Πανεπιστήμιο Δυτικής Μακεδονίας

Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Διπλωματική Εργασία

Δημιουργία υπηρεσίας για την εφαρμογή μεθόδων ανάλυσης συναισθήματος στον τραπεζικό τομέα

Μαραγκός Γεώργιος

Επιβλέπων: Σαρηγιαννίδης Παναγιώτης, Αναπληρωτής Καθηγητής Π.Δ.Μ.

Σεπτέμβριος 2022, Κοζάνη

University Of Western Macedonia

Department of Electrical and Computer Engineering

Bachelor Thesis

Service creation for the application of sentiment analysis methods in the banking sector

Maragkos Georgios

Supervisor: Sarigiannidis Panagiotis, Associate Professor at U.O.W.M

September 2022, Kozani

Περίληψη

Ο τομέας της ανάλυσης συναισθήματος (sentiment analysis) είναι ένα πεδίο που γνωρίζει ραγδαία αύξηση τα τελευταία χρόνια. Η ανάλυση συναισθήματος αποτελεί ισχυρό εργαλείο στον τομέα της διαφήμισης και του μάρκετινγκ, αφού μέσω αυτού οι εταιρείες του χώρου είναι σε θέση να πληροφορηθούν για την γνώμη του κόσμου σε ποικίλους τομείς. Πέρα από τον τομέα την διαφήμισης η ανάλυση συναισθήματος αποτελεί σημαντικό παράγοντα στην διευκόλυνση εξυπηρέτησης πελατών, αφού κατανοεί μέσω του τόνου αλλά και του ύφους που διαθέτουν τις ανάγκες τους. Η παρούσα διπλωματική εργασία σχετίζεται με την εφαρμογή μεθόδων ανάλυσης συναισθήματος στον τομέα της οικονομίας, με χρήση άρθρων οικονομικού χαρακτήρα. Για την αντιμετώπιση του προβλήματος εφαρμόστηκαν αλγόριθμοι μηχανικής μάθησης όπως τα δένδρα αποφάσεων (decision tree - DT), ο αφελής ταξινομητής Bayes (gaussian naïve classifier - GND), οι μηχανές διανυσμάτων υποστήριξης (support vector machines), γραμμική διακριτική ανάλυση (linear discriminating analysis – LDA), η στοχαστική κλίση καθόδου (stochastic gradient descent – SGD), ο αλγόριθμος k-κοντινότερων γειτόνων (K-nearest neighbors - KNN), αλλά και μοντέλα νευρωνικών δικτύων (Artificial Neural Networks - ANN) . Ωστόσο, πριν την εφαρμογή των παραπάνω αλγορίθμων, τα δεδομένα υποβλήθηκαν σε προ-επεξεργασία , ενώ ταυτόχρονα μετατράπηκαν και σε μορφή κατάλληλη ώστε να είναι κατανοητή από αυτούς. Στην παρούσα διπλωματική εργασία, περιλαμβάνονται διαδικασίες όπως ο καθαρισμός των δεδομένων από λέξεις με ελάσσονα σημασία (stop words), η αφαίρεση των σημείων στίξης (data cleansing), η προσαρμογή όλων των λέξεων σε πεζά γράμματα καθώς και η διαδικασία αναγωγής μιας λέξης στο πρωταρχική της ρίζα. Παράλληλα, η διαδικασία μετατροπής των δεδομένων ονομάζεται εξαγωγή χαρακτηριστικών (feature extraction) ενώ έχουν αναπτυχθεί αρκετές μέθοδοι οι οποίες μπορούν να εφαρμοστούν για την υλοποίηση αυτής. Πιο συγκεκριμένα, εφαρμόζεται η μέθοδος της μέτρησης διανύσματος (count vectorizer) και του term frequency - inverse document frequency (TF-IDF) που εξάγουν χαρακτηριστικά (features) από τα δεδομένα τα οποία καθίστανται έτοιμα προς χρήση μέσω των παραπάνω αλγορίθμων. Όσον αφορά τα αποτελέσματα, η μεθοδολογία που ακολουθείτε επιβεβαιώνει την αξιοπιστία των μοντέλων μηχανικής μάθησης για την εκμαίευση του συναισθηματικού τόνου από κείμενα με θεματολογία σχετική με τον τομέα των οικονομικών. Στο τέλος της παρούσας διπλωματικής εργασίας παρουσιάζονται σχετικά διαγράμματα με την ανάλυση συναισθήματος που έχουν προκύψει από κείμενα οικονομικού χαρακτήρα σε σχέση με τομείς όπως είναι η αγορά μετοχών (stock market), κρυπτό νομίσματα (crypto market) καθώς επίσης και το παγκόσμιο εμπόριο (commodities & futures) . Από τα διαγράμματα παρατηρείται η επιτυχία της διαδικασίας αφού σε ημερομηνίες παγκόσμιας ύφεσης όπως η περίοδος της πανδημίας covid-19 φαίνεται ραγδαία μείωση του παραγόμενου δείκτη συναισθήματος.

Λέξεις Κλειδιά: << Μηχανική Μάθηση, Τεχνητή Νοημοσύνη, Βαθιά Μάθηση, Count Vectorizer, TF-IDF, Ανάλυση Συναισθήματος, Εξαγωγή χαρακτηριστικών>>

Abstract

The field of sentiment analysis is a field that has seen rapid growth in recent years. Sentiment analysis offers an immensely powerful tool in advertising and marketing since, through it, the companies in the field can find out about the world's opinion in various fields. Beyond the field of advertising, sentiment analysis is a crucial factor in customer support, which discerns human language to understand customer requests based on emotion, tone, etc. This thesis is about the application of sentiment analysis methods in the field of economy using economic data from texts. To deal with the problem, various machine learning algorithms were applied such as decision trees (Decision Tree -DT), gaussian naïve classifier (GNV), support vector machines (SVM), linear discriminant analysis (LDA), stochastic gradient descent (SGD), the k-nearest neighbors (KNN) algorithm, but also neural network models. It is essential to mention that pre-processing of the data is needed to successfully use sentiment analysis machine learning models. This thesis includes procedures such as cleaning the data from words with minor meaning (stop words), removing punctuation marks (data cleansing), adjusting all words to lowercase letters as well as the process of reducing a word to its primary root. In addition, it is also necessary to transform the data into a form that machine learning algorithms can understand. The process that solves this specific problem is called feature extraction, while several methods have been developed can be applied. In this thesis, the count vectorizer and TF-IDF methods are applied, where features are extracted in a unique way from the data, which are made ready for use through the above algorithms. In terms of results, the applied method confirms the reliability of machine learning models for extracting emotional tone from texts with topics related to the field of finance. At the end of this thesis, relevant diagrams are presented with the analysis of sentiment that have appeared from economic texts about sectors such as the stock market, cryptocurrencies (crypto market) and the global trade (commodities & futures). The success of the process can be seen from the charts since on dates of global recession such as the period of the covid-19 pandemic we have a rapid decrease in the sentiment index produced.

Keywords: << Machine Learning, Artificial Intelligence, Deep Learning, Count Vectorizer, TF-IDF, Sentiment Analysis, Feature Extraction >>

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Δήλωση Πνευματικών Δικαιωμάτων

Δηλώνω ρητά ότι, σύμφωνα με το άρθρο 8 του Ν. 1599/1986 και τα άρθρα 2,4,6 παρ. 3 του Ν. 1256/1982, η παρούσα Διπλωματική Εργασία με τίτλο “Δημιουργία υπηρεσίας για την εφαρμογή μεθόδων ανάλυσης συναισθήματος στον τραπεζικό τομέα”

καθώς και τα ηλεκτρονικά αρχεία και πηγαίοι κώδικες που αναπτύχθηκαν ή τροποποιήθηκαν στα πλαίσια αυτής της εργασίας και αναφέρονται ρητώς μέσα στο κείμενο που συνοδεύουν, και η οποία έχει εκπονηθεί στο Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Πανεπιστημίου Δυτικής Μακεδονίας, υπό την επίβλεψη του μέλους του Τμήματος κ. Σαρηγιαννίδη Παναγιώτη αποτελεί αποκλειστικά προϊόν προσωπικής εργασίας και δεν προσβάλλει κάθε μορφής πνευματικά δικαιώματα τρίτων και δεν είναι προϊόν μερικής ή ολικής αντιγραφής, οι πηγές δε που χρησιμοποιήθηκαν περιορίζονται στις βιβλιογραφικές αναφορές και μόνον. Τα σημεία όπου έχω χρησιμοποιήσει ιδέες, κείμενο, αρχεία ή / και πηγές άλλων συγγραφέων, αναφέρονται ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή.

Υπογραφή Φοιτητή:

1 - Εισαγωγή

1.1 Κίνητρα συγγραφής της εργασίας

Καθημερινά πραγματοποιούνται εκατοντάδες αναρτήσεις οικονομικών άρθρων στο διαδίκτυο από δημοσιογράφους εφημερίδων. Ωστόσο, ένα σημαντικό ερώτημα παραμένει η εξαγωγή χρήσιμων συμπερασμάτων ,σε σχέση με την άφθονη πληροφορία που παρέχεται από ποικίλες πηγές μέσω του διαδικτύου. Ακολουθώντας την υπόθεση ότι τα άρθρα αυτά αντιπροσωπεύουν την γενική άποψη του κοινού, πληθώρα κειμένων μπορούν να χρησιμοποιηθούν με στόχο την αποτύπωση συναισθημάτων σε διάφορες χρονικές περιόδους μέσω της διαδικασίας της ανάλυσης συναισθήματος. Μέσω αυτής καθορίζεται το ύφος του κειμένου οπού μπορεί να είναι θετικό, αρνητικό ή ουδέτερο. Η διαδικασία της ανάλυσης συναισθήματος προβλέπει την χρήση ποικίλων μεθόδων και τεχνικών που στοχεύουν στην εξαγωγή χρήσιμων συμπερασμάτων.

Οι αλγόριθμοι μηχανικής μάθησης, οι τεχνικές προ- επεξεργασίας και οι τεχνικές εξαγωγής χαρακτηριστικών είναι κάποιες από τις λειτουργίες που χρησιμοποιούνται. Όλα τα βήματα είναι άρρηκτα συνδεδεμένα μεταξύ τους ώστε τελικά να προκύψει ένα επιθυμητό αποτέλεσμα τόσο στο επίπεδο των μετρικών όσο και στο επίπεδο των γραφικών παραστάσεων. Όπως, θα παρατηρήσουμε και παρακάτω ανιχνεύονται επιτυχώς γεγονότα κάθε συναισθηματικής πολικότητας τόσο στην αγορά μετοχών και των κρυπτό νομισμάτων αλλά και στο εμπόριο που παρουσιάζεται μέσω του δείκτη συναισθήματος στις γραφικές παραστάσεις. Ωστόσο, τεχνικές δυσκολίες στην υλοποίηση του συστήματος δεν λείπουν. Πιο συγκεκριμένα, η μειωμένη ακρίβεια και απόδοση που βασίζεται σε ελλιπή χαρακτηρισμένα (labeled) δεδομένα, η αδυναμία αντιμετώπισης σύνθετων προτάσεων, καθώς και αδυναμία αποδοτικότητας σε διαφορετικούς τομείς συνθέτουν κάποιες από αυτές.

1.2 Αντικείμενο διπλωματικής

Συμπληρώνοντας την προηγουμένη ενότητα, εξετάζονται τα προβλήματα της διπλωματικής εργασίας παραθέτοντας λύσεις επί αυτών. Αρχικά, παρατηρείται μειωμένη απόδοση στα αποτελέσματα των δεικτών ανάλυσης συναισθήματος εξ αίτιας των ανεπαρκώς χαρακτηρισμένων δεδομένων σε ένα σύνολο δεδομένων (dataset). Η χρήση άρθρων οικονομικού χαρακτήρα, καθώς και λεξιλογίου με οικονομικούς όρους αποτελούν μια λύση για το παραπάνω πρόβλημα. Παράλληλα, η αδυναμία αντιμετώπισης σύνθετων εννοιών και προτάσεων αποτελεί μια ακόμη δυσκολία. Συγκεκριμένα, η εφαρμογή τεχνικών εξαγωγής χαρακτηριστικών όπως είναι η μέτρηση διανύσματος και η TF-IDF συμβάλλουν στην επίλυση της αδυναμίας αυτής. Οι παραπάνω τεχνικές αναφέρονται στην διαδικασία της μετατροπής δεδομένων σε αριθμητικά χαρακτηριστικά, που μπορούν να υποστούν επεξεργασία, διατηρώντας παράλληλα τις πληροφορίες στο αρχικό σύνολο δεδομένων τους. Ακόμα η διαδικασία της ανάλυσης συναισθήματος δεν παρουσιάζει την ίδια αποτελεσματικότητα σε

όλους τους τομείς. Επομένως κρίνεται σημαντική η δημιουργία ειδικών λεξιλογίων. Στην παρούσα διπλωματική εργασία, παρουσιάζονται οι δείκτες συναισθήματος της αγοράς που προκύπτουν από την επεξεργασία άρθρων οικονομικού χαρακτήρα. Οι τιμές των δεικτών αυτών δικαιολογούνται λόγω της ποικιλίας συναισθημάτων των άρθρων, συνθέτοντας τελικά τις γραφικές παραστάσεις. Στις προαναφερόμενες ,τα άρθρα χωρίζονται μηνιαίως με στόχο την λεπτομερή προσέγγιση των δεικτών συναισθήματος. Ο συσχετισμός των άλλοτε αυξημένων και άλλοτε μειωμένων δεικτών αφορούν γεγονότα αρνητικής ή θετικής φύσεως .

