location_of_-maximum	Η πρώτη θέση της μέγιστης τιμής στη χρονοσειρά.
20	first_location_of_-minimum	Η πρώτη θέση της ελάχιστης τιμής στη χρονοσειρά.
21	fourier_entropy	Η εντροπία Fourier της χρονοσειράς.
22	**has_duplicate**	Ύπαρξη διπλότυπων.
23	**has_duplicate_max**	Ύπαρξη πάνω από 1 ίδιας μέγιστης τιμής.
24	**has_duplicate_min**	Ύπαρξη πάνω από 1 ίδιας ελάχιστης τιμής.
25	kurtosis	Η κυρτότητα της χρονοσειράς.
26	**large_standard_-deviation**	Ύπαρξη μεγάλης τυπικής απόκλισης στη χρονοσειρά $(std(x) > 0.05 \cdot (max(X) - min(X)))$
27	last_location_of_-maximum*	Η τελευταία θέση της μέγιστης τιμής στη χρονοσειρά.
28	last_location_of_-minimum*	Η τελευταία θέση της ελάχιστης τιμής στη χρονοσειρά.
29	sum_of_-reoccurring_values	Το άθροισμα των επαναλαμβανόμενων τιμών.
30	length*	Το μήκος της χρονοσειράς.
31	longest_strike_-above_mean	Η μεγαλύτερη ακολουθία τιμών πάνω από τον μέσο όρο.
32	longest_strike_-below_mean	Η μεγαλύτερη ακολουθία τιμών κάτω από τον μέσο όρο.
33	maximum*	Η μέγιστη τιμή της χρονοσειράς.
34	mean	Ο μέσος όρος της χρονοσειράς.
35	mean_change	Η μέση αλλαγή στη χρονοσειρά $(\frac{1}{n-1}(x_n - x_1))$.
36	mean_abs_change	Η μέση απόλυτη αλλαγή στη χρονοσειρά.
37	mean_n_absolute_-max	Ο μέσος όρος των απόλυτων μέγιστων τιμών.
38	mean_second_-derivative_central	Η μέση δεύτερη παράγωγος (κεντρική).
39	median*	Η διάμεσος της χρονοσειράς.
40	minimum	Η ελάχιστη τιμή της χρονοσειράς.
41	number_crossing_0	Ο αριθμός που η χρονοσειρά είτε ξεπερνάει είτε πέφτει κάτω από το 0.
42	time_reversal_-asymmetry_-statistic*	Η στατιστική ασυμμετρία αντιστροφής χρόνου $(\mathbb{E}[L^2(X)^2 \cdot L(X) - L(X) \cdot X^2])$.
43	number_peaks	Ο αριθμός κορυφών στη χρονοσειρά. Κορυφή θεωρείται μία τιμή στη χρονοσειρά αν οι προηγούμενες και οι επόμενες 5 τιμές στη χρονοσειρά είναι χαμηλότερες.
44	zero_value_count*	Ο αριθμός μηδενικών τιμών στη χρονοσειρά.

No	**Feature Name**	**Περιγραφή**
45	percentage_of_-reoccurring_-values_to_all_values	Το ποσοστό των επαναλαμβανόμενων τιμών σε σχέση με όλες τις τιμές.
46	permutation_-entropy	Η εντροπία της χρονοσειράς, με βάση τις μεταθέσεις.
47	sum_values*	Το άθροισμα των τιμών της χρονοσειράς.
48	variance*	Η διακύμανση της χρονοσειράς.
49	ratio_beyond_0.5_-sigma	Ο λόγος των τιμών πέρα από 0.5 φορές την τυπική απόκλιση.
50	ratio_value_-number_to_time_-series_length*	Ο λόγος του αριθμού των μοναδικών τιμών προς το συνολικό αριθμό όλων των τιμών.
51	root_mean_square*	Η ρίζα μέσου τετραγώνου της χρονοσειράς.
52	sample_entropy	Η δειγματική εντροπία της χρονοσειράς.

