ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ ΚΑΙ ΥΠΗΡΕΣΙΕΣ

ΚΑΤΕΥΘΥΝΣΗ : ΜΕΓΑΛΑ ΔΕΔΟΜΕΝΑ ΚΑΙ ΑΝΑΛΥΤΙΚΗ

Μεταπτυχιακή Διπλωματική Εργασία

Θέμα: ***Εκτίμηση Επενδυτικών Αποφάσεων που Βασίζονται σε Τεχνικές Ανάλυσης Συναισθήματος***

ΝΤΙΝΤΙΦΑ ΓΕΩΡΓΙΑ ΑΜ: ΜΕ1938

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ: ΜΙΧΑΗΛ ΦΙΛΙΠΠΑΚΗΣ

Μάρτιος 2022

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον καθηγητή μου κ. Φιλιππάκη Μιχαήλ για την ευκαιρία που μου έδωσε να ασχοληθώ με ένα τόσο ενδιαφέρον θέμα, για τη συνεχή του καθοδήγηση, για την διαρκή του συμμετοχή παρέχοντας συμβουλές καθ' όλη τη διάρκεια της εκπόνησης και διεκπεραίωσης αυτής της διπλωματικής εργασίας. Επίσης θα ήθελα να ευχαριστήσω τη ΔΡ. Μαρία Ελένη Πούλου για την πολύτιμή βοήθεια της στην επίβλεψη της διπλωματικής.

Τέλος, θα ήθελα να εκφράσω την ευγνωμοσύνη μου στην οικογένειά μου και τους φίλους μου για την συμπαράσταση τους και την στήριξη που μου παρείχαν καθ'όλη την διάρκεια της σταδιοδρομίας μου μέχρι και σήμερα.

Περίληψη

Τα τελευταία χρόνια, λόγω της διείσδυσης του Διαδικτύου στο σύνολο σχεδόν του παγκόσμιου πληθυσμού, τα κοινωνικά μέσα χρησιμοποιούνται όλο και περισσότερο από τις επιχειρήσεις και το ευρύ κοινό, με αποτέλεσμα να παράγεται ένας μεγάλος όγκος δεδομένων, ο οποίος είναι διαθέσιμος σε μορφή κειμένου για την εξόρυξη απόψεων ή αλλιώς την ανάλυση συναισθημάτων που περιέχουν. Η ανάλυση συναισθήματος είναι μια μέθοδος μηχανικής μάθησης για την εξαγωγή του προσανατολισμού απόψεων (θετικών, αρνητικών, ουδέτερων) από κείμενα που έχουν γραφτεί για κάποιο προϊόν, οργανισμό, πρόσωπο ή οποιαδήποτε άλλη οντότητα. Ιδιαίτερη ανάπτυξη γνωρίζει η έρευνα των τεχνικών ανάλυσης συναισθήματος για την πρόβλεψη της διάθεσης που δημιουργείται από τις ειδήσεις που αντλούνται από τα κοινωνικά μέσα, προκειμένου αυτή η διάθεση να χρησιμοποιηθεί για την πρόβλεψη της κίνησης των χρηματιστηρίων. Για τους επενδυτές της χρηματιστηριακής αγοράς, η πρόβλεψη των τάσεων της κίνησης των χρηματιστηρίων θεωρείται ζωτικής σημασίας για τη λήψη αποφάσεων, εφόσον στόχος μιας επένδυσης είναι το οικονομικό κέρδος.

Στην παρούσα διπλωματική εργασία διερευνώνται τεχνικές ανάλυσης συναισθήματος από κείμενα τα οποία προέρχονται από ειδήσεις της αμερικανικής ιστοσελίδας συγκέντρωσης ειδήσεων Reddit, με στόχο την πρόβλεψη της κίνησης του χρηματιστηριακού δείκτη Dow Jones εφαρμόζοντας μεθόδους Μηχανικής Μάθησης. Οι τεχνικές ανάλυσης συναισθήματος υλοποιούνται με τη βοήθεια βιβλιοθηκών της python. Οι τεχνικές βασίζονται α) στην χρήση τεχνικών επεξεργασίας κειμένου με τη χρήση της βιβλιοθήκης Μηχανικής Μάθησης scikit-learn και την εφαρμογή αλγόριθμων ταξινόμησης β) στη χρήση της λειτουργίας ανάλυσης συναισθημάτων της βιβλιοθήκης επεξεργασίας κειμένου TextBlob και του λεξικού VADER της βιβλιοθήκης NLTK που βασίζεται σε κανόνες και οι οποίες αποτελούν μεθόδους μη επιβλεπόμενης Μηχανικής Μάθησης. Σκοπός είναι η εύρεση της καλύτερης τεχνικής για την πρόβλεψη της τάσης του χρηματιστηριακού δείκτη.

Τα καλύτερα αποτελέσματα της υλοποίησης προέκυψαν με την εφαρμογή του αλγόριθμου ταξινόμησης της λογιστικής παλινδρόμησης, εφαρμόζοντας τεχνικές επεξεργασίας κειμένου με τη χρήση της βιβλιοθήκης Μηχανικής Μάθησης scikit-learn.

Λέξεις – Κλειδιά: Μηχανική Μάθηση, Επεξεργασία Φυσικής Γλώσσας, Ανάλυση Συναισθήματος, VADER, TextBlob, Πρόβλεψη Χρηματιστηρίου

Πίνακας Περιεχομένων

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Συντομογραφίες & Ακρωνύμια

AI	Artificial Intelligence
API	Application Programming Interface
ANN	Artificial Neural Network
AUC	Area Under the Curve
BoW	Bag-of-Words
CV	Cross Validation
DJIA	Dow Jones Industrial Average
DL	Deep Learning
EDA	Exploratory Data Analysis
LG	Logistic Regression
MCC	Matthews Correlation Coefficient
ML	Machine Learning
MLP	Multi Layer Perceptron
NB	Naïve Bayes
NLP	Natural Language Processing
NLTK	Natural Language ToolKit
RF	Random Forest
ROC	Receiving Operating Characteristic
SA	Sentiment Analysis
SVM	Support Vector Machine
TF-IDF	Term Frequency – Inverse Document Frequency
VADER	Valence Aware Dictionary and sEntiment Reasoner
ΒΜ	Βαθιά Μάθηση
ΗΠΑ	Ηνωμένες Πολιτείες Αμερικής
ΜΜ	Μηχανική Μάθηση
ΤΝ	Τεχνητή Νοημοσύνη
ΤΝΔ	Τεχνητό Νευρωνικό Δίκτυο

1 Εισαγωγή

1.1 Εισαγωγή

Η πρόβλεψη των χρηματιστηρίων είναι ένας τομέας, ο οποίος παρουσιάζει ιδιαίτερο ενδιαφέρον για την ακαδημαϊκή έρευνα και τις επιχειρήσεις. Για τους επενδυτές της χρηματιστηριακής αγοράς, η πρόβλεψη των τάσεων της κίνησης των χρηματιστηρίων θεωρείται ζωτικής σημασίας για τη λήψη αποφάσεων, εφόσον στόχος μιας επένδυσης είναι το οικονομικό κέρδος. Υπάρχουν πολλές στατιστικές μέθοδοι για την πρόβλεψη, οι οποίες βασίζονται σε ιστορικά δεδομένα. Όμως, η πρόβλεψη του χρηματιστηρίου εμφανίζει πολλές προκλήσεις λόγω του ότι, τα χρηματιστηριακά δεδομένα είναι από τη φύση τους σύνθετα, δυναμικά, μη σταθερά και με μεγάλο βαθμό αβεβαιότητας. Η πρόβλεψη γίνεται ακόμη πιο δύσκολη επειδή έχει αποδειχθεί ότι, υπάρχουν διάφοροι παράγοντες, όπως οικονομικοί, πολιτικοί και κοινωνικοί, οι οποίοι επηρεάζουν την χρηματοοικονομική αγορά και διαμορφώνουν την κίνηση των χρηματιστηρίων [1], [2].

Για την ανάπτυξη μιας αποτελεσματικής στρατηγικής για τη λήψη επενδυτικών αποφάσεων θεωρείται ζωτικής σημασίας η διερεύνηση των πολλαπλών παραγόντων που επηρεάζουν τη χρηματιστηριακή αγορά. Τα τελευταία χρόνια, ιδιαίτερη ανάπτυξη στον τομέα της ακαδημαϊκής έρευνας παρουσιάζει ο τομέας των τεχνικών εξόρυξης απόψεων ή αλλιώς της ανάλυσης συναισθήματος που προέρχεται από τις ειδήσεις -ιδιαίτερα αυτών που προέρχονται από τα κοινωνικά μέσα, καθώς θεωρείται ότι, επηρεάζει τις προβλέψεις για τα χρηματιστήρια [3], [4]. Η ανάλυση συναισθήματος είναι μια μέθοδος μηχανικής μάθησης για την εξαγωγή του προσανατολισμού απόψεων (θετικών, αρνητικών, ουδέτερων) από κείμενα που έχουν γραφτεί για κάποιο προϊόν, οργανισμό, πρόσωπο ή οποιαδήποτε άλλη οντότητα και βασίζεται στην επεξεργασία φυσικής γλώσσας. Η ανάπτυξη της έρευνας οφείλεται στο γεγονός ότι, καθημερινά όλο και περισσότεροι χρήστες του Διαδικτύου χρησιμοποιούν τα κοινωνικά μέσα, με αποτέλεσμα να παράγεται ένας τεράστιος όγκος δεδομένων σε μορφή κειμένου, από τον οποίο μπορεί με την εφαρμογή μεθόδων επεξεργασίας φυσικής γλώσσας να αντληθεί πολύτιμη πληροφορία για διάφορους τομείς δραστηριοτήτων, συμπεριλαμβανομένου και του οικονομικού τομέα στον οποίο ανήκουν τα χρηματιστήρια [5].

Χαρακτηριστικό είναι το γεγονός ότι, σύμφωνα με στατιστικά στοιχεία [6], τον Ιούλιο του 2021 οι χρήστες των κοινωνικών μέσων ανέρχονται σε 4,48 δισεκατομμύρια,

κάτι που αντιστοιχεί περίπου στο 57% του παγκόσμιου πληθυσμού, παρουσιάζοντας μια ετήσια αύξηση που αριθμεί σε 520 εκατομμύρια χρήστες σε σχέση με τον Ιούλιο του 2020. Εάν ληφθεί υπόψη το γεγονός ότι, πολλές εταιρείες απαγορεύουν τη χρήση των κοινωνικών μέσων σε χρήστες ηλικίας μικρότερης των 13 ετών, τότε το ποσοστό των χρηστών ανάγεται στο 70% του παγκόσμιου πληθυσμού. Επίσης, ενδιαφέρον είναι το γεγονός ότι, κάθε χρήστης δαπανά κατά μέσο όρο 2,5 ώρες καθημερινά για τη χρήση των κοινωνικών μέσων. Γνωστά κοινωνικά μέσα είναι το Facebook[1], το Twitter[2], το YouTube[3], το Reddit[4] κλπ. και 17 από αυτά, που θεωρούνται τα δημοφιλέστερα, συγκεντρώνουν 300 εκατομμύρια χρήστες σε μηνιαία βάση. Κυριότερες πηγές ειδήσεων από τα δημοφιλέστερα κοινωνικά μέσα είναι το Twitter και το Reddit. Από τα στατιστικά στοιχεία μπορεί να γίνει άμεσα αντιληπτός ο λόγος που οι ερευνητές έχουν στραφεί στην άντληση πληροφορίας από τα κοινωνικά μέσα.

Ο μεγαλύτερος όγκος των ερευνών για το κατά πόσο οι ειδήσεις που προέρχονται από τα κοινωνικά μέσα και κατά πόσο το συναίσθημα που δημιουργείται μέσα από αυτές επηρεάζει την πρόβλεψη του χρηματιστηρίου, αφορά στην άντληση των ειδήσεων από το Twitter. Θεμελιώδης εργασία για το θέμα θεωρείται η εργασία του Bollen [7] που δημοσιεύτηκε το 2011 ο οποίος έδειξε ότι, η διάθεση του κοινού που αναλύεται μέσω των ροών του twitter συσχετίζεται με τον βιομηχανικό μέσο όρο του χρηματιστηριακού δείκτη Dow Jones (Dow Jones Industrial Average - DJIA). Μετά την εργασία του Bollen ακολούθησαν πολλές μελέτες που αφορούν την εφαρμογή τεχνικών ανάλυσης συναισθήματος για την επιρροή της στην πρόβλεψη των χρηματιστηρίων και πολλές από αυτές αναφέρονται στις έρευνες [1]–[4], [8], [9].

Ωστόσο, υπάρχουν πολύ λίγες μελέτες οι οποίες αφορούν την ανάλυση συναισθήματος από άλλες πηγές, όπως είναι η ιστοσελίδα συγκέντρωσης ειδήσεων Reddit. Έτσι, για την παρούσα εργασία επιλέχθηκε η χρήση των νέων που προέρχονται από το Reddit για την πρόβλεψη του χρηματιστηρίου. Αυτό που διαφοροποιεί το Reddit από τα άλλα κοινωνικά μέσα είναι ότι, οι αναρτήσεις κατηγοριοποιούνται σε υποπίνακες (που ονομάζονται «subreddit») και που επικεντρώνονται σε συγκεκριμένα θέματα.

[1] https://www.facebook.com/

[2] https://twitter.com/

[3] https://www.youtube.com/

[4] https://www.reddit.com/

1.2 Αντικείμενο Διπλωματικής Εργασίας

Αντικείμενο της παρούσας διπλωματικής εργασίας είναι η εκτίμηση επενδυτικών αποφάσεων που βασίζονται σε τεχνικές ανάλυσης συναισθήματος.

Για τον σκοπό αυτό διερευνώνται τεχνικές ανάλυσης συναισθήματος από κείμενα τα οποία προέρχονται από ειδήσεις της αμερικανικής ιστοσελίδας συγκέντρωσης ειδήσεων Reddit, με στόχο την πρόβλεψη της ανόδου ή καθόδου του χρηματιστηριακού δείκτη Dow Jones εφαρμόζοντας μεθόδους επιβλεπόμενης Μηχανικής Μάθησης για δυαδική ταξινόμηση, καθώς και μεθόδους μη επιβλεπόμενης Μηχανικής Μάθησης που βασίζεται σε λεξικά και σε κανόνες.

Το σύνολο δεδομένων που επιλέχθηκε να χρησιμοποιηθεί έχει τίτλο «Daily News for Stock Market Prediction» [9]. Αν και αναφέρεται στο kaggle ως ένα σύνολο δεδομένων, περιέχει ουσιαστικά τρία σύνολα δεδομένων σε μορφή .csv. Πιο συγκεκριμένα, περιέχει ένα σύνολο με ιστορικά δεδομένα οκτώ ετών (2008-2016), το RedditNews.csv που αφορά διεθνείς ειδήσεις από ένα subreddit του Reddit, το Reddit WorldNews Channel (/r/worldnews), ένα σύνολο με χρηματιστηριακά δεδομένα για το ίδιο χρονικό διάστημα, το DJIA_table.csv που αντλήθηκαν από τον ιστότοπο Yahoo Finance[5] και ένα τρίτο σύνολο, το CombinedNewsDJIA.csv που συνδυάζει τα δεδομένα των δύο συνόλων. Το σύνολο αυτό, το οποίο θα χρησιμοποιηθεί στην παρούσα εργασία, περιέχει τις ημερομηνίες, τις ειδήσεις και μία στήλη που προέκυψε από την ανάλυση των χρηματιστηριακών δεδομένων από τον συγγραφέα για τον προσδιορισμό της τάσης του δείκτη, έτσι ώστε το σύνολο αυτό να χρησιμοποιηθεί για την εύρεση του καλύτερου προγνωστικού μοντέλου. Με την επιλογή αυτού του συνόλου, δίνεται η δυνατότητα πειραματισμού των τεχνικών ανάλυσης συναισθήματος.

Οι τεχνικές ανάλυσης συναισθήματος υλοποιούνται με τη βοήθεια βιβλιοθηκών της python. Οι τεχνικές βασίζονται στην χρήση τεχνικών επεξεργασίας κειμένου που προκύπτει από τις ειδήσεις α) με τη βοήθεια της βιβλιοθήκης scikit-learn και β) με τη χρήση της λειτουργίας ανάλυσης συναισθήματος που προσφέρεται από τις βιβλιοθήκη επεξεργασίας κειμένου TextBlob και τη χρήση της λειτουργίας ανάλυσης συναισθήματος VADER που προσφέρεται από την βιβλιοθήκη επεξεργασίας κειμένου NLTK και η οποία βασίζεται σε λεξικό και κανόνες.

[5] https://finance.yahoo.com/

Στόχος της παρούσας διπλωματικής εργασίας είναι να βρεθεί η καλύτερη τεχνική ανάλυσης συναισθήματος για την πρόβλεψη της ανόδου ή καθόδου του χρηματιστηριακού δείκτη DJIA για το επιλεγμένο σύνολο δεδομένων.

1.3 Δομή Διπλωματικής Εργασίας

Το Κεφάλαιο 1 της εργασίας αποτελεί την εισαγωγή στο θέμα της εργασίας.

Στο Κεφάλαιο 2, αρχικά δίνεται συνοπτικά το απαιτούμενο θεωρητικό υπόβαθρο. Αναφέρονται οι βασικές έννοιες που αφορούν το χρηματιστήριο, τους χρηματιστηριακούς δείκτες και τις μετοχές. Στη συνέχεια παρουσιάζονται βασικά στοιχεία για τη Μηχανική Μάθηση, την Ανάλυση Συναισθήματος και δημοφιλείς αλγόριθμοι δυαδικής ταξινόμησης. Τέλος, αναφέρονται τα μέτρα για την αποτίμηση της απόδοσης των αλγόριθμων.

Στο Κεφάλαιο 3 γίνεται η βιβλιογραφική επισκόπηση που αφορά ακαδημαϊκές εργασίες σχετικές με την παρούσα διπλωματική εργασία.

Στο Κεφάλαιο 4 παρουσιάζεται το περιβάλλον υλοποίησης σε Python, και η μεθοδολογία για την ανάπτυξη του έργου.

Στο Κεφάλαιο 5 παρουσιάζονται τα συγκριτικά αποτελέσματα της υλοποίησης των αλγόριθμων και συζητούνται τα αποτελέσματα.

Στο Κεφάλαιο 6 αναφέρονται τα συμπεράσματα και οι μελλοντικές επεκτάσεις.

2 Υπόβαθρο

Στο παρόν κεφάλαιο παρουσιάζονται οι βασικές έννοιες που αφορούν το χρηματιστήριο, τους χρηματιστηριακούς δείκτες και τις μετοχές. Στη συνέχεια παρουσιάζονται βασικά στοιχεία για τη Μηχανική Μάθηση, την Ανάλυση Συναισθήματος και δημοφιλείς αλγόριθμοι δυαδικής ταξινόμησης. Τέλος, αναφέρονται τα μέτρα για την αποτίμηση της απόδοσης των αλγόριθμων.

2.1 Χρηματιστήριο και Χρηματιστηριακοί Δείκτες

Ως Χρηματιστήριο ή Χρηματιστηριακή Αγορά (stock market) ορίζεται η οργανωμένη αγορά, η οποία συνήθως είναι επίσημα αναγνωρισμένη από το κράτος, όπου συναντώνται οι ενδιαφερόμενοι για την διενέργεια αγοροπωλησιών κινητών αξιών (όπως μερίδια κεφαλαίου ανωνύμων εταιρειών, δηλαδή μετοχών), τραπεζικά, κρατικά ή άλλα ομόλογα) ή/και εμπορευμάτων. Μετοχή (stock) σημαίνει μια μερική ιδιοκτησία σε μια εταιρεία ή έναν κλάδο, με δικαίωμα συμμετοχής στα κέρδη της. Ένα άτομο, που επενδύει σε μετοχές ή αγοράζει μετοχές μιας εταιρείας, ονομάζεται μέτοχος αυτής της εταιρείας. Τα χρηματιστήρια αποτελούν ιδιόμορφες αγορές με την έννοια της ταυτόχρονης ύπαρξης της προσφοράς και της ζήτησης. Αποτελούν οικονομικό θεσμό που κατά κανόνα αναγνωρίζεται από τα κράτη όπου λειτουργούν με νομοθετικά και διοικητικά μέτρα, τα οποία καθορίζουν το πλαίσιο μέσα στο οποίο διαμορφώνονται οι αγορές και θεσπίζουν τις προϋποθέσεις και τους όρους λειτουργίας τους [10].

Τα χρηματιστήρια δημιουργήθηκαν από την ανάγκη για την εξεύρεση βραχυπρόθεσμων μακροπρόθεσμων κεφαλαίων και τη σύναψη αγοροπωλησιών μεγάλων ποσοτήτων εμπορευμάτων που βρίσκονται μακριά από τον τόπο διαπραγμάτευσης τους. Είναι σε οργανωμένη μορφή προκειμένου να διενεργούνται άμεσα και γρήγορα οι συναλλαγές, να φαίνονται δημόσια και με διαύγεια τα χαρακτηριστικά τους, όπως για παράδειγμα η προσφορά και η ζήτηση. Τα χρηματιστήρια αφενός δίνουν την ευκαιρία στις επιχειρήσεις για την εύρεση κεφαλαίων και αφετέρου σε επενδυτές που ενδιαφέρονται να διαθέσουν χρήματα σε τίτλους με την προσδοκία του οικονομικού κέρδους. Γενικά θεωρείται ότι, τα χρηματιστήρια συμβάλλουν στην τόνωση της παραγωγικότητας και γενικότερα στην ανάπτυξη της χώρας όπου λειτουργούν [10].

Στην Ελλάδα λειτουργεί το Χρηματιστήριο Αθηνών[6] - ΧΑ, ΧΑΑ (Athens Stock Exchange -ASE, ATHEX), όπου λειτουργούν πέντε αγορές: η οργανωμένη αγορά αξιών, η οργανωμένη αγορά παραγώγων, η εναλλακτική αγορά, η αγορά άνθρακα (για EUAs) και η εξωχρηματιστηριακή αγορά.

2.1.1 Χρηματιστηριακοί Δείκτες

Ο χρηματιστηριακός δείκτης (stock market index), ή αλλιώς δείκτης μετοχών, είναι μια μέτρηση της συνολικής αξίας ενός χρηματιστηρίου, ή ενός τομέα του. Υπολογίζεται από τις τιμές επιλεγμένων μετοχών και αποτελεί έναν σταθμισμένο μέσο όρο, με αποτέλεσμα οι πιο ακριβές μετοχές τον επηρεάζουν περισσότερο από ότι οι μετοχές με χαμηλότερες τιμές. Οι χρηματιστηριακοί δείκτες χρησιμοποιούνται αφενός από τους επενδυτές για την περιγραφή της κατάστασης της αγοράς και τη σύγκριση των αποδόσεων των επενδύσεων τους και, αφετέρου, για την περιγραφή της κατάστασης της οικονομίας μιας χώρας. Συνήθως θεωρείται ότι, οι χρηματιστηριακοί δείκτες βοηθούν τους επενδυτές να συγκρίνουν τα τρέχοντα επίπεδα τιμών με τις προηγούμενες τιμές για τον υπολογισμό της απόδοσης της αγοράς. Σκοπός των δεικτών είναι η ύπαρξη ενός αξιόπιστου μέτρου καταγραφής των τάσεων των μετοχών των εισηγμένων εταιριών που διαπραγματεύονται στα διάφορα χρηματιστήρια [11].

Οι χρηματιστηριακοί δείκτες τμηματοποιούνται με διάφορους τρόπους. Ένας τρόπος είναι με βάση κάποιους κανόνες για τον τρόπο κατανομής των μετοχών στον δείκτη, ανεξάρτητα από την κάλυψη των μετοχών του. Ένας άλλος τρόπος είναι με βάση το σύνολο των μετοχών της κάλυψης του δείκτη. Οι μετοχές ομαδοποιούνται ανάλογα με τα υποκείμενα οικονομικά τους ή τη βασική ζήτηση των επενδυτών που ο δείκτης προσπαθεί να αντιπροσωπεύσει ή να παρακολουθήσει. Για παράδειγμα, ένας παγκόσμιος χρηματιστηριακός δείκτης είναι ο MSCI World[7] περιλαμβάνει μετοχές από όλο τον κόσμο και ικανοποιεί τη ζήτηση των επενδυτών για δείκτες για ευρείες παγκόσμιες μετοχές [11].

Οι χρηματιστηριακοί δείκτες κάλυψης χωρών αφορούν την απόδοση της χρηματιστηριακής αγοράς ενός συγκεκριμένου κράτους και αντικατοπτρίζει το συναίσθημα των επενδυτών για την κατάσταση της οικονομίας του. Οι πιο συχνά αναφερόμενοι δείκτες είναι οι εθνικοί δείκτες, οι οποίοι συνιστώνται από μετοχές μεγάλων εταιρειών εισηγμένων στα μεγαλύτερα χρηματιστήρια μιας χώρας, όπως για παράδειγμα ο

[6] https://www.athexgroup.gr/el/

[7] https://www.msci.com/

δείκτης Standard and Poor's 500 (S&P 500)[8] και ο Dow Jones στις Ηνωμένες Πολιτείες Αμερικής (ΗΠΑ) και αφορούν μόνο εταιρείες υπό διαπραγμάτευση σε μια συγκεκριμένη χώρα.

Στην Ελλάδα ο αντίστοιχος χρηματιστηριακός δείκτης είναι ο FTSE/X.A.20[9], που έχει ως σκοπό την καταγραφή, σε πραγματικό χρόνο, των τάσεων των τιμών των μετοχών των είκοσι μεγαλύτερων σε κεφαλαιοποίηση εισηγμένων εταιρειών του Χρηματιστηρίου Αθηνών. Για περισσότερες πληροφορίες που αφορούν το Χρηματιστήριο, τους Χρηματιστηριακούς δείκτες του Χρηματιστηρίου Αθηνών, καθώς και για τον τρόπο που υπολογίζονται οι χρηματιστηριακοί δείκτες, παραπέμπουμε τον αναγνώστη στην [12], καθώς το θέμα είναι εκτός σκοπού της παρούσας εργασίας.

2.1.2 Ο Χρηματιστηριακός Δείκτης Dow Jones Industrial Average (DJIA)

Ο χρηματιστηριακός δείκτης Dow Jones Industrial Average (DJIA) [13], ή απλά Dow Jones, είναι δείκτης σταθμισμένων τιμών μετοχών 30 πασίγνωστων εταιριών που υπάρχουν στα χρηματιστήρια των ΗΠΑ. Είναι από τους παλαιότερους (δημιουργήθηκε στις 26 Μαΐου του 1896) και πλέον χρησιμοποιούμενους δείκτες, αν και θεωρείται ότι δεν είναι από τους πλέον αντιπροσωπευτικούς για την αναπαράσταση της χρηματιστηριακής αγοράς των ΗΠΑ, όπως για παράδειγμα ο S&P 500 [14].

Ο DJIA περιλαμβάνει μόνο 30 μεγάλες εταιρείες εξαιτίας του γεγονότος ότι, όταν δημιουργήθηκε οι εισηγμένες εταιρείες στα χρηματιστήρια ήταν λιγοστές. Οι τιμές του δείκτη είναι το άθροισμα των τιμών των μετοχών, διαιρεμένες με έναν συντελεστή, ο οποίος αλλάζει όταν υφίσταται διάσπαση (split) κάποιας μετοχής, έτσι ώστε η αξία του δείκτη να μην επηρεάζεται από τη διάσπαση της συγκεκριμένης μετοχής [14]. Στον Πίνακα 1 δίνονται οι εταιρείες και το αντίστοιχο βάρος τους, οι οποίες συμπεριλαμβανόταν στον DJIA στις 29/7/2021, σύμφωνα με την [15].

2.1.3 Τα Οικονομικά Δεδομένα ως Χρονοσειρές

Οι χρονοσειρές ή χρονικές σειρές (time series) είναι μια καλώς οργανωμένη διάταξη δεδομένων ή ένα σύνολο σημείων δεδομένων που λαμβάνει μια μεταβλητή σε ίσα χρονικά διαστήματα. Ο στόχος της ανάλυσης των χρονοσειρών είναι η ανάπτυξη μοντέλων που είναι σε θέση να περιγράψουν τις δεδομένες χρονοσειρές με λογική ακρίβεια. Η συστηματική

[8] https://www.spglobal.com/spdji/en/indices/equity/sp-500/#overview

[9] https://www.athexgroup.gr/el/index-profile/-/select-index/62

έρευνα των χρονοσειρών είναι μεγάλης σημασίας ώστε να προβλεφθεί η μεταβλητότητα των δεδομένων στο μέλλον, με βάση τις προηγούμενες παρατηρήσεις των δεδομένων. Έτσι, βασικά η πρόβλεψη της χρονοσειράς μπορεί να χαρακτηριστεί ως μια προσέγγιση για την εκτίμηση του μέλλοντος κατανοώντας το παρελθόν και η ανάλυση χρονοσειρών μπορεί να θεωρηθεί ως παράγοντας λήψης αποφάσεων για μελλοντικά επενδυτικά σχέδια και εκτιμήσεις. Ως εκ τούτου, πρέπει να δοθεί η κατάλληλη προσοχή κατά την επιλογή ενός μοντέλου χρονοσειρών, ώστε να επιτευχθούν καλύτερα αποτελέσματα για τις προβλέψεις [16].

Οι τιμές των μετοχών και των χρηματιστηριακών δεικτών αντιμετωπίζονται ως χρονοσειρές. Οι πληροφορίες που παρέχονται για τις μετοχές ή τους δείκτες από διάφορους ιστότοπους (όπως για παράδειγμα ο Yahoo Finance[10]) για διάφορα χρονικά διαστήματα, αφορούν αριθμητικά δεδομένα, όπως φαίνεται για παράδειγμα στο Σχήμα 1 για τις τιμές του DJIA στις 2/8/2021.

Πίνακας 1. Οι Εταιρίες του DJIA στις 29/7/2021

Α/Α	**Εταιρία**	**Σύμβολο**	**Βάρος**
1	UnitedHealth Group Incorporated	UNH	7,745717
2	Goldman Sachs Group Inc.	GS	7,040709
3	Home Depot Inc.	HD	6,203092
4	Microsoft Corporation	MSFT	5,413985
5	Visa Inc. Class A	V	4,687251
6	salesforce.com inc.	CRM	4,637616
7	Amgen Inc.	AMGN	4,590415
8	McDonald's Corporation	MCD	4,569812
9	Honeywell International Inc.	HON	4,293915
10	Boeing Company	BA	4,230232
11	Caterpillar Inc.	CAT	3,952649
12	3M Company	MMM	3,777334
13	Walt Disney Company	DIS	3,347849
14	American Express Company	AXP	3,233032
15	Johnson & Johnson	JNJ	3,219172

[10] https://finance.yahoo.com/

16	NIKE Inc. Class B	NKE	3,09218
17	JPMorgan Chase & Co.	JPM	2,840446
18	Apple Inc.	AAPL	2,790623
19	Travelers Companies Inc.	TRV	2,787439
20	International Business Machines	IBM	2,674121
21	Walmart Inc.	WMT	2,671499
22	Procter & Gamble Company	PG	2,627483
23	Chevron Corporation	CVX	1,890821
24	Merck & Co. Inc.	MRK	1,446726
25	Dow Inc.	DOW	1,161464
26	Coca-Cola Company	KO	1,068749
27	Verizon Communications Inc.	VZ	1,044775
28	Cisco Systems Inc.	CSCO	1,038968
29	Intel Corporation	INTC	1,017241
30	Walgreens Boots Alliance Inc	WBA	0,885005

Όπως παρατηρούμε στο Σχήμα 1, οι τιμές που δίνονται αφορούν [17]:

- Την τιμή ανοίγματος *(Open)*: είναι η πρώτη τιμή κάθε εισηγμένης μετοχής στην αρχή μιας συναλλαγής την ημέρα διαπραγμάτευσης.
- Την υψηλότερη και την χαμηλότερη τιμή κατά την διάρκεια της ημέρας διαπραγμάτευσης *(High* και *Low* αντίστοιχα): αυτά τα δεδομένα χρησιμοποιούνται για τη μέτρηση της μεταβλητότητας των μετοχών.
- Την τιμή κλεισίματος *(Close*): είναι η τιμή της μετοχής στο τέλος της ημέρας διαπραγμάτευσης
- Την προσαρμοσμένη τιμή κλεισίματος (*Adj*[11] *Close*): θεωρείται ως η πραγματική τιμή της μετοχής και δείχνει την αξία της μετοχής μετά τη διανομή μερισμάτων.
- Τον όγκο συναλλαγών (*Volume*): είναι ο αριθμός των μετοχών ή των συμβολαίων που διαπραγματεύονται για τίτλους σε όλες τις αγορές κατά τη διάρκεια μιας δεδομένης χρονικής περιόδου.