1.3 Συνεισφορά

Η ανάλυση συναισθήματος αποτελεί ~~ένα~~ σημαντικό εργαλείο για την παρακολούθηση και την κατανόηση συναισθημάτων. Συγκεκριμένα, στον οικονομικό τομέα συμβάλλει θετικά ως προς την αντίληψη της διακύμανσης των οικονομικών δεικτών της αγοράς. Το γεγονός αυτό πραγματοποιείται μέσω δύο μεθόδων, είτε μέσω ειδικά διαμορφωμένων λεξικών για την πρόβλεψη συναισθήματος, είτε μέσω αλγορίθμων μηχανικής μάθησης. Στην παρούσα διπλωματική εργασία εφαρμόστηκαν 7 αλγόριθμοι μηχανικής μάθησης. Έπειτα από αξιολόγηση των επιδόσεων τους δημιουργήθηκαν οι αντίστοιχες γραφικές παραστάσεις (συναισθηματικοί δείκτες). Σε αυτές απεικονίζεται η απόδοση των τριών πιο αποτελεσματικών αλγορίθμων στον οικονομικό τομέα. Συμπερασματικά, σκοπός της διπλωματικής εργασίας είναι η πρόβλεψη συναισθήματος οικονομικών άρθρων μέσω αλγορίθμων μηχανικής μάθησης.

1.4 Οργάνωση κειμένου

Η συγκεκριμένη εργασία αποτελείται από 8 κεφάλαια . Στο δεύτερο κεφάλαιο, περιγράφονται βασικοί ορισμοί που αφορούν την ανάλυση συναισθήματος καθώς και την συσχέτιση με άλλες εργασίες του χώρου παρουσιάζοντας τις ομοιότητες και τις διαφορές με την συγκεκριμένη εργασία. Το τρίτο κεφάλαιο, αποτελείται από την συλλογή και την προ επεξεργασία δεδομένων που πραγματοποιείται με σκοπό την εξαγωγή χαρακτηριστικών. Το τέταρτο κεφάλαιο αφορά τις μεθόδους που χρησιμοποιήθηκαν για την ανάλυση των δεδομένων δηλαδή τους αλγορίθμους μηχανικής μάθησης, καθώς και την διαδικασία εξαγωγής χαρακτηριστικών (feature extraction) . Το πέμπτο κεφάλαιο πραγματεύεται την διαδικασία που ακολουθείται ώστε να προκύψει ο δείκτης συναισθήματος για τα νέα δεδομένα. Έπειτα, στο έκτο κεφάλαιο συνοψίζονται τα αποτελέσματα της διπλωματικής εργασίας και περιγράφονται με σαφήνεια τα συμπεράσματα που προέκυψαν. Τα τελευταία δυο κεφάλαια αφορούν τις μελλοντικές επεκτάσεις και την βιβλιογραφία .

2 – Θεωρητικό υπόβαθρο

Η ανάλυση συναισθήματος αποτελεί ένα παράρτημα της μηχανικής μάθησης. Σε συνδυασμό με την επεξεργασία φυσικής γλώσσας [1] συντίθενται τα δομικά στοιχεία της διαδικασίας ανάλυσης συναισθήματος. Η μηχανική μάθηση αποτελεί ένα τομέα της τεχνητής νοημοσύνης και ένα αναπόσπαστο κομμάτι της ανάλυσης συναισθήματος αφού αυτόματα εντοπίζεται και ταξινομείται η συναισθηματική πολικότητα ενός κειμένου [2]. Ειδικότερα τεχνητή νοημοσύνη είναι εκείνος ο κλάδος της επιστήμης των υπολογιστών που ασχολείται με το σχεδιασμό ευφυών υπολογιστικών συστημάτων, δηλαδή συστημάτων με χαρακτηριστικά τα οποία σχετίζονται με την ευφυΐα στην ανθρώπινη συμπεριφορά (μάθηση, αιτίαση, επίλυση προβλημάτων, κατανόηση φυσικής γλώσσας, αναγνώριση αντικειμένων κτλ.) [3]. Επίσης, η μηχανική μάθηση είναι ένα πεδίο έρευνας αφιερωμένο στην κατανόηση και τη δημιουργία μεθόδων που αξιοποιούν δεδομένα για τη βελτίωση της απόδοσης σε κάποιο σύνολο εργασιών. Παραδείγματα της επιστήμης αυτής είναι η διαδικασία του spam filtering ,οπτική αναγνώριση χαρακτήρων, μηχανές αναζήτησης κλπ. Σε συνδυασμό με την μηχανική μάθηση η επεξεργασία φυσικής γλώσσας (natural language processing - NLP) αποτελεί ένα υπό πεδίο της γλωσσολογίας, της επιστήμης των υπολογιστών και της τεχνητής νοημοσύνης που ασχολείται με τις αλληλεπιδράσεις μεταξύ των υπολογιστών και της ανθρώπινης γλώσσας. Στόχος της μεθόδου είναι η κατανόηση των γλωσσικών περιεχομένων των εγγράφων από ένα υπολογιστικό σύστημα. Εξίσου σημαντική κρίνεται η χρήση των δεδομένων εκπαίδευσης (training set) και των δεδομένων δοκιμών (test set) δυο εννοιών άρρηκτα συνδεδεμένων με την μηχανική μάθηση. Οι δυο έννοιες αυτές συνεισφέρουν από κοινού, τόσο στην εκπαίδευση και αξιολόγηση των μοντέλων, όσο και στην πρόβλεψη νέων δεδομένων. Τα δεδομένα εκπαίδευσης χρησιμοποιούνται για την πρόβλεψη ή την ταξινόμηση τιμών γνωστών στο σύνολο εκπαίδευσης. Τα δεδομένα δοκιμών αποτελούν ένα δευτερεύον σύνολο δεδομένων που χρησιμοποιούνται για την δοκιμή μοντέλων μηχανικής μάθησης σε νέες εισόδους για νέα δεδομένα. Παρακάτω αναλύονται σχετικές εργασίες στις οποίες επισημαίνονται ομοιότητες ή διαφορές με την προκείμενη διπλωματική εργασία.

2.1 A Holistic lexicon-based approach to opinion mining

Στο άρθρο το οποίο εξετάζεται οι διαφορές είναι σαφής αφού η χρήση της παραγωγής συναισθήματος πραγματοποιείται αποκλειστικά και μόνο με την χρήση λεξικού. Οι Xiaowen Ding, Bing Liu και Philip S. Yu [4] αναφέρονται αναλυτικά στη διαδικασία της ανάλυσης συναισθήματος μέσω lexicons, ωστόσο τονίζεται ο προβληματισμός που αφορά το προσδιορισμό των σημασιολογικών προσανατολισμών (θετικών, αρνητικών ή ουδέτερων) των απόψεων που εκφράζονται με βάση τις κριτικές διάφορων προϊόντων. Λύση αποτελεί η αξιοποίηση των φυσικών γλωσσικών εκφράσεων, ιδιωμάτων που χρησιμοποιούνται στην ανθρώπινη γλώσσα. Όλες αυτές οι λειτουργίες συντίθεται σε ένα σύστημα που ονομάζεται παρατηρητής γνώμης (opinion observer). Επομένως, με βάση την συγκριτική αξιολόγηση που παρουσιάζεται το σύστημα opinion observer κρίνεται αρκετά αποτελεσματικό.

2.2 Evaluation of sentiment analysis in finance: from lexicons to transformers

Στο άρθρο αυτό οι Konstadin Mishev, Ana Gjorgjevikj, Irena Vodenska [5] κλπ εξετάζουν την ανάλυση συναισθήματος στον τομέα της οικονομίας κάνοντας χρήση πολυπληθών τεχνικών και μεθόδων προσεγγίζοντας το αντικείμενο της ανάλυσης συναισθήματος πολύπλευρα. Στην ανάλυση συναισθήματος στον τομέα των οικονομικών ωστόσο παρουσιάζεται μια πρόκληση η οποία αφορά τις γλωσσικές εκφράσεις που χρησιμοποιούνται στον συγκεκριμένο τομέα. Ο σχεδιασμός μιας πλατφόρμας που συνδυάζει τεχνικές αναπαράστασης κειμένου με αλγορίθμους μηχανικής μάθησης κρίνεται απαραίτητος. Ειδικότερα, εξετάζονται συγκεκριμένα λεξικά για την ανάλυση συναισθήματος και σταδιακά γίνεται χρήση κωδικοποιητών λέξεων και προτάσεων (word and sentence encoders) χρησιμοποιώντας τελικά τους μετασχηματιστές NLP (NLP-transformers). Σύμφωνα με τα αποτελέσματα της εργασίας, κρίθηκε αποτελεσματικότερη η χρήση των NLP Transformers σε σχέση με τις άλλες προσεγγίσεις αφού παρουσίασαν ανώτερη απόδοση.

2.3 Sentiment analysis of student's comment using lexicon-based approach

Στο εκπαιδευτικό σύστημα η ανατροφοδότηση (feedback) των μαθητών αποτελεί ένα αναπόσπαστο κομμάτι που ως στόχο έχει την εκτίμηση της ποιότητας της διδασκαλίας. Στην εργασία τους οι Khin Zezawar Aung, Nyein Nyein Myo [6] αναλύουν το feedback μέσω της χρήσης λεξικών (lexicons) χρησιμοποιώντας δεδομένα από τα σχόλια των μαθητών με βάση την διδασκαλία των καθηγητών. Στο μεγαλύτερο μέρος των προσεγγίσεων με βάση lexicons δεν δίνεται ιδιαίτερη βαρύτητα στις ενισχυτικές λέξεις (intensifier words) και στις λέξεις τυφλής άρνησης (blind negation). Έτσι, υιοθετείται η προσέγγιση της χρήσης λεξικού συμπεριλαμβανομένων της χρήσης intensifier words και blind negation words με στόχο την εξαγωγή της πολικότητας των συναισθημάτων των κειμένων. Τα επίπεδα της πολικότητας που παρουσιάζονται στην μέθοδο αυτή είναι έντονα θετικό, μέτρια θετικό, ασθενώς θετικό, έντονα αρνητικό, μέτρια αρνητικό, ασθενώς αρνητικό ή ουδέτερο. Στα αποτελέσματα που προέκυψαν παρατηρείται πανομοιότυπη συναισθηματική πολικότητα των κειμένων με βάση τους δείκτες που προκύπτουν, σε σύγκριση με διακεκριμένο lexicon του χώρου καθιστώντας την χρήση της μεθόδου αυτής αξιόπιστη.

2.4 Application of lexicon-based approach in sentiment analysis for short tweets

Τα μέσα κοινωνικής δικτύωσης αποτελούν ολοένα και μεγαλύτερο χώρο εφαρμογής της ανάλυσης συναισθήματος. Στο άρθρο των Amit Agarwal, Durga Toshniwal [7] ,συλλέχθηκαν δημοσιεύσεις από το μέσο κοινωνικής δικτύωσης twitter με επίκεντρο ένα αθλητικό γεγονός. Ειδικότερα, χρησιμοποιείται η μέθοδος του lexicon με στόχο τον υπολογισμό του συναισθήματος των οπαδών σε έναν αγώνα κρίκετ. Παράλληλα με την χρήση lexicons, εξίσου σημαντική είναι η χρήση μιας βάσης δεδομένων που αποτελείται από emoticons, τα οποία αφορούν την αναπαράσταση μιας έκφρασης προσώπου ,που ως στόχο έχουν την εξαγωγή συναισθήματος. Τα αποτελέσματα της ανάλυσης στο συγκεκριμένο άρθρο δεν ενισχύονται από την αξιοπιστία των μεθόδων, αλλά εκφράζονται μέσω της ποικιλίας συναισθημάτων που παράγονται για την αγαπημένη ομάδα/παίκτη.

2.5 Integrating a lexicon-based approach and k-nearest

Η ανάλυση συναισθήματος αναφέρεται στην αυτόματη εξαγωγή συναισθημάτων από ένα κείμενο φυσικής γλώσσας. Παρόλα αυτά, ελάχιστες έρευνες διεξάγονται στην ανάλυση συναισθήματος στην γλώσσα Μάλι (Malay). Στο άρθρο των συγγραφέων Ahmed Alsaffar, Nazlia Omar [8] προτείνονται, εφαρμόζονται και αξιολογούνται νέες προσεγγίσεις που αφορούν την ανάλυση συναισθημάτων με θέμα τον κινηματογράφο κάνοντας χρήση της γλώσσας αυτής. Σε αντίθεση με τις περισσότερες τεχνικές που επικεντρώνονται στην επιβλεπομένη (supervised) ή μη επιβλεπόμενη (unsupervised) μηχανική μάθηση, στην έρευνα αυτή προτείνεται ο συνδυασμός των δυο αυτών προσεγγίσεων. Ειδικότερα, χρησιμοποιήθηκαν lexicons με στόχο την δημιουργία ενός νέου συνόλου χαρακτηριστικών (features) για την εκπαίδευση ενός αλγορίθμου k – κοντινότερων γειτόνων. Σύμφωνα με το συμπέρασμα που διεξάγεται ο συνδυασμός της χρήσης του αλγορίθμου και του lexicon υπερτερεί σε αξιοπιστία σε σχέση με την κάθε μέθοδο ξεχωριστά.