Bibliography

Akgün, E., & Demir, M. (2018, 01). Modeling course achievements of elementary education teacher candidates with artificial neural networks. *International Journal of Assessment Tools in Education*, *5*. doi: 10.21449/ijate.444073

Amari, S. (1967). A theory of adaptive pattern classifiers. *IEEE Transactions on Electronic Computers*, *1*(3), 299–307.

Arora, S., Liang, Y., & Ma, T. (2017). A simple but tough-to-beat baseline for sentence embeddings. In *International conference on learning representations*.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... others (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Bommasani, R., Klyman, K., Longpre, S., Kapoor, S., Maslej, N., Xiong, B., ... Liang, P. (2023). The foundation model transparency index. *arXiv preprint arXiv:2310.12941*.

Box, G. E. P., & Cox, D. R. (1964, 12). An Analysis of Transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, *26*(2), 211-243. doi: 10.1111/j.2517-6161.1964.tb00553.x

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*, 123–140.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–32.

Brownlee, J. (2018). *Deep learning for time series forecasting: predict the future with mlps, cnns and lstms in python*. Machine Learning Mastery.

Cao, J., Li, Z., & Li, J. (2019). Financial time series forecasting model based on ceemdan and lstm. *Physica A: Statistical mechanics and its applications*, *519*, 127–139.

Castells, M. (1996). The information age: Economy, society and culture (3 volumes). *Blackwell, Oxford*, *1997*, 1998.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., St. John, R., ... Kurzweil, R. (2018). Universal sentence encoder for English. In E. Blanco & W. Lu (Eds.), *Proceedings of the 2018 conference on empirical methods in natural language processing: System demonstrations* (pp. 169–174). Brussels, Belgium: Association for Computational Linguistics. doi: 10.18653/v1/D18-2029

Cleveland, R. B., Cleveland, W. S., McRae, J. E., Terpenning, I., et al. (1990). Stl: A seasonal-trend decomposition. *J. Off. Stat*, *6*(1), 3–73.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, *13*(1), 21–27.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd* (Vol. 96, pp. 226–231).

Freund, Y., Schapire, R. E., et al. (1996). Experiments with a new boosting algorithm. In *icml* (Vol. 96, pp. 148–156).

Georgouli, A. (2015). *Τεχνητή νοημοσύνη*. Kallipos, Open Academic Editions. doi: 10 .57713/kallipos-666

Glorot, X., Bordes, A., & Bengio, Y. (2011). Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics* (pp. 315–323).

Goodwin, P., & Lawton, R. (1999). On the asymmetry of the symmetric mape. *International Journal of Forecasting*, *15*(4), 405-408. doi: https://doi.org/10.1016/S0169-2070(99)00007-2

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: principles and practice*. OTexts.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679-688. doi: https://doi.org/10.1016/j.ijforecast.2006.03.001

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., ... others (2021). Advances and open problems in federated learning. *Foundations and trends® in machine learning*, *14*(1–2), 1–210.

Kang, Y., Hyndman, R. J., & Smith-Miles, K. (2017). Visualising forecasting algorithm performance using time series instance spaces. *International Journal of Forecasting*, *33*(2), 345–358.

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In *International conference on machine learning* (pp. 1188–1196).

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, *1*(4), 541–551.

Linnainmaa, S. (1970). *Algoritmin kumulatiivinen pyöristysvirhe yksittäisten pyöristysvirheiden taylor-kehitelmänä* (Thesis).

Lloyd, S. (1982). Least squares quantization in pcm. *IEEE transactions on information theory*, *28*(2), 129–137.

Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development*, *1*(4), 309-317. doi: 10.1147/rd.14.0309

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). The m4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, *36*(1), 54-74. (M4 Competition) doi: https://doi.org/10.1016/j.ijforecast.2019.04.014

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2022). M5 accuracy competition: Results, findings, and conclusions. *International Journal of Forecasting*, *38*(4), 1346-

1364. (Special Issue: M5 competition) doi: https://doi.org/10.1016/j.ijforecast.2021.11.013

Makridakis, S., Spiliotis, E., Hollyman, R., Petropoulos, F., Swanson, N., & Gaba, A. (2023). The m6 forecasting competition: Bridging the gap between forecasting and investment decisions. *arXiv preprint arXiv:2310.13357*.

Makridakis, S., Wheelwright, S. C., & Hyndman, R. J. (2008). *Forecasting methods and applications*. John wiley & sons.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, *5*, 115–133.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Pearson, K. (1901). Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, *2*(11), 559–572.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, *14*(3), 130–137.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, *1*, 81–106.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, *66*(336), 846–850.

Rasul, K., Ashok, A., Williams, A. R., Khorasani, A., Adamopoulos, G., Bhagwatkar, R., ... others (2023). Lag-llama: Towards foundation models for time series forecasting. *arXiv preprint arXiv:2310.08278*.

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, *65*(6), 386.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, *20*, 53–65.

Schwab, K. (2017). *The fourth industrial revolution*. Crown Currency.

Semenoglou, A.-A., Spiliotis, E., Makridakis, S., & Assimakopoulos, V. (2021). Investigating the accuracy of cross-learning time series forecasting methods. *International Journal of Forecasting*, *37*(3), 1072-1084. doi: https://doi.org/10.1016/j.ijforecast.2020.11.009

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379-423. doi: 10.1002/j.1538-7305.1948.tb01338.x

Soenksen, L. R., Ma, Y., Zeng, C., Boussioux, L., Villalobos Carballo, K., Na, L., ... Bertsimas, D. (2022). Integrated multimodal artificial intelligence framework for healthcare applications. *NPJ digital medicine*, *5*(1), 149.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, *28*(1), 11–21. doi: 10.1108/eb026526

Spiliotis, E., Kouloumos, A., Assimakopoulos, V., & Makridakis, S. (2020). Are forecasting

competitions data representative of the reality? *International Journal of Forecasting*, *36*(1), 37–53.

Staněk, F. (2023). A note on the m6 forecasting competition: Designing parametric models with hypernetworks. *Available at SSRN 4355794*.

Theodorou, E., Wang, S., Kang, Y., Spiliotis, E., Makridakis, S., & Assimakopoulos, V. (2022). Exploring the representativeness of the competition data. *International Journal of Forecasting*, *38*(4), 1500-1506. doi: https://doi.org/10.1016/j.ijforecast.2021.07.006

URL. (n.d.-a). *Amazon web services*. https://docs.aws.amazon.com/machine-learning/latest/dg/model-fit-underfitting-vs-overfitting.html. (Ημερομηνία πρόσβασης: 28-06-2024)

URL. (n.d.-b). *Fred api documentation*. https://fred.stlouisfed.org/docs/api/fred/. (Ημερομηνία πρόσβασης: 03-11-2023)

URL. (n.d.-c). *Gpt-2 output dataset*. https://github.com/openai/gpt-2-output-dataset. (Ημερομηνία πρόσβασης: 30-04-2024)

URL. (n.d.-d). *Hugging face models*. https://huggingface.co/models. (Ημερομηνία πρόσβασης: 03-06-2024)

URL. (n.d.-e). *Nltk documentation*. https://www.nltk.org/. (Ημερομηνία πρόσβασης: 03-06-2024)

URL. (n.d.-f). *Reddit*. https://www.reddit.com/. (Ημερομηνία πρόσβασης: 29-04-2024)

URL. (n.d.-g). *Word vectors figure*. https://www.cs.cmu.edu/~dst/WordEmbeddingDemo/tutorial.html. (Ημερομηνία πρόσβασης: 23-04-2024)

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, *58*(301), 236–244.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., ... others (2022). Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big data*, *3*, 1–40.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... others (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.