[11] Συντομογραφία του Adjusted

Dow Jones Industrial Average (^DJI) Add to watchlist

DJI - DJI Real Time Price. Currency in USD

Advertisement

34,935.47 -149.03 (-0.42%)

At close: July 30 5:06PM EDT

Summary | Chart | Conversations | **Historical Data** | Options | Components

Time Period: Aug 02, 2020 - Aug 02, 2021 | Show: Historical Prices | Frequency: Daily | Apply

Currency in USD

Date	Open	High	Low	Close*	Adj Close**	Volume
Jul 30, 2021	35,013.26	35,106.30	34,871.13	34,935.47	34,935.47	276,410,000
Jul 29, 2021	34,985.99	35,171.52	34,985.99	35,084.53	35,084.53	222,680,000
Jul 28, 2021	35,109.95	35,116.37	34,876.84	34,930.93	34,930.93	347,170,000
Jul 27, 2021	35,078.90	35,078.90	34,878.07	35,058.52	35,058.52	326,610,000
Jul 26, 2021	35,055.86	35,150.37	34,950.19	35,144.31	35,144.31	259,790,000
Jul 23, 2021	34,855.11	35,095.33	34,855.11	35,061.55	35,061.55	314,040,000
Jul 22, 2021	34,799.68	34,879.28	34,673.03	34,823.35	34,823.35	291,610,000
Jul 21, 2021	34,556.96	34,820.24	34,556.96	34,798.00	34,798.00	317,090,000

Σχήμα 1. Εμφάνιση Ιστορικών Δεδομένων του DJIA από τον ιστότοπο Yahoo Finance

Παράλληλα, δίνεται η δυνατότητα γραφικής απεικόνισης της χρονικής εξέλιξης των τιμών- συνήθως για την τιμή Close- επιλέγοντας τα επιθυμητά χρονικά διαστήματα, πχ για 2 χρόνια, όπως φαίνεται στο Σχήμα 2 για την ίδια ημερομηνία (2/8/2021).

Σχήμα 2. Γραφική Παράσταση της Εξέλιξης Τιμών του DJIA από τον ιστότοπο Yahoo Finance

Οι τιμές που ανακτώνται μπορούν να χρησιμοποιηθούν απευθείας σε προγνωστικά μοντέλα ως εξαρτημένες ή ανεξάρτητες μεταβλητές. Επίσης, μπορούν να υπολογιστούν διάφορα νέα χαρακτηριστικά, όπως για παράδειγμα το κέρδος που θα προκύψει, όπως αναφέρεται στην [4], σύμφωνα με τον τύπο:

$$Gain = Close\_Price_t - Close\_Price_{t-1} / Close\_Price_{t-1}$$

όπου $Close\_Price_t$ και $Close\_Price_{t-1}$ είναι οι τιμές κλεισίματος για την τρέχουσα και την προηγούμενη ημερομηνία αντίστοιχα.

Η τάση ανόδου ή καθόδου του δείκτη είναι ένα άλλο μέγεθος το οποίο προκύπτει από αυτές τις τιμές και εξάγεται με βάση τον τύπο :

$$Trend = \begin{cases} 0, \text{εάν } Close\_price_{diff} \leq 0 \\ 1, \quad \text{αλλιώς} \end{cases}$$

όπου $Close\_price_{diff} = Close\_price_{for\_current\_day} - Close\_price_{for\_previous\_day}$ και αναφέρονται στις αντίστοιχες τιμές για την τρέχουσα ημέρα και την προηγούμενη ημέρα [4], [18].

2.2 Μηχανική Μάθηση

Ζούμε στην εποχή των Μεγάλων Δεδομένων, όπου ιδιαίτερη αξία έχει αποκτήσει η εξαγωγή, ή αλλιώς η εξόρυξη χρήσιμης πληροφορίας από τους τεράστιους όγκους των δεδομένων που παράγονται καθημερινά από το Διαδίκτυο. Η Μηχανική Μάθηση -ΜΜ (Machine Learning – ML) παρέχει την τεχνική βάση για την Εξόρυξη Δεδομένων (Data Mining) και χρησιμοποιείται για την εξαγωγή πληροφορίας από ακατέργαστα δεδομένα που προέρχονται από διάφορες πηγές δεδομένων. Η πληροφορία που εξάγεται μπορεί να εκφραστεί σε κατανοητή μορφή και να χρησιμοποιηθεί για διάφορους σκοπούς. Η ΜΜ είναι υποσύνολο της Τεχνητής Νοημοσύνης -ΤΝ (Artificial Intelligence - AI), μιας ευρύτερης έννοιας της επιστήμης των υπολογιστών που πρωτοεμφανίστηκε στη δεκαετία του 1950 και αφορά τη δημιουργία έξυπνων μηχανών που μπορούν να προσομοιώσουν τις ανθρώπινες ικανότητες.

Ένας απλός ορισμός της ΤΝ δόθηκε το 1990 από τους Rich & Knight: *«Τεχνητή Νοημοσύνη είναι η μελέτη του πώς να κάνουμε τους υπολογιστές ικανούς να κάνουν πράγματα στα οποία προς το παρόν οι άνθρωποι τα καταφέρνουν καλύτερα»* [19].

Για ένα μεγάλο χρονικό διάστημα υπήρχε η πεποίθηση ότι, η ΤΝ μπορεί να επιτευχθεί με τη συγγραφή προγραμμάτων, όπου περιλαμβάνεται ένα σύνολο ρητών κανόνων για τη διαχείριση της γνώσης. Αυτή η προσέγγιση είναι γνωστή ως συμβολική ΤΝ

ή ΤΝ βασισμένη στη γνώση. Η συμβολική ΤΝ είναι κατάλληλη για την επίλυση σαφώς καθορισμένων λογικών προβλημάτων, όπως για παράδειγμα ενός παιχνιδιού όπως το σκάκι, όπου οι κανόνες του παιχνιδιού είναι συγκεκριμένοι και μπορούν εύκολα να διατυπωθούν [20].

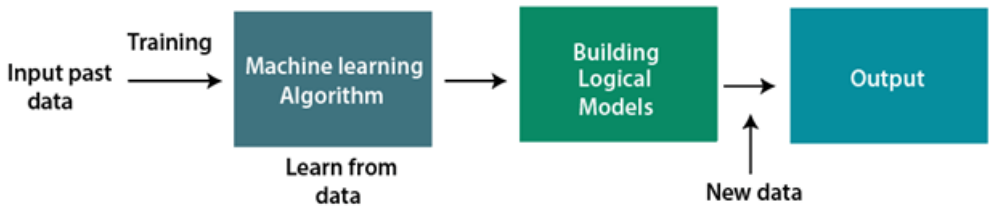
Το ερώτημα εάν ένας υπολογιστής μπορεί να προγραμματιστεί έτσι ώστε, να εκτελεί μια εργασία, όπως για παράδειγμα την αναγνώριση μιας εικόνας ή την μετάφραση ενός κειμένου, μαθαίνοντας τους κανόνες μόνο μέσα από την εξέταση των δεδομένων οδήγησε στην ΜΜ, μια άλλη προσέγγιση για την ΤΝ. Το 1959 δίνεται ο γενικός ορισμός της ΜΜ από τον Arthur Samuel: *«Η Μηχανική Μάθηση είναι το πεδίο μελέτης το οποίο δίνει στους υπολογιστές τη δυνατότητα να μαθαίνουν χωρίς να έχουν ρητά προγραμματιστεί».* Ένας πιο περιεκτικός ορισμός είναι αυτός που δίνει Tom Mitchell το 1997, και ο οποίος ουσιαστικά προσδιορίζει τι είναι μάθηση: *«Ένα πρόγραμμα υπολογιστή λέγεται ότι μαθαίνει από την εμπειρία E (Experience) σε σχέση με κάποια εργασία T (Task) και κάποιο μέτρο απόδοσης P (Performance), εάν βελτιωθεί η απόδοσή του στο T, όπως μετράται από το P, με εμπειρία E»* [20].

Για παράδειγμα, η εργασία *T*, μπορεί να είναι η πρόβλεψη της τιμής μιας μετοχής. Οι ειδικοί στον τομέα συλλέγουν ιστορικά δεδομένα που αφορούν στοιχεία τιμών ή οικονομικών δεικτών, των λεγόμενων *χαρακτηριστικών (features)* ή αλλιώς *γνωρισμάτων (attributes)* που, σύμφωνα με τη γνώση που διαθέτουν οι ειδικοί, θα βοηθούσαν στην πρόβλεψη της τιμής της μετοχής. Αυτά τα δεδομένα ονομάζονται παραδείγματα ή στιγμιότυπα. Για κάθε παράδειγμα υπάρχει μια *ετικέτα (label)* όπου δηλώνεται η τιμή της μετοχής. Η τιμή αυτή αναφέρεται και ως *στόχος (target)*. Το σύνολο των παραδειγμάτων σε αυτή τη μορφή αποτελούν ένα *σύνολο δεδομένων (dataset)* ή αλλιώς *σύνολο εκπαίδευσης (training set)* μέσω του οποίου θα αποκτηθεί η εμπειρία *E*. Ο αλγόριθμος ΜΜ με συνεχείς δοκιμές και επαναλήψεις καλείται να βρει μια όσο το δυνατόν απλούστερη σχέση με βάση τα χαρακτηριστικά που του δίνονται από τα παραδείγματα, για να εξάγει το ορθό αποτέλεσμα για τον στόχο. Το απαιτούμενο επίπεδο ορθότητας του αλγόριθμου εξαρτάται από την εκάστοτε εργασία και είναι το μέτρο απόδοσης *P* του αλγόριθμου.

Συνήθως, το ενδιαφέρον επικεντρώνεται στο πόσο καλά αποδίδει ένας αλγόριθμος ΜΜ σε δεδομένα που δεν έχει ξαναδεί, γιατί έτσι προσδιορίζεται πόσο αποτελεσματικά θα χρησιμοποιηθεί σε εφαρμογές του πραγματικού κόσμου. Έτσι, η απόδοση του αλγόριθμου αποτιμάται με την χρήση ενός *συνόλου δοκιμής (test set),* το οποίο διαχωρίζεται από τα δεδομένα που χρησιμοποιούνται για την εκπαίδευση του συστήματος ΜΜ. Η δυνατότητα

της καλής απόδοσης ενός αλγόριθμου σε νέα δεδομένα, τα οποία το σύστημα δεν έχει ξαναδεί ονομάζεται *γενίκευση (generalization)* [20].

Το σύστημα ΜΜ μαθαίνει από ιστορικά δεδομένα, δημιουργεί μοντέλα πρόβλεψης με σκοπό όταν θα λάβει νέα δεδομένα τα οποία δεν έχει ξαναδεί, να προβλέψει την έξοδο για αυτά τα δεδομένα. Η ακρίβεια της προβλεπόμενης εξόδου εξαρτάται από την ποσότητα των δεδομένων, καθώς η τεράστια ποσότητα δεδομένων βοηθά στη δημιουργία ενός καλύτερου μοντέλου που προβλέπει την έξοδο με μεγαλύτερη ακρίβεια. Έτσι, σε περίπλοκα προβλήματα, αντί να γραφεί ένα πρόγραμμα για την επίλυσή του, χρειάζεται μόνο η τροφοδοσία με δεδομένα σε γενικούς αλγόριθμους και με τη βοήθεια αυτών των αλγορίθμων, η μηχανή δημιουργεί τα λογικά μοντέλα και προβλέπει την έξοδο [20]. Η διαδικασία της ΜΜ απεικονίζεται στο Σχήμα 3.

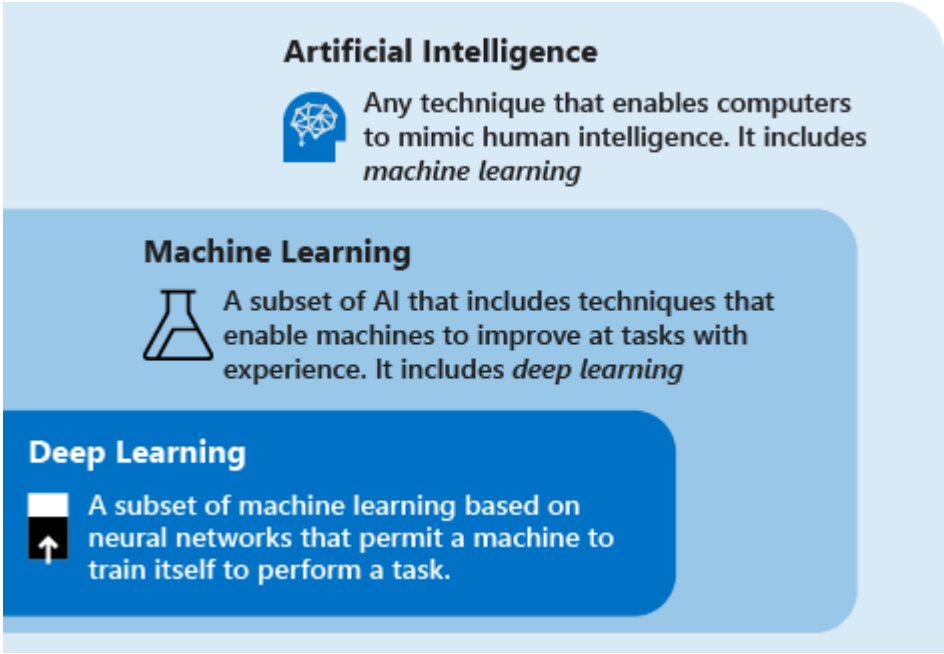


Σχήμα 3. Η Διαδικασία Μάθησης στην Μηχανική Μάθηση [21]

Η Βαθιά Μάθηση – ΒΜ (Deep Learning- DL) είναι μια κατηγορία αλγορίθμων ΜΜ εμπνευσμένη από τη δομή του ανθρώπινου εγκεφάλου, τα Τεχνητά Νευρωνικά Δίκτυα - ΤΝΔ (Artificial Neural Networks - ANN) που δίνουν τη δυνατότητα στους υπολογιστές να βελτιώνονται με την εμπειρία και τα δεδομένα. Η βασική αρχή για την ΒΜ είναι η μάθηση μέσω πολλαπλών επιπέδων σύνθεσης από μη γραμμικούς μετασχηματισμούς δεδομένων εισόδου. Οι βασικές θεμελιώδεις ιδέες για την ΒΜ διατυπώθηκαν στη δεκαετία του 1980, όπως για παράδειγμα ο αλγόριθμος οπισθοδιάδοσης που διατυπώθηκε το 1986 και η υπολογιστική όραση με τη βοήθεια των Συνελικτικών Νευρωνικών Δικτύων που υλοποιήθηκε το 1989, αλλά η ΒΜ άρχισε να αναπτύσσεται μετά το 2006 και ιδιαίτερα μετά το 2012 λόγω της προόδου στο υλικό (hardware) των υπολογιστών, της αύξησης των συνόλων των δεδομένων, των αλγοριθμικών προόδων και της ανάπτυξης έτοιμων βιβλιοθηκών λογισμικού [20]. Ωστόσο, οι αλγόριθμοι ΒΜ για να αποδώσουν καλά, απαιτούν πολύ περισσότερα δεδομένα από τους παραδοσιακούς αλγόριθμους ΜΜ και συνήθως εφαρμόζονται σε εργασίες που αφορούν μικρά σύνολα δεδομένων. Επιπλέον, ο τεράστιος όγκος των δεδομένων που απαιτούνται για τους αλγόριθμους ταξινόμησης ΒΜ

επιδεινώνει περαιτέρω την υπολογιστική πολυπλοκότητα κατά τη διάρκεια του σταδίου εκπαίδευσης [22].

Στο Σχήμα 4 απεικονίζεται η συσχέτιση της ΤΝ με την ΜΜ και την ΒΜ.



Σχήμα 4. Τεχνητή Νοημοσύνη- Μηχανική Μάθηση – Βαθιά Μάθηση [23]

2.2.1 Κατηγορίες Μηχανικής Μάθησης

Η βασική κατηγοριοποίηση των αλγόριθμων ΜΜ γίνεται ανάλογα με το είδος της εμπειρίας που έχουν κατά την εκπαίδευση. Οι αλγόριθμοι αποκτούν την εμπειρία, δηλαδή εκπαιδεύονται, με βάση το σύνολο εκπαίδευσης. Έτσι, κατηγοριοποιούνται ανάλογα με το εάν η εκπαίδευση γίνεται με ανθρώπινη επίβλεψη ή όχι. Σύμφωνα με αυτό το κριτήριο, υπάρχουν τρεις κύριες κατηγορίες αλγορίθμων μάθησης [20]:

- *Επιβλεπόμενη ή εποπτευόμενη μάθηση (supervised learning)*
- *Μη επιβλεπόμενη ή μη εποπτευόμενη μάθηση (unsupervised learning)*
- *Ενισχυτική μάθηση (reinforcement learning)*

Στην *επιβλεπόμενη μάθηση* στο σύστημα ΜΜ παρέχεται ένα σύνολο ακατέργαστων δεδομένων με βάση το οποίο θα αποκτήσει εμπειρία ο αλγόριθμος. Το σύνολο των δεδομένων περιέχει τα παραδείγματα με τα χαρακτηριστικά τους σε μορφή διανύσματος $\boldsymbol{x}$ και κάθε παράδειγμα του συνόλου συνδέεται με ένα στόχο (target) ή αλλιώς ετικέτα (label) y. Τέτοιου είδους δεδομένα, ονομάζονται *δεδομένα με ετικέτα (labeled)*. Στόχος του συστήματος είναι η ανακάλυψη της σχέσης ανάμεσα στον στόχο y, ο οποίος αναφέρεται

και ως εξαρτημένη μεταβλητή, και τα χαρακτηριστικά ***x,*** τα οποία αναφέρονται ως ανεξάρτητες μεταβλητές**.**

Οι κύριες εργασίες της επιβλεπόμενης μάθησης, είναι δύο:

- *Η ταξινόμηση ή κατηγοριοποίηση (classification):* από το σύστημα ΜΜ αναμένεται η πρόβλεψη για την ταξινόμηση των δεδομένων σε διάφορες εκ των προτέρων γνωστές κατηγορίες. Σε περίπτωση που οι κατηγορίες προς ταξινόμησης είναι δύο, τότε έχουμε *τη δυαδική ταξινόμηση (binary classification)*, ενώ εάν είναι περισσότερες έχουμε την ταξινόμηση *πολλαπλών κλάσεων (multiclass classification).*
- *Η παλινδρόμηση (regression):* από το σύστημα ΜΜ αναμένεται η πρόβλεψη μιας τιμής.

Στην *μη-επιβλεπόμενη μάθηση* ο αλγόριθμος εκπαιδεύεται στο σύνολο των δεδομένων που περιέχει τα χαρακτηριστικά, μαθαίνοντας τις χρήσιμες ιδιότητες της δομής του συνόλου δεδομένων. Τα παραδείγματα δεν περιέχουν ετικέτες (unlabeled). Η μη επιβλεπόμενη μάθηση αφορά την παρατήρηση διάφορων παραδειγμάτων και προσπαθεί έμμεσα ή άμεσα να μάθει την κατανομή πιθανότητας *p(x)* ή τις ενδιαφέρουσες ιδιότητες αυτής της κατανομής.

Οι πιο συνηθισμένες εργασίες μη επιβλεπόμενης μάθησης είναι [24]:

- *Η συσταδοποίηση (clustering)*: από το σύστημα ΜΜ αναμένεται ο διαχωρισμός των δεδομένων σε ομοειδείς συστάδες.
- *Η μείωση διαστάσεων (dimensionality reduction)*: από το σύστημα ΜΜ αναμένεται η εξαγωγή χρήσιμων ιδιοτήτων των δεδομένων με απλοποίηση των δεδομένων χωρίς να χαθεί όμως χρήσιμη πληροφορία.
- *Η εκμάθηση κανόνα συσχέτισης (association rule learning)*: στόχος του συστήματος ΜΜ είναι η ανακάλυψη συσχετίσεων μεταξύ των χαρακτηριστικών ενός πολύ μεγάλου όγκου δεδομένων.

Η *ενισχυτική μάθηση* είναι μια τελείως διαφορετική κατηγορία αλγορίθμων. Υπάρχει ένα σύστημα εκμάθησης, ο πράκτορας (agent) που μπορεί να παρατηρεί το περιβάλλον, να επιλέγει και να εκτελεί ενέργειες. Εάν επιλέγει την ορθή ενέργεια επιβραβεύεται και στην αντίθετη περίπτωση τιμωρείται. Σε αυτή την περίπτωση ο πράκτορας θα πρέπει να μαθαίνει μόνος του και ο στόχος είναι η συνεχής επιβράβευση.

Επίσης, στις κατηγορίες της ΜΜ εντάσσεται και η *ημι-επιβλεπόμενη μάθηση (semi-supervised learning),* η οποία αποτελεί συνδυασμό επιβλεπόμενης και μη επιβλεπόμενης μάθησης, όπου το σύστημα ΜΜ καλείται να εκπαιδευτεί σε σύνολο δεδομένων που περιέχει δομημένα παραδείγματα, αλλά και αδόμητα.

2.2.2 Η Μηχανική Μάθηση στον Χρηματιστηριακό Τομέα

Η εφαρμογή της ΜΜ στον χρηματηστηριακό τομέα, είναι ένα πεδίο που ελκύει την ακαδημαϊκή κοινότητα και τις επιχειρήσεις από τη δεκαετία του 1990, όταν ξεκίνησε η μεγάλη τεχνολογική ανάπτυξη και η ευρεία διάδοση των προσωπικών υπολογιστών. Έκτοτε, έχουν προταθεί πολυάριθμες προσεγγίσεις για την αντιμετώπιση του προβλήματος της πρόβλεψης των τιμών στο χρηματιστήριο (stock market prediction) και των τάσεων των δεικτών των μετοχών και των χρηματιστηριακών δεικτών [8].

Στις στατιστικές μεθόδους που χρησιμοποιούνται για τις προβλέψεις εφαρμόζονται ολοένα και περισσότερο μέθοδοι ΜΜ και μάλιστα κατά κανόνα επιβλεπόμενης ΜΜ. Σε συνδυασμό με όσα αναφέρθηκαν στην παράγραφο 2.1.3 και στην προηγούμενη παράγραφο, το πρόβλημα της πρόβλεψης της τιμής ενός δείκτη ή μιας μετοχής ανάγεται στη δημιουργία ενός μοντέλου παλινδρόμησης, ενώ το πρόβλημα της πρόβλεψης της τάσης ανόδου ή καθόδου ενός δείκτη ή μιας μετοχής ανάγεται στη δημιουργία ενός μοντέλου δυαδικής ταξινόμησης.

Οι στατιστικές μέθοδοι για την πρόβλεψη του χρηματιστηρίου βασίζονται σε ιστορικά δεδομένα τιμών, δηλαδή στις χρονοσειρές και αφορούν την τεχνική ανάλυση (technical analysis). Ωστόσο, έχει αποδειχθεί ότι, υπάρχουν πολλοί παράγοντες που επηρεάζουν τα χρηματιστήρια, όπως οικονομικοί, πολιτικοί και κοινωνικοί. Λαμβάνοντας αυτούς τους παράγοντες, υπάρχει μια άλλη κατεύθυνση ανάλυσης, η θεμελιώδης ανάλυση (fundamental analysis) [17].

Γενικά, το θέμα των προβλέψεων τιμών ή τάσεων στον χρηματιστηριακό τομέα δημιουργεί μεγάλες προκλήσεις, λόγω της δυναμικής και της πολυπλοκότητας που παρουσιάζει η χρηματιστηριακή αγορά. Παράλληλα με την τεχνολογική πρόοδο και την διείσδυση του Διαδικτύου σε κάθε τομέα του πραγματικού κόσμου- συνεπώς και στον οικονομικό, παράγονται τεράστιες ποσότητες δεδομένων που αφορούν την πληροφόρηση μέσω των ειδήσεων σε μορφή κειμένου και που προέρχονται από διάφορες πηγές, όπως είναι οι ιστότοποι οικονομικών νέων, γενικών ειδήσεων, αλλά και των κοινωνικών μέσων [5]. Πολλοί ερευνητές έχουν πειραματιστεί στην επιρροή που ασκούν οι ειδήσεις και οι γνώμες που εκφράζονται όσον αφορά τις τάσεις στη διαμόρφωση της χρηματιστηριακής

αγοράς. Από διάφορες ακαδημαϊκές μελέτες έχει αποδειχθεί ότι, υπάρχει συσχέτιση της χρηματιστηριακής αγοράς από τις ειδήσεις και μάλιστα είναι ένα πεδίο που ελκύει όλο και περισσότερο το ενδιαφέρον των ερευνητών [8].

Η τεράστια ποσότητα των παραγόμενων δεδομένων κειμένου σε συνδυασμό με τις προόδους στην ΜΜ οδήγησε στην ανάγκη εφαρμογών της *Επεξεργασίας Φυσικής Γλώσσας (Natural Language Processing – NLP)*. Η NLP ως τομέας της ΜΜ, ορίζεται ευρέως ως ο αυτόματος χειρισμός της φυσικής γλώσσας, όπως η ομιλία και το κείμενο, με τη βοήθεια του λογισμικού. Η μελέτη της επεξεργασίας φυσικής γλώσσας υπάρχει εδώ και περισσότερα από 50 χρόνια, αλλά αποκτά όλο και περισσότερο ενδιαφέρον λόγω της διαθεσιμότητας δεδομένων και των διαφόρων τεχνικών ΜΜ που αναπτύχθηκαν την τελευταία δεκαετία.

Η έρευνα σε τεχνικές NLP βρίσκει εφαρμογή σε πολλούς τομείς και ο οικονομικός τομέας δεν αποτελεί εξαίρεση. Σύμφωνα με τους Zing et al. [9], έχει δημιουργηθεί πλέον ένας νέος τομέας, οι οικονομικές προβλέψεις που βασίζονται στη φυσική γλώσσα (Natural Language based Financial Forecasting – NLFF) ή από την άποψη εφαρμογής ο τομέας πρόβλεψης του χρηματιστηρίου. Στο Σχήμα 5, που παρέχεται από τους συγγραφείς, απεικονίζεται η NLFF ως η τομή της NLP και της χρηματοοικονομικής πρόβλεψης.

Η ανάγκη για την εξόρυξη πληροφορίας από κείμενα (text mining) που προέρχονται από ειδήσεις, σε συνδυασμό με το γεγονός ότι οι ειδήσεις επηρεάζουν τον οικονομικό τομέα και ιδιαίτερα τον χρηματιστηριακό, οδήγησε στην αναζήτηση μεθόδων για την εξόρυξη άποψης (opinion mining) για τις προβλέψεις. Ο ισοδύναμος όρος με την εξόρυξη άποψης ο οποίος έχει επικρατήσει είναι η Ανάλυση Συναισθήματος (Sentiment Analysis – SA).

Σχήμα 5. Η Επεξεργασία Φυσικής Γλώσσας στον Οικονομικό Τομέα (NLFF) [9]

2.3 Ανάλυση Συναισθήματος

Γενικά, ως Ανάλυση Συναισθήματος (ή ισοδύναμα εξόρυξη άποψης) ορίζεται η υπολογιστική μελέτη των απόψεων, των συναισθημάτων και των διαθέσεων. Όσον αφορά εφαρμογές που εμπίπτουν στην επεξεργασία φυσικής γλώσσας, η ανάλυση συναισθήματος βοηθά στον εντοπισμό της διάθεσης που εκφράζεται μέσα από ένα κείμενο για μια συγκεκριμένη οντότητα ή αντικείμενο. Η διάθεση αυτή μπορεί να είναι θετική, αρνητική ή ουδέτερη. Απαραίτητη προϋπόθεση για την ανάλυση συναισθήματος είναι η συλλογή δεδομένων κειμένου που είναι γραμμένα σε φυσική γλώσσα. Έτσι, είναι απαραίτητη η εφαρμογή μεθόδων NLP [25].

Στον οικονομικό τομέα, το πρωταρχικό κίνητρο για εργασίες που αφορούν την ανάλυση συναισθήματος είναι η συσχέτιση των απόψεων με την πρόβλεψη της μελλοντικής κατεύθυνσης των τιμών των μετοχών και των χρηματιστηριακών δεικτών [5]. Καθώς υπάρχει η άποψη από πολλούς ερευνητές ότι, οι ειδήσεις είναι βασικός και καθοριστικός παράγοντας στη διαμόρφωση των τάσεων για την κίνηση των χρηματιστηριακών αγορών, τα τελευταία χρόνια γνωρίζει ιδιαίτερη ανάπτυξη η πρόβλεψη των χρηματιστηρίων με βάση το συναίσθημα που δημιουργείται από τις ειδήσεις. Η ανάπτυξη αυτή βασίζεται την ιδέα ότι, οι ειδήσεις και οι γνώμες καθοδηγούν τους επενδυτές, δημιουργούν τάσεις στις τιμές των μετοχών και έτσι παρέχουν τη δυνατότητα σε ένα επενδυτή να αποκομίσει οικονομικό κέρδος από τις χρηματιστηριακές συναλλαγές [8].

Το συναίσθημα μπορεί να μοντελοποιηθεί ως πρόβλημα ταξινόμησης. Η ταξινόμηση επιτυγχάνεται με δύο προσεγγίσεις, οι οποίες θα παρουσιαστούν αναλυτικότερα σε επόμενα κεφάλαια. Οι δύο αυτές προσεγγίσεις είναι [18]:

- *Με τη χρήση μεθόδων επιβλεπόμενης ΜΜ:* εφαρμόζονται τεχνικές NLP για την επεξεργασία δεδομένων κειμένου με ετικέτες [22].
- *Με τη χρήση μεθόδων μη- επιβλεπόμενης ΜΜ:* εφαρμόζονται λεξιλόγια (lexicon). Τα δεδομένα κειμένου κατηγοριοποιούνται χρησιμοποιώντας λεξικά (dictionaries) από λέξεις που έχουν θετικό ή αρνητικό σημασιολογικό προσανατολισμό. Στη συνέχεια, ο αλγόριθμος αναζητά στο κείμενο τις γνωστές λέξεις και συνδυάζει τους σημασιολογικούς προσανατολισμούς εξάγοντας τον μέσο όρο ή αθροίζοντας τις συσχετιζόμενες τιμές ή τα αποτελέσματα. Η προσέγγιση αυτή αναφέρεται και ως βασισμένη σε κανόνες (rule based) [26], [27].

Η δημιουργία μοντέλων πρόβλεψης χρηματιστηρίου με χρήση μεθόδων επιβλεπόμενης μάθησης μπορεί να αφορά την πρόβλεψη τιμής, οπότε το θέμα αφορά τη

δημιουργία ενός μοντέλου παλινδρόμησης ή την πρόβλεψη της τάσης, οπότε το θέμα αφορά τη δημιουργία ενός μοντέλου ταξινόμησης. Συνεπώς, επιλέγεται η τεχνική με βάση την οποία θα εξαχθεί το συναίσθημα και στη συνέχεια επιλέγεται η τεχνική ΜΜ με βάση την οποία θα προβλεφθεί ο στόχος.

Στην επόμενη παράγραφο παρουσιάζονται επιλεγμένοι αλγόριθμοι ταξινόμησης, οι οποίοι παράλληλα βρίσκουν εφαρμογή και στις τεχνικές NLP για κείμενο και κατ' επέκταση στην ανάλυση συναισθήματος.

2.4 Αλγόριθμοι Ταξινόμησης

Οι αλγόριθμοι ΜΜ για την ταξινόμηση είναι πολυάριθμοι. Ωστόσο, για θέματα που αφορούν τις τεχνικές ανάλυσης συναισθήματος με μεθόδους ΜΜ, η οποία εμπίπτει στην ευρύτερη κατηγορία του προβλήματος της ταξινόμησης κειμένου με σκοπό τις προβλέψεις, ιδιαίτερα δημοφιλείς είναι συγκεκριμένοι αλγόριθμοι, τους κυριότερους από οποίους θα παρουσιάσουμε στη συνέχεια. Αξίζει να σημειωθεί ότι, ο κυρίαρχος του παιχνιδιού τα τελευταία χρόνια σε ότι αφορά γενικότερα εφαρμογές NLP είναι αλγόριθμοι που βασίζονται στα ΤΝΔ και ιδιαίτερα στην ΒΜ. Στα πλαίσια της παρούσας εργασίας επιλέχθηκε να παρουσιαστούν παραδοσιακοί αλγόριθμοι ΜΜ, εξαιρώντας αυτούς της ΒΜ.

Ορισμένες πρώιμες μέθοδοι ΜΜ, όπως η λογιστική παλινδρόμηση και ο απλοϊκός Bayes θεωρούνται πιο παραδοσιακές, αλλά εξακολουθούν να χρησιμοποιούνται συνήθως στην επιστημονική κοινότητα, ειδικότερα για μικρά σύνολα δεδομένων. Νεότερες, επίσης παραδοσιακές μέθοδοι, είναι οι μηχανές υποστήριξης διανυσμάτων και τα τυχαία δάση. Πριν την παρουσίαση των αλγόριθμων, παρουσιάζεται συνοπτικά το θέμα της δυαδικής ταξινόμησης.