2.6 Machine learning-based sentiment analysis for twitter accounts

Η ανάπτυξη στον τομέα της εξόρυξης γνώμης και της ανάλυσης συναισθήματος πραγματοποιείται με ταχείς ρυθμούς και στοχεύει στη διερεύνηση των απόψεων σε διαφορετικές πλατφόρμες των μέσων κοινωνικής δικτύωσης μέσω τεχνικών μηχανικής μάθησης. Ωστόσο, πάρα την πληθώρα εργαλείων και μεθόδων που παρουσιάζονται, η απόλυτη ανάγκη για μια προσέγγιση αιχμής κρίνεται απαραίτητη. Στο άρθρο αυτό μέσω των συγγραφέων Ali Hasan, Sana Moin, Ahmed Karim, Shahaboddin Shamshirband [9] περιλαμβάνεται η υιοθέτηση μιας υβριδικής προσέγγισης στην οποία η μηχανική μάθηση και τα lexicons συνδυάζονται. Ακόμα παρέχεται μια σύγκριση τεχνικών ανάλυσης συναισθήματος μέσω της εφαρμογής εποπτευόμενων αλγορίθμων μηχανικής μάθησης [10], όπως ο αφελής Bayes και

μηχανές διανυσμάτων υποστήριξης. Παράλληλα με τους παραπάνω αλγορίθμους γίνεται χρήση συγκεκριμένων λεξικών. Ανάλογα με τον συνδυασμό των μεθόδων που πραγματοποιήθηκαν καλύτερα αποτελέσματα παρουσιάζονται μέσω της χρήσης του λεξικού Text Blob.

2.7 Sentiment analysis of movie reviews using machine learning techniques

Ανάλυση συναισθήματος ή εξόρυξη γνώμης είναι η έκφραση συναισθήματος σε οποιαδήποτε μορφή κειμένου. Η ανάλυση συναισθήματος των δεδομένων αποτελεί ένα χρήσιμο εργαλείο για την έκφραση της γνώμης μιας ομάδας ή οποιουδήποτε ατόμου. Συγκεκριμένα σε πληθώρα διαδικτυακών ιστοσελίδων και εφαρμογών παρέχεται τεράστιος όγκος δεδομένων όπου η ανάλυση συναισθήματος βρίσκει έδαφος. Στο άρθρο αυτό οι συγγραφείς Palak Baid, Neelam Chaplot, Apoova Gupta [11] αναλύουν διάφορες κριτικές με θέμα τον κινηματογράφο χρησιμοποιώντας τεχνικές όπως τον αφελή Bayes ,K – κοντινότερων γειτόνων και Τυχαίου Δάσους (random forest-RF). Σύμφωνα με τα αποτελέσματα που προέκυψαν ο αφελής Bayes διαθέτει καλύτερες μετρικές συγκριτικά με τις άλλες μεθόδους μηχανικής μάθησης.

2.8 A novel lexicon-based approach in determining sentiment in financial data using learning automata

Η ανάλυση συναισθήματος αποτελεί μια δύσκολη εργασία πού ως στόχο έχει την αναγνώριση και την εξαγωγή υποκειμενικών πληροφοριών από πηγές κειμένου. Ειδικότερα, οι συγγραφείς Αντώνιος Σαρηγιαννίδης, Πάρης-Αλέξανδρος Καρυπίδης, Παναγιώτης Σαρηγιαννίδης, Ιωάννης-Χρυσόστομος Πραγίδης [12] παρουσιάζουν μια προσέγγιση που βασίζεται στην εποπτευόμενη μάθηση και ειδικότερα στη χρήση λεξικού για την πρόβλεψη νέων δεδομένων οικονομικού χαρακτήρα. Ειδικότερα, τα δεδομένα πρέπει πρώτα να εκπαιδευτούν ώστε να είναι ικανά να πραγματοποιήσουν προβλέψεις. Επομένως, διανύουν μια διαδικασία προ επεξεργασίας κατά την οποία τα δεδομένα καθαρίζονται και φιλτράρονται από ανεπιθύμητους παράγοντες. Μέσω του διανύσματος της πολικότητας μιας λέξης (word polarity vector) εξάγονται τα απαραίτητα δεδομένα ώστε να δημιουργηθεί το lexicon. Συμπερασματικά, τα αποτελέσματα που παρουσιάζονται είναι θετικά με ακρίβεια που ξεπερνάει το 60%.

3 - Μέθοδοι ανάλυσης δεδομένων

3.1 Συλλογή και προ επεξεργασία δεδομένων

Η συλλογή και η προ επεξεργασία δεδομένων αποτελεί ένα αναπόσπαστο κομμάτι και αφετηρία για την διαδικασία της ανάλυσης συναισθήματος. Η συλλογή δεδομένων

περιλαμβάνει την κατασκευή ειδικών βάσεων δεδομένων (dataset) με κείμενα οικονομικού ενδιαφέροντος. Τα κείμενα τα οποία λαμβάνονται χωρίζονται με βάση την θεματολογία τους στην αγορά μετοχών (stock market), τα κρυπτό νομίσματα (cryptocurrencies ή crypto) και το εμπόριο. Επομένως, αποθηκεύοντας την κάθε κατηγορία σε ένα υπολογιστικό φύλλο δημιουργούνται συνολικά τρία dataset. Στην παρούσα διπλωματική εργασία περιλαμβάνεται ο καθαρισμός και το φιλτράρισμα των κειμένων με στόχο την ύπαρξη μόνο ουσιωδών όρων. Επομένως μέσω της συρρίκνωσης του κειμένου και της απομάκρυνσης μη χρήσιμων όρων προκύπτει ένα σύνολο δεδομένων αυξημένων δυνατοτήτων σε σχέση με το αρχικό. Τα κείμενα των άρθρων που καλούμαστε να επεξεργαστούμε διακρίνονται από τον θόρυβο (σημεία στίξης, υπέρ σύνδεσμοι, συνδετικές λέξεις) που εμφανίζουν καθιστώντας απαραίτητη την προ επεξεργασία τους. Επίσης η αφαίρεση όρων που δεν επηρεάζουν την εννοιολογική σημασία του κειμένου στοχεύουν στην αύξηση της αποδοτικότητας του μοντέλου. Ως εκ τούτου, προτείνονται διάφορες τεχνικές που ακολουθούνται κατά την διαδικασία αυτή. Η αφαίρεση σημείων στίξης -που χρησιμοποιούνται για τη διαίρεση του κειμένου σε προτάσεις, παραγράφους και φράσεις- αποτελεί μια από τις τεχνικές προ επεξεργασίας στοχεύοντας στην αύξηση της αποδοτικότητας του μοντέλου. Η μετατροπή όλων των κεφαλαίων γραμμάτων σε πεζά (lowercase) συμβάλλει στο φιλτράρισμα του κειμένου μειώνοντας το κόστος εφαρμογής των αλγορίθμων, αφού κοινές λέξεις δεν αντιπροσωπεύονται πλέον ως διαφορετικές στο μοντέλο διανυσματικού χώρου. Η αφαίρεση λέξεων ελάσσονας σημασίας (stop words) προβλέπει την αφαίρεση λέξεων που δεν προσθέτουν ιδιαίτερη σημασία στο κείμενο παρουσιάζοντας μεγάλη συχνότητα εμφάνισης. Παραδείγματα τέτοιων λέξεων αποτελούν τα άρθρα, συνδετικές λέξεις κλπ. Προκειμένου, να επιτευχθούν καλύτερα αποτελέσματα πραγματοποιείται αφαίρεση των προθεμάτων και των επιθεμάτων των λέξεων με στόχο την αναγωγή τους σε λήμματα. Η παραπάνω διαδικασία συνθέτει τις τεχνικές των stemming και lemmatization. Οι μέθοδοι που αναφέρθηκαν εκτελέστηκαν μέσω συναρτήσεων της γλώσσας προγραμματισμού Python με στόχο την πρακτική τους εφαρμογή.

3.2 Εξαγωγή χαρακτηριστικών

Είναι γνωστό ότι οι μέθοδοι ανάλυσης δεδομένων και συγκεκριμένα η εξαγωγή χαρακτηριστικών, τα μοντέλα εκπαίδευσης και τα μοντέλα εκτιμήσεων αποτελούν ένα αναπόσπαστο κομμάτι της ανάλυσης συναισθήματος στο κομμάτι της επεξεργασίας κειμένου μέσω της μηχανικής μάθησης [13]. Εφόσον, η διαδικασία της προ-επεξεργασίας έχει προηγηθεί, επόμενο βήμα αποτελεί η εξαγωγή χαρακτηριστικών. Μέσω της εξαγωγής χαρακτηριστικών ελαχιστοποιείται ο τεράστιος όγκος δεδομένων της βάσης , αφού τα παραπάνω μετατρέπονται σε αριθμητικά χαρακτηριστικά κατάλληλα για την χρήση τους από τα μοντέλα μηχανικής μάθησης. Η διαδικασία αυτή παρέχει αρκετά πλεονεκτήματα όπως η ταχύτητα στην εκπαίδευση των δεδομένων, αυξημένη απόδοση και μείωση του παράγοντα της υπερ. προσαρμογής [14]. Υπερ. προσαρμογή πραγματοποιείται όταν η ακρίβεια του μοντέλου είναι εξαιρετικά υψηλή στο σύνολο των δεδομένων εκπαίδευσης αλλά το μοντέλο αδυνατεί να γενικεύσει την αποκτημένη

γνώση του σε άγνωστα δεδομένα. Παρακάτω, περιγράφονται δυο τεχνικές εξαγωγής χαρακτηριστικών υπό μορφή αριθμητικών διανυσμάτων.

3.2.1 CV –μέτρηση διανύσματος

Η τεχνική της μέτρησης διανύσματος αποτελεί εξαιρετικό εργαλείο που παρέχεται από την βιβλιοθήκη scikit-learn στην γλώσσα προγραμματισμού python. Συμβάλει στην μετατροπή ενός κειμένου σε διάνυσμα με βάση την συχνότητα (πλήθος) εμφάνισης κάθε λέξης που συναντάται σε όλη την έκταση του κειμένου. Οι θέσεις στο διάνυσμα αντιπροσωπεύονται από τον αριθμό εμφάνισης κάθε λέξης στην πρόταση. Η τεχνική της μέτρησης διανύσματος πραγματοποιεί εξαγωγή χαρακτηριστικών μέσω ενός λεξιλογίου που κατασκευάζεται είτε από το ίδιο σώμα κειμένου είτε εισάγεται εξωγενώς. Ωστόσο η προσέγγιση της μέτρησης διανύσματος εμφανίζει ορισμένα μειονεκτήματα. Μέσω της τεχνικής αυτής επισκιάζονται σημαντικές λέξεις, που διαθέτουν χαρακτηριστικά λήψης αποφάσεων για αλγορίθμους, σε σχέση με λέξεις οι οποίες δεν προσδίδουν κάποιο ιδιαίτερο νόημα. Ως αποτέλεσμα, των μειονεκτημάτων αυτών σημαντικές πληροφορίες για την ανάλυση συναισθημάτων ενδέχεται να μην είναι εμφανής, γεγονός που απαιτεί την χρήση πιο εξελιγμένων μεθόδων. Παρακάτω χρησιμοποιείται ένα παράδειγμα χρήσης της τεχνικής με την χρήση κώδικα για περαιτέρω κατανόηση.

Οι παρακάτω προτάσεις χρησιμοποιούνται για την περιγραφή της μεθόδου της μέτρησης διανύσματος

```
corpus = [
                    'The sky is blue and beautiful',
                'The king is old, and the queen is beautiful',
                    'Love this beautiful blue sky',
                'The beautiful queen and the old king'
]
```

Κώδικας:

```
from sklearn.feature_extraction.text import CountVectorizer
import pandas as pd
vectorizer = CountVectorizer()
X = vectorizer.fit_transform(corpus)
print(vectorizer.get_feature_names())
Doc_Term_Matrix = pd.DataFrame(X.toarray(),columns= vectorizer.get_feature_names())
Doc_Term_Matrix
```

Ως αποτέλεσμα, προκύπτει ο παρακάτω πίνακας με τα βάρη.

	and	beautiful	blue	is	king	love	old	queen	sky	the	this
0	1	1	1	1	0	0	0	0	1	1	0
1	1	1	0	2	1	0	1	1	0	2	0
2	0	1	1	0	0	1	0	0	1	0	1
3	1	1	0	0	1	0	1	1	0	2	0