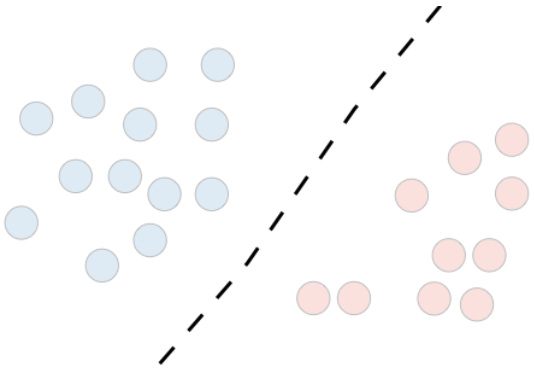
2.4.1 Η Δυαδική Ταξινόμηση

Γενικά, σε μια εργασία δυαδικής ταξινόμησης δίνεται ένα σύνολο δεδομένων, έστω *D,* με *K* παραδείγματα με ετικέτα $\{(\boldsymbol{x}_k, y_k)\}_{k=1}^{K}$ όπου κάθε στοιχείο $\boldsymbol{x}_k$ είναι ένα διάνυσμα *n* χαρακτηριστικών του οποίου κάθε διάσταση *i= 1,2, … n* περιέχει μια τιμή που περιγράφει το παράδειγμα και το y_k είναι η ετικέτα που παίρνει τιμές από το σύνολο {0,1}, που αντιστοιχούν στις τάξεις για τις οποίες είναι επιθυμητό από τον αλγόριθμο να μάθει πως θα τις διαχωρίζει. Συνεπώς, με την εφαρμογή ενός αλγορίθμου ΜΜ, είναι επιθυμητή η δημιουργία ενός μοντέλου που θα μάθει να προβλέπει το y από το διάνυσμα $\boldsymbol{x}$. Μία μέθοδος για την επίλυση του προβλήματος είναι η προσέγγιση με μια συνάρτηση $y = f(\boldsymbol{x})$

με μια άγνωστη συνάρτηση f και ο στόχος είναι η εκμάθηση της συνάρτησης για το δοσμένο σύνολο και στι συνέχεια να γίνουν προβλέψεις χρησιμοποιώντας την $\hat{y} = \hat{f}(\boldsymbol{x})$ ως εκτίμηση [20].

Μια κατηγορία μοντέλων ΜΜ είναι τα διακριτά (discriminative) μοντέλα, όπου από το μοντέλο εκτιμάται η πιθανότητα της περίπτωσης στιγμιότυπου να ανήκει στην κλάση 0 ή 1 δοθέντος του *x*, δηλαδή η πιθανότητα $p(y|x)$. Εάν η εκτίμηση της πιθανότητας είναι μεγαλύτερη του 50%, τότε το μοντέλο προβλέπει ότι το στιγμιότυπο ανήκει στην κλάση 1 (ή αλλιώς στη θετική κλάση), αλλιώς προβλέπει ότι ανήκει στην κλάση 0 (αρνητική κλάση), όπως φαίνεται στο Σχήμα 6 [28].

Η κατανομή της υπό συνθήκη πιθανότητας σε πιθανές ετικέτες για το διάνυσμα εισόδου $\boldsymbol{x}$ και το σύνολο εκπαίδευσης D για τη δυαδική ταξινόμηση δίνεται από τη σχέση $p(y = 1 \mid x, D)$ καθώς $p(\, y = 1 \mid x, D\,) + p(\, y = 0 \mid x, D\,) = 0$.



Σχήμα 6. Ο διαχωρισμός Κλάσεων με Διακριτά Μοντέλα [25]

Δοθείσας μιας πιθανοτικής εξόδου, υπολογίζεται η αρχική πρόβλεψη σε σχέση με την πραγματική τιμή χρησιμοποιώντας τη σχέση

$$\hat{y} = \hat{f}(\boldsymbol{x}) = \arg\max_{c} \; p(y = c \mid x, D)$$

που είναι η μέγιστη a posteriori[12] (Maximum A Posteriori - MAP) εκτίμηση.

Τα διακριτά μοντέλα καλούνται να μάθουν ένα όριο απόφασης (decision boundary) που θα διαχωρίζει τις τάξεις των δεδομένων. Οι δύο τάξεις των δεδομένων διαχωρίζονται ιδανικά με γραμμικό (linear) όριο απόφασης, ένα υπερεπίπεδο, που η εξίσωσή του δίνεται από τη σχέση:

$$w_0 x_0 + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n = 0$$

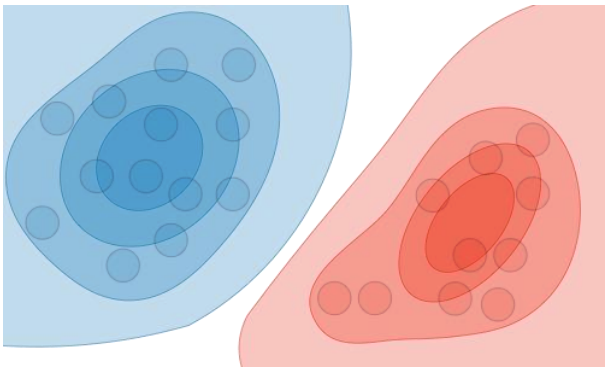
[12] Με τον όρο a prosteriori αναφέρεται η γνώση που προέρχεται από την εμπειρία, ενώ με τον όρο a priori χαρακτηρίζεται η έμφυτη γνώση- η εκ των προτέρων γνώση

η οποία σε μορφή πινάκων γράφεται ως $\boldsymbol{w^T x} = 0$.

Ο όρος w_0 αντιπροσωπεύει την μετατόπιση ως προς την πηγή (0,0) του καρτεσιανού συστήματος συντεταγμένων και αναφέρεται συνήθως ως πόλωση (bias) και κατά σύμβαση τίθεται το $x_0 = 1$. Το στοιχείο w_i είναι ο συντελεστής που πολλαπλασιάζεται με το χαρακτηριστικό x_i. Οι συντελεστές αυτοί ονομάζονται επίσης βάρη (weights) γιατί ορίζουν πως το κάθε χαρακτηριστικό επηρεάζει την πρόβλεψη για το y και είναι οι παράμετροι που πρέπει να μάθει το σύστημα. Τέτοιου είδους μοντέλα δημιουργούνται με την εφαρμογή του αλγόριθμου της λογιστικής παλινδρόμησης.

Στην πράξη όμως, κατά κανόνα, το όριο απόφασης είναι μη γραμμικό. Το γραμμικό μοντέλο μπορεί να μετασχηματιστεί σε μη γραμμικό, με τη βοήθεια των μαθηματικών και συγκεκριμένα των μη γραμμικών συναρτήσεων που θα εφαρμόζονται σε κάθε είσοδο $\boldsymbol{x}$. Τέτοιου είδους μοντέλα μπορούν να δημιουργηθούν με τη βοήθεια του αλγόριθμου της μηχανής υποστήριξης διανυσμάτων.

Μία άλλη κατηγορία μοντέλων είναι τα παραγωγικά (generative) μοντέλα. Ένα παραγωγικό μοντέλο πρώτα προσπαθεί να μάθει πως παράγονται τα δεδομένα με εκτίμηση της πιθανότητας $p(x|y)$, η ο οποία χρησιμοποιείται για την εκτίμηση της πιθανότητας $p(y|x)$ με τη χρήση του κανόνα του Bayes και στη συνέχεια γίνεται ο διαχωρισμός των κλάσεων, όπως φαίνεται στο Σχήμα 7. Τέτοιου είδους μοντέλα μπορούν να δημιουργηθούν με τη βοήθεια του αλγόριθμου του απλοϊκού Bayes.



Σχήμα 7. Ο διαχωρισμός Κλάσεων με Παραγωγικά Μοντέλα [25]

Άλλες κατηγορίες μοντέλων είναι τα δέντρα απόφασης και τα συλλογικά (ensemble) μοντέλα. Στα συλλογικά μοντέλα ανήκουν τα τυχαία δάση.

Ανεξάρτητα από την κατηγορία του μοντέλου που θα επιλεγεί, εφαρμόζονται κάποιες βασικές έννοιες για την εκτίμηση του λάθους στην πρόβλεψη.

Η συνάρτηση λάθους (error function) L: $(\hat{y}, y) \in \mathbb{R} \times Y \mapsto L(\hat{y}, y) \in \mathbb{R}$ παίρνει ως είσοδο την τιμή πρόβλεψης $\hat{y}$ και την αντίστοιχη πραγματική τιμή y και εξάγει την διαφορά τους. Ανάλογα με τον αλγόριθμο, ορίζεται και η αντίστοιχη συνάρτηση κόστους.

Η συνάρτηση κόστους (cost function) J χρησιμοποιείται για να αξιολογηθεί η απόδοση ενός μοντέλου και ορίζεται σε σχέση με την συνάρτηση λάθους ως

$$J(w) = \sum_{i=1}^{k} L((\hat{y}(i)), y(i))$$

Ο σκοπός του συστήματος ΜΜ είναι να ελαχιστοποιηθεί η συνάρτηση κόστους μετά από διαδοχικές δοκιμές και να αποτιμηθεί το σύστημα με βάση το μέτρο ή τα μέτρα απόδοσης που προσδιορίζονται. Τα μέτρα απόδοσης για την αποτίμηση των μοντέλων που δημιουργούνται από τους αλγόριθμους ταξινόμησης, παρουσιάζονται στην επόμενη ενότητα.

2.4.2 Λογιστική Παλινδρόμηση

Η λογιστική παλινδρόμηση (Logistic Regression - LG) είναι ένας από τους παλαιότερους αλγόριθμους ΜΜ και παρουσιάστηκε το 1958 από τον Cox [29]. Με τον αλγόριθμο της λογιστικής παλινδρόμησης δημιουργούνται γραμμικά μοντέλα.

Στο μοντέλο που δημιουργείται οι πιθανότητες που περιγράφουν τις πιθανές εξόδους μοντελοποιούνται χρησιμοποιώντας τη λογιστική συνάρτηση, που είναι γνωστή και ως σιγμοειδής συνάρτηση (sigmoid function). Η σιγμοειδής συνάρτηση, δέχεται ως παράμετρο μια πραγματική τιμή z και εξάγει ως αποτέλεσμα μία πραγματική τιμή που ανήκει στο διάστημα [0,1]. Η σιγμοειδής συνάρτηση δίνεται από τον τύπο:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Στο μοντέλο της λογιστικής παλινδρόμησης, υπολογίζεται το σταθμισμένο άθροισμα
$z = w_0 x_0 + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n$ των χαρακτηριστικών της εισόδου, και στο z εφαρμόζεται η σιγμοειδής συνάρτηση, η οποία επιστρέφει τιμές στο εύρος [0,1], οπότε υπολογίζεται η πιθανότητα $p = \sigma(z)$, όπου z το σταθμισμένο άθροισμα κάθε εισόδου [20]. Για ένα στιγμιότυπο του συνόλου εκπαίδευσης x η πρόβλεψη της εξόδου $\hat{y}$ είναι:

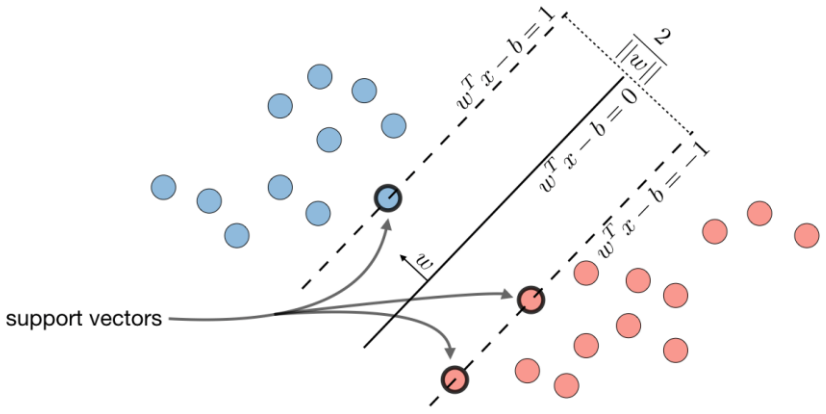
$$\hat{y} = \begin{cases} 0 \text{ εάν } \sigma(z) < 0.5 \\ 1 \text{ εάν } \sigma(z) \geq 0.5 \end{cases}$$

2.4.3 Μηχανές Διανυσμάτων Υποστήριξης

Οι Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines- SVM) είναι από τους αλγόριθμους που χρησιμοποιούνται ευρέως σε προβλήματα κατηγοριοποίησης. Αρχικά προτάθηκε το 1963 από τον Vapnik και η θεωρία του επεκτάθηκε ο 1995 από τον ίδιο και τον Cortes [30].

Θεωρώντας ένα πρόβλημα δυαδικής ταξινόμησης, η βασική ιδέα των SVM είναι να κατασκευαστεί ένα υπερεπίπεδο (hyperplane) το οποίο θα λειτουργεί ως σύνορο μεταξύ των δύο κλάσεων που διαχωρίζει τα αρνητικά με τα θετικά παραδείγματα, δηλαδή να βρεθεί μια συνάρτηση απόφασης διαχωρισμού των δύο κλάσεων ταξινόμησης. Διαισθητικά, ο διαχωρισμός επιτυγχάνεται από το υπερεπίπεδο που έχει τη μεγαλύτερη απόσταση από τα πλησιέστερα σημεία δεδομένων εκπαίδευσης οποιασδήποτε κατηγορίας και το οποίο ονομάζεται λειτουργικό περιθώριο, καθώς γενικά όσο μεγαλύτερο είναι το περιθώριο τόσο μικρότερο είναι το σφάλμα γενίκευσης του ταξινομητή. Τα σημεία που βρίσκονται πάνω στο όριο του περιθωρίου ονομάζονται *διανύσματα υποστήριξης (support vectors)*, όπως φαίνεται στο Σχήμα 8.

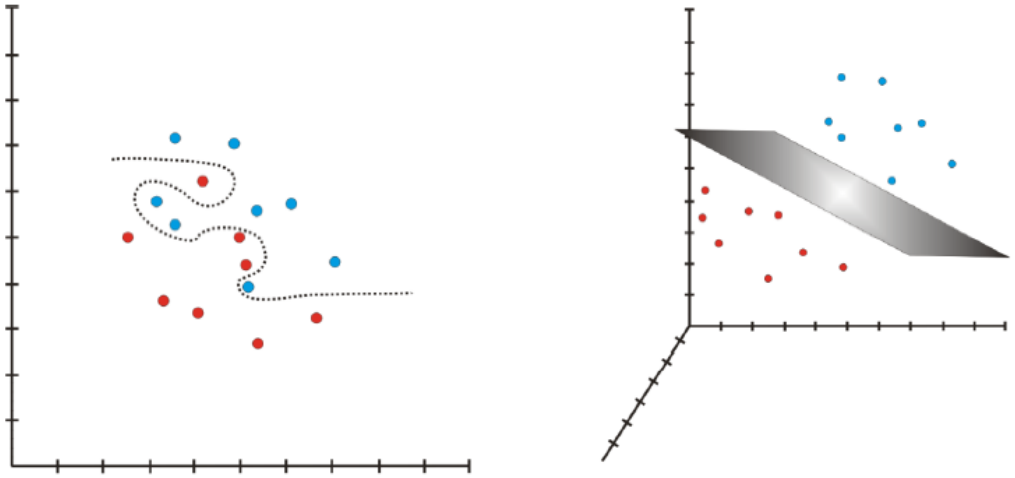
Δοθέντων των διανυσμάτων υποστήριξης για τις δύο κλάσεις, με μαθηματικούς υπολογισμούς δημιουργείται το υπερεπίπεδο του μέγιστου περιθωρίου. Με βάση τα διανύσματα υποστήριξης γίνονται οι κατάλληλοι υπολογισμοί, έτσι ώστε να ταξινομηθούν τα υπόλοιπα στοιχεία του συνόλου δεδομένων σε μία από τις δύο κλάσεις .



Σχήμα 8. Μηχανές Διανυσμάτων Υποστήριξης [25]

Εάν τα δεδομένα δεν διαχωρίζονται γραμμικά, μέσω κατάλληλων συναρτήσεων, που ονομάζονται συναρτήσεις πυρήνα (kernel functions), οι SVM μπορούν να

μετασχηματίσουν τον αρχικό χώρο, έτσι ώστε μη-γραμμικά διαχωρίσιμα προβλήματα να μετατραπούν σε γραμμικά διαχωρίσιμα και τελικά να λυθούν με την ίδια μεθοδολογία. Η συνάρτηση του πυρήνα είναι μία συνάρτηση που χαρτογραφεί τα διανύσματα των χαρακτηριστικών σε μια μεγαλύτερη διάσταση, προκειμένου να γίνουν γραμμικά διαχωρίσιμα, όπως φαίνεται στο Σχήμα 9.



Σχήμα 9. Γραμμικός διαχωρισμός κλάσεων με Εφαρμογή Συνάρτησης Πυρήνα [31]

2.4.4 Απλοϊκός Bayes

Ο απλοϊκός Bayes (Naive Bayes - NB) εμφανίζεται στη δεκαετία του 1990 και βασίζεται στην εφαρμογή του θεωρήματος του Bayes (το οποίο διατυπώθηκε από τον 18ο αιώνα) με την «αφελή» υπόθεση της υπό όρους ανεξαρτησίας μεταξύ του διανύσματος των χαρακτηριστικών $\boldsymbol{x}$ και της αντίστοιχης μεταβλητής κλάσης y. Αξιόπιστες πηγές για την περιγραφή του αλγορίθμου είναι οι [32], [33].

Το θεώρημα του Bayes δηλώνει την ακόλουθη σχέση πιθανότητας, με δεδομένη της μεταβλητής κλάσης ταξινόμησης y και το διάνυσμα με τις τιμές των n χαρακτηριστικών x_1 έως x_n:

$$P(y \mid x_1, \dots, x_n) = \frac{P(y)P(x_1, \dots, x_n \mid y)}{P(x_1, \dots, x_n)}$$

Χρησιμοποιώντας την απλοϊκή υπόθεση της υπό όρους ανεξαρτησίας η οποία είναι:

$$P(x_i|y, x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) = P(x_i|y),$$

για όλα τα i η παραπάνω σχέση απλοποιείται σε:

$$P(y \mid x_1, \dots, x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i \mid y)}{P(x_1, \dots, x_n)}$$

Καθώς η $P(x_1, \dots, x_n)$ είναι σταθερή δοθείσης της εισόδου $\boldsymbol{x}$, εφαρμόζεται ο ακόλουθος κανόνας ταξινόμησης για την πρόβλεψη $\hat{y}$:

$$\hat{y} = \arg\ \max_{y} P(y) \prod_{i=1}^{n} P(x_i \mid y)$$

Στη συνέχεια, χρησιμοποιείται η μέγιστη εκ των υστέρων (Maximum A Posteriori-MAP) εκτίμηση για την εκτίμηση της πιθανότητας $P(y)$ που είναι η σχετική συχνότητα της κλάσης y στο σύνολο εκπαίδευσης και την εκτίμηση της $P(x_i \mid y)$.

Ο NB επιτυγχάνει στην ταξινόμηση, εφόσον πραγματικά ισχύει η υπόθεση ότι, με δεδομένη την κλάση, οι μεταβλητές εισόδου είναι ανεξάρτητες. Οι διαφορετικοί NB ταξινομητές διαφέρουν κυρίως με βάση την υπόθεση σε σχέση με την κατανομή της πιθανότητας $P(x_i \mid y)$, δηλαδή εάν είναι Gaussian, Multinomial ή Bernulli [33]. Σε προβλήματα ταξινόμησης κειμένου χρησιμοποιούνται συνήθως ο Multinomial NB, ο οποίος απαιτεί ως είσοδο ακέραια δεδομένα και ο Bernulli NB ο οποίος απαιτεί ως είσοδο δυαδικά δεδεμένα.

2.4.5 Τυχαίο Δάσος

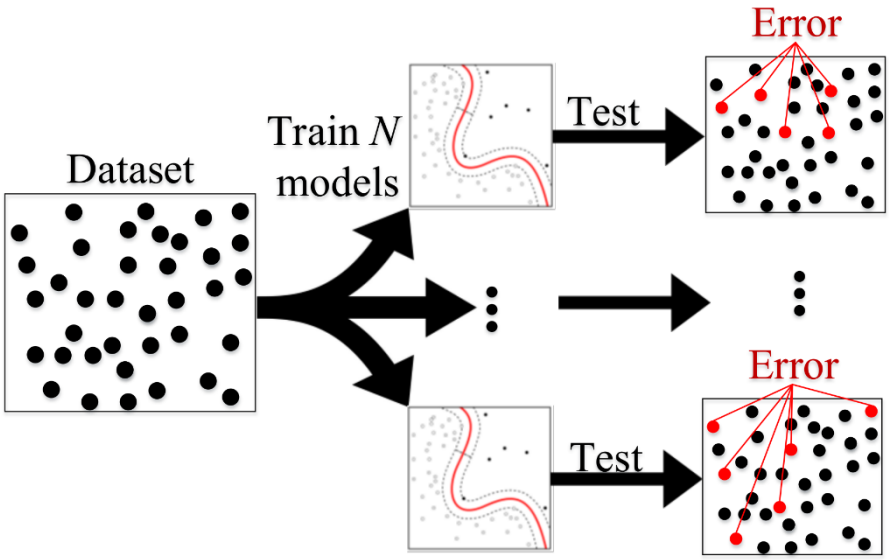
Το τυχαία δάσος (Random Forest – RF) είναι μία μέθοδος συλλογικής (ensemble) μάθησης που βασίζεται στον αλγόριθμο του δέντρου αποφάσεων (Decision Tree - DT) και στην τεχνική του bagging. Ο αλγόριθμος προτάθηκε αρχικά από τον Ho [24] το 1996 και στη συνέχεια επεκτάθηκε και βελτιώθηκε από τον Breiman [34].

Ο αλγόριθμος DT είναι ένας από τους παλαιότερους αλγόριθμους ταξινόμησης. Αρχικά αναπτύχθηκε από τον Quinlan [35], βελτιώθηκε από τον ίδιο το 1993 [36] και αποτελεί μία μη παραμετρική μέθοδο επιβλεπόμενης ΜΜ. Τα δέντρα απόφασης ονομάζονται έτσι επειδή, παρόμοια με ένα δέντρο, ο αλγόριθμος ξεκινά από έναν κόμβο-ρίζα που αναπτύσσεται σε κλαδιά και έτσι δημιουργείται μια δομή που μοιάζει με δέντρο, όπως φαίνεται στο Σχήμα . Ο στόχος του αλγόριθμου είναι η δημιουργία ενός μοντέλου που προβλέπει την τιμή της μεταβλητής στόχου μαθαίνοντας απλούς κανόνες απόφασης που

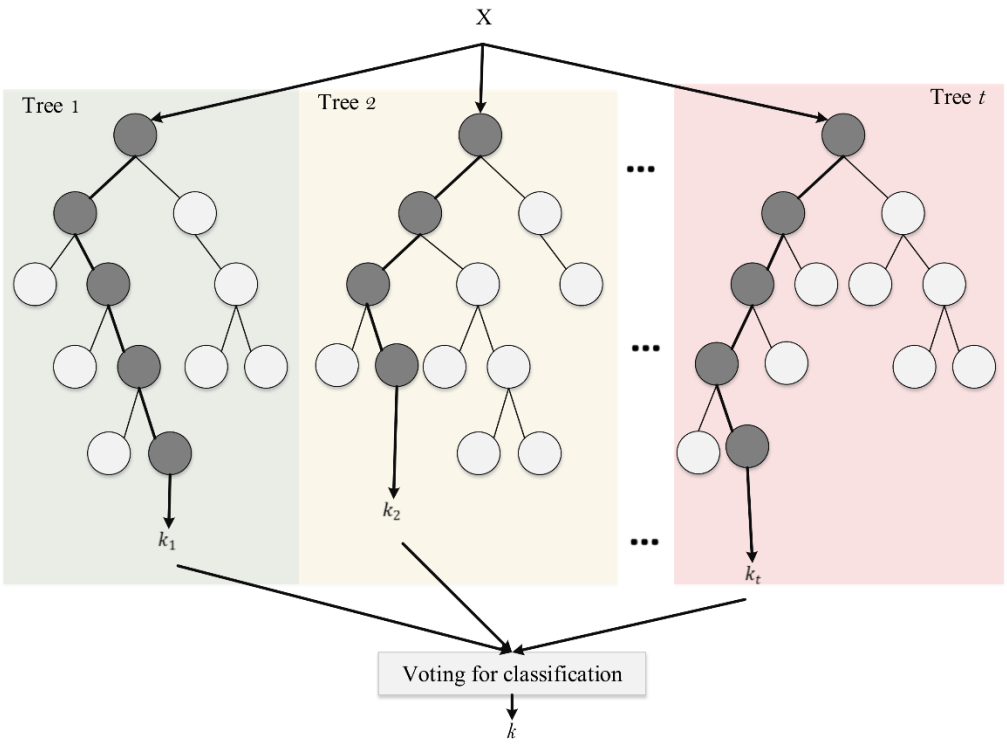
συμπεραίνονται από τα χαρακτηριστικά των δεδομένων. Με δοσμένα τα διανύσματα εκπαίδευσης $x_i \in R^n$, $i=1,\ldots, l$ και το διάνυσμα ετικέτας $y \in R^l$, το δέντρο απόφασης διαχωρίζει αναδρομικά τον χώρο των χαρακτηριστικών έτσι ώστε δείγματα με την ίδια ετικέτα να ομαδοποιούνται μαζί. Υπάρχουν πολλοί διαφορετικοί αλγόριθμοι δέντρων απόφασης, όπως για παράδειγμα ο ID3, ο CART κλπ. [37].

Ο όρος bagging χρησιμοποιείται ως συντόμευση για την αθροιστική εκκίνηση (bootstrap aggregating). Η τεχνική του bagging εφαρμόζεται για την βελτίωση της σταθερότητας και της ακρίβειας αλγόριθμων μηχανικής μάθησης.

Ο αλγόριθμος του bagging προτάθηκε από τον Breiman [38] το 1996 ως μέθοδος ταξινόμησης με ψηφοφορία (voting classifier). Ο αλγόριθμος δημιουργείται από διαφορετικά δείγματα bootstrap [39]. Δοθέντος ενός συνόλου δεδομένων εκπαίδευσης D με K παραδείγματα, το bagging δημιουργεί N νέα σύνολα εκπαίδευσης D_i με δειγματοληψία από το D με ομοιόμορφη κατανομή και με αντικατάσταση. Κατά τη δειγματοληψία κάποια παραδείγματα μπορεί να επαναλαμβάνονται σε κάθε D_i . Κάθε ένα από αυτά τα δείγματα είναι ένα δείγμα εκκίνησης (bootstrap sample). Για κάθε ένα από αυτά τα δείγματα εφαρμόζεται ένας ταξινομητής C_i για την εκπαίδευσή τους. Τελικά, ο ταξινομητής C περιέχει ή δημιουργείται από τους $C_1, C_2, \ldots, C_N$ των οποίων η έξοδος είναι η κλάση που προβλέπεται συχνότερα από τους υπο-κατηγοριοποιητές. Η τεχνική του bagging φαίνεται στο Σχήμα 10.



Σχήμα 10. Η Τεχνική του Bagging [22]



Σχήμα 11. Ο Αλγόριθμος του Τυχαίου Δάσους [22]

Στην περίπτωση των τυχαίων δασών, ο ταξινομητής που εφαρμόζεται σε όλα τα δείγματα είναι τα δέντρα απόφασης και τα οποία δημιουργούν ένα «δάσος». Μετά την εκπαίδευση, λαμβάνεται η απόφαση από κάθε δέντρο και με τη μέθοδο της ψηφοφορίας κατά πλειοψηφία, εφόσον πρόκειται για εργασία ταξινόμησης, προβλέπεται η τελική απόδοση και γίνεται η ταξινόμηση. Στο Σχήμα 11 απεικονίζεται ο τρόπος που λειτουργεί ο αλγόριθμος του τυχαίου δάσους για την πρόβλεψη της κλάσης.

2.5 Μετρικές Απόδοσης Ταξινόμησης

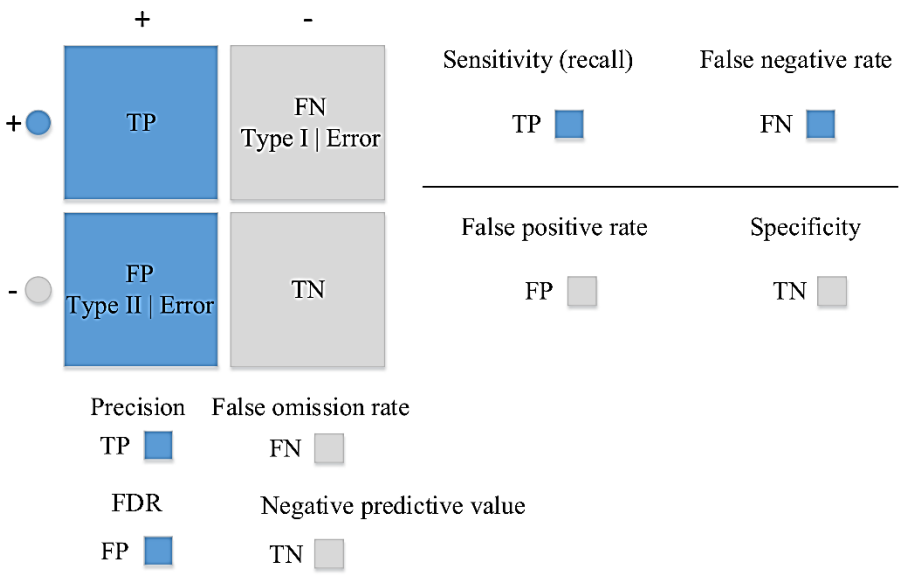
Η αποτίμηση ενός μοντέλου ταξινόμησης γίνεται με ορισμένα μέτρα απόδοσης. Γενικά, σε μια εργασία ταξινόμησης, χρησιμοποιούνται οι παρακάτω συμβάσεις για την ονοματολογία και τους συμβολισμούς [22], [40]:

- Οι δύο κλάσεις χαρακτηρίζονται ως *θετική (Positive – P)* και *αρνητική (Negative – N)*.
- Ο αριθμός των προβλέψεων που ανήκουν στη *θετική* κλάση, ονομάζονται *αληθώς θετικά (True Positives- TP)*, δηλαδή είναι ο αριθμός των σωστών προβλέψεων για ασθένεια.

- Ο αριθμός των προβλέψεων που ανήκουν στην αρνητική κλάση, ονομάζονται *αληθώς αρνητικά (True Negatives- TN)*, δηλαδή είναι ο αριθμός των σωστών προβλέψεων για μη ασθένεια.
- Ο αριθμός προβλέψεων που ταξινομήθηκαν στη θετική κλάση, ενώ στην πραγματικότητα ανήκουν στην αρνητική, ονομάζονται *ψευδώς αρνητικά (False Positive- FP)* ή αλλιώς λάθος τύπου II (Type II Error)
- Ο αριθμός προβλέψεων που ταξινομήθηκαν στην αρνητική κλάση ενώ στην πραγματικότητα ανήκουν στη θετική, ονομάζονται *ψευδώς αρνητικά (False Negatives - FN)* ή αλλιώς ή αλλιώς λάθος τύπου I (Type I Error)

Οι προβλέψεις μιας δυαδικής ταξινόμησης απεικονίζονται σε ένα μητρώο 2x2 που ονομάζεται *πίνακας σύγχυσης (confusion matrix)* και συνήθως αναφέρεται ως πίνακας ταξινόμησης, όπως φαίνεται στο Σχήμα 12. Ο πίνακας ταξινόμησης παρέχει τη δυνατότητα οπτικοποίησης της απόδοσης ενός μοντέλου [22], [40].

Κάθε γραμμή παριστάνει τα παραδείγματα της κλάσης που έχει προβλέψει ο ταξινομητής (κλάση πρόγνωσης), ενώ κάθε στήλη παριστάνει τα παραδείγματα της κλάσης που ανήκουν πραγματικά τα παραδείγματα (πραγματική κλάση). Ο όρος σύγχυση προκύπτει από το γεγονός ότι, μπορούμε να δούμε εύκολα εάν το μοντέλο συγχέει τις δύο κλάσεις, δηλαδή εάν ταξινομεί λάθος σε μία κλάση από αυτή που αντιστοιχεί πραγματικά σε ένα παράδειγμα. Από τον πίνακα ταξινόμησης εξάγονται με κατάλληλους υπολογισμούς διάφορα μέτρα απόδοσης για ένα μοντέλο, όπως περιγράφονται στην αμέσως επόμενη παράγραφο.



Σχήμα 12. Ο Πίνακας Ταξινόμησης

2.5.1 Μέτρα απόδοσης

Γενικά, τα μέτρα απόδοσης αξιολογούν συγκεκριμένες πτυχές της απόδοσης των εργασιών ταξινόμησης. Συνεπώς δεν παρουσιάζουν πάντα ταυτόσημες πληροφορίες. Δεδομένου ότι, οι υποκείμενες μηχανικές των διαφορετικών μετρήσεων αξιολόγησης μπορεί να διαφέρουν, η κατανόηση του τι ακριβώς αντιπροσωπεύει κάθε μία από αυτές τις μετρήσεις και τι είδους πληροφορίες προσπαθούν να μεταδώσουν είναι ζωτικής σημασίας για τη συγκρισιμότητα των μοντέλων που εφαρμόζονται για την εργασία της ταξινόμησης.