Εικόνα 1 - Πίνακας χαρακτηριστικών μέτρησης διανύσματος

3.2.2 TF-IDF (term frequency- inverse document frequency)

Ο αλγόριθμος TF-IDF (term frequency – inverse document frequency) είναι αλγόριθμος στατιστικής φύσεως μέσω του οποίου αξιολογείται η συνάφεια μιας λέξης σε ένα έγγραφο σε σχέση με ένα σύνολο εγγράφων [15]. Η λειτουργία του αλγορίθμου περιλαμβάνει τον πολλαπλασιασμό δύο μετρήσεων, την συχνότητα όρου (term frequency) που υπολογίζει την συχνότητα εμφάνισης του όρου αυτού σε μια ακολουθία και την αντίστροφη συχνότητα εγγράφου (inverse document frequency) που εκτιμά τον ρυθμό εμφάνισης μιας λέξης σε ένα κείμενο σε σχέση με το σύνολο των κειμένων. Από την ανάλυση της παραπάνω μέτρησης προκύπτει υψηλότερος δείκτης (IDF) μέσω της χρήσης σπάνιας λέξης στο σύνολο των κειμένων, ενώ αντίστοιχα προκύπτει χαμηλότερη τιμή του δείκτη (IDF) για την υψηλή συχνότητα εμφάνισης των λέξεων στο σύνολο των κειμένων. Η μέθοδος της TF-IDF πλεονεκτεί σε σχέση με την μέτρηση διανύσματος όχι μόνο επειδή εστιάζει στην συχνότητα της μεθόδου αλλά μέσω της παραπάνω τεχνικής δίνεται σημασία στις λέξεις αυτές καθαυτές. Στο παρακάτω παράδειγμα περιγράφεται η χρήση της μεθόδου μέσω της χρήσης κώδικα για καλύτερη κατανόηση.

```
from sklearn.feature_extraction.text import TfidfTransformer
transformer = TfidfTransformer()
tfidf = transformer.fit_transform(X)
Doc_Term_Matrix=pd.DataFrame(tfidf.toarray(),columns=vectorizer.get_feature_names())
pd.set_option("display.precision", 2)
Doc_Term_Matrix
```

Πίνακας βαρών:

Εικόνα 2 - Πίνακας χαρακτηριστικών TF-IDF

	beautiful	beautiful blue	beautiful queen	blue	blue beautiful	blue sky	king	king old	love	love beautiful	old	old king	old queen	queen	queen beautiful	queen old	sky	sky blue
0	0.28	0.00	0.00	0.42	0.53	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.42	0.53
1	0.23	0.00	0.00	0.00	0.00	0.00	0.35	0.44	0.00	0.00	0.35	0.00	0.44	0.35	0.44	0.00	0.00	0.00
2	0.22	0.43	0.00	0.34	0.00	0.43	0.00	0.00	0.43	0.43	0.00	0.00	0.00	0.00	0.00	0.00	0.34	0.00
3	0.23	0.00	0.44	0.00	0.00	0.00	0.35	0.00	0.00	0.00	0.35	0.44	0.00	0.35	0.00	0.44	0.00	0.00

$$tf\ idf(t,d,D) = tf(t,d)\ .idf(t,D)$$

$$tf(t,d) = \log(1 + freq\ (t,d))$$

$$idf\ (t,D) = \log\left(\frac{N}{count\ (d\ \mathcal{E}\ D: t\ \mathcal{E}\ d)}\right)$$

Ο όρος $tf(t,d)$ περιλαμβάνει την συχνότητα ενός όρου (t) σε ένα κείμενο (d) . Ο συνολικός αριθμός των κειμένων αναπαρίσταται με τον όρο N και ο όρος $count\ (d\ \mathcal{E}\ D: t\ \mathcal{E}\ d)$ αφορά τον αριθμό των κειμένων που περιέχουν τουλάχιστον μια φορά τον όρο t.

Μετά την εξαγωγή των χαρακτηριστικών των μοντέλων σημαντική ενέργεια αποτελεί η εφαρμογή των αλγορίθμων μηχανικής μάθησης. Ωστόσο, για να οριστούν οι παράμετροι των μοντέλων μηχανικής μάθησης που έχουν επιλεχθεί σημαντικό βήμα αποτελεί η εφαρμογή της τεχνικής της αναζήτησης πλέγματος και της διασταυρωμένης επικύρωσης. Παρακάτω περιγράφονται οι δυο τεχνικές αυτές μέσω των οποίων προβλέπεται αύξηση της αποδοτικότητας των μοντέλων.

3.2.3 Αναζήτηση πλέγματος

Η αναζήτηση πλέγματος (grid search) είναι μία τεχνική μέσω της οποίας υπολογίζονται οι βέλτιστες τιμές των υπερ-παραμέτρων [16]. Ειδικότερα, πρόκειται για εξαντλητική αναζήτηση που εκτελείται σε συγκεκριμένες τιμές παραμέτρων ενός μοντέλου μηχανικής μάθησης. Η αναζήτηση πλέγματος χρησιμοποιεί όλους τους διαφορετικούς συνδυασμούς των καθορισμένων υπερ-παραμέτρων υπολογίζοντας τις καλύτερες τιμές. Μέσω της διαδικασίας της αναζήτησης πλέγματος εξοικονομούνται υπολογιστικός χρόνος και πόροι με στόχο την αύξηση της αποτελεσματικότητας των μεθόδων.

3.2.4 Διασταυρωμένη επικύρωση

Η διασταυρωμένη επικύρωση (cross validation) αποτελεί τεχνική επαναλαμβανόμενης δειγματοληψίας που στοχεύει στον διαχωρισμό του συνόλου δεδομένων σε δυο μέρη – δεδομένα εκπαίδευσης και δεδομένα δοκιμών [17]. Τα δεδομένα εκπαίδευσης στοχεύουν στην εκπαίδευση του μοντέλου και τα δεδομένα δοκιμών χρησιμοποιούνται για προβλέψεις. Η αποδοτικότητα των μοντέλων στα δεδομένα δοκιμών ενδέχεται να χαρακτηριστεί ως υψηλή με αποτέλεσμα το μοντέλο να μην εμφανίζει δείγματα υπερ-προσαρμογή και η πρόβλεψη

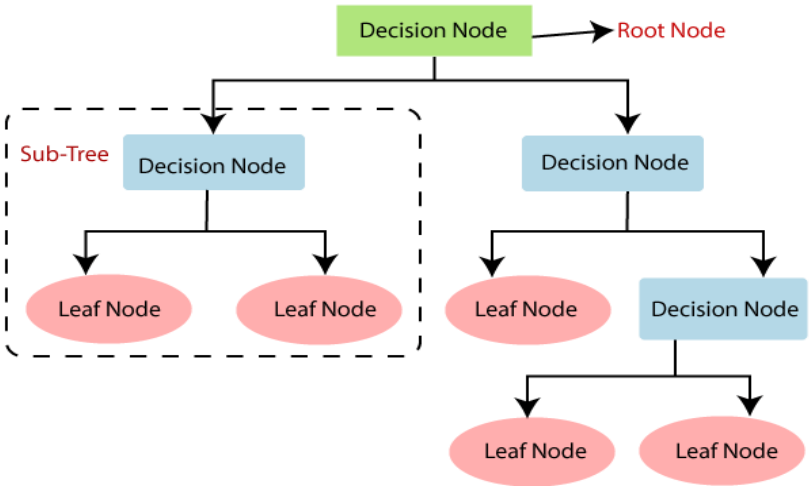
δεδομένων να είναι εφικτή. Ωστόσο, υπάρχουν πολλές διαφορετικές τεχνικές διασταυρωμένης επικύρωσης με την επικρατέστερη να είναι η k-fold cross-validation. Στην μέθοδο αυτή η τιμή του αριθμού k καθορίζεται από τον χρήστη και αφορά τα μέρη πτυχώσεων που θα εφαρμοστούν με συνηθέστερες τιμές να είναι αυτές των πέντε και δέκα. Ειδικότερα, κατά την εκπαίδευση των μοντέλων η πρώτη πτυχή αφορά τα δεδομένα δοκιμών και οι υπόλοιπες αφορούν τα δεδομένα εκπαίδευσης. Η διαδικασία επαναλαμβάνεται με τον ίδιο τρόπο βάση της τιμής του k, με την διαφορά ότι τα δεδομένα δοκιμών και τα δεδομένα εκπαίδευσης εναλλάσσονται. Τελικά, υπολογίζονται τόσες μετρικές απόδοσης όσες και οι πτυχώσεις που έχουν οριστεί.

3.3 Εκπαίδευση μοντέλων

Όπως αναφέρθηκε και σε προηγούμενο κεφάλαιο τα σύνολα εκπαίδευσης και τα σύνολα δοκιμών συνεισφέρουν στην πρόβλεψη νέων δεδομένων. Ειδικότερα, χωρίζεται το αρχικό σύνολο δεδομένων σε στήλες εισόδου και στήλες εξόδου που αποτελούν αντίστοιχα ανεξάρτητες και εξαρτημένες μεταβλητές [18]. Στη συνέχεια, αφού χωριστούν τα δεδομένα προκύπτουν τα δεδομένα εκπαίδευσης και τα δεδομένα δοκιμών. Ο διαχωρισμός ενός συνόλου δεδομένων είναι το βήμα που προηγείται σε σχέση με την εκτέλεση των αλγορίθμων μηχανικής μάθησης [19]. Παρακάτω παρατίθενται οι αλγόριθμοι που χρησιμοποιήθηκαν μετά τον διαχωρισμό του συνόλου δεδομένων.

3.3.1 Δένδρα απόφασης – Decision trees

Η διαδικασία της ταξινόμησης αποτελείται από δύο βήματα, το βήμα μάθησης και το βήμα πρόβλεψης. Στο βήμα μάθησης το μοντέλο αναπτύσσεται με βάση τα δεδομένα εκπαίδευσης ,ενώ στο βήμα πρόβλεψης το μοντέλο είναι ικανό να παρέχει αποτελέσματα με βάση την χρήση νέων δεδομένων. Ο αλγόριθμος των δένδρων απόφασης είναι από τους πιο εύκολους και δημοφιλής αλγορίθμους ώστε οι έννοιες αυτές να κατανοηθούν και ερμηνευτούν με τον καλύτερο δυνατό τρόπο [20]. Ο ίδιος υπάγεται στην κατηγορία των αλγορίθμων εποπτευόμενης μάθησης και μπορεί να χρησιμοποιηθεί τόσο για ταξινόμηση (classification) όσο και για παλινδρόμηση (regression). Ο στόχος ενός δένδρου απόφασης αφορά την δημιουργία ενός μοντέλου εκπαίδευσης που μπορεί να χρησιμοποιηθεί μέσω της εκμάθησης απλών κανόνων απόφασης που συνάγονται από προηγούμενα δεδομένα. Η αφετηρία εφαρμογής του αλγορίθμου είναι η ρίζα του δένδρου. Με στόχο την πρόβλεψη μιας ετικέτας κλάσης συγκρίνονται οι τιμές των χαρακτηριστικών της ρίζας του δένδρου με αυτές των επόμενων κόμβων .Με βάση την σύγκριση αυτή ακολουθείται ο κλάδος που αντιστοιχεί στην επιθυμητή τιμή και επόμενο βήμα αποτελεί η μεταπήδηση σε νέο κόμβο. Ωστόσο χρησιμοποιείται η τεχνική του διαίρει και βασίλευε με στόχο την διάσπαση του χώρου αναζήτησης σε υποσύνολα.



Εικόνα 3 - Κατασκευή δένδρου απόφασης [21]

3.3.2 Αφελής κατηγοροιοποιητής Bayes – Naïve bayes classifier

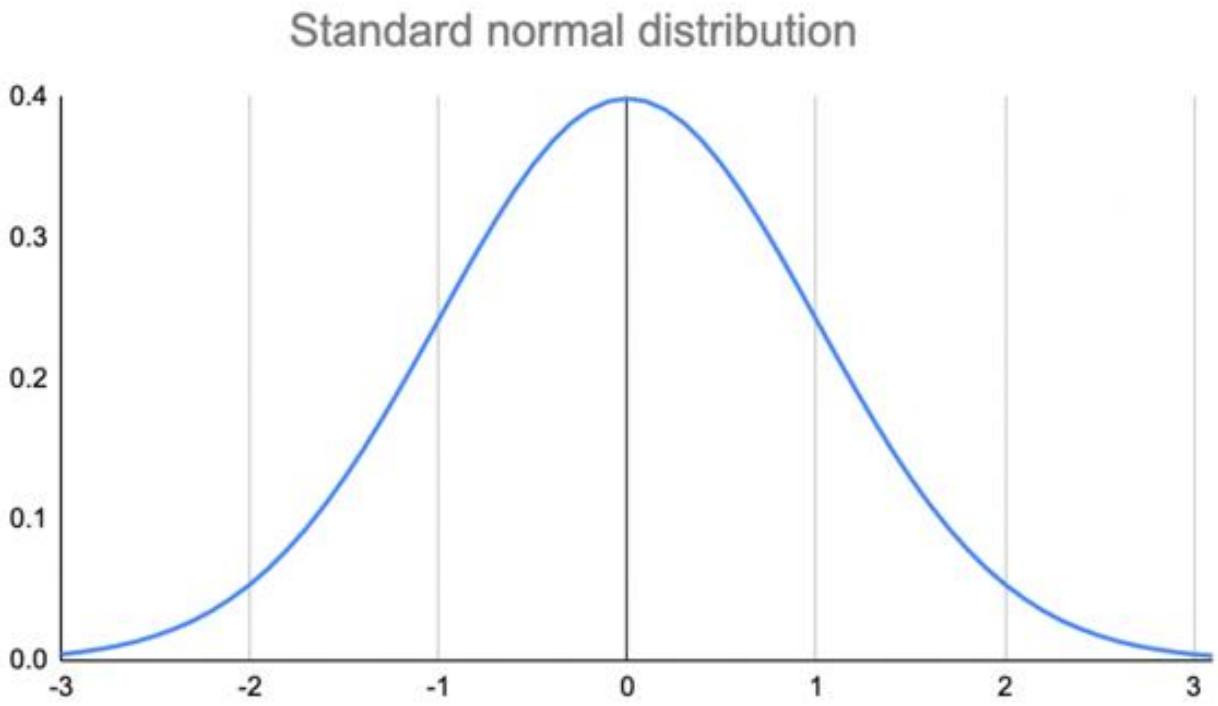
Ο επόμενος αλγόριθμος μηχανικής μάθησης που θα αναλύσουμε είναι ο αφελής κατηγοροιοποιητής Bayes [22]. Ο αλγόριθμος αυτός αποτελεί παραλλαγή του αφελή Bayes ακολουθώντας την κανονική κατανομή Gaussian υποστηρίζοντας παράλληλα συνεχή δεδομένα. Ο αφελής Bayes αποτελεί πρώιμο στάδιο του αλγορίθμου ο οποίος αναλύεται. Ειδικότερα, αφορά μια ομάδα εποπτευομένων αλγορίθμων ταξινόμησης που βασίζονται στο θεώρημα Bayes. Μέσω της τεχνικής προσφέρεται υψηλή λειτουργικότητα με την χρήση της σε προβλήματα υψηλών διαστάσεων. Το θεώρημα Bayes είναι ικανό να χρησιμοποιηθεί για τον υπολογισμό της υπό όρους πιθανότητας. Εξαιτίας, της υψηλής αποδοτικότητας του σε μοντέλα πιθανοτήτων μπορεί να εφαρμοστεί και στην μηχανική μάθηση. Η βασική συνάρτηση πιθανοτήτων με P(A) τη πιθανότητα να συμβεί το γεγονός A, P(B) να συμβεί το γεγονός B, P(B|A) τη πιθανότητα να συμβεί το γεγονός B δεδομένου ότι συμβαίνει το A και P(A|B) η πιθανότητα να συμβεί το γεγονός A δεδομένου ότι συμβαίνει το B, παρουσιάζεται παρακάτω.

$$P(A|B) = \frac{P(A) * P(B|A)}{P(B)}$$

Η πιθανότητα των χαρακτηριστικών που κατανέμονται με την κανονική (Gaussian) κατανομή, δηλαδή η συνάρτηση πυκνότητας πιθανότητας είναι.

$$P(x_i|\, y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

Όπου τα μ_y και σ_y^2 υπολογίζονται με την μέθοδο μέγιστης πιθανοφάνειας.



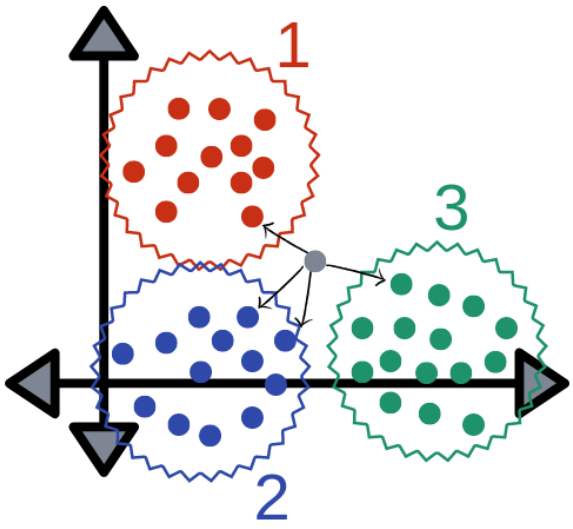
Εικόνα 4 – Γκαουσιανή κατανομή[23]

Η δημιουργία ενός απλού μοντέλου περιλαμβάνει την υπόθεση ότι τα δεδομένα περιγράφονται μέσω της κατανομής Gauss χωρίς συν διακύμανση (ανεξάρτητες διαστάσεις). Το μοντέλο αυτό μπορεί να προσαρμοστεί μέσω του υπολογισμού του μέσου όρου και της τυπικής απόκλισης των σημείων σε κάθε ετικέτα.

3.3.2 Κ-κοντινότερων γειτόνων – K-nearest neighbours

Ο αλγόριθμος πλησιέστερων γειτόνων είναι μια μέθοδος ταξινόμησης δεδομένων για την εκτίμηση της πιθανότητας ενός στοιχείου να αποτελέσει τμήμα μιας κλάσης πλησιέστερα σε αυτό [24]. Ο KNN είναι ένας τύπος εποπτευομένου αλγορίθμου μηχανικής μάθησης που χρησιμοποιείται τόσο για προβλήματα ταξινόμησης όσο και προβλήματα παλινδρόμησης. Ο ίδιος χαρακτηρίζεται ως μη-παραμετρικός αφού δεν πραγματοποίει υποθέσεις σχετικά με την κατανομή δεδομένων αλλά προσδιορίζεται μέσω των ταξινομημένων στοιχείων πλησιέστερα σε ένα στοιχείο. Η χρήση του αλγορίθμου αποτελεί ιδανικό σενάριο στην περίπτωση που τα δεδομένα είναι καλά καθορισμένα ή μη γραμμικά. Ειδικότερα, ο αλγόριθμος k-κοντινότερων γειτόνων χρησιμοποιεί ένα μηχανισμό «ψηφοφορίας» ώστε να προσδιοριστεί η κατηγορία των μη χαρακτηρισμένων δεδομένων. Επομένως, η κλάση με τις περισσότερες «ψήφους» θα αποτελέσει την κατηγορία του εν λόγω σημείου δεδομένων [25]. Η τιμή της μεταβλητής k

αποτελεί στοιχείο υψηλής σημασίας αφού, ανάλογα της τιμής που λαμβάνεται, επιλέγεται και ο αριθμός των γειτόνων που θα χρησιμοποιηθούν στην διαδικασία ταξινόμησης.



Εικόνα 5 - Παράδειγμα ταξινόμησης νέου στοιχείου στον αλγόριθμο k-κοντινότερων γειτόνων [26]

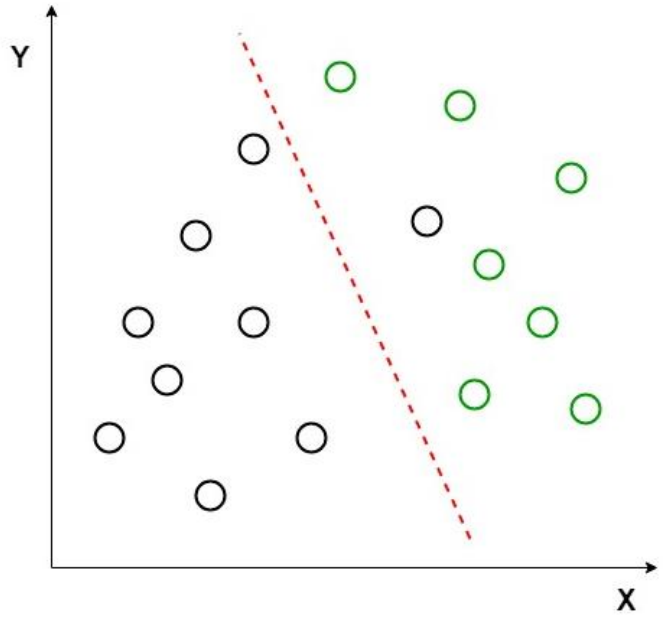
Ωστόσο, η μέτρηση της απόστασης συμβάλλει στον προσδιορισμό, ενός σημείου δεδομένων, της κλάσης στην οποία ανήκει. Ειδικότερα υπάρχουν τέσσερις τρόποι για τον υπολογισμό της μέτρησης της απόστασης μεταξύ του σημείου δεδομένων και του πλησιέστερου γείτονά του. Αυτοί αφορούν τους: Ευκλείδεια απόσταση, απόσταση Μανχάταν, απόσταση Χάμινγκ και απόσταση Μινκόφσκι. Η συνηθέστερα χρησιμοποιούμενη απόσταση είναι η Ευκλείδεια και δεδομένου ότι οι μεταβλητές x και y δίνονται μέσω του χρήστη η εξίσωση που προκύπτει είναι η εξής [27].

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2}$$

3.3.4 Γραμμική Διακριτική Ανάλυση - Linear discriminant analysis

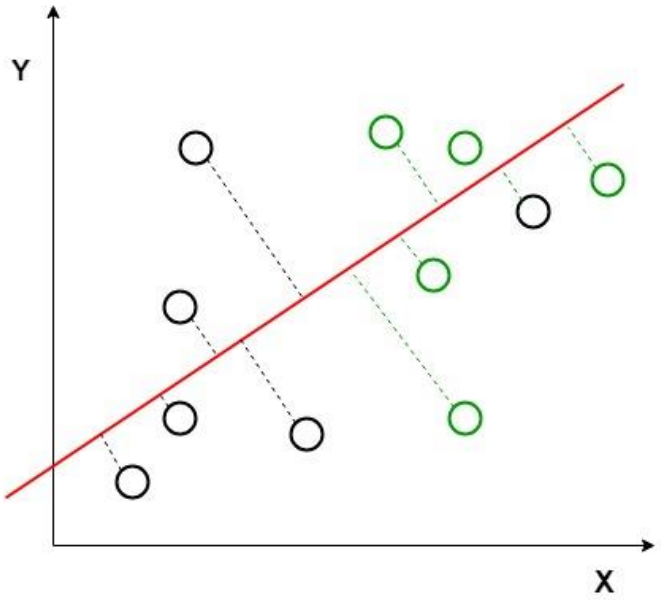
Η ανάλυση του γραμμικού διαχωρισμού αποτελεί ένα γραμμικό μοντέλο ταξινόμησης και μείωσης διαστάσεων [28]. Παρά την απλότητα του, η γραμμική διακριτική ανάλυση παράγει ισχυρά αποτελέσματα ταξινόμησης. Ο αλγόριθμος αυτός χρησιμοποιείται μέσω της προβολής των χαρακτηριστικών του σε έναν χώρο χαμηλότερης διάστασης με στόχο την ταξινόμηση των δεδομένων σε δυο ή παραπάνω κλάσεις. Ειδικότερα, στην ανάλυση του γραμμικού διαχωρισμού, ως αφετηρία παρουσιάζεται η δημιουργία ενός νέου άξονα που στόχο έχει να προβάλει όλα τα στοιχεία των δεδομένων των κλάσεων στον άξονα αυτό [29]. Ωστόσο, είναι

σημαντικό να αναφερθεί ότι η δημιουργία του άξονα προϋποθέτει την παρουσία ευθύγραμμου τμήματος μέσω του οποίου διαχωρίζονται οι δυο κατηγορίες των σημείων δεδομένων.



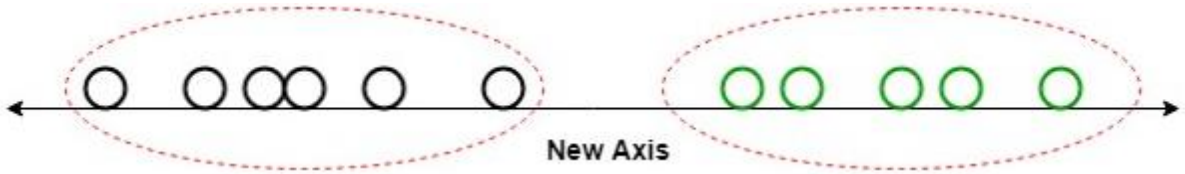
Εικόνα 6 - Γραμμικά διαχωριζόμενες κλάσεις για γραμμική διακριτική ανάλυση [28]

Η χρήση ενός παραδείγματος θα ήταν αποτελεσματική για την καλύτερη κατανόηση της μεθόδου. Η γραμμική διακριτική ανάλυση μεταχειρίζεται δύο άξονες (Χ και Υ) με στόχο την δημιουργία ενός νέου με αποτέλεσμα την μεγιστοποίηση του διαχωρισμού των κατηγοριών και ως εκ τούτου την μείωση της διάστασης του γραφήματος. Για την εκπλήρωση των παραπάνω λειτουργιών η τήρηση δύο κριτηρίων είναι απαραίτητη. Η μεγιστοποίηση της απόστασης των μέσων όρων των στοιχείων των κλάσεων και η ελαχιστοποίηση της διακύμανσης μέσα στην κλάση αποτελούν τις προϋποθέσεις για την ολοκλήρωση της διαδικασίας. Το πρώτο αφορά την μέγιστη απόσταση που θα πρέπει να έχουν τα στοιχεία των κλάσεων ώστε η ταξινόμηση να είναι αποτελεσματική και το δεύτερο αφορά την ελάχιστη απόσταση που πρέπει να έχουν τα δεδομένα μέσα στην ίδια κλάση.



Εικόνα 7 - Δημιουργία νέου άξονα για διαχωρισμό δεδομένων [28]

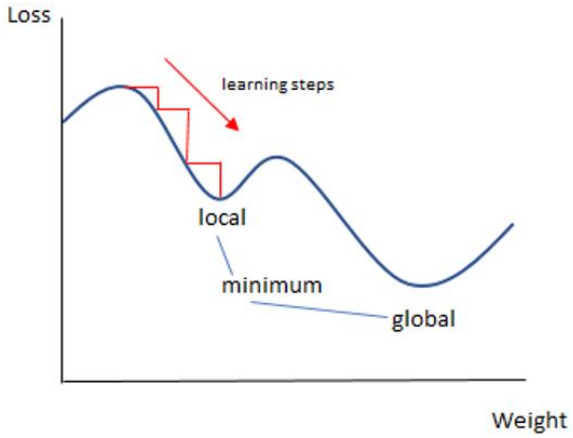
Επομένως, μετά την δημιουργία του νέου άξονα, χρησιμοποιώντας τα προαναφερθέντα κριτήρια, ο σχεδιασμός των στοιχείων δεδομένων στον νέο άξονα είναι ο εξής.