Η *ακρίβεια (accuracy)* ή αλλιώς *ορθότητα,* είναι το πιο συνηθισμένο μέτρο απόδοσης ενός μοντέλου. Η ορθότητα είναι η αναλογία των παραδειγμάτων για τα οποία έχει γίνει η σωστή πρόβλεψη επί του συνόλου των προβλέψεων και με βάση τις παραπάνω συμβάσεις και ισοδυναμεί με τον δείκτη λάθους (error rate) του μοντέλου [20] και δίνεται από τον τύπο:

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

Σε πολλές περιπτώσεις οι κλάσεις των δεδομένων στα σύνολα δεδομένων δεν κατανεμημένες με ισορροπία. Για παράδειγμα, σε ένα σύνολο δεδομένων το 95% των παραδειγμάτων μπορεί να ανήκουν στην αρνητική κλάση και μόνο το 5 % στη θετική κλάση. Σε τέτοια περίπτωση, μπορεί να έχουμε ένα μοντέλο με υψηλή ορθότητα, αλλά το μοντέλο δεν έχει προγνωστική δύναμη. Συνεπώς, η ορθότητα ως μετρική απόδοσης δεν αρκεί. Για τον σκοπό αυτό, στην ταξινόμηση χρησιμοποιούνται και άλλες μετρικές απόδοσης [24], [41], [22], [40].

Η *ακρίβεια*[13] (*precision*) ή αλλιώς *θετική προγνωστική αξία*, είναι η αναλογία του κάθε παραδείγματος που προβλέπεται ως θετικό και είναι πραγματικά θετικό. Η ακρίβεια υπολογίζεται από τον τύπο:

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

Η *ανάκληση* (*recall*), γνωστή και ως *ευαισθησία (sensitivity)* ή δείκτης αληθώς θετικών (True Positive Rate – TPR), είναι η αναλογία κάθε θετικού παραδείγματος που

[13] Για να μην υπάρξει σύγχυση μεταξύ των εννοιών accuracy και precision οι όροι θα αναφέρονται ως ορθότητα και ακρίβεια αντίστοιχα.

είναι πραγματικά θετικό. Αφορά την ικανότητα του μοντέλου να αναγνωρίζει ένα παράδειγμα της θετικής κλάσης και υπολογίζεται από τον τύπο:

$$Recall = TPR = \frac{TP}{TP + FN}$$

Σε πολλές περιπτώσεις, εάν θέλουμε να συνοψίσουμε την απόδοση του μοντέλου, επιδιώκοντας ένα είδος ισορροπίας μεταξύ της ακρίβειας και την ανάκλησης, χρησιμοποιείται ένα μέτρο απόδοσης που συνδυάζει τα δύο αυτά μέτρα, τον αρμονικό μέσο όρο της ακρίβειας και της ανάκλησης που ονομάζεται *F1 αποτέλεσμα* (*F1* score) και υπολογίζεται από τον τύπο:

$$F1 = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{TP}{TP + \frac{FN + FP}{2}}$$

Η *προσδιοριστικότητα (specificity)*, γνωστή και ως *δείκτης αληθώς αρνητικών (True Negative Rate – TNR)* μετρά τις ορθά αρνητικές προβλέψεις στο σύνολο των ορθών αρνητικών παραδειγμάτων και υπολογίζεται από τον τύπο:

$$Specificity = TNR = \frac{TN}{TN + FP}$$

Το *ψευδώς θετικό ποσοστό (False Positive Rate- FPR)* αντιστοιχεί στην αναλογία των αρνητικών παραδειγμάτων που θεωρήθηκαν ως θετικά, σε σχέση με όλα τα αρνητικά παραδείγματα. Όσο μεγαλύτερος είναι ο *FPR*, πόσα περισσότερα αρνητικά παραδείγματα έχουν ταξινομηθεί λάθος. Ο *FPR* υπολογίζεται από τον τύπο:

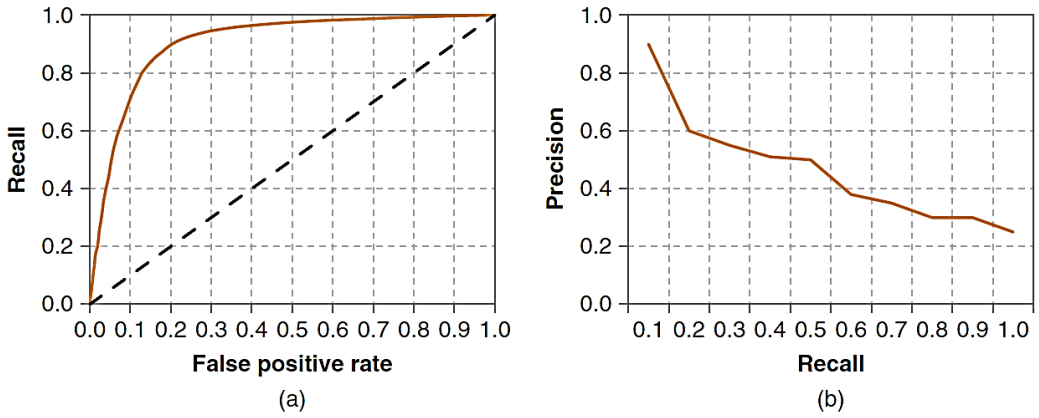
$$FPR = \frac{FP}{FP + TN} = 1 - specificity = 1 - TNR$$

2.5.2 Ποιοτικοί Δείκτες

Ο πίνακας ταξινόμησης δίνει μόνο μια μερική εικόνα της συνολικής απόδοσης ενός μοντέλου. Αυτό γιατί, σε ταξινομητές που βασίζονται στον υπολογισμό πιθανοτήτων, οι αλγόριθμοι βασίζονται σε ένα κατώφλι για να μετατρέψουν την πιθανότητα σε μία από τις κλάσεις 0 ή 1. Το κατώφλι αυτό είναι συνήθως 0,5. Έτσι, για παράδειγμα αν υπολογιστεί η πιθανότητα για ένα παράδειγμα 0,45 το παράδειγμα θα ταξινομηθεί στην αρνητική κλάση,

ενώ εάν υπολογιστεί 0,55 το παράδειγμα θα ταξινομηθεί στην θετική κλάση. Αυτή η περικοπή πολλές φορές μπορεί να οδηγήσει σε απώλεια πληροφορίας. Επίσης, μεταβάλλοντας το κατώφλι, μεταβάλλονται και τα αποτελέσματα του πίνακα ταξινόμησης, οπότε διαφοροποιούνται όλες οι μετρικές που αναφέρθηκαν στην προηγούμενη παράγραφο [5]. Για να αντιμετωπιστεί αυτό το μειονέκτημα, έχουν προταθεί δείκτες οι οποίοι χαρακτηρίζονται στην βιβλιογραφία και ως ποιοτικοί δείκτες [24].

Ένας εύκολος σχετικά γραφικός τρόπος και ο πιο συνηθισμένος για την ποιοτική εκτίμηση της απόδοσης ενός ταξινομητή είναι η *λήψη χαρακτηριστικής καμπύλης λειτουργίας (Receiving Operating Characteristic curve – ROC curve)*. Η ROC καμπύλη είναι η σχεδίαση της μετρικής *TPR*, σε σχέση με την μετρική *FPR* για διάφορα κατώφλια. Η ROC καμπύλη χρησιμοποιείται ως μια γενική μετρική του μοντέλου. Όσο καλύτερο είναι ένα μοντέλο, τόσο υψηλότερα είναι η καμπύλη και συνεπώς τόσο μεγαλύτερη η επιφάνεια κάτω από την καμπύλη. Η επιφάνεια κάτω από την ROC καμπύλη (Area Under the Curve – AUC) ονομάζεται ROC AUC ή απλά AUC. Ο τέλειος ταξινομητής έχει ROC AUC ίση με 1, ενώ ένας κακός ταξινομητής θα έχει 0.5, ίσως και λιγότερο [24]. Αυτό γιατί, όταν η τιμή είναι ίση με 1, αυτό σημαίνει ότι ο ταξινομητής μπορεί να διαχωρίσει πλήρως τις θετικές από τις αρνητικές κλάσεις. Εάν έχει την τιμή 0,5, τότε ο ταξινομητής δεν είναι σε θέση να διακρίνει μεταξύ θετικών και αρνητικών σημείων κλάσης και τότε ο ταξινομητής προβλέπει είτε τυχαία κλάση, είτε σταθερή κλάση για όλα τα σημεία δεδομένων. Συνεπώς, όσο πιο κοντά η ROC AUC στο 1, τόσο καλύτερος είναι ο ταξινομητής.



Σχήμα 13. Παράδειγμα ROC Καμπύλης (a) και PR Καμπύλης (b) [5]

Ένας άλλος γραφικός τρόπος για την ποιοτική εκτίμηση είναι η καμπύλη ακρίβειας-ανάκλησης (Precision Recall – PR). Η καμπύλη PR βοηθά στην αντιστάθμιση της ακρίβειας

του ταξινομητή σε σχέση με την ανάκληση της θετικής κλάσης, καθώς μεταβάλλεται το κατώφλι [5], [24].

Περισσότερες πληροφορίες για τις καμπύλες PR και ROC δίνονται στην [42]. Παραδείγματα των καμπυλών ROC και PR φαίνονται στο Σχήμα 13.

Επίσης, ως μέτρο της ποιότητας ενός μοντέλου ταξινόμησης, θεωρείται ο συντελεστής συσχέτισης του Matthews (Matthews Correlation Coefficient - MCC). Λαμβάνει υπόψη τα αληθώς και ψευδώς θετικά και αρνητικά και θεωρείται γενικά ως ένα ισορροπημένο μέτρο απόδοσης, το οποίο μπορεί να χρησιμοποιηθεί ακόμη και αν οι τάξεις έχουν πολύ διαφορετικά μεγέθη [43]. Ο MCC είναι μια τιμή συντελεστή συσχέτισης μεταξύ -1 και +1. Ο συντελεστής +1 αντιπροσωπεύει μια τέλεια πρόβλεψη, ο συντελεστής 0 μια μέση τυχαία πρόβλεψη και ο συντελεστής -1 μια αντίστροφη πρόβλεψη . Ο MCC υπολογίζεται από τον τύπο:

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

3 Βιβλιογραφική Επισκόπηση

Το παρόν κεφάλαιο αφορά πρόσφατες σχετικές εργασίες με την παρούσα εργασία. Αρχικά αναφέρονται εργασίες που αφορούν προβλέψεις με τεχνικές ανάλυσης συναισθήματος για διάφορους δείκτες χρηματιστηρίων και προβλέψεις μετοχών και στη συνέχεια αναφέρονται εργασίες που αφορούν ειδικά τον δείκτη Dow Jones.

Ωστόσο, αξίζει να αναφερθεί ότι, στην πρόσφατη βιβλιογραφία υπάρχουν πολλές αξιόλογες έρευνες, όπου συγκεντρώνονται άρθρα σχετικά με την ανάλυση συναισθήματος για την πρόβλεψη του χρηματιστηρίου. Ενδεικτικά αναφέρουμε τις πρόσφατες έρευνες [1]–[4], [8], [44], [45].

3.1 Σχετικές Εργασίες για πρόβλεψη Δεικτών και Μετοχών

Οι Bouktif et al. στην εργασία τους με τίτλο “Augmented Textual Features-Based Stock Market Prediction” [18] συλλέγουν και επεξεργάζονται ιστορικά χρηματιστηριακά δεδομένα μετοχών δέκα εταιρειών που δραστηριοποιούνται σε διαφορετικούς τομείς και θεωρείται ότι επηρεάζουν τον χρηματηστηριακό δείκτη NASDAG για χρονικό διάστημα δέκα ετών (2010-2018). Για το ίδιο χρονικό διάστημα συλλέγουν tweets σχετικά με τις εταιρείες που επέλεξαν και πραγματοποιούν σε αυτά ανάλυση συναισθήματος με τεχνικές επιβλεπόμενης και μη επιβλεπόμενης μάθησης. Συνενώνουν τα σύνολα των ιστορικών δεδομένα με τα αποτελέσματα της ανάλυσης συναισθήματος πραγματοποιώντας διάφορα σενάρια συνένωσης και στη συνέχεια, με βάση το εκάστοτε νέο σύνολο δεδομένων εφαρμόζουν έξι αλγόριθμους δυαδικής ταξινόμησης (NB, LR, SVM, ΤΝΔ, RF, XGBoost) για την πρόβλεψη της τάσης. Επίσης, δημιουργούν επιπλέον συνδυαστικά σενάρια με μείωση των χαρακτηριστικών των συνόλων δεδομένων. Πετυχαίνοντας ορθότητα της τάξης του 60% για την πρόβλεψη της τάσης των τιμών της κάθε μετοχής και συμπεραίνουν ότι, η άνοδος ή η κάθοδος της τιμής μιας μετοχής επηρεάζεται από τις δημόσιες γνώμες και τα συναισθήματα που εκφράζονται στο twitter. Κατά την άποψη των συγγραφέων, εκείνο που απαιτείται είναι η εύρεση καλύτερων μεθόδων για την ανάλυση συναισθήματος.

Ο Agarwal στην εργασία του με τίτλο “Sentiment Analysis of Financial News” [46] συλλέγει και επεξεργάζεται οικονομικές ειδήσεις από τον ιστότοπο Finwiz που αφορούν δύο εταιρείες, την Apple και την Tesla για ένα διάστημα μίας εβδομάδας του 2019. Η ανάλυση συναισθήματος πραγματοποιείται με το VADER και το αποτέλεσμα συγκρίνεται με τις τάσεις της εκάστοτε μετοχής, σύμφωνα με ιστορικά δεδομένα τιμών των μετοχών που συλλέχθηκαν από το Yahoo Finance. Η ιδιαιτερότητα αυτής της εργασίας έγκειται στο ότι, εμπλουτίζει το λεξιλόγιο του VADER. Ο συγγραφέας συμπεραίνει ότι, για συγκεκριμένες ημέρες που πραγματοποίησε την ανάλυση των συναισθημάτων από τα οικονομικά άρθρα που αφορούν τις δύο μετοχές, υπάρχει μια ισχυρή συσχέτιση του αποτελέσματος που εξάγεται από το συναίσθημα με τις τάσεις των τιμών των δύο μετοχών.

Οι . Bourezk et al. στην εργασία τους με τίτλο “Analyzing Moroccan Stock Market using Machine Learning and Sentiment Analysis” [47] συλλέγουν και επεξεργάζονται δεδομένα από ιστότοπους του Μαρόκου, όπου δημοσιεύονται οικονομικά νέα από τον Ιούνιο έως τον Οκτώβριο του 2019. Τα οικονομικά δεδομένα που συλλέγουν για το ίδιο χρονικό διάστημα αφορούν μετοχές εταιρειών του χρηματιστηρίου της Καζαμπλάνκα. Στα δεδομένα κειμένου πραγματοποιούν ανάλυση συναισθήματος με τεχνικές επιβλεπόμενης και μη επιβλεπόμενης μάθησης. Οι συγγραφείς καταλήγουν σε δύο συμπεράσματα: αφενός ότι μπορεί να εξαχθεί η δημόσια διάθεση από τα οικονομικά νέα και, αφετέρου ότι η αρνητική διάθεση έχει μεγάλη επιρροή στους δείκτες του χρηματιστηρίου.

Οι Kalra και Prasad στην εργασία τους με τίτλο “Efficacy of News Sentiment for Stock Market Prediction” [48] συλλέγουν και επεξεργάζονται δεδομένα από ιστότοπους της Ινδίας, όπου δημοσιεύονται οικονομικά νέα για διάστημα ενός μηνός. Τα οικονομικά δεδομένα που συλλέγουν για το ίδιο χρονικό διάστημα αφορούν μετοχές πέντε τραπεζών. Στα δεδομένα κειμένου που προκύπτουν από τις ειδήσεις πραγματοποιούν ανάλυση συναισθήματος με τεχνικές επιβλεπόμενης μάθησης. Δημιουργούν νέα σύνολα δεδομένων που συνδυάζουν το συναίσθημα με τα οικονομικά μεγέθη και πειραματίζονται με τέσσερις αλγόριθμους δυαδικής ταξινόμησης για να εξάγουν την τάση της τιμής της εκάστοτε μετοχής (KNN, ΤΝΔ, SVM, NB). Η ορθότητα των πειραμάτων τους κυμαίνεται σε διάφορα επίπεδα της τάξης του 65% έως 86%, λόγω του ότι για την πρόβλεψή τους χρησιμοποιούν ιστορικά δεδομένα ενός μήνα.

3.2 Σχετικές Εργασίες με τον Δείκτη Dow Jones

Οι Velay και Daniel στην εργασία τους με τίτλο "Using NLP on news headlines to predict index trends" [49] εφαρμόζουν διάφορους αλγόριθμους δυαδικής ταξινόμησης και χρησιμοποιούν το ίδιο σύνολο δεδομένων με της παρούσας εργασίας. Τα πειράματα των συγγραφέων επικεντρώνονται στους τρόπους της επεξεργασίας του κειμένου, προκειμένου να επιτευχθεί η μέγιστη ακρίβεια στην πρόβλεψη της δυαδικής ταξινόμησης. Για την ταξινόμηση οι συγγραφείς χρησιμοποιούν διάφορες μεθόδους ΜΜ και ΒΜ, καταλήγοντας στο ότι τα καλύτερα αποτελέσματα δίνει η LR με 57% ορθότητα. Οι συγγραφείς συμπεραίνουν ότι, υπάρχει μια μικρή επιρροή των ειδήσεων στην τάση του DJIA.

Οι Vicari και Gaspari στην εργασία τους με τίτλο "Analysis of news sentiments using natural language processing and deep learning" [50] χρησιμοποιούν μεθόδους ΒΜ για την πρόβλεψη της τάσης του DJIA. Αρχικά πειραματίζονται με το ίδιο σύνολο δεδομένων με της παρούσας εργασίας με διάφορα σενάρια χρονικής πρόβλεψης πετυχαίνοντας ορθότητα που κυμαίνεται περίπου από 53,5% έως 58% Θεωρώντας την ορθότητα μη ικανοποιητική, επεκτείνουν το σύνολο δεδομένων προσθέτοντας δεδομένα που άντλησαν από την ίδια πηγή και με τον ίδιο τρόπο μέχρι και τον Αύγουστο του 2020. Τα αποτελέσματα, εφαρμόζοντας τα ίδια σενάρια για την πρόβλεψη, εμφανίζονται περίπου τα ίδια για κάθε σενάριο, άλλοτε ελαφρώς μειωμένα, άλλοτε ελαφρώς αυξημένα, παραμένουν όμως στο ίδιο εύρος τιμών. Οι συγγραφείς συμπεραίνουν ότι, τα χαμηλά ποσοστά ακρίβειας οφείλονται κατά κύριο λόγο στην πηγή των ειδήσεων.

Οι Liu et al. στην εργασία τους με τίτλο "Machine Learning for Predicting Stock Market Movement using News Headlines" [51] συλλέγουν και επεξεργάζονται δεδομένα των πέντε κορυφαίων τίτλων ειδήσεων από το Reddit για το χρονικό διάστημα 7/2/2008-7/6/2020 για την πρόβλεψη της ανόδου ή της καθόδου του DJIA. Σκοπός των συγγραφέων είναι να συγκρίνουν μοντέλα ΜΜ με μοντέλα ΒΜ, εφαρμόζοντας συνδυαστικά σενάρια για διαφορετικές μεθόδους αναπαράστασης δεδομένων για κάθε αλγόριθμο και για προβλέψεις που αφορούν τον δείκτη την ίδια ημέρα με τη δημοσίευση των ειδήσεων και για προβλέψεις πέντε ημέρες μετά. Επιπλέον, αποτιμούν τα μοντέλα τους σε δύο σύνολα που προκύπτουν από το αρχικό σύνολο δεδομένων. Το ένα σύνολο αφορά δεδομένα πριν την πανδημία του COVID-19 (μέχρι 31/12/2019) και το άλλο μετά την πανδημία. Τα καλύτερα αποτελέσματα επιτυγχάνονται με τη βοήθεια της ΒΜ. Για το πρώτο σύνολο η μέγιστη ορθότητα είναι 59,6% και για το δεύτερο σύνολο 62,9%. Οι αλγόριθμοι ΜΜ που εφαρμόζονται για τη δημιουργία των μοντέλων τους είναι οι LR, SVM και RF. Το καλύτερο μοντέλο

προβλέψεων και για τα δύο σύνολα προκύπτει με την εφαρμογή του SVM με ορθότητα 54,3% για το πρώτο σύνολο και 59,5% για το δεύτερο. Οι συγγραφείς συμπεραίνουν ότι, το μοντέλο τους μπορεί να χρησιμοποιηθεί για προβλέψεις σε εποχές όπου συμβαίνουν αναπάντεχα γεγονότα, όπως αυτό της πανδημίας του COVID-19.

Οι Hassanzadeh Kalshani et al. στην εργασία τους με τίτλο "Stock Market Prediction using Daily News Headlines" [52] χρησιμοποιούν το ίδιο σύνολο δεδομένων με της παρούσας εργασίας και πειραματίζονται με διάφορους συνδυασμούς που αφορούν την προεπεξεργασία των δεδομένων κειμένου, την εξαγωγή της ανάλυσης συναισθήματος και τα δεδομένα των χρονοσειρών. Για την ανάλυση συναισθήματος χρησιμοποιούν την τεχνική που έχει προταθεί στην [53]. Για τους διάφορους συνδυασμούς ως προγνωστικό μοντέλο χρησιμοποιούν έναν αλγόριθμο BM (LSTM). Ως μέτρο απόδοσης χρησιμοποιούν την ορθότητα, το F1 αποτέλεσμα και τη μετρική ROC AUC με καλύτερα αποτελέσματα 52%, 67% και 62% αντίστοιχα. Οι συγγραφείς συμπεραίνουν ότι, αφενός χρειάζονται καλύτεροι μέθοδοι για την ανάλυση συναισθήματος και αφετέρου ότι, το σύνολο δεδομένων είναι σχετικά μικρό και όχι τόσο κατάλληλο για πρόβλεψη της τάσης του DJIA.

4 Μεθοδολογία

Το παρόν κεφάλαιο αφορά τη μεθοδολογία υλοποίησης του έργου. Αρχικά παρουσιάζεται το περιβάλλον υλοποίησης και οι σχετικές βιβλιοθήκες της Python. Στη συνέχεια, παρουσιάζεται η μεθοδολογία της υλοποίησης.

4.1 Περιβάλλον Υλοποίησης

Για τη δημιουργία του έργου επιλέχθηκε η γλώσσα προγραμματισμού Python, η οποία στις μέρες μας είναι η δημοφιλέστερη γλώσσα για εφαρμογές ΜΜ και ανάλυσης δεδομένων.

Η Python[14] είναι διερμηνευόμενη υψηλού επιπέδου και γενικού σκοπού γλώσσα προγραμματισμού, η οποία δημιουργήθηκε από τον Guido van Rossum το 1991. Είναι ισχυρή, γρήγορη, ανοιχτού κώδικα και υποστηρίζεται από την μεγαλύτερη ίσως κοινότητα

[14] https://www.python.org/

προγραμματιστών. Η έκδοση 2 της python σταμάτησε να υποστηρίζεται το 2020 και αυτή τη στιγμή υπάρχει υποστήριξη για τις εκδόσεις 3.6.x και τις μεταγενέστερες.

Η υλοποίηση έγινε στο περιβάλλον της ελεύθερης διανομής του Anaconda[15] (anaconda individual edition 2021.05-64bits), το οποίο είναι περιβάλλον βασισμένο στην python 3.8 και είναι από τις δημοφιλέστερες πλατφόρμες για τη δημιουργία έργων ΜΜ, καθώς περιέχει πολλές προεγκατεστημένες βιβλιοθήκες και τη δυνατότητα άμεσης και γρήγορης εγκατάστασης διάφορων βιβλιοθηκών. Οι σημαντικότερες βιβλιοθήκες της γλώσσας προγραμματισμού Python για την επιστήμη των δεδομένων και την ΜΜ, οι οποίες περιλαμβάνονται στη διανομή του Anaconda είναι ελεύθερες και ανοιχτού κώδικα (free and open-source), υποστηρίζονται από μεγάλες κοινότητες προγραμματιστών και διαθέτουν πολύ καλή τεκμηρίωση (documentation). Το ίδιο ισχύει και για τις βιβλιοθήκες που επιλέχθηκαν για τη δημιουργία του έργου και εγκαταστάθηκαν επιπρόσθετα στο περιβάλλον της Anaconda.

Τα αρχεία δημιουργήθηκαν στο Jupyter[16] ως notebooks. Το Jupyter είναι ένα έργο ανοιχτού κώδικα και διαθέτει ένα web-based περιβάλλον για εργασία με σημειωματάρια (notebooks) που περιέχουν κώδικα, δεδομένων και κείμενο. Τα Jupyter notebooks έχουν καθιερωθεί ως πρότυπα workspace για τα έργα που υλοποιούνται σε python σε ακαδημαϊκές εργασίες.

4.1.1 Βασικές Βιβλιοθήκες Python

Η βασική βιβλιοθήκη για επιστημονικούς και τεχνικούς υπολογισμούς είναι η SciPy[17] [54], η οποία
περιέχει λειτουργικές μονάδες για υπολογισμούς μαθηματικών και στατιστικής, καθώς και σχεδίασης.

Στα βασικά πακέτα- βιβλιοθήκες του οικοσυστήματος της SciPy περιλαμβάνονται η NumPy[18] και η pandas[19]. Η NumPy (συντόμευση της Numerical Python) θεωρείται το θεμελιώδες πακέτο για επιστημονικούς υπολογισμούς με την Python. Παρέχει υποστήριξη για μεγάλους, πολυδιάστατους πίνακες και μητρώα και διαθέτει μια μεγάλη συλλογή

[15] https://www.anaconda.com/

[16] https://jupyter.org/

[17] https://www.scipy.org/

[18] https://numpy.org/

[19] https://pandas.pydata.org/

μαθηματικών συναρτήσεων υψηλού επιπέδου για λειτουργίες γραμμικής άλγεβρας. Η pandas χρησιμοποιείται για τον χειρισμό και την ανάλυση δεδομένων. Προσφέρει δομές δεδομένων και λειτουργίες για χειρισμό αριθμητικών πινάκων και χρονοσειρών.

Η βασική βιβλιοθήκη για δισδιάστατη σχεδίαση- η δημοφιλέστερη της Python- είναι η Matplotlib[20]. Η matplotlib δεν υποστηρίξει την Python 2 μετά το 2020. Ενότητα της Matplotlib είναι η Pyplot που παρέχει διεπαφή τύπου MATLAB[21] και έχει σχεδιαστεί έτσι ώστε, να μπορεί να χρησιμοποιηθεί όπως το MATLAB. Στην Matplotlib βασίζεται και η βιβλιοθήκη σχεδίασης Seaborn που παρέχει τη δυνατότητα για τη σχεδίαση τρισδιάστατων, καθώς και πιο ελκυστικών και ενημερωτικών στατιστικών γραφικών. Μια άλλη νεότερη βιβλιοθήκη σχεδίασης είναι η Plotly [55].

Ειδικότερα για εφαρμογές επεξεργασίας κειμένου, υπάρχει η βιβλιοθήκη WordCloud [56]. Με την WordCloud δίνεται η δυνατότητα απεικόνισης των λέξεων που εμφανίζονται μέσα σε ένα κείμενο. Οι λέξεις απεικονίζονται με διαφορετικά χρώματα και μεγέθη, ανάλογα με τη συχνότητά τους ή τη σημαντικότητά τους.

Η βιβλιοθήκη scikit learn[22] [57] είναι η κυριότερη βιβλιοθήκη για την ΜΜ. Η βιβλιοθήκη διαθέτει διάφορους αλγόριθμους επιβλεπόμενης και μη επιβλεπόμενης ΜΜ, όπως για παράδειγμα για ταξινόμηση, παλινδρόμηση, συσταδοποίηση κλπ., καθώς και εργαλεία για προεπεξεργασία δεδομένων κειμένου, εξαγωγή χαρακτηριστικών και κανονικοποίηση. Έχει σχεδιαστεί για να λειτουργεί με τις αριθμητικές και επιστημονικές βιβλιοθήκες NumPy και SciPy.

4.1.2 Βιβλιοθήκες Επεξεργασίας Φυσικής Γλώσσας και Ανάλυσης Συναισθήματος

Η πρώτη βιβλιοθήκη της Python που δημιουργήθηκε για εφαρμογές επεξεργασίας κειμένου είναι η βιβλιοθήκη NLTK[23] [58] (Natural Language ToolKit). Η δημιουργία της ξεκίνησε το 2001ως μέρος ενός μαθήματος υπολογιστικής γλωσσολογίας (computational linguistics) στο τμήμα Computer and Information Science του πανεπιστημίου της Pennsylvania. Από τότε έχει αναπτυχθεί και επεκταθεί με τη συνεισφορά πολλών προγραμματιστών και έχει

[20] https://matplotlib.org/

[21] https://www.mathworks.com/products/matlab.html

[22] https://scikit-learn.org/stable/

[23] https://www.nltk.org/

υιοθετηθεί ως μία από τις πιο βασικές βιβλιοθήκες για εφαρμογές επεξεργασίας κειμένου. Νεότερη βιβλιοθήκη με παρεμφερείς και εκτεταμένες δυνατότητες είναι η spaCy[24].

Μια άλλη δημοφιλής βιβλιοθήκη είναι η TextBlob [59]. Η TextBlob βασίζεται στην NLTK και στην βιβλιοθήκη Pattern [60]. Παρέχει μια διεπαφή προγραμματισμού εφαρμογών (Application Programming Interface - API) για την πραγματοποίηση πολλών εργασιών επεξεργασίας κειμένου.

Η NLTK και η TextBlob διαθέτουν λειτουργίες ανάλυσης συναισθήματος, ενώ η spaCy δεν διαθέτει. Πιο συγκεκριμένα, η NLTK διαθέτει την λειτουργία VADER (Valence Aware Dictionary and sEntiment Reasoner) [26] που βασίζεται σε λεξιλόγιο και κανόνες, ενώ η λειτουργία ανάλυσης συναισθήματος της TextBlob παρέχεται μέσω του API της [61] και συνδυάζει λειτουργίες της NLTK και της Pattern.

4.2 Μεθοδολογία Υλοποίησης Έργου

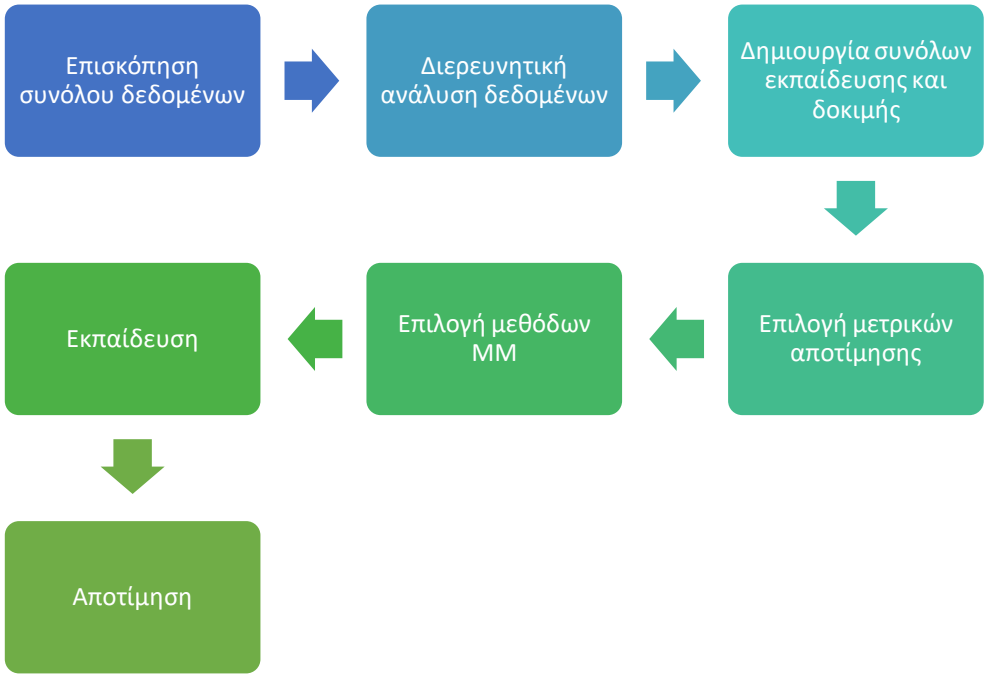
Το πρόβλημα που καλείται να διερευνήσει η παρούσα διπλωματικής εργασία είναι η επιρροή που ασκούν οι ειδήσεις στις τάσεις ανόδου ή καθόδου δεικτών του χρηματιστηρίου. Η επιρροή των ειδήσεων εκτιμάται με τεχνικές ανάλυσης συναισθήματος και οι τάσεις των δεικτών προκύπτουν από την επεξεργασία ιστορικών δεδομένων των τιμών του δείκτη. Έτσι, το πρόβλημα ανάγεται σε πρόβλημα δυαδικής ταξινόμησης και καλούμαστε να δημιουργήσουμε ένα σύστημα ΜΜ που θα δέχεται ως χαρακτηριστικά τις ειδήσεις και θα προβλέπει μια ετικέτα 0 ή 1, η οποία θα αντιπροσωπεύει την κάθοδο ή την άνοδο του δείκτη αντίστοιχα.

Το σύνολο δεδομένων που επιλέχθηκε να χρησιμοποιηθεί είναι το σύνολο CombinedNewsDJIA.csv που προέρχεται από την kaggle και είναι ένα από τα τρία σύνολα που παρέχονται κάτω από τον τίτλο «Daily News for Stock Market Prediction» [9]. Το επιλεγμένο σύνολο δεδομένων προέρχεται από συνδυασμό δύο συνόλων από δεδομένα που άντλησε ο συγγραφέας: ενός συνόλου με ιστορικά δεδομένα οκτώ ετών (2008-2016) που αφορά μόνο τις 25 κορυφαίες ημερήσιες διεθνείς ειδήσεις από ένα subreddit του Reddit, το Reddit WorldNews Channel (/r/worldnews) και ενός συνόλου με χρηματιστηριακά δεδομένα για το ίδιο χρονικό διάστημα, που αντλήθηκαν από τον ιστότοπο Yahoo Finance για τον DJIA. Τα δεδομένα του δεύτερου συνόλου επεξεργάστηκαν από τον συγγραφέα, έτσι ώστε να εντοπιστεί η τάση του δείκτη με σκοπό να δημιουργηθεί ένα τρίτο σύνολο δεδομένων που θα περιέχει τις ημερομηνίες, τις κορυφαίες ειδήσεις και την τάση του δείκτη

[24] https://spacy.io/

σε δυαδική μορφή και το πρόβλημα να αφορά εργασία δυαδικής ταξινόμησης. Πιο συγκεκριμένα, ο συγγραφέας αποδίδει την τιμή 0 για τις περιπτώσεις καθόδου του δείκτη και την τιμή 1 για τις περιπτώσεις όπου ο δείκτης ανεβαίνει ή μένει σταθερός με βάση την προσαρμοσμένη τιμή κλεισίματος. Με την επιλογή αυτού του συνόλου, δίνεται η δυνατότητα πειραματισμού για την εύρεση του καλύτερου συνδυαστικού μοντέλου τεχνικών ανάλυσης συναισθήματος και δυαδικής ταξινόμησης, καθώς και εφαρμογής μεθόδων μη επιβλεπόμενης ΜΜ.

Εφόσον εντοπίστηκε το πρόβλημα και επιλέχθηκε το σύνολο δεδομένων, το σύνολο εισάγεται στο σύστημα και πραγματοποιείται μια σειρά ενεργειών, η οποία απεικονίζεται διαγραμματικά στο Σχήμα 14 και περιγράφεται αναλυτικά στις επόμενες παραγράφους.



Σχήμα 14. Βήματα Υλοποίησης Έργου

Σημαντική παρατήρηση – σημείωση για τη συνέχεια: Τα δεδομένα κειμένου στο επιλεγμένο σύνολο δεδομένων είναι στην αγγλική γλώσσα. Ως εκ τούτου, σε ότι αναφέρεται για λεξιλόγια, λεξικά, επεξεργασία κειμένου και τα σχετικά παραδείγματα θα αφορούν την αγγλική γλώσσα.

4.3 Βήματα Υλοποίησης

4.3.1 Επισκόπηση Συνόλου Δεδομένων

Η επισκόπηση του συνόλου δεδομένων αφορά το περιεχόμενό του και τη δομή του. Πρέπει να δούμε πόσα χαρακτηριστικά διαθέτει κάθε παράδειγμα, τους τύπους των δεδομένων των χαρακτηριστικών και τις ετικέτες που αφορούν τον στόχο της ταξινόμησης. Επίσης, θα πρέπει να αποκτήσουμε μια εικόνα για το περιεχόμενο του κάθε χαρακτηριστικού για να διαπιστώσουμε εάν είναι ενδιαφέρον ως χαρακτηριστικό. Για παράδειγμα, εάν σε ένα σύνολο υπάρχει η πληροφορία για τον χρήστη που καταχώρησε μία εγγραφή, προφανώς σε μια εργασία ανάλυσης συναισθήματος αυτή η πληροφορία είναι περιττή.

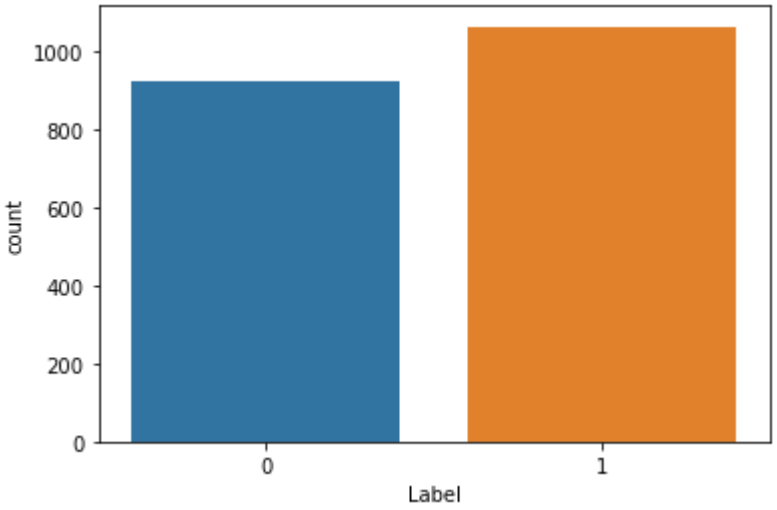
Η δομή του συνόλου μετά την εισαγωγή του .csv αρχείου ως πλαίσιο δεδομένων της Pandas φαίνεται στο Σχήμα 15. Στο σύνολο υπάρχουν 1989 γραμμές και 27 στήλες. Για τα σημερινά δεδομένα της ΜΜ, θεωρείται ένα πολύ μικρό σύνολο δεδομένων. Η στήλη Date η οποία περιέχει την ημερομηνία, η στήλη Label η οποία αποτελεί την ετικέτα για κάθε στιγμιότυπο και είναι ο στόχος της εκπαίδευσης και των προβλέψεων και οι επόμενες 25 στήλες αφορούν τους τίτλους των ειδήσεων ανάλογα με την κατάταξη (Top1 έως Top25).

Date	Label	Top1	Top2	Top3	Top4	Top5	Top6	Top7	Top8	...	Top16	Top17	Top18	
2008-08-08	0	b"Georgia 'downs two Russian warplanes' as cou...	b'BREAKING: Musharraf to be impeached.'	b'Russia Today: Columns of troops roll into So...	b'Russian tanks are moving towards the capital...	b"Afghan children raped with 'impunity,' U.N. ...	b'150 Russian tanks have entered South Ossetia...	b"Breaking: Georgia invades South Ossetia, Rus...	b"The 'enemy combatent' trials are nothing but...	...	b'Georgia Invades South Ossetia - if Russia ge...	b'Al-Qaeda Faces Islamist Backlash'	b'Condoleezza Rice: "The US would not act to p...	b' b E Unic
2008-08-11	1	b'Why wont America and Nato help us? If they w...	b'Bush puts foot down on Georgian conflict'	b"Jewish Georgian minister: Thanks to Israeli ...	b'Georgian army flees in disarray as Russians ...	b"Olympic opening ceremony fireworks 'faked'"	b'What were the Mossad with fraudulent New Zea...	b'Russia angered by Israeli military sale to G...	b'An American citizen living in S.Ossetia blam...	...	b'Israel and the US behind the Georgian aggres...	b'"Do not believe TV, neither Russian nor Geor...	b'Riots are still going on in Montreal (Canada...	b' ove a manu
2008-08-12	0	b'Remember that adorable 9-year-old who sang a...	b"Russia 'ends Georgia operation'"	b'"If we had no sexual harassment we would hav...	b"Al-Qa'eda is losing support in Iraq because ...	b'Ceasefire in Georgia: Putin Outmaneuvers the...	b'Why Microsoft and Intel tried to kill the XO...	b'Stratfor: The Russo-Georgian War and the Bal...	b"I'm Trying to Get a Sense of This Whole Geor...	...	b'U.S. troops still in Georgia (did you know t...	b'Why Russias response to Georgia was right'	b'Gorbachev accuses U.S. of making a "serious ...	b Geo NA V

Σχήμα 15. Επισκόπηση Συνόλου Δεδομένων

Με την επισκόπηση διαπιστώνεται εάν το σύνολο δεδομένων περιέχει κενές εγγραφές, οπότε θα πρέπει να αποφασιστεί πως θα γίνει η διαχείριση αυτών των κενών εγγραφών, δηλαδή εάν θα διαγραφούν ή θα συμπληρωθούν με βάση κάποια κριτήρια. Διαπιστώνεται ότι υπάρχουν κενά κελιά σε 3 εγγραφές του συνόλου και αποφασίζεται η διαγραφή των γραμμών που περιέχουν κελιά με κενό περιεχόμενο, εφόσον αντιπροσωπεύουν ένα ελάχιστο ποσοστό επί του συνόλου και το σύνολο περιέχει πλέον 1986 γραμμές.

Από τα 1986 παραδείγματα τα 1062 αντιστοιχούν σε περιπτώσεις όπου ο DJIA παρέμεινε σταθερός ή είχε άνοδο (Label= 1) και 924 αντιστοιχούν σε περιπτώσεις όπου ο DJIA είχε κάθοδο (Label= 0), όπως φαίνεται στο Σχήμα 16. Πρόκειται για ένα σχετικά ισορροπημένο σύνολο δεδομένων, με τα θετικά δείγματα να υπερτερούν, αλλά η ύπαρξη της σχετικής ισορροπίας θα πρέπει να ελεγχθεί στη συνέχεια όταν θα δημιουργηθεί το σύνολο της εκπαίδευσης.



Σχήμα 16. Κατανομή Θετικών και Αρνητικών Στιγμιοτύπων του Συνόλου Δεδομένων

Εφόσον περιλαμβάνονται για κάθε ημερομηνία και κάθε παράδειγμα 25 τίτλοι ειδήσεων, οι τίτλοι ειδήσεων συνενώνονται ανά ημερομηνία σε μια νέα στήλη με την ονομασία Combined_News, έτσι ώστε για κάθε παράδειγμα να υπάρχει ένα τμήμα κειμένου ανά ημερομηνία και οι στήλες Top1 έως Top25 δεν χρειάζονται πλέον και διαγράφονται. Το πλαίσιο των δεδομένων έχει πλέον τρεις στήλες: Date, Label και Combined_News.

Ειδικά για τα χαρακτηριστικά που αφορούν κείμενο, θα πρέπει να εξεταστεί το περιεχόμενό τους, έστω και δειγματοληπτικά, προκειμένου να διαπιστωθεί εάν υπάρχει θόρυβος και να απομακρυνθεί για να γίνει το κείμενο πιο καθαρό. Για παράδειγμα, χαρακτήρες όπως το backslash (\ ) ή τα HTML tags θεωρούνται ως θόρυβος. Ενδεικτικά, εξετάζεται το κείμενο της πρώτης γραμμής του πλαισίου δεδομένων της στήλης Combined_News, το οποίο μετά τη συνένωση έχει ως εξής:

```
 'b"Georgia \'downs two Russian warplanes\' as countries move to brink of
war" b\'BREAKING: Musharraf to be impeached.\' b\'Russia Today: Columns o
f troops roll into South Ossetia; footage from fighting (YouTube)\' b\'Ru
ssian tanks are moving towards the capital of South Ossetia, which has re
portedly been completely destroyed by Georgian artillery fire\' b"Afghan
children raped with \'impunity,\' U.N. official says - this is sick, a th
ree year old was raped and they do nothing" b\'150 Russian tanks have ent
ered South Ossetia whilst Georgia shoots down two Russian jets.\' b"Break
ing: Georgia invades South Ossetia, Russia warned it would intervene on S
```

O\'s side" b"The \'enemy combatent\' trials are nothing but a sham: Salim Haman has been sentenced to 5 1/2 years, but will be kept longer anyway j ust because they feel like it." b\'Georgian troops retreat from S. Osetta in capital, presumably leaving several hundred people killed. [VIDEO]\' b \'Did the U.S. Prep Georgia for War with Russia?\' b\'Rice Gives Green Li ght for Israel to Attack Iran: Says U.S. has no veto over Israeli militar y ops\' b\'Announcing:Class Action Lawsuit on Behalf of American Public A gainst the FBI\' b"So---Russia and Georgia are at war and the NYT\'s top story is opening ceremonies of the Olympics? What a fucking disgrace and yet further proof of the decline of journalism." b"China tells Bush to st ay out of other countries\' affairs" b\'Did World War III start today?\' b\'Georgia Invades South Ossetia - if Russia gets involved, will NATO abs orb Georgia and unleash a full scale war?\' b\'Al-Qaeda Faces Islamist Ba cklash\' b\'Condoleezza Rice: "The US would not act to prevent an Israeli strike on Iran." Israeli Defense Minister Ehud Barak: "Israel is prepared for uncompromising victory in the case of military hostilities."\' b\'Thi s is a busy day: The European Union has approved new sanctions against I ran in protest at its nuclear programme.\' b"Georgia will withdraw 1,000 soldiers from Iraq to help fight off Russian forces in Georgia\'s breakaw ay region of South Ossetia" b\'Why the Pentagon Thinks Attacking Iran is a Bad Idea - US News & World Report\' b\'Caucasus in crisis: Georgia invades South Ossetia\' b\'Indian shoe manufactory - And again in a seri es of "you do not like your work?"\' b\'Visitors Suffering from Mental Il lnesses Banned from Olympics\' b"No Help for Mexico\'s Kidnapping Surge"'

Διαπιστώνεται άμεσα ότι, υπάρχει χαρακτηριστικός θόρυβος: οι χαρακτήρες b΄, b΄΄, b΄΄΄ και τα backslash, οπότε σε πρώτη φάση πραγματοποιείται ένας καθαρισμός και αυτοί οι χαρακτήρες απομακρύνονται.

4.3.2 Διερευνητική Ανάλυση Δεδομένων

Η διερευνητική ανάλυση δεδομένων (Exploratory Data Analysis – EDA) είναι η προσέγγιση για την ανάλυση των συνόλων δεδομένων και τη σύνοψη των κύριων χαρακτηριστικών των δεδομένων με τη χρήση γραφικών παραστάσεων και διάφορες μεθόδους οπτικοποίησης δεδομένων [24].

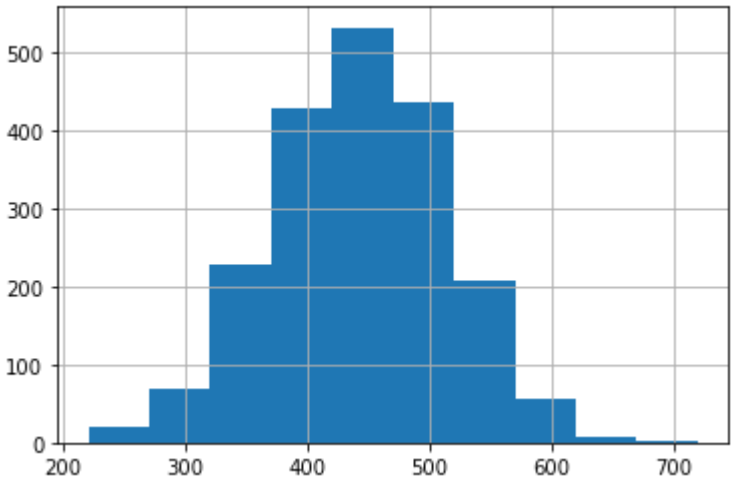
Από πολλούς ερευνητές, η EDA θεωρείται θεμελιώδους σημασίας για την ανάπτυξη ενός συστήματος ΜΜ και ιδιαίτερα μιας εφαρμογής NLP, όπως είναι η ανάλυση συναισθήματος. Μέσω της EDA αποκτάται μια εικόνα των δεδομένων του κειμένου, η οποία θα βοηθήσει, για παράδειγμα, στις αποφάσεις που θα ληφθούν για την προεπεξεργασία του κειμένου, την επιλογή των κυριότερων χαρακτηριστικών κλπ.

Στην παρούσα διπλωματική εργασία χρησιμοποιούνται οι βιβλιοθήκες σχεδίασης της Python που αναφέρθηκαν στο περιβάλλον υλοποίησης.

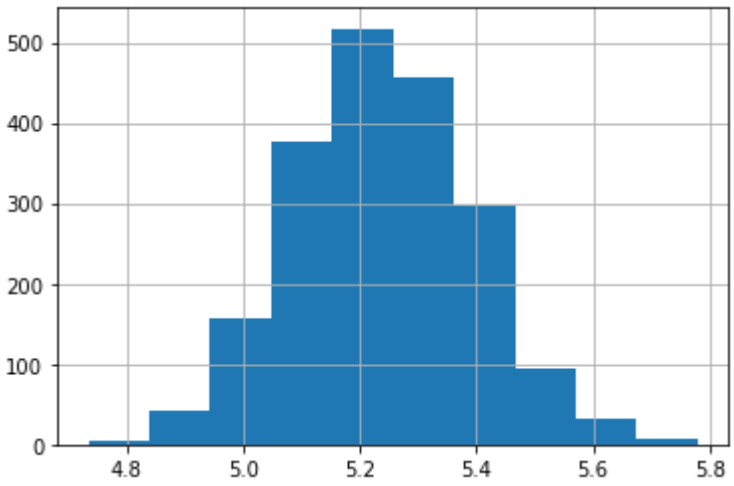
Κατά την EDA πραγματοποιείται συνήθως μια στοιχειώδης προεπεξεργασία με τη βοήθεια απλών λειτουργιών που παρέχονται από βασικές βιβλιοθήκες της Python, καθώς και την βιβλιοθήκη NLTK. Οι λειτουργίες αυτές θα αναλυθούν περισσότερο σε επόμενη παράγραφο. Για τους σκοπούς της EDA η προεπεξεργασία αφορά αρχικά την μετατροπή

όλων των γραμμάτων του κειμένου σε πεζά γράμματα. Από αυτή την επεξεργασία προκύπτει ότι, αρχικά οι λέξεις των δεδομένων του κειμένου είναι 878.697. Από το Σχήμα 17 φαίνεται ότι, ο αριθμός των λέξεων σε κάθε παράδειγμα κυμαίνεται από 250 έως 750 λέξεις με την πλειοψηφία των παραδειγμάτων να περιέχουν 400 έως 500 λέξεις. Το μέσο μήκος των λέξεων είναι περίπου 5 γράμματα, όπως φαίνεται στο Σχήμα 18, πράγμα που σημαίνει ότι, στις κορυφαίες ειδήσεις οι λέξεις που χρησιμοποιούνται είναι μικρές.

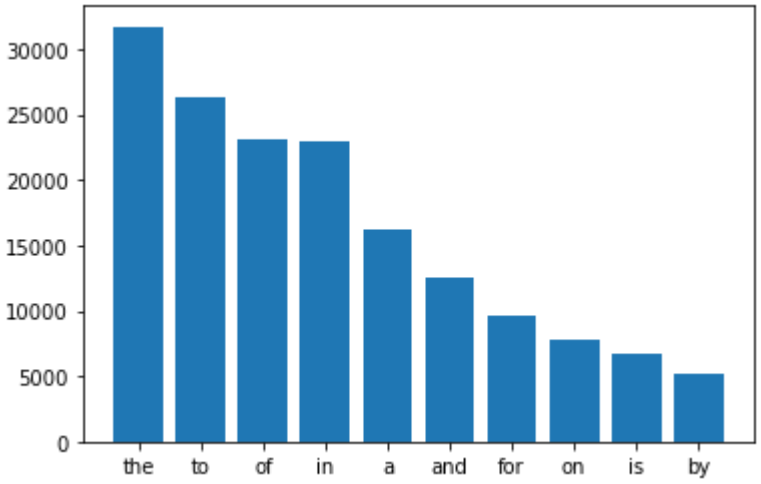
Από αυτή τη διαπίστωση οδηγούμαστε στη διερεύνηση της συχνότητας της χρήσης κοινών και πολύ συχνά χρησιμοποιούμενων λέξεων, των ονομαζόμενων stop words, τα οποία διατίθενται ως σύνολο από την NLTK, καθώς τη συχνότητα της χρήσης των υπόλοιπων λέξεων.



Σχήμα 17. Αριθμός Λέξεων ανά Παράδειγμα

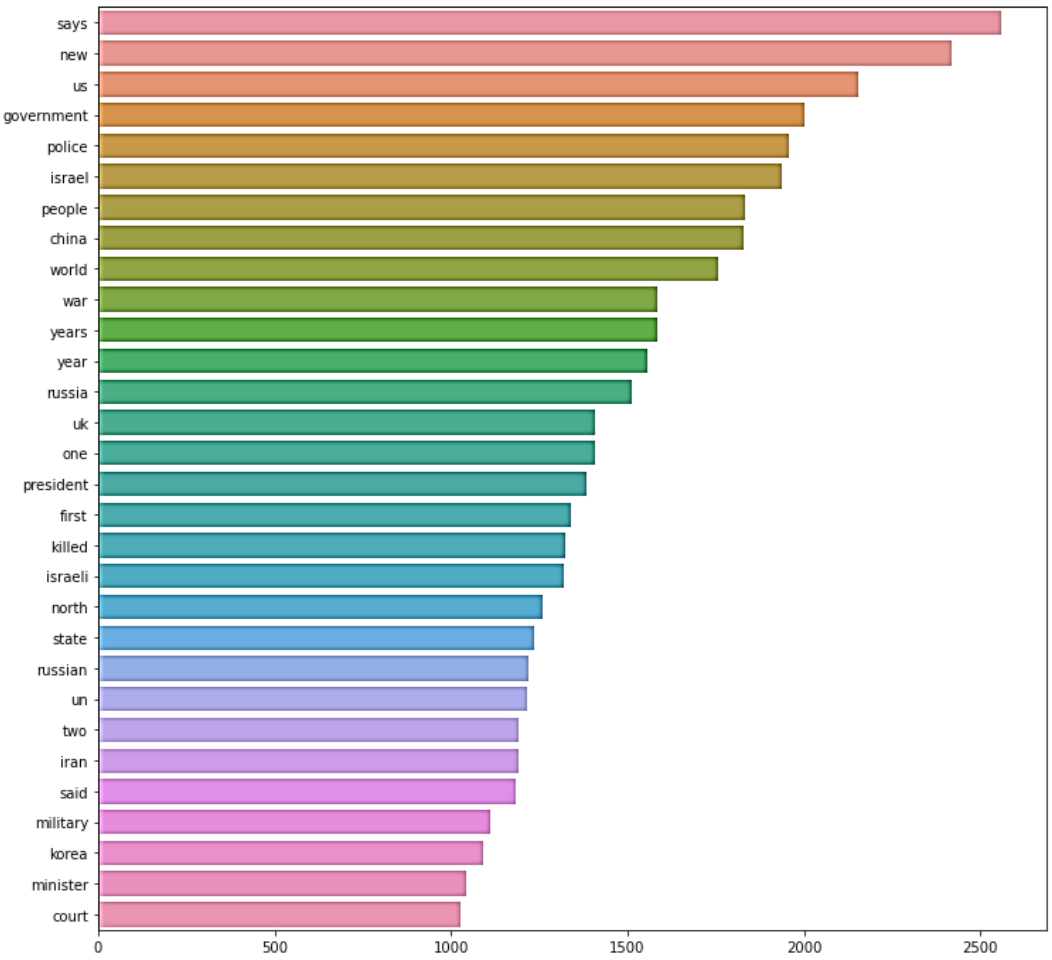


Σχήμα 18. Μέσο Μήκος Λέξεων ανά Παράδειγμα



Σχήμα 19. Οι 10 Κορυφαίες Stop Words των Ειδήσεων

Από το Σχήμα 19 διαπιστώνουμε ότι οι stop words χρησιμοποιούνται αρκετά μέσα στα κείμενα των ειδήσεων. Όσον αφορά τις υπόλοιπες λέξεις του κειμένου, μετά από την απομάκρυνση των σημείων στίξης, των αριθμών και των stop words, διαπιστώνεται ότι οι υπόλοιπες λέξεις που χρησιμοποιούνται είναι 582.980. Οι τριάντα πλέον χρησιμοποιούμενες λέξεις, καθώς και ο αριθμός για την κάθε μία από αυτές φαίνονται στο Σχήμα 20.



Σχήμα 20. Οι Συχνότερες Λέξεις των Ειδήσεων

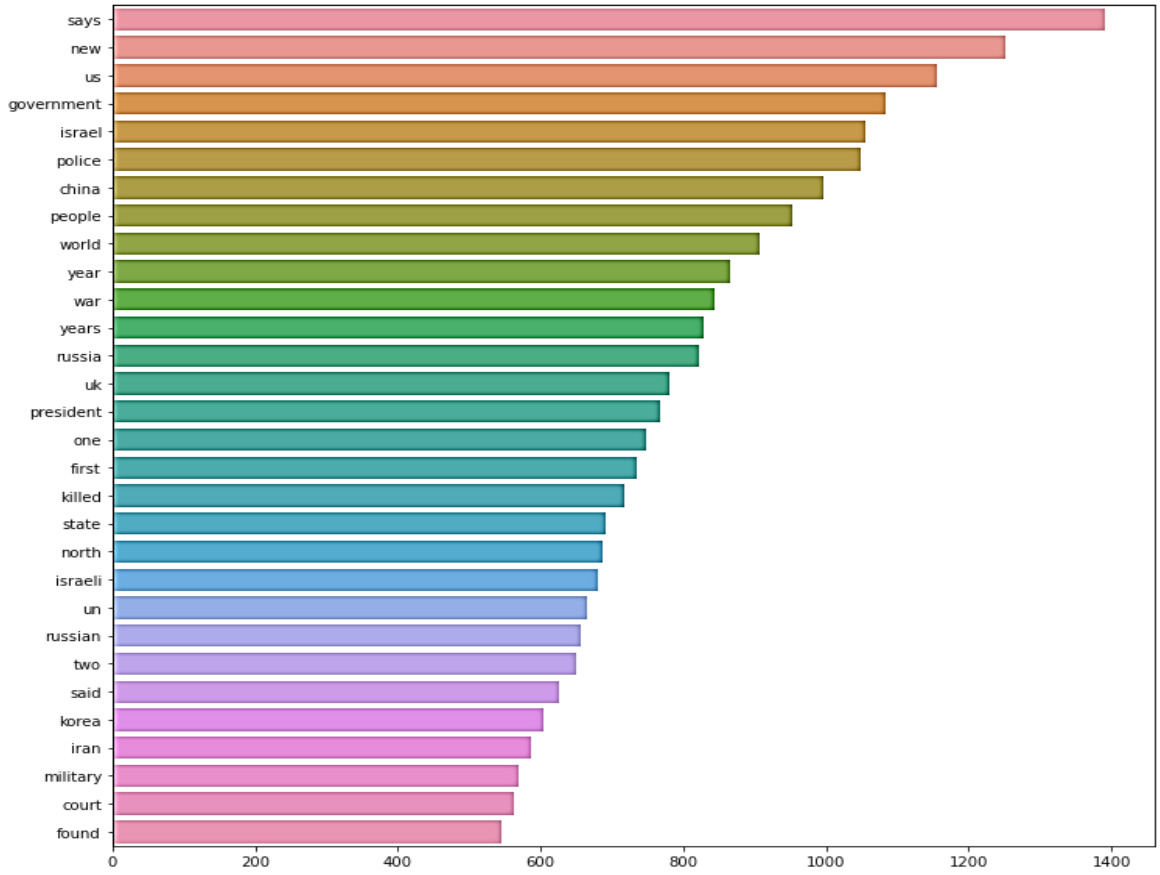
Μία πιο γενική εικόνα για τις λέξεις που κυριαρχούν στο κείμενο των ειδήσεων μπορούμε να έχουμε μέσω της βιβλιοθήκης WordCloud, όπως φαίνεται στο Σχήμα 21. Αξίζει να σημειωθεί ότι, για την οπτικοποίηση των λέξεων στο WordCloud γίνεται προεπεξεργασία κειμένου εσωτερικά, με διάφορες τεχνικές που θα δούμε στις επόμενες ενότητες. Οι λέξεις που εμφανίζονται προκύπτουν μετά από την εφαρμογή αυτών των τεχνικών και κατόπιν στατιστικής επεξεργασίας. Απλά το WorldCloud μας δίνει μια διαίσθηση για τις λέξεις που επικρατούν στο κείμενο σύμφωνα με τους δικούς του κανόνες.

Σχήμα 21. WordCloud Λέξεων των Ειδήσεων

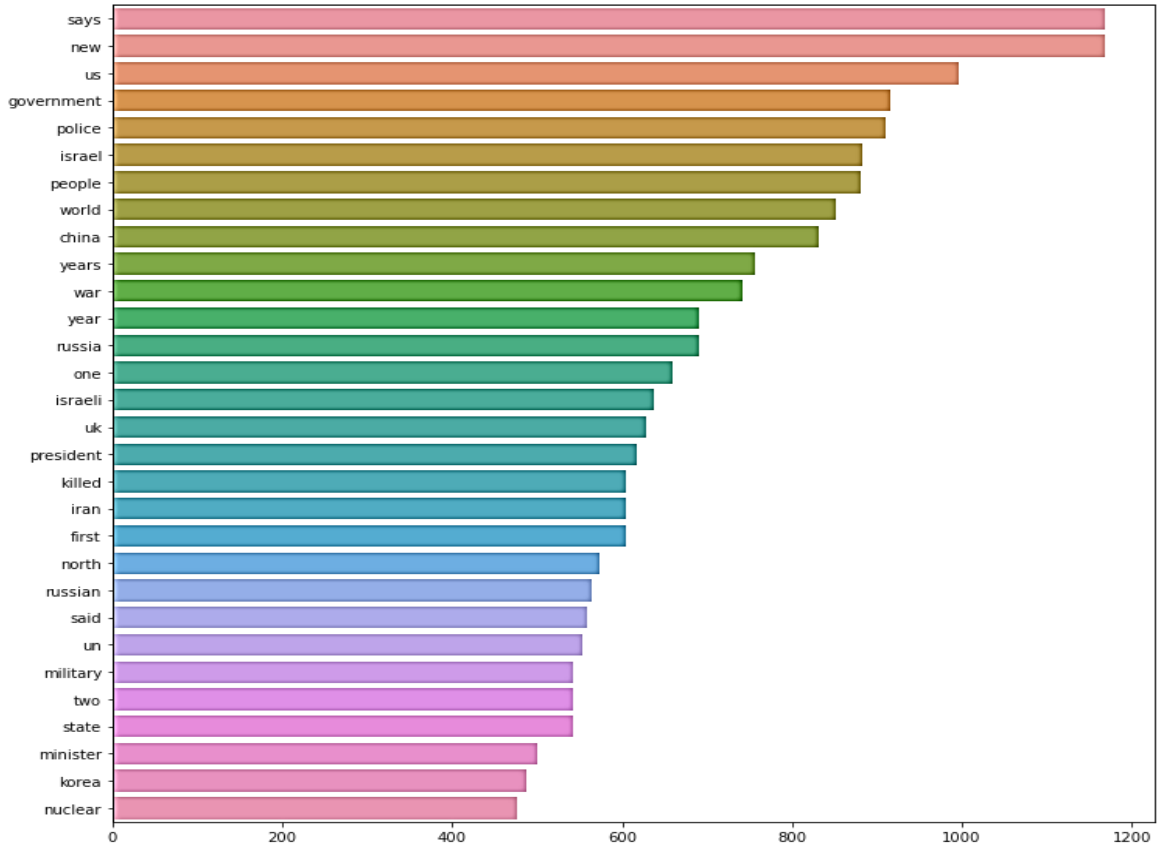
Θεωρούμε ότι, θα πρέπει να αποκτήσουμε εικόνα και διαίσθηση για τις λέξεις που περιλαμβάνονται στις ειδήσεις όταν ο δείκτης παραμένει σταθερός ή ανέρχεται και για τις λέξεις που περιλαμβάνονται στις ειδήσεις όταν ο δείκτης παρουσιάζει κάθοδο.

Από τα Σχήματα 22 και 23, όπου απεικονίζονται οι λέξεις για Label= 1 και Label= 0 αντίστοιχα, διαπιστώνουμε ότι, οι περισσότερες λέξεις στα δύο υποσύνολα είναι κοινές

και μάλιστα περίπου με τον ίδιο περίπου πλήθος. Διαισθητικά το ίδιο συμπέρασμα προκύπτει και από τα word clouds των δύο υποσυνόλων που φαίνονται στα Σχήματα 24 και 25 για τα δύο υποσύνολα Label= 1 και Label= 0 αντίστοιχα. Αυτά τα στοιχεία μας δημιουργούν ήδη την αίσθηση ότι, θα είναι δύσκολο να δημιουργήσουμε ένα μοντέλο με ικανοποιητικές μετρικές απόδοσης. Επίσης, παρατηρούμε ότι κυριαρχούν λέξεις που αφορούν χώρες (πχ russia, georgia).