Εικόνα 8 - Τελικό αποτέλεσμα γραμμικής διακριτικής ανάλυσης [28]

Ωστόσο, η χρήση της παραπάνω μεθόδου είναι αδύνατη, όταν ο μέσος όρος των κατανομών δεν είναι διαχωρίσιμος ανάμεσα στις δυο κλάσεις, καθιστώντας παράλληλα αδύνατη την δημιουργία του νέου άξονα με στόχο τον γραμμικό διαχωρισμό των δυο κλάσεων.

3.3.5 Στοχαστική κλίση καθόδου - Gradient descent



Εικόνα 9 - Παράδειγμα 2 διαστάσεων στοχαστική κλίση καθόδου [30]

Η στοχαστική κλίση καθόδου είναι μια εξαιρετικά δημοφιλής και κοινή μέθοδος που χρησιμοποιείται σε διάφορους αλγορίθμους μηχανικής μάθησης αποτελώντας παράλληλα την βάση των νευρωνικών δικτύων. Ωστόσο η αναφορά στον αλγόριθμο κλίσης καθόδου κρίνεται απαραίτητη με στόχο την καλύτερη κατανόηση της μεθόδου. Η κλίση καθόδου αποτελεί ένα αλγόριθμο βελτιστοποίησης που χρησιμοποιείται κατά την εκπαίδευση ενός μοντέλου μηχανικής μάθησης [31]. Ειδικότερα, βασίζεται στην χρήση μια κυρτής συνάρτησης μέσω της οποίας τροποποιούνται οι παράμετροι επαναληπτικά με στόχο την ελαχιστοποίηση της συνάρτησης κόστους. Επομένως, με την κλίση καθόδου επιδιώκεται η επαναληπτική παραμετροποίηση του μοντέλου με στόχο την εύρεση ολικού ελαχίστου. Η συνάρτηση που χρησιμοποιείται για την περιγραφή της μεθόδου της κλίσης καθόδου είναι η εξής.

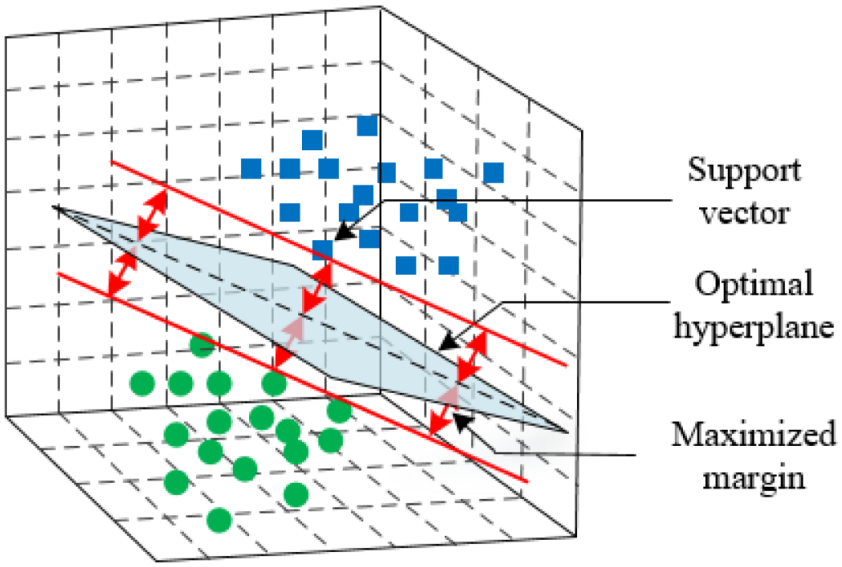
$$P_{n+1} = p_n - \eta \nabla f(p_n)$$

Η P_{n+1} αποτελεί την επόμενη θέση που θα λάβει η συνάρτηση, η p_n αντιπροσωπεύει την τρέχουσα θέση και το $\nabla f(p_n)$ αφορά την κατεύθυνση της πιο απότομης κατάβασης. Η παράμετρος η αφορά τον ρυθμό εκμάθησης της συνάρτησης ώστε να προχωρήσει στο επόμενο βήμα. Πρέπει να σημειωθεί, ότι ο ρυθμός εκμάθησης αποτελεί ένα σημαντικό βήμα της διαδικασίας εφόσον μικρές μεταβολές μπορούν να επηρεάσουν σημαντικά τον χρόνο και την ποιότητα της εκπαίδευσης. Ειδικότερα, μια μικρή τιμή μπορεί να σημαίνει επιβράδυνση της σύγκλισης, ενώ μια μεγάλη να οδηγήσει σε απόκλιση μέσω σημαντικής αύξησης των επαναλήψεων και του χρόνου εκπαίδευσης. Η στοχαστική κλίση καθόδου βασίζεται στην μεθοδολογία που αναφέρθηκε παραπάνω με την διαφορά ότι η κλίση υπολογίζεται με βάση ενός υποσυνόλου του σετ δεδομένων. Η λειτουργία αυτή επιταχύνει σημαντικά τον αλγόριθμο καθιστώντας την απαραίτητη σε προβλήματα υψηλών διαστάσεων. Εκ πρώτης όψεως, η Στοχαστική κλίση καθόδου παρουσιάζεται ως αλγόριθμος χαμηλής ποιότητας, ωστόσο

επιλύονται σημαντικά προβλήματα του κλίσης καθόδου όπως είναι η προσκόλληση της συνάρτησης σε τοπικά ελάχιστα.

3.3.6 Μηχανές διανυσμάτων υποστήριξης – Support vector machines

Οι μηχανές διανυσμάτων υποστήριξης αποτελούν έναν από τους πιο δημοφιλής αλγορίθμους εποπτευομένης μάθησης οι οποίοι χρησιμοποιούνται τόσο σε προβλήματα ταξινόμησης όσο και σε προβλήματα παλινδρόμησης. Οι μηχανές διανυσμάτων υποστήριξης προτιμώνται ιδιαίτερα καθώς συνδυάζουν την υψηλή ακρίβεια των αποτελεσμάτων τους με την λιγότερη δυνατή υπολογιστική ισχύ που απαιτείται [32]. Ο στόχος του αλγορίθμου είναι η δημιουργία ενός ευθύγραμμου τμήματος ή ενός ορίου απόφασης μέσω του οποίου διαχωρίζεται ένας χώρος n-διαστάσεων σε κλάσεις. Το όριο το οποίο στοχεύει στον καλύτερο διαχωρισμό των κλάσεων και στην έγκυρη ταξινόμηση των μελλοντικών δεδομένων ονομάζεται υπέρ επίπεδο (hyperplane). Ο διαχωρισμός δυο κατηγοριών σημείων δεδομένων προβλέπει την ύπαρξη πολλών υπέρ επιπέδων. Ωστόσο, η τελική επιλογή του υπέρ επιπέδου εξαρτάται από την μέγιστη δυνατή απόσταση μεταξύ των σημείων δεδομένων των δυο κατηγοριών. Τα σημεία δεδομένων που είναι κοντά στο υπέρ επίπεδο και επηρεάζουν την θέση του ονομάζονται διανύσματα υποστήριξης (support vectors). Η μεγιστοποίηση της απόστασης του περιθωρίου παρέχει στο σύστημα την βεβαιότητα κατά την οποία μελλοντικά δεδομένα μπορούν να ταξινομηθούν με μεγαλύτερη ασφάλεια. Το σχήμα το οποίο θα λάβει το υπέρ επίπεδο εξαρτάται από τον αριθμό των χαρακτηρισμένων δεδομένων. Αυτό σημαίνει ότι αν παρέχονται δυο features τότε τα δεδομένα χωρίζονται με μια γραμμή και αν παρέχονται τρία features τότε το υπέρ επίπεδο θα αποτελέσει ένα επίπεδο δυο διαστάσεων.

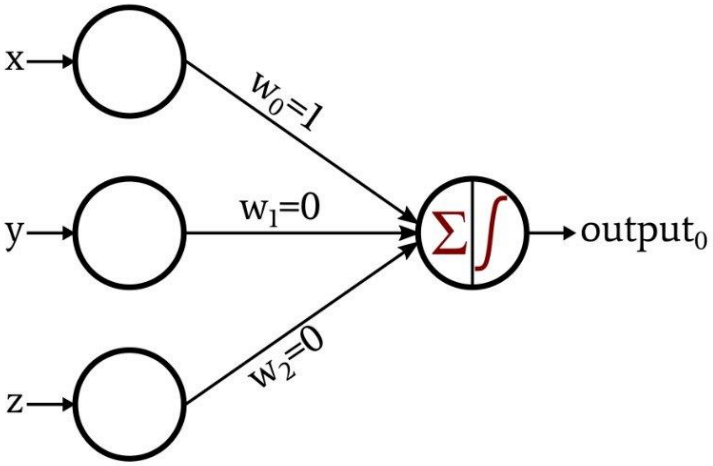


Εικόνα 10 – Γραφική αναπαράσταση μηχανών διανυσμάτων υποστήριξης $3^{ων}$ διαστάσεων [32]

Οι μηχανές διανυσμάτων υποστήριξης χωρίζονται σε δυο κατηγορίες ανάλογα με τα δεδομένα που αναλαμβάνουν. Πρώτη κατηγορία αφορά τα γραμμικώς διαχωρίσιμα δεδομένα και όπως υποδηλώνεται μέσω του ονόματος τους αφορούν δεδομένα που χωρίζονται με μια γραμμή. Η δεύτερη αφορά τα μη-γραμμικώς διαχωριζόμενα δεδομένα στα οποία το όριο διαχωρισμού δεν αποτελεί ευθεία γραμμή. Ειδικότερα, για την ταξινόμηση των μη-γραμμικών δεδομένων χρησιμοποιείται το λεγόμενο “τέχνασμα πυρήνα” (kernel trick) όπου μη-γραμμικά στοιχεία προβάλλονται σε έναν χώρο υψηλότερης διάστασης με στόχο την διευκόλυνση της ταξινόμησης των δεδομένων που θα μπορούσαν να διαιρεθούν γραμμικά με ένα επίπεδο. Η λειτουργία αυτή διαχωρίζεται βάση του πυρήνα των μηχανών διανυσμάτων υποστήριξης σε: πυρήνας gaussian, πυρήνας πολυωνύμου (polynomial kernel), σιγμοειδής πυρήνας (sigmoid kernel) .

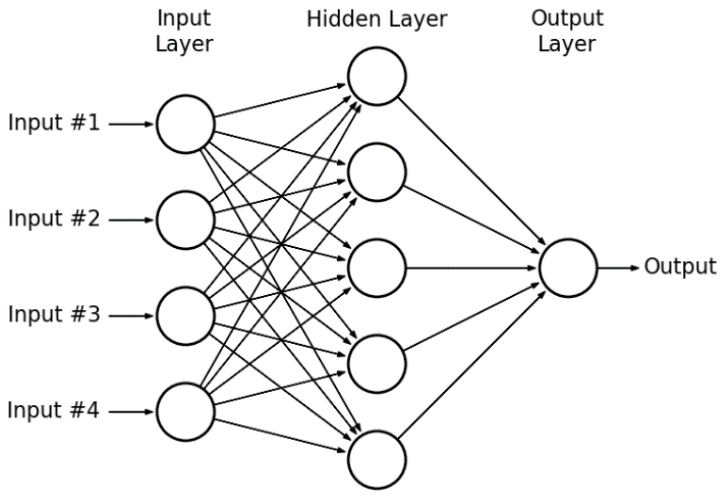
3.3.7 Πολύ στρωματικός ταξινομητής - Multi-layer perceptron classifier)

Ο πολύ στρωματικός αλγόριθμος perceptron είναι ένας αλγόριθμος εποπτευομένης μηχανικής μάθησης που ταξινομεί δεδομένα μέσω των αρχικών δεδομένων εκπαίδευσης με την βοήθεια νευρώνων [33]. Ειδικότερα, τα perceptron χρησιμοποιούνται μέσω της λήψης n χαρακτηριστικών ως είσοδο (x = x1, x2,…, x n) και καθένα από αυτά σχετίζονται με ένα βάρος [34]. Τα δεδομένα τα οποία εισάγονται σε ένα perceptron πρέπει να είναι αριθμητικά ώστε να ληφθούν υπόψη. Τα χαρακτηριστικά εισόδου μεταβιβάζονται σε μια συνάρτηση, μέσω της οποίας υπολογίζεται το σταθμισμένο άθροισμα (weighted sum). Το αποτέλεσμα του υπολογισμού αυτού μεταφέρεται σε μια συνάρτηση ενεργοποίησης f (activation function) μέσω της οποίας εξάγονται τα δεδομένα του perceptron. Μέσω της συνάρτησης ενεργοποίησης αποφασίζεται εάν ένας νευρώνας ενεργοποιείται ή όχι. Ειδικότερα, η απόφαση προκύπτει από την αξιοπιστία του νευρώνα να προβλέψει νέα δεδομένα καθώς και από την αναγκαιότητα του μοντέλου να αποτρέψει την γραμμικότητα.



Εικόνα 11 – Perceptron [35]

Σε μοντέλα υψηλότερης πολυπλοκότητας η χρήση ενός μόνο perceptron είναι αδύνατη για αυτό και η μεταχείριση ενός πολύ στρωματικού μοντέλου καθίσταται απαραίτητη. Η κατασκευή των μοντέλων αυτών παρέχεται μέσω του συνδυασμού πολλαπλών perceptron, των οποίων η δομή παρουσιάζεται σε τρία επίπεδα. Ειδικότερα, παρέχεται ένα στρώμα δεδομένων εισόδου (input layer) το οποίο διανέμει τα χαρακτηριστικά στο πρώτο κρυφό στρώμα [36]. Στη συνέχεια, ένα ή περισσότερα κρυφά στρώματα (hidden layer) perceptron λαμβάνουν ως είσοδο τα χαρακτηριστικά του προηγούμενου στρώματος και ένα στρώμα εξόδου (output layer) δέχεται ως είσοδο την έξοδο κάθε perceptron του τελευταίου κρυφού στρώματος. Τα perceptron που χρησιμοποιούνται μέσω του πολύ στρωματικού ταξινομητή perceptron εκμεταλλεύονται πολλούς διαφορετικούς τύπους συναρτήσεων ενεργοποίησης των οποίων η επιλογή εξαρτάται από το πλαίσιο χρήσης τους. Η μη γραμμικότητα αποτελεί έναν πολύ σημαντικό παράγοντα για το μοντέλο ,αφού στοχεύει στην εκμάθηση πολύπλοκων μοτίβων για την επίλυση προβλημάτων πραγματικού χρόνου. Η απόδοση ενός μοντέλου νευρωνικού δικτύου ποικίλλει σημαντικά ανάλογα με τον τύπο της συνάρτησης ενεργοποίησης που χρησιμοποιείται στα κρυφά επίπεδα και στα επίπεδα εξόδου. Οι συναρτήσεις ενεργοποίησης που χρησιμοποιούνται στο πλαίσιο του μοντέλου αυτού είναι οι Sigmoid, Tanh, ReLU, Leaky ReLU, Parametric ReLU (PReLU), Softmax, Binary step, Identity, Swish. Η μάθηση στα πολύ στρωματικά perceptron προβλέπει επίσης την προσαρμογή των βαρών των perceptron ώστε το σφάλμα στα δεδομένα εκπαίδευσης να είναι χαμηλό και η γενίκευση σε μελλοντικά δεδομένα να είναι υψηλότερη. Αυτό επιτυγχάνεται μέσω της χρήσης του αλγορίθμου back propagation.



Εικόνα 12 - Παράδειγμα Νευρωνικού Δικτύου [37]

Στο παραπάνω κεφάλαιο αναφέρθηκαν όλα τα βήματα τα οποία συνθέτουν την εκπαίδευση ενός μοντέλου μηχανικής μάθησης. Η εκπαίδευση μοντέλου περιλαμβάνει τον καλύτερο συνδυασμό βαρών ώστε να ελαχιστοποιηθούν οι απώλειες με αποτέλεσμα η πρόβλεψη αλλά και η ταξινόμηση των δεδομένων να πραγματοποιηθεί με την μεγαλύτερη δυνατή επιτυχία. Τα βήματα τα οποία αναφέρθηκαν αφορούν την εξαγωγή αριθμητικών χαρακτηριστικών μέσω του συνόλου δεδομένων και της τροφοδότησης των αριθμητικών στοιχείων σε αλγορίθμους μηχανικής μάθησης. Στο επόμενο κεφάλαιο αναφέρονται τα αποτελέσματα των αλγορίθμων καθώς και οι μέθοδοι που οδήγησαν στην επαλήθευση τους.

4 - Αξιολόγηση μοντέλων και επιδόσεις

Σε αυτό το κεφάλαιο θα αναλυθούν εκτενώς τα αποτελέσματα που προέκυψαν, οι μετρικές και τα εργαλεία που χρησιμοποιήθηκαν με στόχο την απόδειξη της αξιοπιστίας των αποτελεσμάτων. Είναι σημαντικό πριν αναφερθούν όλες οι μετρικές που χρησιμοποιήθηκαν να αναλυθούν τα στοιχεία που τις συνθέτουν. Τα στοιχεία αυτά αφορούν τις εξής τέσσερις μετρικές: “True Positive“, “True Negative“, “False Positive“, “False Negative“ [38]. Η χρήση της “True” αφορά τις περιπτώσεις που η πρόβλεψη των αποτελεσμάτων είναι έγκυρη ενώ η ”False” το αντίθετο ,ενώ οι τιμές των ”Positive” και ”Negative” αναφέρονται στις πραγματικές τιμές των αποτελεσμάτων μας. Η ”True Positive” δηλώνει το πλήθος των περιπτώσεων ορθής ταξινόμησης, ενός θετικού παραδείγματος στην κατηγορία των θετικών παραδειγμάτων. Ακόμα, η ”True Negative” δηλώνει το πλήθος περιπτώσεων ορθής ταξινόμησης ενός αρνητικού παραδείγματος στην κατηγορία των αρνητικών παραδειγμάτων. Έπειτα, η ”False Negative” δηλώνει το πλήθος των περιπτώσεων εσφαλμένης ταξινόμησης, ενός θετικού παραδείγματος στην κατηγορία των αρνητικών παραδειγμάτων. Τέλος, η ”False Positive” δηλώνει το πλήθος των περιπτώσεων εσφαλμένης ταξινόμησης ,ενός αρνητικού παραδείγματος στην κατηγορία των θετικών παραδειγμάτων. Οι τέσσερις μετρικές αυτές συνθέτουν τον πίνακα σύγχυσης (confusion matrix) που αποτελεί ένα εργαλείο μέτρησης της αποδοτικότητας των αποτελεσμάτων των

αλγορίθμων μηχανικής μάθησης. Οι τιμές της διαγώνιου αφορούν αυτές των true positive και true negative και η υψηλή τιμή τους προδιαθέτει ένα μοντέλο υψηλής ακρίβειας. Έπειτα, μετά την επεξήγηση των μεθόδων αυτών μπορούμε να αναφερθούμε σε άλλες μετρικές που οδήγησαν στην επιλογή των καλύτερων μοντέλων μηχανικής μάθησης [39] [16].

Η πρώτη μετρική αφορά την ακρίβεια (accuracy) που ορίζεται ως το ποσοστό σωστών προβλέψεων που πραγματοποιούνται από το μοντέλο ταξινόμησης [40]. Παρακάτω παρατίθεται ο τύπος σύμφωνα με τον οποίο εφαρμόζεται.

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{True\ Negatives} + \boldsymbol{True\ Positives}}{\boldsymbol{True\ Positives} + \boldsymbol{False\ Positive} + \boldsymbol{True\ Negative} + \boldsymbol{False\ Negative}}$$

Η επόμενη μετρική που θα αναφερθεί είναι αυτή του precision (ακρίβεια) ,και αναφέρεται στον λόγο των σωστών θετικών προβλέψεων προς τις συνολικές θετικές προβλέψεις [41]. Παρακάτω φαίνεται ο τύπος που αναφέρεται στην μετρική αυτή.

$$\boldsymbol{Precision} = \frac{\boldsymbol{True\ Positive}}{\boldsymbol{True\ Positive} + \boldsymbol{False\ Positive}}$$

Στη συνέχεια, η μετρική της ανάκλησης (recall) παρουσιάζει την αναλογία των σωστών θετικών προβλέψεων προς τον συνολικό αριθμό θετικών περιπτώσεων στο σύνολο δεδομένων. Ο τύπος της ανάκλησης είναι ο εξής.

$$\boldsymbol{Recall} = \frac{\boldsymbol{True\ Positive}}{\boldsymbol{True\ Positive} + \boldsymbol{False\ Negative}}$$

Η ταυτόχρονη αύξηση των μετρικών recall και precision είναι αδύνατη αφού οι δυο τιμές είναι αντιστρόφως ανάλογες μεταξύ τους. Σε αυτή την περίπτωση, χρησιμοποιούμε την μετρική του f1-score ως τον αρμονικό μέσο όρο μεταξύ της ακρίβειας και της ανάκλησης [42]. Ο τύπος που την συνθέτει είναι ο εξής.

$$\boldsymbol{F1 - Score} = \frac{\boldsymbol{2}}{\frac{\boldsymbol{1}}{\boldsymbol{precision}} + \frac{\boldsymbol{1}}{\boldsymbol{recall}}}$$

Παρακάτω, παρουσιάζονται τα αποτελέσματα των μετρικών της διπλωματικής που βοήθησαν στην επιλογή των κατάλληλων αλγορίθμων για την μετέπειτα πρόβλεψη των νέων δεδομένων. Όπως παρατηρείται στις μετρικές συμπεριλήφθηκαν οι τιμές των αποτελεσμάτων με βάση την χρήση της διασταυρωμένης επικύρωσης, που προβλέπει την περαιτέρω επαλήθευση των αποτελεσμάτων.

Τα δεδομένα που χρησιμοποιήθηκαν για την εκπαίδευση και αξιολόγηση των μοντέλων μηχανικής μάθησης που δοκιμάστηκαν συλλέχθηκαν από το investing.com. Το Investing.com αποτελεί μια πλατφόρμα χρηματοοικονομικών αγορών που παρέχει δεδομένα σε πραγματικό χρόνο, τιμές, γραφήματα, χρηματοοικονομικά εργαλεία, έκτακτες ειδήσεις και αναλύσεις για 250 χρηματιστήρια σε όλο τον κόσμο ενώ παρέχει το περιεχόμενο του σε 44 γλώσσες.

4.1 Αγορά μετοχών

Στους παρακάτω δύο πίνακες παρουσιάζονται τα αποτελέσματα για τα δεδομένα της κατηγορίας αγοράw μετοχών. Στον πρώτο πίνακα παρουσιάζονται οι μετρικές των μοντέλων που αναλύθηκαν στην αρχή του κεφαλαίου 5 ενώ χρησιμοποιείται η μέθοδος TF-IDF για την εξαγωγή των χαρακτηριστικών των δεδομένων. Στον δεύτερο πίνακα παρουσιάζονται τα ίδια μοντέλα ανάλυσης του κεφαλαίου 5 με την διαφορά ότι χρησιμοποιείται η μέθοδος της μέτρησης διανύσματος. Για την εκπαίδευση των μοντέλων χρησιμοποιήθηκαν 200 άρθρα τα οποία χαρακτηρίστηκαν σε τρεις κλάσεις (αρνητικό συναίσθημα, ουδέτερο συναίσθημα και θετικό συναίσθημα), ενώ 200 άρθρα χρησιμοποιήθηκαν για την αξιολόγηση των μοντέλων.

Table 1 - TF-IDF (Αγορά μετοχών)

TF-IDF	Ακρίβεια δεδομένων εκπαίδευσης	Ακρίβεια δεδομένων δοκιμών	F1-σκορ	Ακρίβεια (Precision)	Ανάκληση	Ακρίβεια (Cross-Val)	F1-σκορ (Cross-Val)	Ακρίβεια-Precision(Cross-Val)	Ανάκληση(Cross-Val)
Δέντρα Απόφασης	1.00	0.65714	0.529	0.528	0.533	0.6	0.4831	0.4782	0.5065
Αφελής κατηγοριοποιητής Bayes	0.836	0.6428	0.445	0.513	0.517	0.665	0.4859	0.4494	0.5377
K-κοντινότερων γειτόνων	0.71	0.621428	0.552	0.547	0.583	0.55	0.4270	0.4504	0.4560

Ανάλυση γραμμικού διαχωρισμού	0.71	0.57142	0.487	0.459	0.550	0.545	0.4085	0.4367	0.4452
Στοχαστική κλίση καθόδου	1.00	0.7142	0.598	0.619	0.617	0.67	0.5456	0.6002	0.5604
Διανύσματα μηχανών υποστήριξης	1.00	0.7142	0.486	0.537	0.550	0.69	0.5325	0.5279	0.5687
Πολύστρωματικός ταξινομητής perceptron	1.00	0.6785	0.594	0.587	0.617	0.65	0.5033	0.5412	0.5361

Table 2 - Count-Vectorizer (Αγορά μετοχών)

Count-V	Ακρίβεια δεδομένων εκπαίδευσης	Ακρίβεια δεδομένων δοκιμών	F1-σκορ	Ακρίβεια (Precision)	Ανάκληση	Ακρίβεια (Cross-Val)	F1-σκορ (Cross-Val)	Ακρίβεια-Precision(Cross-Val)	Ανάκληση(Cross-Val)
Δέντρα Απόφασης	1.00	0.65714	0.490	0.482	0.500	0.57	0.4979	0.5016	0.5032
Αφελής κατηγοριοποιητής Bayes	1.00	0.62666	0.385	0.371	0.400	0.575	0.4843	0.5781	0.4914
K-κοντινότερων γειτόνων	0.60	0.55714	0.466	0.434	0.517	0.58	0.4599	0.62474	0.4863

Ανάλυση γραμμικού διαχωρισμού	0.71	0.571428	0.487	0.459	0.550	0.495	0.3739	0.4507	0.4109
Στοχαστική κλίση καθόδου	1.00	0.70	0.604	0.623	0.600	0.625	0.5661	0.5904	0.5776
Διανύσματα μηχανών υποστήριξης	1.00	0.71428	0.353	0.415	0.450	0.675	0.5927	0.6095	0.6006
Πολύ στρωματικός ταξινομητής perceptron	1.00	0.7285	0.602	0.560	0.650	0.66	0.5105	0.5495	0.5449

Βάση των παραπάνω πινάκων, φαίνεται ότι το μοντέλο που εμφανίζει την υψηλότερη ακρίβεια (precision) είναι ο πολύ στρωματικός ταξινομητής perceptron.