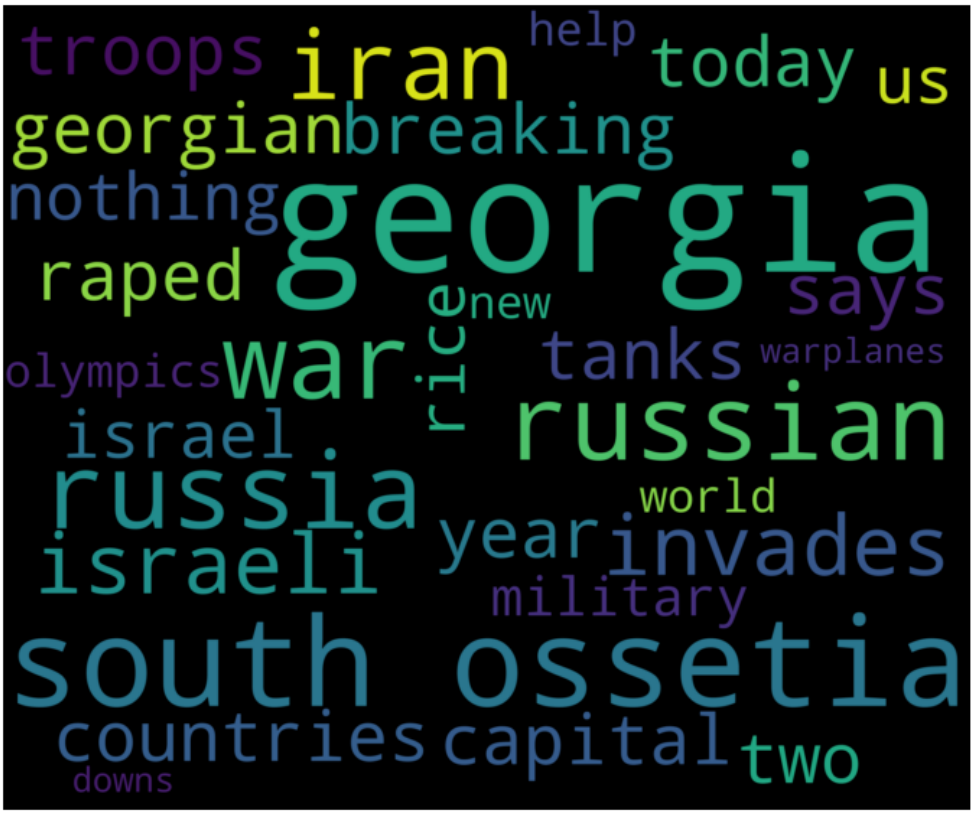


Σχήμα 22. Λέξεις Ειδήσεων για Άνοδο ή Σταθερότητα του DJIA



Σχήμα 23. Λέξεις Ειδήσεων για Κάθοδο του DJIA

Σχήμα 24. WordCloud για Άνοδο ή Σταθερότητα του DJIA

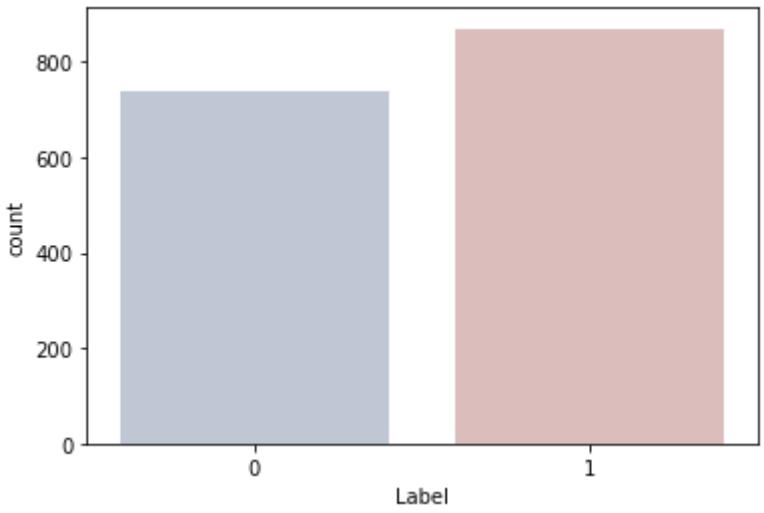


Σχήμα 25. WordCloud για Κάθοδο του DJIA

4.3.3 Δημιουργία Συνόλων Εκπαίδευσης και Δοκιμής

Για το σύνολο των δεδομένων που έχει επιλεγεί, ο συγγραφέας προτείνει ως σύνολο δεδομένων εκπαίδευσης να χρησιμοποιηθούν τα δεδομένα που υπάρχουν στο σύνολο για τις ημερομηνίες από 8/8/2008 έως και 31/12/2014 και ως σύνολο εκπαίδευσης τα δεδομένα που υπάρχουν για τις ημερομηνίες από 2/1/2015 έως και 1/7/2016. Είναι ένας διαχωρισμός του συνόλου που αντιστοιχεί περίπου στο 80%-20% και που συνήθως εφαρμόζεται σε μικρά έτοιμα σύνολα δεδομένων.

Στην παρούσα διπλωματική εργασία ακολουθείται αυτός ο διαχωρισμός. Ωστόσο, πρέπει να εξεταστεί εάν το σύνολο εκπαίδευσης είναι ισορροπημένο όσον αφορά τις ετικέτες. Το σύνολο εκπαίδευσης περιέχει 1608 παραδείγματα, όπου τα 870 αντιστοιχούν σε περιπτώσεις όπου ο DJIA παρέμεινε σταθερός ή είχε άνοδο (Label= 1) και 738 αντιστοιχούν σε περιπτώσεις όπου ο DJIA είχε κάθοδο (Label= 0), οπότε είναι ένα σχετικά ισορροπημένο σύνολο όπως φαίνεται και στο Σχήμα 26, με τα θετικά δείγματα και πάλι να υπερτερούν.



Σχήμα 26. Κατανομή Θετικών και Αρνητικών Στιγμιοτύπων του Συνόλου Εκπαίδευσης

4.3.4 Επιλογή Μετρικών Αποτίμησης

Όπως ήδη έχει αναφερθεί στην παράγραφο 2.5 (Μετρικές Απόδοσης Ταξινόμησης), για την περίπτωση της δυαδικής ταξινόμησης υπάρχουν διάφορα μέτρα για την αποτίμηση ενός μοντέλου τα οποία είναι χρήσιμα όχι μόνο για την απόδοση του μοντέλου, αλλά και τη σύγκριση διαφορετικών μοντέλων μεταξύ τους.

Στην παρούσα διπλωματική εργασία ως μετρικές αποτίμησης επιλέγονται:

- Η ορθότητα (accuracy)
- Το F1 αποτέλεσμα (F1 score)
- Η ROC AUC[25]

Οι μετρικές αυτές επιλέχθηκαν γιατί η ορθότητα είναι ένα μέτρο που πάντα αναφέρεται σε αποτιμήσεις μοντέλων ΜΜ στην σχετική βιβλιογραφία, το F1 αποτέλεσμα γιατί λαμβάνει υπόψη και την ακρίβεια και την ανάκληση σε σύνολα όπου δεν υπάρχει απόλυτη ισορροπία μεταξύ των κλάσεων και συνοψίζει την απόδοση ενός ταξινομητή [20], [24], και η ROC AUC επειδή αποτελεί ποιοτικό δείκτη του μοντέλου [5], [24]. Ωστόσο, για κάθε μοντέλο που δημιουργείται μέσα στο αντίστοιχο jupiter notebook παρουσιάζεται η ακρίβεια, η ανάκληση και ο πίνακας ταξινόμησης.

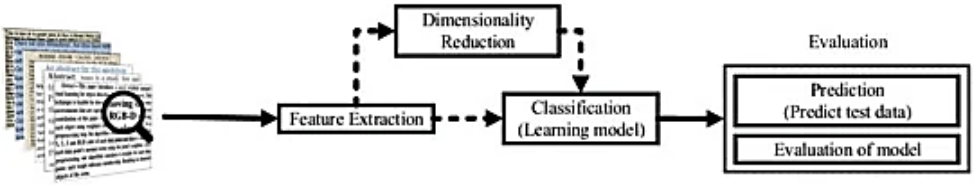
[25] Είναι η μετρική που προτείνεται από τον συγγραφέα του συνόλου δεδομένων

4.3.5 Επιλογή Μεθόδων ΜΜ

Για την ανάλυση συναισθήματος που προκύπτει από τα δεδομένα του κειμένου επιλέχθηκαν δύο μέθοδοι μάθησης: α) επιβλεπόμενης μάθησης με τη χρήση της βιβλιοθήκης ΜΜ scikit-learn και β) μη επιβλεπόμενης μάθησης.

4.3.5.1 Μέθοδοι Επιβλεπόμενης Μάθησης

Απαραίτητη προϋπόθεση για τη δημιουργία μοντέλων επιβλεπόμενης ΜΜ είναι η προεπεξεργασία των δεδομένων του κειμένου σε κατάλληλη αριθμητική μορφή για να καταστεί δυνατή η επεξεργασία τους από τον εκάστοτε αλγόριθμο. Για την ταξινόμηση του κειμένου- και κατ' επέκταση της ανάλυσης συναισθήματος που είναι μία από τις εφαρμογές της- με μεθόδους επιβλεπόμενης ΜΜ, η γενική ροή των εργασιών φαίνεται στο Σχήμα 15.



Σχήμα 27. Γενική Ροή Ταξινόμησης Δεδομένων κειμένου [22]

Γενικά, τα δεδομένα κειμένου που εισάγονται στο σύστημα περιέχουν ακολουθίες εγγράφων $D = \{X_1, X_2, \dots, X_N\}$, όπου το X_i αναφέρεται σε ένα παράδειγμα του συνόλου δεδομένων που περιέχει s αριθμό προτάσεων, κάθε πρόταση περιέχει w_s λέξεις και κάθε λέξη περιέχει l_w γράμματα . Σε κάθε παράδειγμα αντιστοιχεί μία ετικέτα στόχος, που στην περίπτωση της δυαδικής ταξινόμησης είναι 0 ή 1 [22].

Όπως ήδη έχει αναφερθεί, οι τίτλοι ειδήσεων συνενώνονται ανά ημερομηνία, έτσι ώστε για κάθε παράδειγμα να υπάρχει ένα τμήμα κειμένου ανά ημερομηνία.

Ακολουθεί η διαδικασία της *εξαγωγής των χαρακτηριστικών (feature extraction),* η οποία αφορά την μετατροπή των χαρακτηριστικών του κειμένου σε αριθμητική μορφή. Προαιρετικά, μπορεί να γίνει *μείωση των διαστάσεων (dimensionality reduction)* των χαρακτηριστικών με διάφορες μεθόδους για την περίπτωση που το σύνολο των χαρακτηριστικών είναι πολύ μεγάλο και απαιτεί μεγάλους υπολογιστικούς πόρους και πολύ χρόνο για την εκπαίδευση.

Όπως ήδη έχει αναφερθεί σε προηγούμενη παράγραφο, πριν την διαδικασία εξαγωγής χαρακτηριστικών, στις εφαρμογές ταξινόμησης κειμένου απαιτείται μία προεργασία για το κείμενο, την οποία θα περιγράψουμε αναλυτικότερα.

Τα περισσότερα κείμενα που υπάρχουν στα σύνολα δεδομένων περιέχουν λέξεις οι οποίες μπορεί να μην είναι απαραίτητες, όπως για παράδειγμα άρθρα (‘the’) επιρρήματα (‘above’, ‘across’), συχνά χρησιμοποιούμενα ρήματα ('has', ‘have’), αντωνυμίες (‘i’), κλπ. . Στις εφαρμογές NLP οι λέξεις αυτές ονομάζονται stop words και συνήθως παρέχονται ως σύνολα από τις βιβλιοθήκες της python. Επίσης, μπορεί να υπάρχουν λέξεις αργκό, ανορθογραφίες, συντομεύσεις και ακρωνύμια. Γενικά, η προεπεξεργασία κειμένου ακολουθεί τον καθαρισμό του κειμένου από τον θόρυβο και περιλαμβάνει τις ακόλουθες ενέργειες [22]:

- Μετατροπή των κεφαλαίων γραμμάτων σε μικρά (capitalization): Για παράδειγμα, οι λέξεις The και THE μετατρέπονται σε the.
- Διάσπαση του κειμένου σε λέξεις- τμήματα που ονομάζονται tokens (tokenization): Για παράδειγμα για την πρόταση "God is Great! I won a lottery." με tokenization αποδίδει τη λίστα ['God', 'is', 'Great', '!', 'I', 'won', 'a', 'lottery', '.']. Το σύνολο των tokens ενός κειμένου ονομάζεται *σώμα (corpus)*.
- Εντοπισμός stop words.
- Διαχείριση λέξεων αργκό και ακρωνυμίων ή συντομεύσεων: Για παράδειγμα, συχνά οι ΗΠΑ στο αγγλικό κείμενο αναφέρονται ως U.S.A, U.S., u.s.
- Αφαίρεση σημείων στίξης: Για παράδειγμα στην Python σημεία στίξης θεωρούνται τα !"#$%&'()*+, -./:;<=>?@[\]^_`{|}~
- Διόρθωση ορθογραφίας.
- Στελεχοποίηση (Stemming): Οι λέξεις μετατρέπονται στο στέλεχός τους. Για παράδειγμα, οι λέξεις program, programs, programmer, programming, programmers, μετατρέπονται σε program.
- Λημματοποίηση (Lemmatization): Οι λέξεις αναπαριστάνονται με το λήμμα τους ανάλογα με το μέρος του λόγου που ανήκουν. Για παράδειγμα οι λέξεις am, are, is μετατρέπονται σε be

Το σύνολο των λέξεων ενός κειμένου που προέρχεται από την προεπεξεργασία ονομάζεται *σώμα (corpus)*. Όλες οι παραπάνω λειτουργίες κατά κανόνα καλύπτονται από τις βιβλιοθήκες επεξεργασίας κειμένου που διαθέτει η Python.

Αξίζει να αναφερθεί ότι, στις περισσότερες περιπτώσεις ταξινόμησης κειμένου η απομάκρυνση των stop words δεν βελτιώνει την απόδοση του ταξινομητή, οπότε μπορεί να παραληφθεί αυτό το βήμα [22].

Μετά την προεπεξεργασία του κειμένου, για την εξαγωγή των χαρακτηριστικών απαιτείται η επιλογή της τεχνικής με την οποία θα γίνει η επιλογή των λέξεων, ώστε να μετατραπούν σε αριθμητικά διανύσματα. Η επιλογή μπορεί να γίνει με βάση τη συντακτική απεικόνιση των λέξεων και η πιο συχνά χρησιμοποιούμενη τεχνική είναι η τεχνική των $n-gram$. Τα $n-grams$ είναι όμορες ακολουθίες n λέξεων. Η τιμή $n = 1$ αφορά κάθε λέξη του κειμένου, αλλά χάνεται η συντακτική απεικόνιση. Εάν $n = 2$ οι ακολουθίες ονομάζονται διγράμματα (bigrams) και για $n = 3$ τριγράμματα (trigrams) [22].

Άλλες τεχνικές που εφαρμόζονται αφορούν την απόδοση βαρύτητας στις λέξεις ανάλογα με τη συχνότητα της εμφάνισής τους στο κείμενο. Υπάρχει η τεχνική του «σάκου των λέξεων» (Bag-of-Words – BoW), όπου απλά το κείμενο διαχωρίζεται σε λέξεις και δεν λαμβάνονται υπόψη οι λέξεις που επαναλαμβάνονται. Η πιο δημοφιλής τεχνική όμως, είναι η συχνότητα όρων - αντίστροφη συχνότητα εγγράφου (Term Frequency – Inverse Document Frequency - TF- IDF) Με αυτή την τεχνική προσδιορίζεται πόσο σημαντική είναι μια λέξη σε ένα κείμενο. Ο στόχος της χρήσης της TF- IDF σε ένα έγγραφο είναι να μειώσει τον αντίκτυπο των λέξεων που εμφανίζονται πολύ συχνά σε ένα δεδομένο σώμα (corpus) και τα οποία είναι εμπειρικά λιγότερο ενημερωτικά από τα χαρακτηριστικά που εμφανίζονται σε ένα μικρό κλάσμα του συνόλου της εκπαίδευσης του σώματος [22].

Για την προεπεξεργασία του κειμένου η scikit-learn προσφέρει όλες τις προαναφερθείσες λειτουργίες μέσω δύο κλάσεων: της CountVectorizer και της TfidfTransformer. Όμως προσφέρει ακόμη μία επιλογή που αποτελεί τον συνδυασμό των δύο κλάσεων: την TfidfVectorizer [62].

Στις πολλαπλές δυνατότητες που δίνει η TfidfVectorizer περιλαμβάνονται η tokenization, η capitalization, η αφαίρεση των stop words, η αγνόηση των σημείων στίξης, η επιλογή n-gram, η αγνόηση όρων που εμφανίζονται με συχνότητα μεγαλύτερη ή μικρότερη από ένα δοσμένο κατώφλι και η επιλογή μέγιστου αριθμού των χαρακτηριστικών. Οι δύο τελευταίες αναφερθείσες δυνατότητες μπορούν να χρησιμοποιηθούν ως μέθοδος μείωσης των διαστάσεων των χαρακτηριστικών. Επίσης, αξίζει να αναφερθεί ότι, ενώ εξ ’ορισμού δεν δίνεται από την κλάση η δυνατότητα stemming και lemmatization, μπορεί να γίνει μέσω υπερπαραμέτρου της κλάσης με επέμβαση του χρήστη.

Έτσι, μετά τον καθαρισμό του κειμένου, για την εξαγωγή των χαρακτηριστικών και τη μείωση των διαστάσεων επιλέχθηκε να χρησιμοποιηθεί η κλάση TfidfVectorizer χωρίς stemming ή lemmatization γιατί έχουμε ένα σχετικά μικρό corpus.

Το επόμενο βήμα είναι η επιλογή του αλγόριθμου ταξινόμησης και στη συνέχεια η αποτίμησή του.

Για τη δημιουργία προγνωστικών μοντέλων, επιλέχθηκαν οι πιο αντιπροσωπευτικοί από τους δημοφιλέστερους αλγόριθμους για προβλήματα δυαδικής ταξινόμησης, δηλαδή ταξινομητές, που διαθέτει η βιβλιοθήκη scikit-learn [57].

Πιο συγκεκριμένα, επιλέχθηκαν για τις προβλέψεις τέσσερις αλγόριθμοι:

1. Η λογιστική παλινδρόμηση (LR)
2. Ο απλοϊκός Bayes (NB)
3. Η μηχανή υποστήριξης διανυσμάτων (SVM)
4. Το τυχαίο δάσος (RF)

Για κάθε έναν από τους παραπάνω αλγορίθμους στην scikit-learn υπάρχουν εξ ‘ορισμού (default) τιμές για τις υπερπαραμέτρους του κάθε αλγόριθμου. Οι υπερπαράμετροι και το τί αντιπροσωπεύει η κάθε μία από αυτές αναφέρονται συνοπτικά στο documentation του ιστότοπου της βιβλιοθήκης. Ο κώδικας για κάθε αλγόριθμο και τις υπερπαραμέτρους του διατίθεται στο αποθετήριο του github της βιβλιοθήκης [63]. Επίσης, στον ιστότοπο της βιβλιοθήκης, δίνεται για κάθε αλγόριθμο το απαιτούμενο θεωρητικό υπόβαθρο με βάση το οποίο έχει υλοποιηθεί ο κάθε αλγόριθμος στην scikit-learn.

Επίσης, εκμεταλλευόμαστε τη δυνατότητα που δίνει η scikit-learn για την εκτέλεση διαδοχικών λειτουργιών μέσω της κλάσης Pipeline[26], που δημιουργεί ένα αγωγό δεδομένων (pipeline), δηλαδή ένα σύνολο στοιχείων επεξεργασίας δεδομένων που συνδέονται σε σειρά, όπου η έξοδος ενός στοιχείου είναι η είσοδος του επόμενου. Μέσω αυτής της δυνατότητας εφαρμόζεται διαδοχικά μια λίστα μετασχηματισμών και ένας τελικός ταξινομητής. Δηλαδή ορίζουμε πρώτα στον pipeline τη μέθοδο εξαγωγής χαρακτηριστικών, ώστε να πραγματοποιηθούν όλοι οι απαραίτητοι μετασχηματισμοί και η έξοδος της μεθόδου είναι η είσοδος των αριθμητικών πλέον δεδομένων στον ταξινομητή.

[26] https://scikit-learn.org/stable/modules/generated/sklearn.pipeline.Pipeline.html

4.3.5.2 Μέθοδοι Μη- Επιβλεπόμενης Μάθησης

Για την εφαρμογή μεθόδων μη-επιβλεπόμενης ΜΜ επιλέχθηκαν δύο αντιπροσωπευτικές βιβλιοθήκες της python που διαθέτουν λειτουργίες ανάλυσης συναισθήματος, η NLTK που διαθέτει την VADER και η TextBlob.

Μέσω των λειτουργιών της VADER και της TextBlob εξάγεται ένα αποτέλεσμα για το συναίσθημα με απευθείας εφαρμογή των λειτουργιών στο σύνολο της δοκιμής, χωρίς να προαπαιτείται εκπαίδευση. Με κατάλληλες μετατροπές των αποτελεσμάτων που εξάγονται οι τιμές μπορούν να μετατραπούν με βάση κριτήρια που τίθενται από τον χρήστη σε δυαδικές τιμές και στη συνέχεια να γίνει η αποτίμηση στο σύνολο δοκιμής.

Η VADER είναι ένα λεξιλόγιο (lexicon) και ένα εργαλείο ανάλυσης συναισθήματος βασισμένο σε κανόνες (rule-based sentiment analysis tool), το οποίο είναι προσαρμοσμένο κυρίως για τα συναισθήματα που εκφράζονται στα κοινωνικά μέσα, αλλά λειτουργεί ικανοποιητικά για κείμενα που προέρχονται και από άλλους τομείς [64]. Με το VADER δίνεται η δυνατότητα εύρεσης της πολικότητας (polarity) του συναισθήματος σε τρεις κατηγορίες- αρνητικό, ουδέτερο και θετικό που αποδίδονται με μορφή κειμένου (neg, neu, pos αντίστοιχα), καθώς και η ένταση της κάθε κατηγορίας, η οποία εκφράζεται ως πραγματικός αριθμός. Επίσης, υπολογίζεται ένα αποτέλεσμα συνδυασμού (compound score) ανάλογα με το σθένος κάθε λέξης. Το σθένος κάθε λέξης είναι προσδιορισμένο μέσα στο λεξιλόγιο του VADER. Το compound score κανονικοποιείται μεταξύ του -1 (που αντιπροσωπεύει το άκρως αρνητικό συναίσθημα) και του +1 (που αντιπροσωπεύει το άκρως θετικό συναίσθημα) και εξάγεται ως πραγματικός αριθμός. Σύμφωνα με την [64], είναι η πιο χρήσιμη και η πλέον χρησιμοποιούμενη μετρική για την εξαγωγή συναισθήματος μιας πρότασης ενός κειμένου, οπότε για την εξαγωγή του αποτελέσματος της ανάλυσης συναισθήματος στην παρούσα εργασία χρησιμοποιείται το compound score.

Η λειτουργία ανάλυσης συναισθήματος της TextBlob δίνει τη δυνατότητα προσδιορισμού της πολικότητας και την υποκειμενικότητα; (subjectivity) εξάγοντας ένα αποτέλεσμα. Η τιμή της πολικότητας είναι ένας πραγματικός αριθμός στο διάστημα [-1, 1], όπου οι τιμές αντιπροσωπεύουν ότι ακριβώς και το compound score του VADER, δηλαδή το -1 αντιστοιχεί στο άκρως αρνητικό και το +1 στο άκρως θετικό συναίσθημα. Η τιμή της υποκειμενικότητας ένας πραγματικός αριθμός στο διάστημα [0, 1], όπου το 0 αντιπροσωπεύει το άκρως αντικειμενικό και το 1 το άκρως υποκειμενικό. Για την εξαγωγή του αποτελέσματος της ανάλυσης συναισθήματος στην παρούσα εργασία χρησιμοποιείται η τιμή polarity.

Για την εφαρμογή των μεθόδων VADER και TextBlob, εκτός από τον καθαρισμό του κειμένου από τον θόρυβο, δεν απαιτείται καμία άλλη προεπεξεργασία στο κείμενο, σε αντίθεση με τις μεθόδους επιβλεπόμενης μάθησης. Αξίζει επίσης να σημειωθεί ότι, οι δημιουργοί και των δύο εργαλείων, προτείνουν την πραγματοποίηση της ανάλυσης συναισθήματος σε επίπεδο πρότασης του κειμένου. Για τις περιπτώσεις παραγράφων, το κείμενο χωρίζεται σε προτάσεις, υπολογίζεται το αποτέλεσμα του συναισθήματος και στη συνέχεια εξάγεται ο μέσος όρος. Στην παρούσα διπλωματική εργασία, εφόσον έχουν συνενωθεί οι κορυφαίες ειδήσεις σε μία παράγραφο, ακολουθείται αυτή την προσέγγιση.

4.3.6 Αποτίμηση

Η αποτίμηση του κάθε μοντέλου γίνεται στο σύνολο δοκιμής με τις μετρικές που έχουν επιλεγεί.

Για την περίπτωση εφαρμογής μεθόδων επιβλεπόμενης ΜΜ, αφού δημιουργηθεί ο pipeline, προηγείται η εκπαίδευση στο σύνολο εκπαίδευσης και στη συνέχεια γίνεται η αποτίμηση.

Για την περίπτωση εφαρμογής μεθόδων μη επιβλεπόμενης ΜΜ η αποτίμηση γίνεται απευθείας στο σύνολο δοκιμής.

5 Αποτελέσματα και Συζήτηση

Το παρόν κεφάλαιο αφορά την παρουσίαση των αποτελεσμάτων τα συγκριτικά αποτελέσματα της υλοποίησης των μεθόδων ΜΜ για την ανάλυση συναισθημάτων και την πρόβλεψη της τάσης του δείκτη DJIA που εφαρμόστηκαν στο σύνολο δεδομένων Combined_News_DJIA.csv από το σετ συνόλων «Daily News for Stock Market Prediction» και η συζήτηση για τα επιμέρους αποτελέσματα.

5.1 Αποτελέσματα Υλοποίησης με Μεθόδους Επιβλεπόμενης ΜΜ

Οι πειραματισμοί για την εύρεση του καλύτερου δυνατού προγνωστικού μοντέλου με μεθόδους επιβλεπόμενης ΜΜ έγιναν εφαρμόζοντας διαφορετικούς συνδυασμούς προεπεξεργασίας κειμένου και εξαγωγής χαρακτηριστικών με τη δημιουργία των αντίστοιχων pipelines. Σε κάθε pipeline ορίστηκαν δύο διαδοχικές εργασίες:

- Η εφαρμογή της κλάσης TfidfVectorizer [62]
- Η εφαρμογή της κλάσης του αλγόριθμου ΜΜ

Οι κλάσεις των αλγόριθμων που επιλέχθηκαν από την sklearn, διατηρώντας ως υπερπαραμέτρους τις εξ' ορισμού τιμές της βιβλιοθήκης, είναι:

- Για τον αλγόριθμο SVM οι LinearSVC και SVC [65]
- Για τον αλγόριθμο LR η LogisticRegression [66]
- Για τον αλγόριθμο NB οι BernoulliNB και MultinomialNB [33]
- Για τον αλγόριθμο RF η RandomForestClassifier [67]

Η πρώτη φάση των πειραματισμών έγινε διατηρώντας στην κλάση TfidfVectorizer τις εξ' ορισμού υπερπαραμέτρους της για την προεπεξεργασία και την εξαγωγή των χαρακτηριστικών, οι οποίες είναι:

class sklearn.feature_extraction.text.**TfidfVectorizer**(*, *input='content'*, *encoding='utf-* *decode_error='strict'*, *strip_accents=None*, *lowercase=True*, *preprocessor=None*, *tokenizer=None*, *anal*

yzer='word', stop_words=None, token_pattern='(?u)\b\w\w+\b', ngram_range=(1, 1), max_df=1.0, mi n_df=1, max_features=None, vocabulary=None, binary=False, dtype=<class 'numpy.float64'>, norm='l2', use_idf=True, smooth_idf=True, sublinear_tf=False)

Ουσιαστικά, διατηρώντας τις εξ' ορισμού παραμέτρους, εκτός από την στοιχειώδη επεξεργασία που αφορά capitalization, tokenization και αφαίρεση σημείων στίξης, στα χαρακτηριστικά περιλαμβάνονται όλες ανεξαιρέτως οι λέξεις ξεχωριστά με τις παραμέτρους max_features= None και ngram_range= (1, 1), ανεξάρτητα από τη συχνότητα εμφάνισης στο κείμενο με τις παραμέτρους max_df= 1.0 και min_df= 1. Αυτή η επιλογή είχε ως αποτέλεσμα τη δημιουργία ενός πίνακα χαρακτηριστικών εκπαίδευσης με 1608 γραμμές και 31743 στήλες.

Τα αριθμητικά αποτελέσματα των μετρικών του κάθε μοντέλου παρουσιάζονται στον Πίνακα 2 και στο Σχήμα 28 απεικονίζονται γραφικά. Επίσης, στο Σχήμα 29 συγκρίνονται οι επιμέρους μετρικές για κάθε μοντέλο.

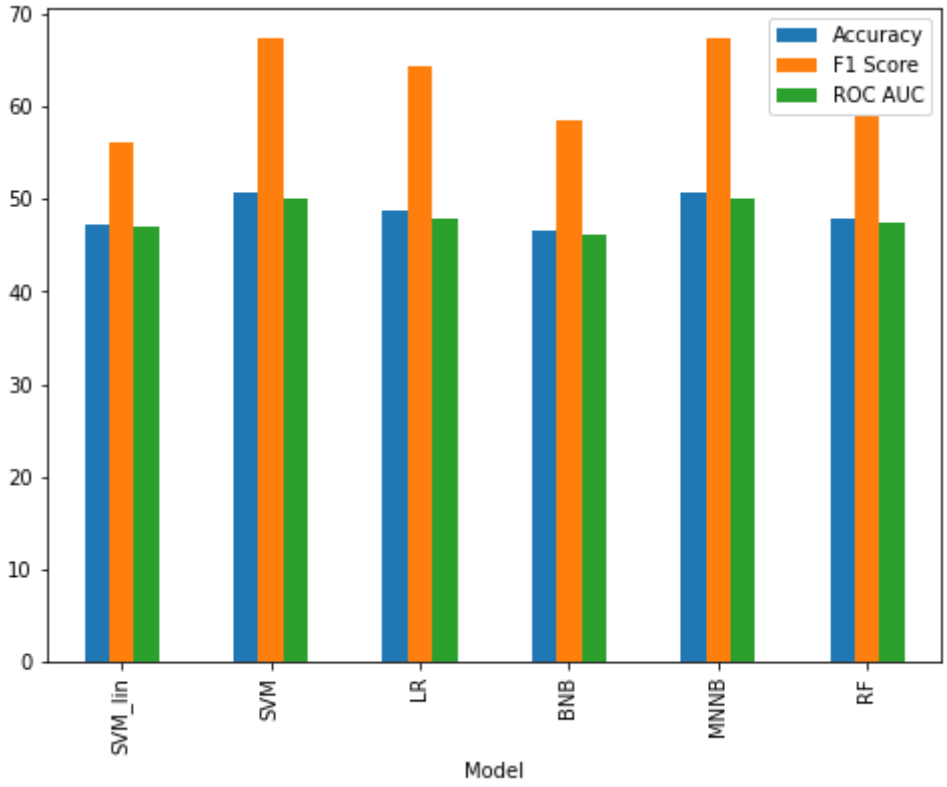
Πίνακας 2. Αποτελέσματα Μοντέλων για ngram (1,1)

Μοντέλο	Accuracy (%)	F1 Score (%)	ROC AUC (%)
LinearSVC	47,35	56,26	47,04
SVC	50,79	67,37	50,00
LogisticRegression	48,68	64,47	47,98
BernoulliNB	46,56	58,44	46,12
MultinomialNB	50,79	67,37	50,00
RandomForestClassifier	47,88	58,87	47,47

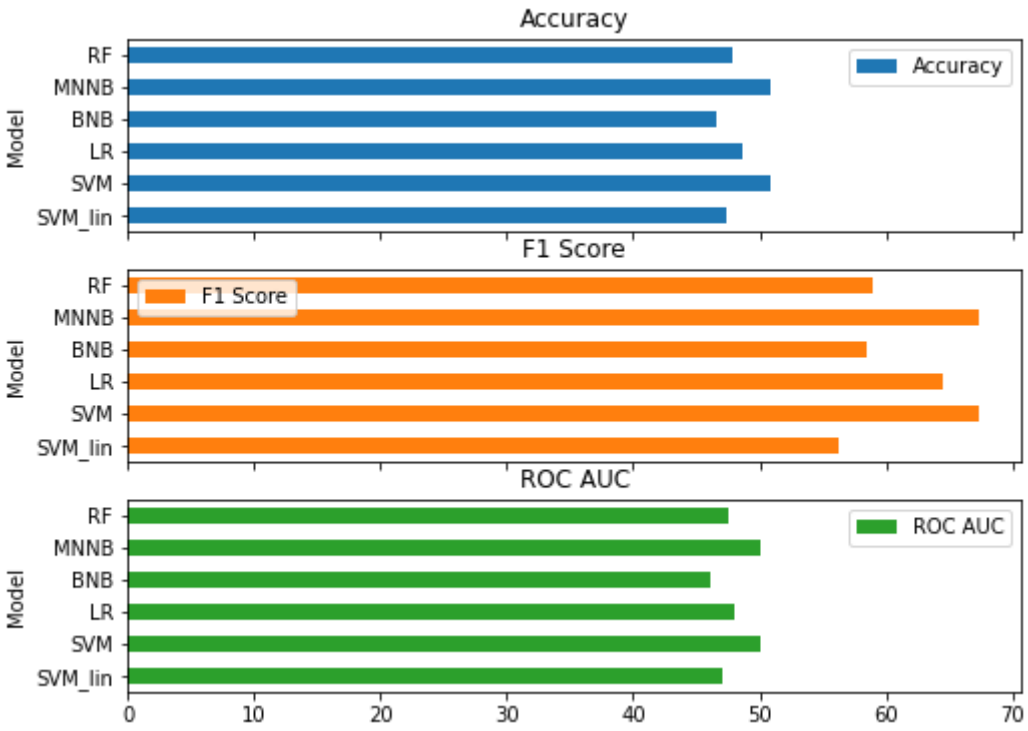
Από τα αποτελέσματα βλέπουμε ότι κανένα μοντέλο δεν αποδίδει ικανοποιητικά, κάτι το οποίο ήταν αναμενόμενο από τα αποτελέσματα της EDA, όπου είχε διαπιστωθεί ότι υπάρχουν πολλές κοινές λέξεις για τις περιπτώσεις όπου ο δείκτης εμφανίζει άνοδο ή μένει σταθερός και για τις περιπτώσεις όπου ο δείκτης εμφανίζει κάθοδο. Επίσης, όπως ήδη έχει αναφερθεί, παίρνοντας την κάθε λέξη του κειμένου ξεχωριστά, χάνεται η συντακτική απεικόνιση.

Ωστόσο, είναι εμφανές ότι, για τις συγκεκριμένες επιλογές των παραμέτρων σε κάθε pipeline, τα καλύτερα αποτελέσματα- και μάλιστα ακριβώς τα ίδια- έχει η εφαρμογή των ταξινομητών SVC και MultinomialNB, των οποίων ο πίνακας ταξινόμησης και η ROC καμπύλη δίνονται στο Σχήμα 30.

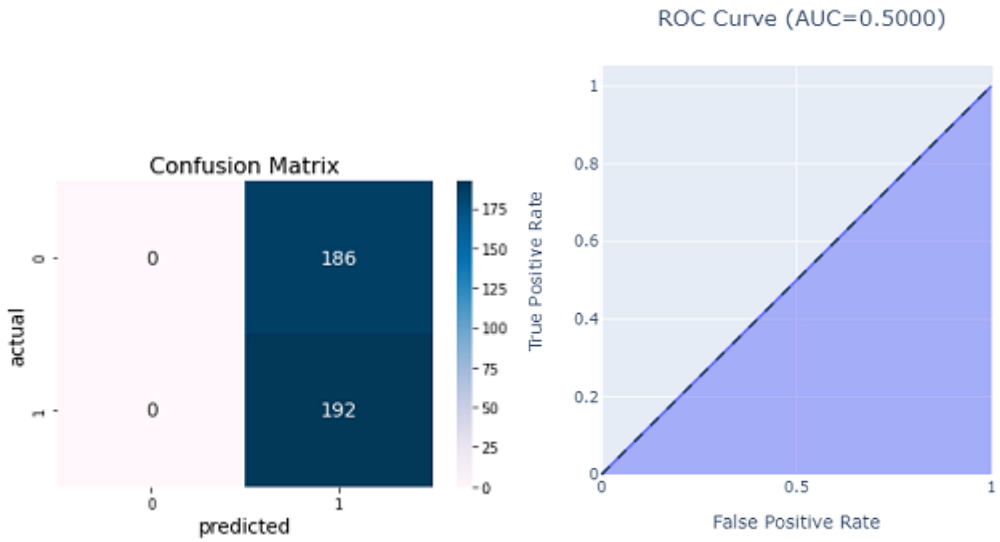
Εφόσον οι ταξινομητές έχουν τιμή ROC-AUC= 0,5, σημαίνει ότι δεν είναι σε θέση να διακρίνουν μεταξύ θετικών και αρνητικών σημείων και προβλέπουν είτε τυχαία κλάση, είτε σταθερή κλάση για όλα τα σημεία δεδομένων.



Σχήμα 28. Σύγκριση Μοντέλων για ngram (1,1)



Σχήμα 29. Σύγκριση Μετρικών Μοντέλων για ngram (1,1)



Σχήμα 30. Πίνακας Ταξινόμησης και ROC Καμπύλη

Από τον Πίνακα ταξινόμησης φαίνεται ξεκάθαρα η ικανότητα των μοντέλων να αναγνωρίζουν μόνο τη θετική κλάση. Η ανάκληση (recall), δηλαδή η αναλογία κάθε θετικού παραδείγματος που είναι πραγματικά θετικό, που αφορά την ικανότητα του μοντέλου να αναγνωρίζει ένα παράδειγμα της θετικής κλάσης και στα δύο μοντέλα είναι

100%. Η ακρίβεια (precision), δηλαδή η αναλογία του κάθε παραδείγματος που προβλέπεται ως θετικό και είναι πραγματικά θετικό για κάθε μοντέλο είναι 50,79%. Έτσι προκύπτει ένας χαμηλός μέσος αρμονικός όρος που αντικατοπτρίζει το F1 αποτέλεσμα.

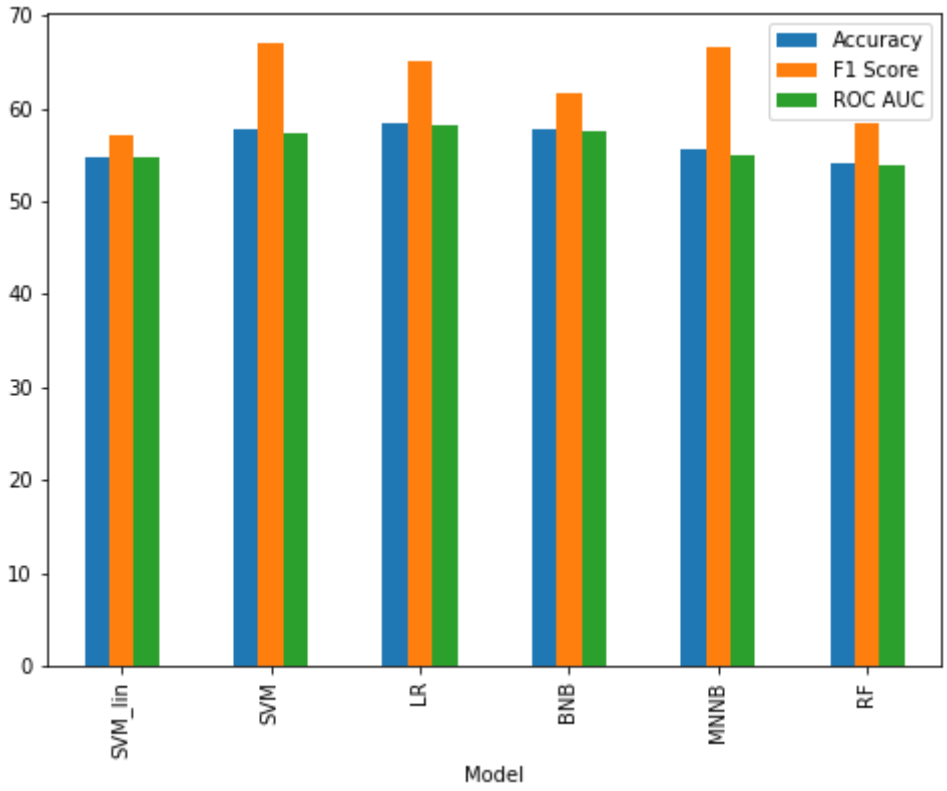
Τέλος, να αναφέρουμε ότι, σε αυτή την πρώτη φάση για τους δύο αυτούς αλγόριθμους έγιναν δοκιμές για διαφορετικές υπερπαραμέτρους τους με τη βοήθεια της κλάσης GridSearchCV της sklearn που αυτοματοποιεί τη διαδικασία εύρεσης των καλύτερων υπερπαραμέτρων ενός μοντέλου [68], χωρίς να υπάρξει καμία απολύτως βελτίωση στα αποτελέσματα. Επίσης, το αξιοπερίεργο ήταν ότι, με την απομάκρυνση των stop words και των αριθμητικών ψηφίων, τα αποτελέσματα ήταν σε κάποιες περιπτώσεις χειρότερα. Συνεπώς, θεωρήθηκε ως μη απαραίτητη η παρουσίαση των αποτελεσμάτων από τις δύο παραπάνω δοκιμές.

Η δεύτερη φάση των πειραματισμών έγινε αλλάζοντας στην κλάση TfidfVectorizer τις υπερπαραμέτρους που αφορούν τα n-grams, αλλάζοντας την παράμετρο ngram_range σε (2, 2) και θέτοντας ως όριο ελάχιστης συχνότητας εμφάνισης των λέξεων min_df= 0,025. Αυτή η επιλογή είχε ως αποτέλεσμα τη δημιουργία ενός πολύ μικρού σχετικά πίνακα χαρακτηριστικών εκπαίδευσης με 1608 γραμμές και 869 στήλες.

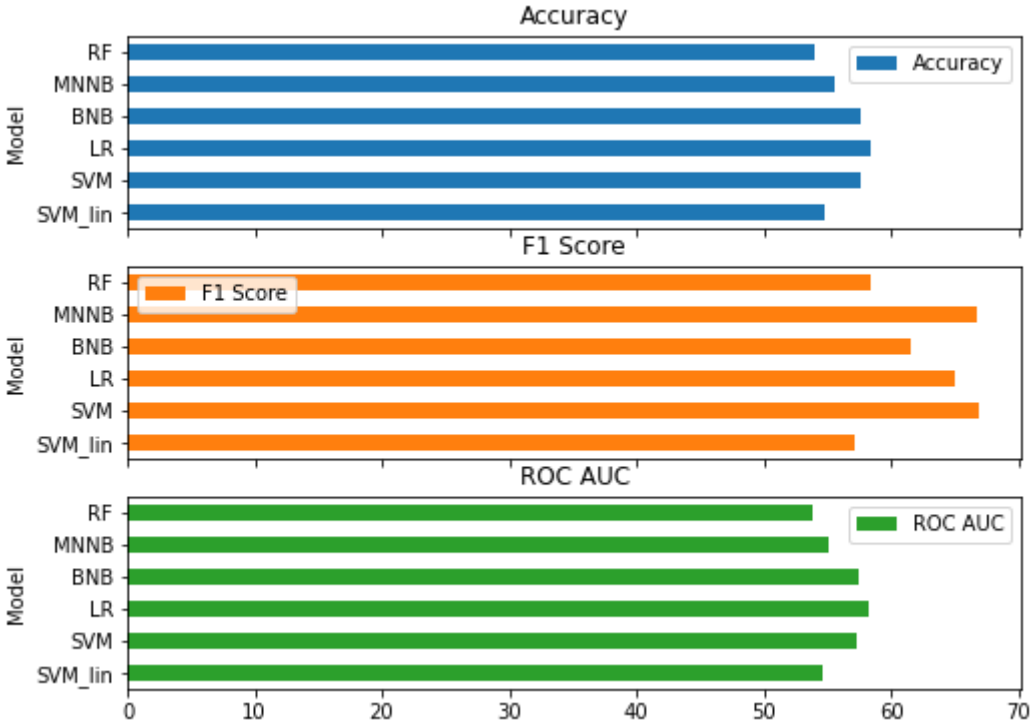
Τα αριθμητικά αποτελέσματα των μετρικών του κάθε μοντέλου παρουσιάζονται στον Πίνακα 3 και στο Σχήμα 31 απεικονίζονται γραφικά. Επίσης, στο Σχήμα 32 συγκρίνονται οι επιμέρους μετρικές για κάθε μοντέλο.

Πίνακας 3. Αποτελέσματα Μοντέλων για ngram (2, 2)

Μοντέλο	Accuracy (%)	F1 Score (%)	ROC AUC (%)
LinearSVC	54,76	57,14	54,69
SVC	57,67	66,94	57,24
LogisticRegression	58,47	65,03	58,18
BernoulliNB	57,67	61,54	57,53
MultinomialNB	55,56	66,67	55,04
RandomForestClassifier	53,97	58,37	53,81



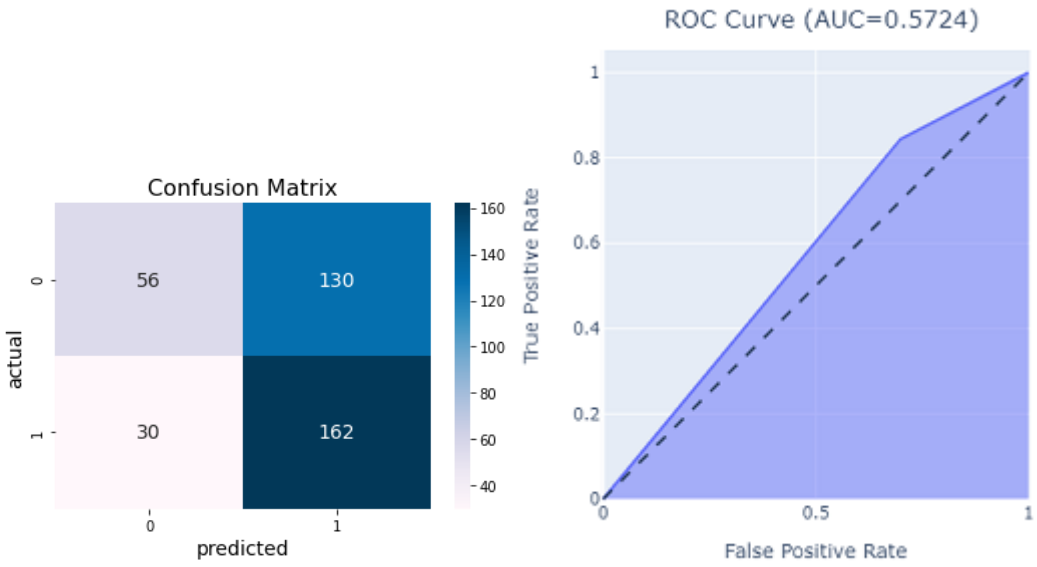
Σχήμα 31. Σύγκριση Μοντέλων για ngram (2,2)



Σχήμα 32. Σύγκριση Μετρικών Μοντέλων για ngram (2,2)

Από τα αποτελέσματα βλέπουμε ότι, βελτιώθηκαν αισθητά οι μετρικές της ορθότητας και της ROC AUC για όλα τα μοντέλα, ενώ το F1 αποτέλεσμα παρέμεινε στα

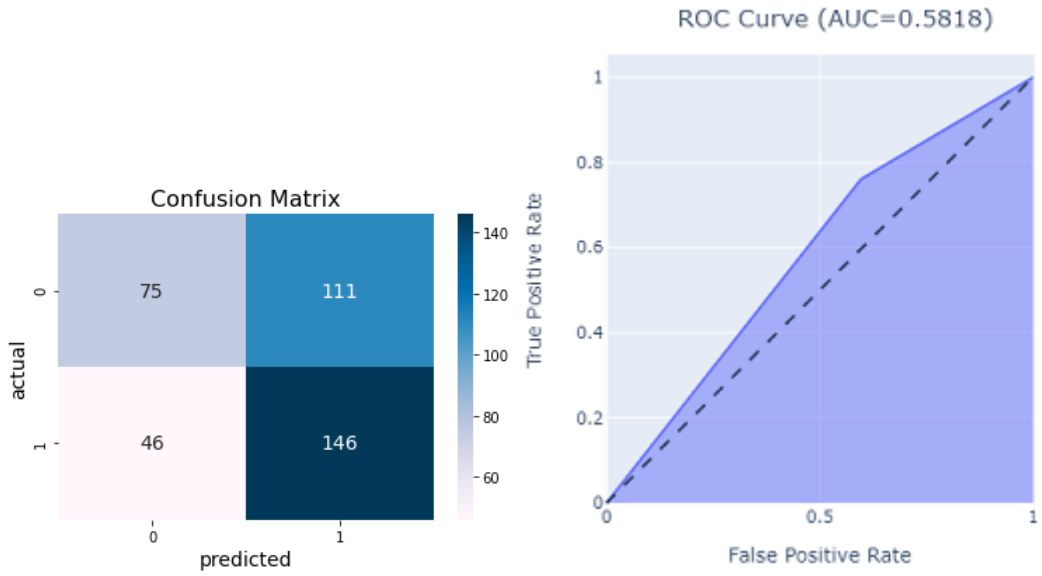
ίδια περίπου επίπεδα. Εάν για την επιλογή του καλύτερου μοντέλου λάβουμε υπόψη το F1 αποτέλεσμα, την καλύτερη επίδοση εμφανίζει ο ταξινομητής SVC, ενώ εάν λάβουμε υπόψη είτε την ορθότητα, είτε την ROC AUC τιμή, την καλύτερη επίδοση έχει ο ταξινομητής LogisticRegression, δηλαδή ένα απλό γραμμικό μοντέλο.



Σχήμα 33. Πίνακας Ταξινόμησης και ROC Καμπύλη SVC

Ο πίνακας ταξινόμησης και η καμπύλη ROC του ταξινομητή SVC φαίνονται στο Σχήμα 33. Ο ταξινομητής έχει ανάκληση 84,38% και ακρίβεια 55,48%, οπότε κατά κάποιο τρόπο βελτιώνεται η πρόβλεψη, με αποτέλεσμα στο σύνολο να έχουμε 218 σωστές προβλέψεις. Επίσης, είναι εμφανής η βελτίωση του ταξινομητή στην τιμή AUC, η οποία βελτιώθηκε κατά 7,24%.

Ο πίνακας ταξινόμησης και η καμπύλη ROC του ταξινομητή LogisticRegression φαίνονται στο Σχήμα 34. Ο ταξινομητής έχει ανάκληση 76,04% και ακρίβεια 56,81%, οπότε και εδώ βελτιώνεται η πρόβλεψη, με αποτέλεσμα στο σύνολο να έχουμε 221 σωστές προβλέψεις. Επίσης, είναι εμφανής η βελτίωση του ταξινομητή στην τιμή AUC, η οποία βελτιώθηκε κατά 10,2%.



Σχήμα 34. Πίνακας Ταξινόμησης και ROC Καμπύλη LogisticRegression

Όπως στην πρώτη, έτσι και στη δεύτερη φάση εφαρμόστηκε για τα καλύτερα μοντέλα η μέθοδος GridSearchCV για την εύρεση τυχόν καλύτερων υπεπαραμέτρων από αυτές που είχαν επιλεγεί, χωρίς το αποτέλεσμα να βελτιωθεί. Το ίδιο συνέβη και με την αφαίρεση των stop words και των αριθμητικών ψηφίων. Ενδεικτικά, παραθέτουμε στον Πίνακα 4 με τα σχετικά αποτελέσματα των ταξινομητών με εκπαίδευση σε κείμενο που αφαιρέθηκαν οι αναφερθείσες λέξεις.

Πίνακας 4. Αποτελέσματα Μοντέλων για ngram (2, 2) και Αφαίρεση Stop Words

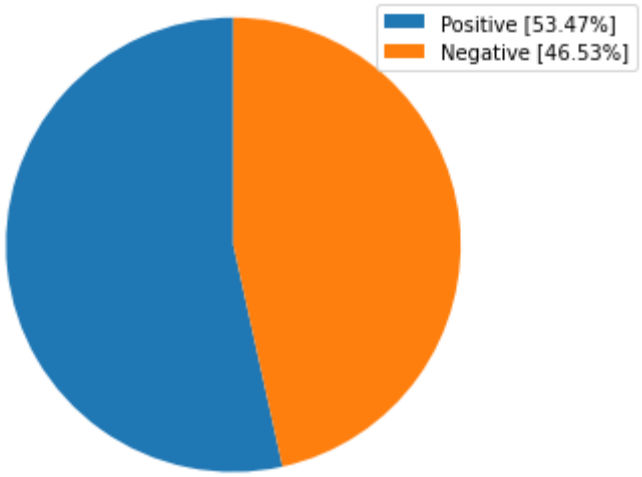
Μοντέλο	Accuracy (%)	F1 Score (%)	ROC AUC (%)
LinearSVC	51,85	56,04	51,71
SVC	54,50	64,46	54,07
LogisticRegression	50,53	57,40	50,29
BernoulliNB	50,00	57,34	49,74
MultinomialNB	51,32	63,20	50,82
RandomForestClassifier	51,59	55,47	51,46

5.2 Αποτελέσματα Υλοποίησης με Μεθόδους μη-Επιβλεπόμενης ΜΜ

Οι πειραματισμοί για την εύρεση του καλύτερου δυνατού προγνωστικού μοντέλου με μεθόδους μη επιβλεπόμενης ΜΜ έγιναν με τη χρήση της VADER και της TextBlob εφαρμόζοντας τις σχετικές λειτουργίες ανάλυσης συναισθημάτων.

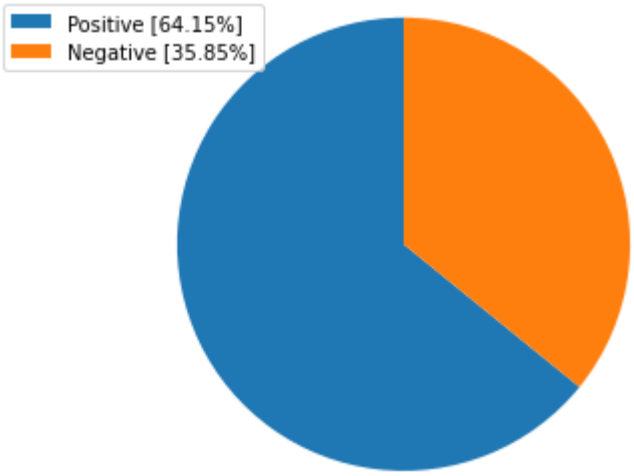
Πριν την αποτίμηση των μεθόδων στο σύνολο δοκιμής, έγινε εφαρμογή των δύο λειτουργιών σε όλο το σύνολο δεδομένων, θεωρώντας ότι εάν ο δείκτης DJIA μένει

σταθερός ή ανεβαίνει το συναίσθημα είναι θετικό, ενώ εάν παρουσιάζει κάθοδο το συναίσθημα είναι αρνητικό. Στο Σχήμα 35 απεικονίζεται η κατανομή των στιγμιοτύπων με αρνητική και θετική ετικέτα στο αρχικό σύνολο, δηλαδή με θετικό και αρνητικό συναίσθημα αντίστοιχα.

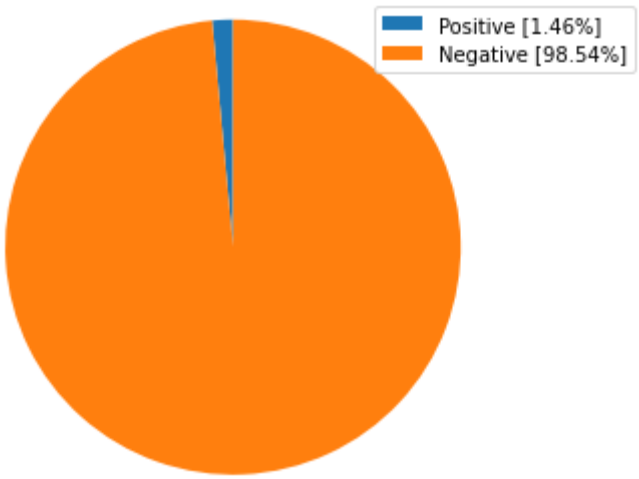


Σχήμα 35.

Στο Σχήμα 36 απεικονίζεται η κατανομή της ανάλυσης συναισθημάτων όπως προέκυψε από την εφαρμογή της TextBlob και στο Σχήμα 37 η κατανομή της ανάλυσης συναισθημάτων όπως προέκυψε από την εφαρμογή της VADER.



Σχήμα 36. Ανάλυση Συναισθήματος TextBlob

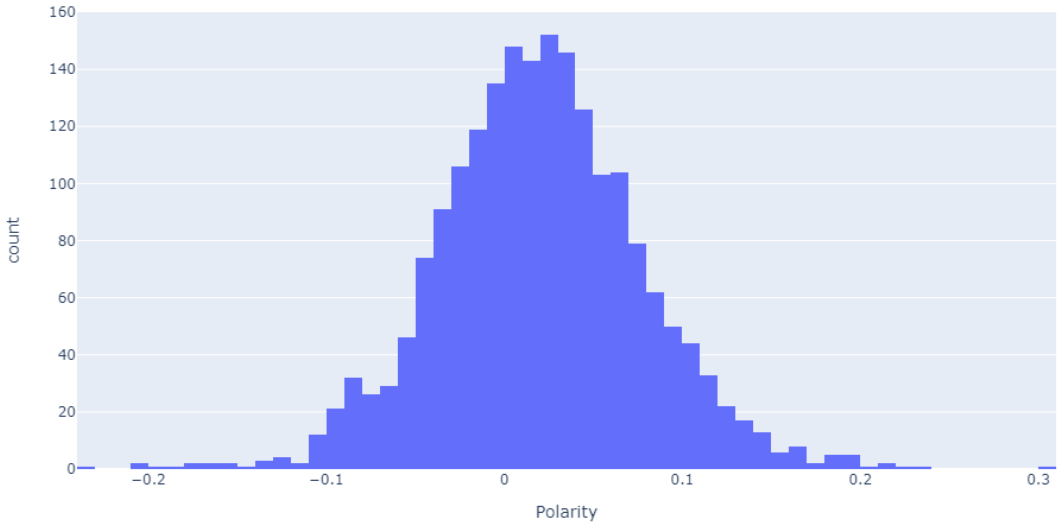


Σχήμα 37. Ανάλυση Συναισθήματος VADER

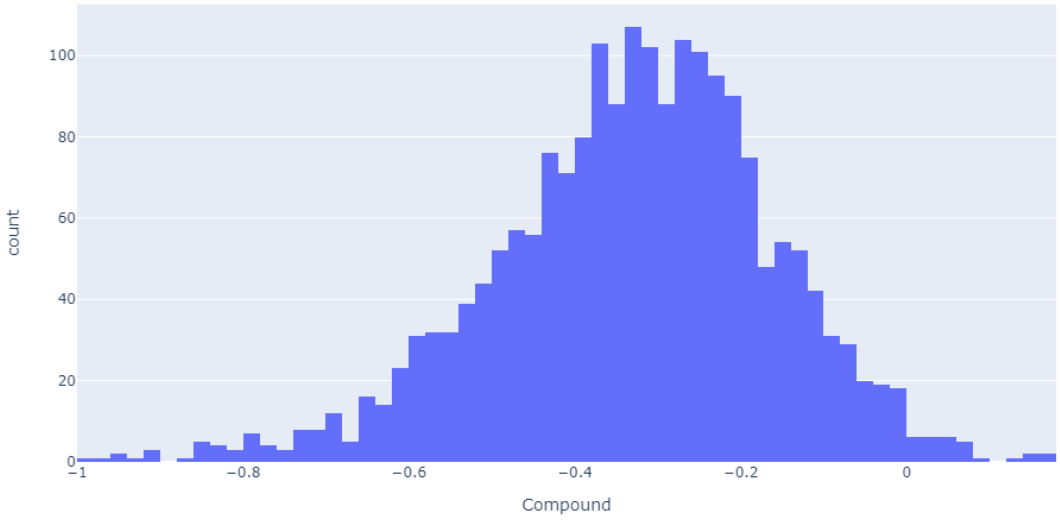
Όπως διαπιστώνουμε, υπάρχει μια πολύ μεγάλη διαφορά στα αποτελέσματα που εξάγονται από τις δύο λειτουργίες επί του συνόλου των δεδομένων. Πιο συγκεκριμένα, στην αποτίμηση της TextBlob υπερισχύουν τα παραδείγματα με θετικό συναίσθημα- όπως στο αρχικό σύνολο δεδομένων, αλλά σε μεγαλύτερη αναλογία (64,15% έναντι 53,47%), ενώ στην αποτίμηση της VADER η συντριπτική πλειοψηφία είναι τα αρνητικά συναισθήματα με τα θετικά συναισθήματα να αναλογούν μόνο στο 1,46%.

Θεωρούμε ότι, αξίζει να δούμε πως κατανέμεται το αποτέλεσμα που εξάγεται από την TextBlob σύμφωνα με τη μετρική της πολικότητας (polarity), όπως απεικονίζεται στο Σχήμα 38 και από την VADER σύμφωνα με τη συνδυαστική (compound) μετρική, όπως απεικονίζεται στο Σχήμα 39. Από το Σχήμα 38 βλέπουμε ότι, η πλειοψηφία των παραδειγμάτων αποτιμάται στο διάστημα -0,01 και 0,04, δηλαδή πολύ κοντά στο ουδέτερο συναίσθημα, μπορεί όμως με βάση το πρόσημο να προκύψει ο διαχωρισμός του θετικού και του αρνητικού συναισθήματος.

Επίσης, στο Σχήμα 40 παρατίθεται συγκριτικά η εξέλιξη των συναισθημάτων ανάλογα με την ημερομηνία για τα αποτελέσματα polarity και compound.



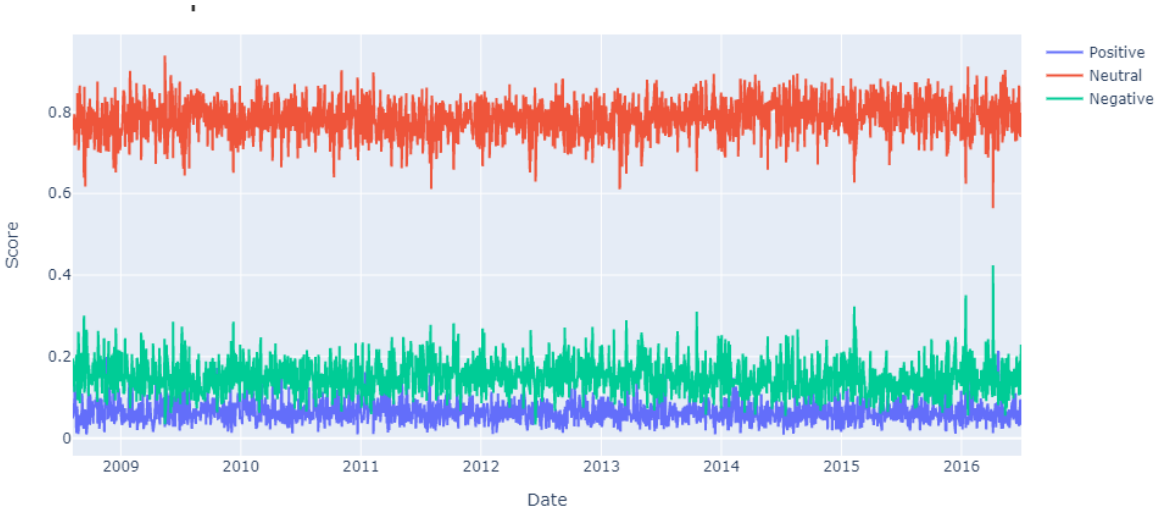
Σχήμα 38. Κατανομή Αποτελέσματος Polarity (TextBlob)



Σχήμα 39. Κατανομή Αποτελέσματος Compound (VADER)

Επιθυμώντας να διερευνήσουμε περισσότερο τα αποτελέσματα της VADER για το αποτέλεσμα compound, εφόσον η VADER μας δίνει τη δυνατότητα να δούμε το σθένος του κάθε συναισθήματος, πραγματοποιήθηκε περαιτέρω ανάλυση στην VADER για το πως εκτιμάται το συναίσθημα ως θετικό, αρνητικό και ουδέτερο και τα αποτελέσματα φαίνονται στο Σχήμα 41.

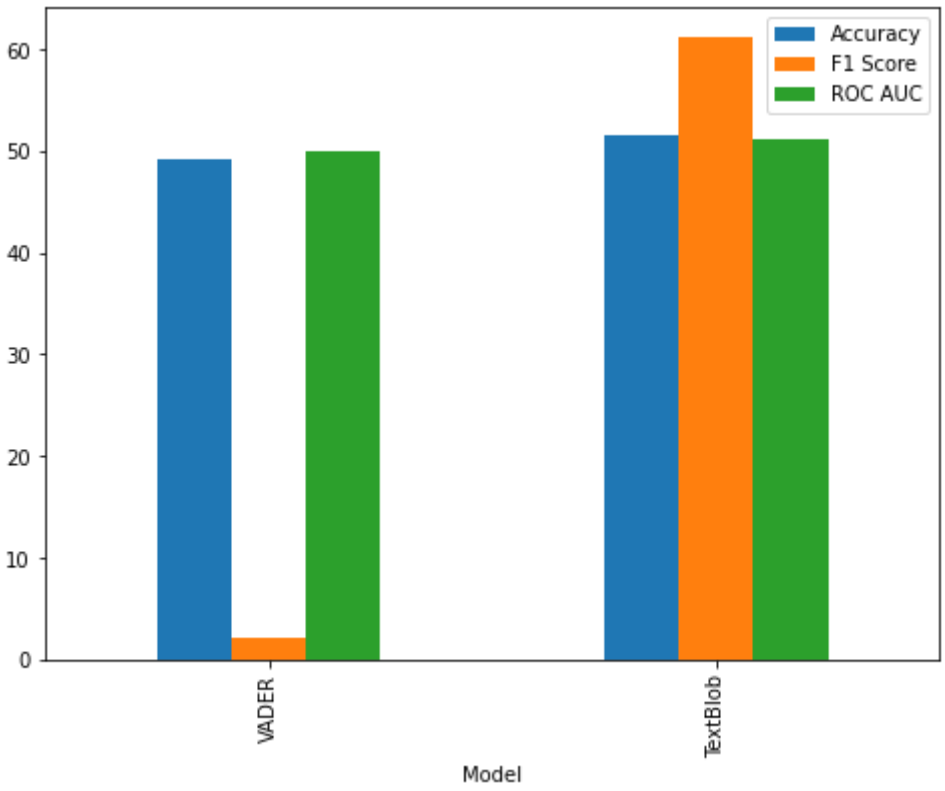
Σχήμα 40. Σύγκριση Εξέλιξης Συναισθημάτων VADER-TextBlob



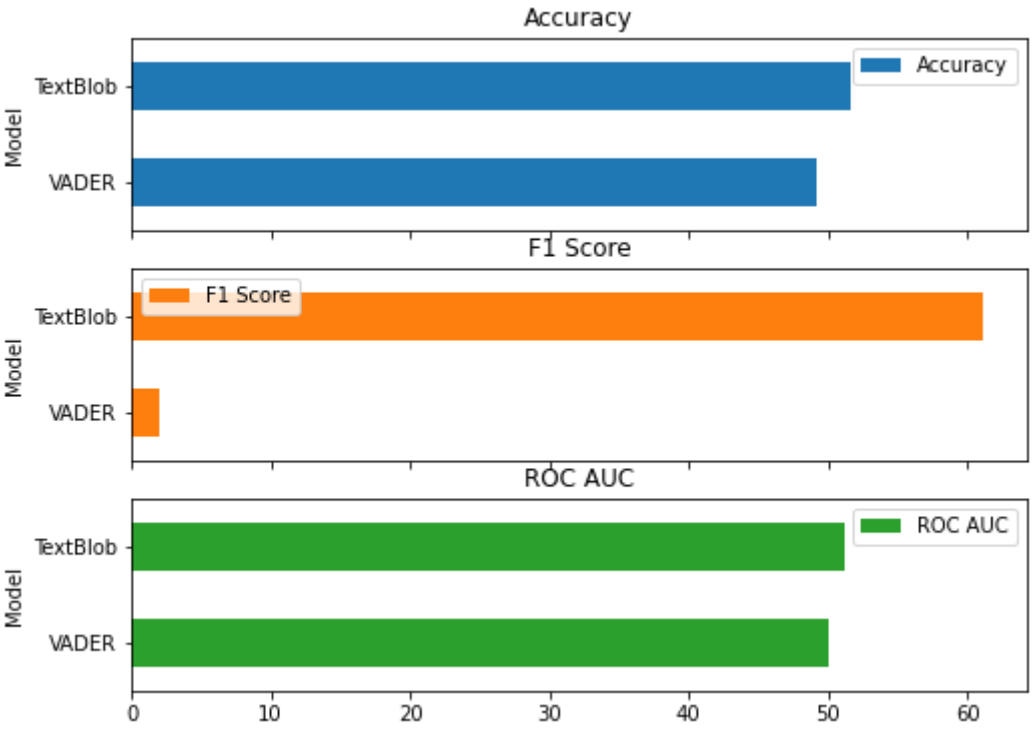
Σχήμα 41. Εξέλιξη Συναισθημάτων VADER

Από την παραπάνω διερεύνηση, γίνεται ξεκάθαρο το γεγονός ότι, η VADER αποτιμά τα συναισθήματα στην πλειοψηφία τους ως ουδέτερα με μεγάλη διαφορά όσον αφορά την ένταση των αρνητικών και των θετικών, και, επίσης, η ένταση των αρνητικών κυμαίνεται σε υψηλότερες τιμές. Αυτό έχει ως αποτέλεσμα- βάσει της μετρικής του compound- η πλειοψηφία των παραδειγμάτων να χαρακτηρίζονται ως αρνητικά.

Ωστόσο, για να συγκρίνουμε τα αποτελέσματα των μεθόδων αυτών με τα αποτελέσματα των μεθόδων επιβλεπόμενης ΜΜ, πρέπει να δούμε την αποτίμηση στο σύνολο δοκιμής. Τα συγκριτικά αποτελέσματα των δύο μεθόδων φαίνονται στο Σχήμα 42 και η σύγκριση των μετρικών τους στο Σχήμα 43. Επίσης, στον Πίνακα 5 παρουσιάζονται τα αριθμητικά αποτελέσματα της αποτίμησης των δύο μοντέλων.



Σχήμα 42. Σύγκριση Μοντέλων VADER-TextBlob



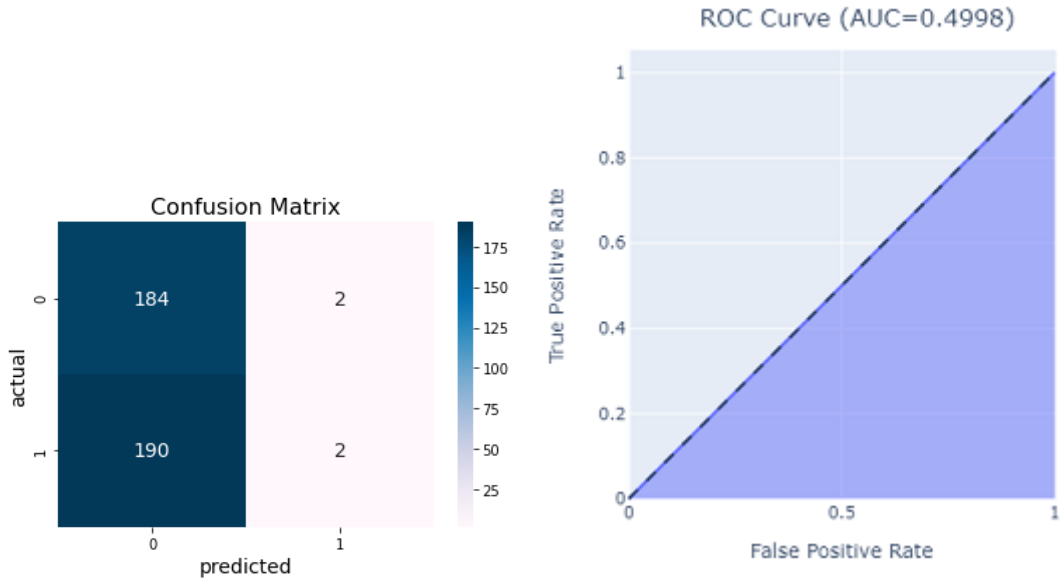
Σχήμα 43. Σύγκριση Μετρικών Μοντέλων VADER-TextBlob

Πίνακας 5. Αποτελέσματα μοντέλων VADER-TextBlob

Μοντέλο	Accuracy (%)	F1 Score (%)	ROC AUC (%)
VADER	49,21	2,04	49,98
TextBlob	51,59	61,15	51,21

Από τα αποτελέσματα διαπιστώνουμε ότι, η VADER αποτυγχάνει στην ταξινόμηση για το συγκεκριμένο σύνολο δεδομένων, ενώ καλύτερα αποτελέσματα δίνει η TextBlob, τα οποία όμως πάλι δεν είναι ικανοποιητικά. Από την EDA διαπιστώθηκε ότι, στο λεξικό που δημιουργείται για το κείμενο κυριαρχούν λέξεις οι οποίες αφορούν χώρες και κύρια ονόματα, οι οποίες προφανώς θεωρούνται ουδέτερες στο λεξικό και επηρεάζουν κατά πολύ το αποτέλεσμα της VADER, καθιστώντας αυτή τη μέθοδο ακατάλληλη για το επιλεγμένο σύνολο δεδομένων. Όσο για τα αποτελέσματα της TextBlob, επειδή χρησιμοποιεί μεθόδους της βιβλιοθήκης NLTK, δίνει αποτελέσματα βασισμένα σε αυτήν και τα οποία είναι ανάλογα με αυτά των μοντέλων που προέκυψαν από τις μεθόδους επιβλεπόμενης ΜΜ.

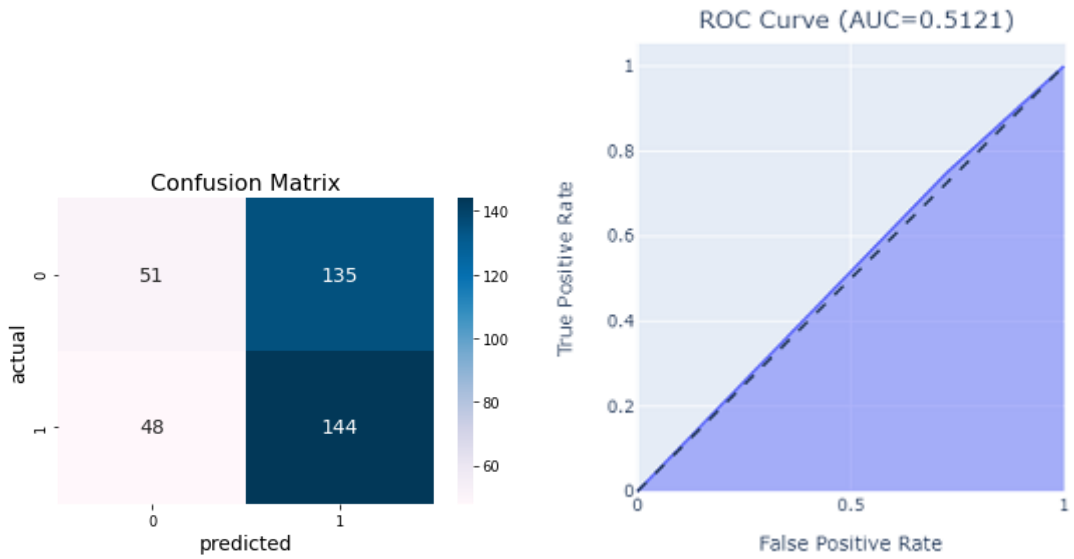
Για λόγους συνέπειας, στα Σχήματα παρουσιάζονται οι πίνακες ταξινόμησης και οι ROC καμπύλες της VADER και της TextBlob αντίστοιχα.



Σχήμα 44. Πίνακας Ταξινόμησης και ROC Καμπύλη VADER

Το μοντέλο της VADER είχε precision 50,00% και recall μόλις 1,04% με αποτέλεσμα το πολύ κακό F1 αποτέλεσμα και μια τιμή AUC τέτοια που δείχνει την τυχαία ταξινόμηση των συναισθημάτων. Αντίστοιχα το μοντέλο της TextBlob είχε precision

51,61% και recall 75,00%, δηλαδή οι επιδόσεις του κινούνται στο ίδιο εύρος τιμών με αυτές των μοντέλων επιβλεπόμενης ΜΜ.



Σχήμα 45. Πίνακας Ταξινόμησης και ROC Καμπύλη TextBlob

Το κέρδος από την εφαρμογή της TextBlob είναι ότι, δεν απαιτεί καμιά προεπεξεργασία δεδομένων (παραμόνο τον καθαρισμό των δεδομένων από τον θόρυβο) και επίσης δεν απαιτείται εκπαίδευση. Συνεπώς, η διαδικασία για τη δημιουργία μοντέλου είναι ταχύτατη.

Τέλος, αξίζει να αναφερθεί ότι, έγινε δοκιμή για τις δύο λειτουργίες για την ανάλυση συναισθήματος σε επίπεδο παραγράφου (και όχι προτάσεων όπως συνιστούν οι δημιουργοί) και τα αποτελέσματα ήταν χειρότερα και για τις δύο λειτουργίες, οπότε δεν κρίθηκε απαραίτητη η παρουσίασή τους.

6 Συμπεράσματα και Μελλοντικές Επεκτάσεις

Το παρόν κεφάλαιο αφορά τα συμπεράσματα από την παρούσα εργασία και τις μελλοντικές επεκτάσεις.

6.1 Συμπεράσματα

Στην παρούσα διπλωματική εργασία διερευνήθηκε το θέμα της εκτίμησης επενδυτικών αποφάσεων που βασίζονται σε τεχνικές ανάλυσης συναισθήματος με τη βοήθεια μεθόδων επιβλεπόμενης και μη επιβλεπόμενης ΜΜ. Για τον σκοπό αυτό, επιλέχθηκε ένα σύνολο δεδομένων το οποίο περιέχει τις κυριότερες διεθνείς ειδήσεις που συλλέχθηκαν από την πλατφόρμα του κοινωνικού μέσου Reddit για ένα διάστημα 8 ετών (2008-2016) από ένα subreddit του Reddit σε συνδυασμό με την τάση της κίνησης του χρηματιστηριακού δείκτη Dow Jones. Η άνοδος και η σταθερότητα του DJIA αντιστοιχούν στη δημιουργία θετικού συναισθήματος (αντίστοιχα ετικέτας με ένδειξη 1), ενώ η κάθοδος στη δημιουργία αρνητικού συναισθήματος (αντίστοιχα ετικέτας με ένδειξη 0).

Αρχικά, παρουσιάστηκε το απαιτούμενο θεωρητικό υπόβαθρο για το χρηματιστήριο και την ΜΜ για επιβλεπόμενη και μη επιβλεπόμενη μάθηση, βασικές τεχνικές και βιβλιοθήκες της Python, όπως και αλγόριθμοι και μετρικές αποτίμησης μοντέλων.
Το θέμα της ανάλυσης συναισθήματος που προέρχεται από τα κοινωνικά μέσα, έχει απασχολήσει την ακαδημαϊκή κοινότητα ιδιαίτερα τα τελευταία χρόνια, συνεπώς κρίθηκε απαραίτητη η αναφορά στη σχετική βιβλιογραφία.

Στη συνέχεια, παρουσιάστηκε η μεθοδολογία για την υλοποίηση του έργου. Πραγματοποιήθηκε εξερεύνηση των δεδομένων μέσω της διαδικασίας EDA (Exploratory Data Analysis) η οποία ήταν εξαιρετικά σημαντική στην εκτίμηση του συνόλου δεδομένων και καθοριστική για την επιλογή των μέτρων απόδοσης, των τεχνικών αποτίμησης και των αλγόριθμων ΜΜ. Πιο συγκεκριμένα, επιλέχθηκαν για τις προβλέψεις μέθοδοι επιβλεπόμενης ΜΜ και μη επιβλεπόμενης ΜΜ.

Με μεθόδους επιβλεπόμενης ΜΜ το πρόβλημα της πρόβλεψης της τάσης του δείκτη ανάγεται σε πρόβλημα δυαδικής ταξινόμησης και για τον σκοπό αυτό επιλέχθηκαν τέσσερις αντιπροσωπευτικοί αλγόριθμοι για τη δημιουργία μοντέλων: οι SVM, η LR, ο NB και ο RF, καθώς και μέθοδοι εξαγωγής χαρακτηριστικών που διαθέτει η βιβλιοθήκη scikit-learn.

Ως μέθοδοι μη επιβλεπόμενης μάθησης επιλέχθηκαν οι λειτουργίες ανάλυσης συναισθήματος που διαθέτουν δύο από τις δημοφιλέστερες βιβλιοθήκες της Python για NLP και ιδιαίτερα επεξεργασία κειμένου: η λειτουργία VADER που διαθέτει η βιβλιοθήκη NLTK, η οποία βασίζεται σε λεξικό και κανόνες και η λειτουργία ανάλυσης συναισθήματος της βιβλιοθήκης TextBlob. Οι λειτουργίες αυτές εξάγουν από το κείμενο το αρνητικό ή το θετικό συναίσθημα, το οποίο δημιουργεί αντίστοιχα την κάθοδο και την άνοδο ή σταθερότητα του δείκτη.

Από τους διάφορους πειραματισμούς και δοκιμές το καλύτερο προγνωστικό μοντέλο δημιουργήθηκε με μεθόδους επιβλεπόμενης ΜΜ με την εξαγωγή χαρακτηριστικών σε διγράμματα (2-grams) και με περιορισμό των χαρακτηριστικών σύμφωνα με την ελάχιστη συχνότητα εμφάνισής τους και στη συνέχεια με την εφαρμογή του αλγόριθμου LR με ορθότητα 58,47%, F1 αποτέλεσμα 65,03% και ROC AUC τιμή 58,18%. Στα ίδια επίπεδα κινήθηκαν και οι επιδόσεις του μοντέλου με εφαρμογή του αλγόριθμου SVM με ορθότητα 57,67% , F1 αποτέλεσμα 66,94% και ROC AUC τιμή 57,24% με τον αλγόριθμο του NB να έπεται σε επιδόσεις.

Όσον αφορά τις μεθόδους μη επιβλεπόμενης μάθησης, η εφαρμογή της VADER για το επιλεγμένο σύνολο δεδομένων αποδείχθηκε τελείως ακατάλληλη, κάτι που ήταν αναμενόμενο λόγω του ότι βασίζεται σε λεξικό και κανόνες. Οι λέξεις που εμφανίζονται στο κείμενο στην πλειοψηφία τους δεν περιλαμβάνονται στο λεξικό και οι υπόλοιπες συχνά εμφανιζόμενες έχουν κατά κύριο αρνητικό σθένος. Αντίθετα, η εφαρμογή της TextBlob είχε καλύτερα αποτελέσματα, αρκετά κοντά με αυτά των εφαρμογών των μεθόδων της επιβλεπόμενης ΜΜ.

Ενδεχόμενα τα αποτελέσματα με την VADER να ήταν καλύτερα, εάν εκμεταλλευόμασταν τη δυνατότητα που δίνει η VADER για επέκταση του λεξικού της και τη δημιουργία νέων κανόνων. Κάτι τέτοιο δεν επιχειρήθηκε γιατί από την EDA ήδη ήταν γνωστή η ύπαρξη πολλών κοινών λέξεων για τις περιπτώσεις όπου ο δείκτης παρουσίαζε άνοδο ή κάθοδο.

Ανατρέχοντας στη σχετική βιβλιογραφία τα αποτελέσματα αυτά για κάποιους ερευνητές θεωρούνται ικανοποιητικά, ενώ για άλλους μη ικανοποιητικά. Θεωρούμε ότι, για το συγκεκριμένο σύνολο δεδομένων, τα αποτελέσματα της παρούσας εργασίας με την εφαρμογή των επιλεγμένων αλγορίθμων είναι ικανοποιητικά. Κρίνοντας όμως τα αριθμητικά αποτελέσματα της αποτίμησης, προφανώς τα αποτελέσματα δεν είναι ικανοποιητικά. Αυτό οφείλεται σε διάφορους λόγους. Ο κυριότερος λόγος είναι η πηγή του κειμένου για το οποίο γίνεται η ανάλυση συναισθήματος. Αφενός, οι ειδήσεις από το

συγκεκριμένο subredit αφορούν διεθνή νέα, ενώ ο DJIA είναι χρηματιστηριακός δείκτης των ΗΠΑ. Αφετέρου, τα νέα δεν αφορούν τον οικονομικό τομέα, αλλά είναι γενικότερου ενδιαφέροντος. Αυτοί οι δύο παράγοντες και μόνο είναι αρκετοί για να αιτιολογήσουν την καταλληλότητα της πηγής των δεδομένων.

6.2 Μελλοντικές επεκτάσεις

Η ανάλυση του συναισθήματος από κείμενα τα οποία προέρχονται από κοινωνικά μέσα με σκοπό την πρόβλεψη των τάσεων των δεικτών του χρηματιστηρίου ή μετοχών με μεθόδους ΜΜ, είναι ένα επιστημονικό πεδίο, το οποίο παρουσιάζει ιδιαίτερο ενδιαφέρον και εξελίσσεται συνεχώς.

Στο μέλλον, επιθυμούμε να επανεξετάσουμε το θέμα χρησιμοποιώντας δεδομένα από άλλες πηγές, είτε από άλλο κοινωνικό μέσο, όπως για παράδειγμα το twitter ή από άλλο subreddit πιο σχετικό με οικονομικές ειδήσεις. Θεωρούμε δε, ότι ιδιαίτερο ενδιαφέρον παρουσιάζει η άντληση των δεδομένων σε πραγματικό χρόνο μέσω των API που διαθέτουν οι ιστότοποι, αντί της χρήσης ενός έτοιμου συνόλου δεδομένων.

Όσον αφορά της μεθόδους μη επιβλεπομενης μάθησης, με δεδομένο ότι, η VADER δίνει τη δυνατότητα επέκτασης του λεξικού και της δημιουργίας κανόνων, μπορούμε να εκμεταλλευτούμε αυτή τη δυνατότητα και να πειραματιστούμε σε κατάλληλα σύνολα δεδομένων.

Επίσης, θεωρούμε ότι, ιδιαίτερο ενδιαφέρον παρουσιάζει η χρήση μεγαλύτερων συνόλων δεδομένων, ώστε να μπορούν να εφαρμοστούν τεχνικές Βαθιάς Μάθησης για την ανάλυση συναισθήματος, οι οποίες γνωρίζουν ιδιαίτερη ανάπτυξη και εξελίσσονται συνεχώς.

Βιβλιογραφία

[1] D. Y. N. Le, A. Maag, and S. Senthilananthan, “Analysing stock market trend prediction using machine deep learning models: A comprehensive review,” *CITISIA 2020 - IEEE Conf. Innov. Technol. Intell. Syst. Ind. Appl. Proc.*, 2020, doi: 10.1109/CITISIA50690.2020.9371852.

[2] S. Singh, T. K. Madan, J. Kumar, and A. K. Singh, “Stock market forecasting using machine learning: Today and tomorrow,” in *2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)*, 2019, vol. 1, pp. 738–745.

[3] F. S. Alzazah and X. Cheng, “Recent Advances in Stock Market Prediction Using Text Mining: A Survey,” *E-Business-Higher Educ. Intell. Appl.*, 2020.

[4] S. Usmani and J. A. Shamsi, “News sensitive stock market prediction: literature review and suggestions,” *PeerJ Comput. Sci.*, vol. 7, p. e490, 2021, doi: 10.7717/peerj-cs.490.

[5] M. B. Sesen, Y. Romahi, and V. Li, “Natural Language Processing of Financial News,” in *Big Data and Machine Learning in Quantitative Investment*, John Wiley & Sons, Ltd, 2018, pp. 185–210.

[6] “GLOBAL SOCIAL MEDIA STATS,” 2021. https://datareportal.com/social-media-users (accessed Jul. 31, 2021).

[7] J. Bollen, H. Mao, and X. Zeng, “Twitter mood predicts the stock market,” *J. Comput. Sci.*, vol. 2, no. 1, pp. 1–8, 2011.

[8] F. G. D. C. Ferreira, A. H. Gandomi, and R. T. N. Cardoso, “Artificial Intelligence Applied to Stock Market Trading: A Review,” *IEEE Access*, vol. 9, pp. 30898–30917, 2021, doi: 10.1109/ACCESS.2021.3058133.

[9] F. Z. Xing, E. Cambria, and R. E. Welsch, “Natural language based financial forecasting: a survey,” *Artif. Intell. Rev.*, vol. 50, no. 1, pp. 49–73, 2018, doi: 10.1007/s10462-017-9588-9.

[10] Wikipedia the free encyclopedia, “Χρηματιστήριο,” 2021. https://el.wikipedia.org/wiki/Χρηματιστήριο (accessed Aug. 28, 2021).

[11] Wikipedia the free encyclopedia, “Stock market index,” 2021. https://en.wikipedia.org/wiki/Stock_market_index (accessed Aug. 28, 2021).

[12] Χρηματιστήριο Αθηνών, “Βασικοί Κανόνες Διαχείρισης & Υπολογισμού των Δεικτών της Αγοράς Μετοχών του ΧΑ.” 2019, [Online]. Available: https://www.athexgroup.gr/documents/10180/5517704/ATHEX+General+Index+Ground+Rules+V2-15+June19+%28F%29.pdf/31838abf-c5bd-468d-847f-2783f5a65196.

[13] S&P Global, “Dow Jones Industrial Average®,” 2021. https://www.spglobal.com/spdji/en/indices/equity/dow-jones-industrial-average/#overview (accessed Aug. 28, 2021).

[14] Wikipedia the free encyclopedia, “Dow Jones Industrial Average,” 2021. https://en.wikipedia.org/wiki/Dow_Jones_Industrial_Average (accessed Aug. 28, 2021).

[15] “Dow Jones Companies,” 2021. https://www.slickcharts.com/dowjones (accessed Jul. 29, 2021).

[16] S. M. Idrees, M. A. Alam, and P. Agarwal, “A Prediction Approach for Stock Market Volatility Based on Time Series Data,” *IEEE Access*, vol. 7, pp. 17287–17298, 2019, doi: 10.1109/ACCESS.2019.2895252.

[17] M. Obthong, N. Tantisantiwong, W. Jeamwatthanachai, and G. Wills, “A survey on machine learning for stock price prediction: Algorithms and techniques,” *FEMIB 2020 - Proc. 2nd Int. Conf. Financ. Econ. Manag. IT Bus.*, pp. 63–71, 2020, doi: 10.5220/0009340700630071.

[18] S. Bouktif, A. Fiaz, and M. Awad, “Augmented Textual Features-Based Stock Market Prediction,” *IEEE Access*, vol. 8, pp. 40269–40282, 2020, doi: 10.1109/ACCESS.2020.2976725.

[19] Α. Γεωργούλη, *Τεχνητή Νοημοσύνη. [ηλεκτρ. βιβλ.]*. Αθήνα: Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών, 2015.

[20] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[21] Javatpoint, “Machine Learning Tutorial.” 2021, [Online]. Available: https://www.javatpoint.com/machine-learning.

[22] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, “Text classification algorithms: A survey,” *Inf.*, vol. 10, no. 4, pp. 1–68, 2019, doi: 10.3390/info10040150.

[23] “Deep learning vs. machine learning in Azure Machine Learning,” 2021. https://docs.microsoft.com/en-us/azure/machine-learning/concept-deep-learning-vs-machine-learning (accessed Aug. 28, 2021).

[24] A. Géron, *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O’Reilly Media, 2019.

[25] B. Liu, “Sentiment Analysis and Subjectivity,” 2010.

[26] C. Hutto and E. Gilbert, “Vader: A parsimonious rule-based model for sentiment analysis of social media text,” in *Proceedings of the International AAAI Conference on Web and Social Media*, 2014, vol. 8, no. 1.

[27] S. Loria, “textblob Documentation,” *Release 0.15*, vol. 2, 2018.

[28] A. Amidi and S. Amidi, “Supervised Learning cheatsheet,” 2019. https://stanford.edu/~shervine/teaching/cs-229/cheatsheet-supervised-learning# (accessed Apr. 25, 2021).

[29] D. R. Cox, “The regression analysis of binary sequences,” *J. R. Stat. Soc. Ser. B*, vol. 20, no. 2, pp. 215–232, 1958.

[30] C. Cortes and V. Vapnik, “Support-vector networks,” *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995, doi: 10.1007/BF00994018.

[31] Ε. Κύρκος, *Επιχειρηματική ευφυΐα και εξόρυξη δεδομένων*, [Ηλεκτρ. β. Αθήνα:Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών, 2015.

[32] H. Zhang, “The optimality of naive Bayes,” 2004.

[33] “Naive Bayes,” 2021. https://scikit-learn.org/stable/modules/naive_bayes.html (accessed Jul. 31, 2021).

[34] L. Breiman, “Random Forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi:

10.1023/A:1010933404324.

[35] J. R. Quinlan, “Induction of decision trees,” *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986, doi: 10.1007/BF00116251.

[36] R. Quinlan, “4.5: Programs for machine learning morgan kaufmann publishers inc,” *San Fr. USA*, 1993.

[37] Wikipedia the free encyclopedia, “Decision tree learning,” 2021. https://en.wikipedia.org/wiki/Decision_tree_learning (accessed Jul. 30, 2021).

[38] L. Breiman, “Bagging predictors,” *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996, doi: 10.1007/BF00058655.

[39] E. Bauer and R. Kohavi, “An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants,” *Mach. Learn.*, vol. 36, no. 1, pp. 105–139, 1999, doi: 10.1023/A:1007515423169.

[40] W.-M. Lee, *Python machine learning*. John Wiley \& Sons, 2019.

[41] A. Burkov, *The Hundred-Page Machine Learning Book*. Andriy Burkov (January 12, 2019), 2019.

[42] J. Davis and M. Goadrich, *The Relationship Between Precision-Recall and ROC Curves*, vol. 06. 2006.

[43] Wikipedia the free encyclopedia, “Matthews correlation coefficient.” 2021, [Online]. Available: https://en.wikipedia.org/wiki/Matthews_correlation_coefficient.

[44] X. Man, T. Luo, and J. Lin, “Financial sentiment analysis(FSA): A survey,” *Proc. - 2019 IEEE Int. Conf. Ind. Cyber Phys. Syst. ICPS 2019*, pp. 617–622, 2019, doi: 10.1109/ICPHYS.2019.8780312.

[45] A. Gupta, V. Dengre, H. A. Kheruwala, and M. Shah, “Comprehensive review of text-mining applications in finance,” *Financ. Innov.*, vol. 6, no. 1, 2020, doi: 10.1186/s40854-020-00205-1.

[46] A. Agarwal, “Sentiment Analysis of Financial News,” *Proc. - 2020 12th Int. Conf. Comput. Intell. Commun. Networks, CICN 2020*, pp. 312–315, 2020, doi: 10.1109/CICN49253.2020.9242579.

[47] H. Bourezk, A. Raji, N. Acha, and H. Barka, “Analyzing Moroccan Stock Market using Machine Learning and Sentiment Analysis,” *2020 1st Int. Conf. Innov. Res. Appl. Sci. Eng. Technol. IRASET 2020*, pp. 24–28, 2020, doi: 10.1109/IRASET48871.2020.9092304.

[48] S. Kalra and J. S. Prasad, “Efficacy of News Sentiment for Stock Market Prediction,” *Proc. Int. Conf. Mach. Learn. Big Data, Cloud Parallel Comput. Trends, Prespectives Prospect. Com. 2019*, pp. 491–496, 2019, doi: 10.1109/COMITCon.2019.8862265.

[49] M. Velay and F. Daniel, “Using NLP on news headlines to predict index trends,” *arXiv Prepr. arXiv1806.09533*, 2018.

[50] M. Vicari and M. Gaspari, “Analysis of news sentiments using natural language processing and deep learning,” *AI Soc.*, no. 0123456789, 2020, doi: 10.1007/s00146-020-01111-x.

[51] Y. Liu, J. Trajkovic, H. G. H. Yeh, and W. Zhang, “Machine Learning for Predicting Stock Market Movement using News Headlines,” *2020 IEEE Green Energy Smart Syst. Conf.*

IGESSC 2020, 2020, doi: 10.1109/IGESSC50231.2020.9285163.

[52] A. Hassanzadeh Kalshani, A. Razavi, and R. Asadi, “Stock Market Prediction using Daily News Headlines,” *SSRN Electron. J.*, no. December, 2020, doi: 10.2139/ssrn.3685530.

[53] N. Pappas, G. Katsimpras, and E. Stamatatos, “Distinguishing the popularity between topics: a system for up-to-date opinion retrieval and mining in the web,” in *International conference on intelligent text processing and computational linguistics*, 2013, pp. 197–209.

[54] P. Virtanen *et al.*, “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python,” *Nat. Methods*, vol. 17, pp. 261–272, 2020, doi: 10.1038/s41592-019-0686-2.

[55] “Plotly Python Open Source Graphing Library,” 2021. https://plotly.com/python/ (accessed Jul. 30, 2021).

[56] A. Mueller, “WordCloud for Python documentation.” 2021, [Online]. Available: https://amueller.github.io/word_cloud/.

[57] F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python,” *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[58] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. O’Reilly Media, Inc., 2009.

[59] S. Loria *et al.*, “Textblob: simplified text processing,” *Second. TextBlob Simpl. text Process.*, vol. 3, 2014.

[60] T. De Smedt and W. Daelemans, “Pattern for Python,” *J. Mach. Learn. Res.*, vol. 13, no. 66, pp. 2063–2067, 2012, [Online]. Available: http://jmlr.org/papers/v13/desmedt12a.html.

[61] S. Loria, “TextBlob: Simplified Text Processing,” 2021. https://textblob.readthedocs.io/en/dev/ (accessed Aug. 10, 2021).

[62] “Text feature extraction,” 2021. https://scikit-learn.org/stable/modules/feature_extraction.html (accessed Aug. 05, 2021).

[63] C. Lorentzen, “scikit-learn: machine learning in Python.” 2021, [Online]. Available: https://github.com/scikit-learn/scikit-learn.

[64] C. Hutto, “VADER-Sentiment-Analysis.” 2021, [Online]. Available: https://github.com/cjhutto/vaderSentiment.

[65] “Support Vector Machines,” 2021. https://scikit-learn.org/stable/modules/svm.html (accessed Jul. 31, 2021).

[66] “Logistic regression,” 2021. https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression (accessed Jul. 31, 2021).

[67] “Random Forests,” 2021. https://scikit-learn.org/stable/modules/ensemble.html (accessed Jul. 31, 2021).

[68] “Tuning the hyper-parameters of an estimator,” 2021. https://scikit-learn.org/stable/modules/grid_search.html (accessed Aug. 10, 2021).