4.2 Αγορά κρυπτό νομισμάτων

Όπως και παραπάνω στους δύο επόμενους πίνακες παρουσιάζονται τα αποτελέσματα και οι μετρικές για την κατηγορία των κρυπτό νομισμάτων. Αρχικά, εμφανίζεται η μέθοδος του TF-IDF που εισάγει δεδομένα στους αλγορίθμους μηχανικής μάθησης που αναφέρθηκαν στο κεφάλαιο 5. Στη συνέχεια, ο επόμενος πίνακας αφορά την μέθοδο της μέτρησης διανύσματος που αποτελεί μια ακόμα τεχνική εξαγωγής αριθμητικών χαρακτηριστικών. Για να πραγματοποιηθεί εκπαίδευση στα μοντέλα μηχανικής μάθησης έγινε χρήση περίπου 200 άρθρων που χαρακτηρίστηκαν με βάση την πολικότητα που αναφέρεται παραπάνω, ενώ ίσος αριθμός από άρθρα χρησιμοποιήθηκαν για την αξιολόγηση των μοντέλων.

Table 3 - TF-IDF (Αγορά κρυπτό νομισμάτων)

TF-IDF	Ακρίβεια δεδομένων εκπαίδευσης	Ακρίβεια δεδομένων δοκιμών	F1-σκορ	Ακρίβεια (Precision)	Ανάκληση	Ακρίβεια (Cross-Val)	F1-σκορ (Cross-Val)	Ακρίβεια-Precision(Cross-Val)	Ανάκληση(Cross-Val)

Δέντρα Απόφασης	1.00	0.57	0.508	0.512	0.519	0.49	0.472	0.568	0.496
Αφελής κατηγοριοποιητής Bayes	0.946	0.44	0.285	0.416	0.370	0.48	0.448	0.518	0.484
Κ-κοντινότερων γειτόνων	0.81	0.5054	0.376	0.435	0.395	0.414	0.37	0.502	0.422
Ανάλυση γραμμικού διαχωρισμού	0.97	0.53	0.502	0.518	0.506	0.47	0.462	0.5056	0.47
Στοχαστική κλίση καθόδου	0.995	0.67	0.375	0.474	0.444	0.54	0.534	0.544	0.448
Διανύσματα μηχανών υποστήριξης	0.98	0.68	0.531	0.533	0.531	0.538	0.532	0.546	0.536
Πολύστρωματικός ταξινομητής perceptron	1.00	0.6669	0.519	0.521	0.519	0.574	0.56	0.572	0.5688

Table 4 - Count-Vectorizer (Αγορά κρυπτό νομισμάτων)

Count-V	Ακρίβεια δεδομένων εκπαίδευσης	Ακρίβεια δεδομένων δοκιμών	F1-σκορ	Ακρίβεια (Precision)	Ανάκληση	Ακρίβεια (Cross-Val)	F1-σκορ (Cross-Val)	Ακρίβεια-Precision(Cross-Val)	Ανάκληση(Cross-Val)
Δέντρα Απόφασης	1.00	0.56	0.422	0.448	0.432	0.488	0.45	0.52	0.4864

Αφελής κατηγοριοποιητής Bayes	0.98	0.595	0.506	0.512	0.507	0.5178	0.5	0.508	0.509
K-κοντινότερων γειτόνων	0.81	0.5054	0.376	0.435	0.395	0.4182	0.3738	0.5054	0.422
Ανάλυση γραμμικού διαχωρισμού	0.97	0.51	0.398	0.413	0.395	0.448	0.4449	0.462	0.446
Στοχαστική κλίση καθόδου	1.00	0.6349	0.515	0.515	0.519	0.53	0.524	0.542	0.53
Διανύσματα μηχανών υποστήριξης	0.978	0.6087	0.481	0.479	0.481	0.552	0.5462	0.5489	0.5471
Πολύ στρωματικός ταξινομητής perceptron	1.00	0.667	0.518	0.518	0.518	0.558	0.562	0.57	0.466

Βάση των παραπάνω πινάκων, φαίνεται ότι το μοντέλο που εμφανίζει την υψηλότερη ακρίβεια (precision) είναι ο πολύ στρωματικός ταξινομητής perceptron.

4.3 Παγκόσμιο εμπόριο

Ως προέκταση των δύο παραπάνω κατηγοριών έχουμε το παγκόσμιο εμπόριο που αφορά το παγκόσμιο εμπόριο. Παρακάτω εμφανίζονται οι δυο πίνακες με τις μετρικές των μοντέλων που αναλύθηκαν εκτενώς στο κεφάλαιο 5. Η ειδοποιός διαφορά των πινάκων έγκειται στην μέθοδο εξαγωγής χαρακτηριστικών. Συγκεκριμένα, στο πρώτο πίνακα έχουμε την μέθοδο του TF-IDF, ενώ στον δεύτερο παρουσιάζεται η μέθοδος της μέτρησης διανύσματος. Για την αξιόπιστη εξαγωγή μετρικών με στόχο την πρόβλεψη νέων δεδομένων χρησιμοποιήθηκαν περισσότερα από 200 άρθρα και ίσος αριθμός άρθρων για την αξιολόγηση των μοντέλων.

Table 5 - TF-IDF (Παγκόσμιο εμπόριο)

TF-IDF	Ακρίβεια δεδομένων εκπαίδευσης	Ακρίβεια δεδομένων δοκιμών	F1-σκορ	Ακρίβεια (Precision)	Ανάκληση	Ακρίβεια (Cross-Val)	F1-σκορ (Cross-Val)	Ακρίβεια-Precision(Cross-Val)	Ανάκληση(Cross-Val)
Δέντρα Απόφασης	1.00	0.594	0.519	0.522	0.523	0.51	0.502	0.516	0.502
Αφελής κατηγοριοποιητής Bayes	1.00	0.591	0.586	0.594	0.589	0.602	0.592	0.622	0.598
K-κοντινότερων γειτόνων	0.69	0.6	0.526	0.567	0.536	0.572	0.562	0.584	0.57
Ανάλυση γραμμικού διαχωρισμού	0.87	0.5	0.441	0.439	0.444	0.44	0.422	0.434	0.438
Στοχαστική κλίση καθόδου	1.00	0.63	0.593	0.598	0.596	0.592	0.58	0.596	0.588
Διανύσματα μηχανών υποστήριξης	1.00	0.628	0.614	0.638	0.623	0.62	0.6152	0.638	0.616
Πολύστρωματικός ταξινομητής perceptron	1.00	0.61	0.603	0.603	0.603	0.6	0.596	0.6	0.596

Table 6 - Count-V (Παγκόσμιο εμπόριο)

Count-V	Ακρίβεια δεδομένων εκπαίδευσης	Ακρίβεια δεδομένων δοκιμών	F1-σκορ	Ακρίβεια (Precision)	Ανάκληση	Ακρίβεια (Cross-Val)	F1-σκορ (Cross-Val)	Ακρίβεια-Precision(Cross-Val)	Ανάκληση(Cross-Val)
Δέντρα Απόφασης	1.00	0.5485	0.533	0.536	0.543	0.52	0.48	0.58	0.62
Αφελής κατηγοριοποιητής Bayes	0.99	0.5457	0.548	0.562	0.543	0.504	0.45	0.508	0.50
K-κοντινότερων γειτόνων	0.45	0.42	0.303	0.315	0.404	0.392	0.332	0.476	0.39
Ανάλυση γραμμικού διαχωρισμού	0.95	0.49	0.436	0.446	0.457	0.428	0.392	0.422	0.424
Στοχαστική κλίση καθόδου	1.00	0.60	0.576	0.576	0.576	0.604	0.6	0.61	0.6
Διανύσματα μηχανών υποστήριξης	0.934	0.6	0.506	0.513	0.523	0.597	0.584	0.62	0.588
Πολύ στρωματικός ταξινομητής perceptron	1.00	0.623	0.621	0.622	0.623	0.574	0.574	0.586	0.574

Βάση των παραπάνω πινάκων, φαίνεται ότι το μοντέλο που εμφανίζει την υψηλότερη ακρίβεια (precision) είναι ο πολύ στρωματικός ταξινομητής perceptron.

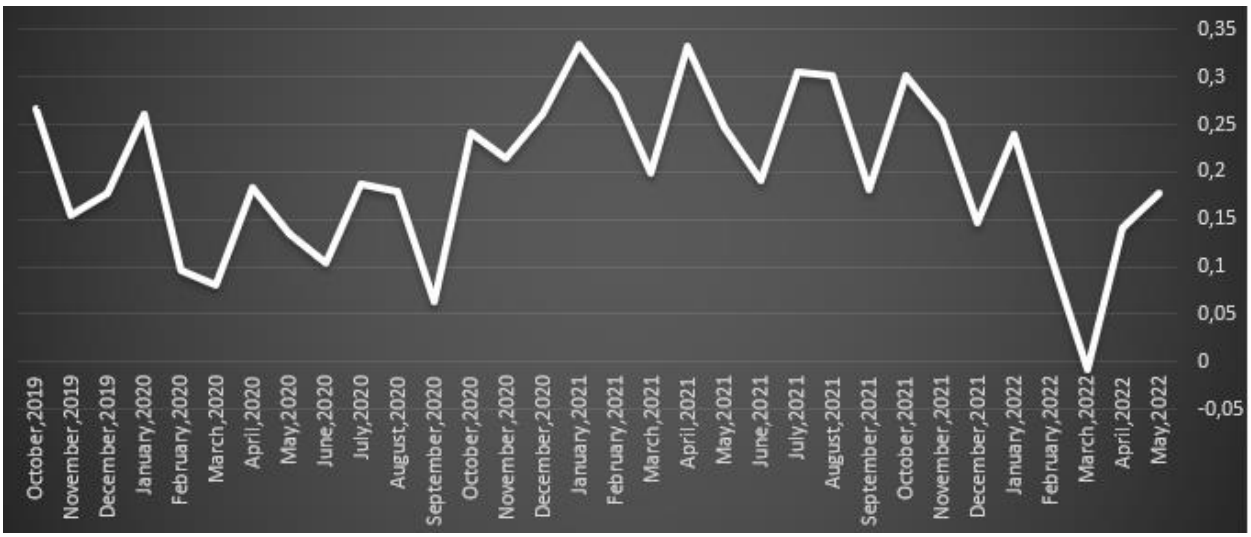
5 - Δείκτες συναισθήματος στον τομέα των οικονομικών

Σε αυτό το κεφάλαιο, γίνεται αναφορά στην απόξεση δεδομένων (data scraping) ,καθώς και σε όλη την διαδικασία και βήματα που ακολουθούνται προκειμένου να πραγματοποιηθεί πρόβλεψη συναισθήματος σε κείμενα οικονομικού χαρακτήρα μέσω αλγορίθμων μηχανικής μάθησης [43]. Η απόξεση δεδομένων αναφέρεται στην διαδικασία μέσω της οποίας ένα πρόγραμμα υπολογιστή εξάγει δεδομένα μέσω ενός άλλου προγράμματος, συνήθως μιας ιστοσελίδας. Στην συγκεκριμένη διπλωματική εργασία η τεχνική του data scraping εφαρμόζεται σε άρθρα οικονομικού χαρακτήρα με τελικό στόχο την εξαγωγή συναισθήματος. Ειδικότερα, τα άρθρα αυτά μετά την απόξεση τους τοποθετούνται σε ένα αρχείο φύλλο μετρητή όπου μέσω εφαρμογής διάφορων τεχνικών επεξεργασίας προκύπτουν οι αναγκαίοι δείκτες συναισθήματος. Αρχικά, εφαρμόζονται τεχνικές προ επεξεργασίας στα ακατέργαστα δεδομένα που εξήχθησαν μέσω του scraping. Η προ επεξεργασία δεδομένων αποτελεί παραδοσιακά ένα προκαταρκτικό βήμα υψηλής σημασίας για την διαδικασία εξόρυξης δεδομένων. Οι τεχνικές και τα εργαλεία που εφαρμόστηκαν σε αυτό το βήμα περιλαμβάνουν τον καθαρισμό κειμένου από σημεία στίξης, την μετατροπή όλων των γραμμάτων σε πεζά, την αφαίρεση λέξεων ελάσσονας σημασίας και την αναγωγή λέξεων στο αρχικό τους θέμα. Ακόμα, εκτελείται η τεχνική της εξαγωγής χαρακτηριστικών (feature extraction) κατά την οποία μετατρέπονται ακατέργαστα δεδομένα σε αριθμητικά στοιχεία τα οποία μπορούν να χρησιμοποιηθούν από τα μοντέλα μηχανικής μάθησης. Συγκεκριμένα, εφαρμόζονται οι τεχνικές των TF-IDF και της μέτρησης διανύσματος οι οποίες έχουν αναλυθεί εκτενώς σε προηγούμενο κεφάλαιο. Στο σημείο αυτό να αναφέρουμε ότι δεν επαναλαμβάνεται η εκπαίδευση των μοντέλων μηχανικής μάθησης, επομένως τα επεξεργασμένα στοιχεία των νέων δεδομένων εφαρμόζονται στο σύνολο εκπαίδευσης (training data) των μοντέλων που είναι ικανά να πραγματοποιήσουν προβλέψεις. Ειδικότερα, χρησιμοποιώντας τις κατάλληλες εντολές της γλώσσας προγραμματισμού Python μέσω χρήσης της βιβλιοθήκης του scikit-learn προκύπτουν οι αναμενόμενοι δείκτες συναισθήματος.

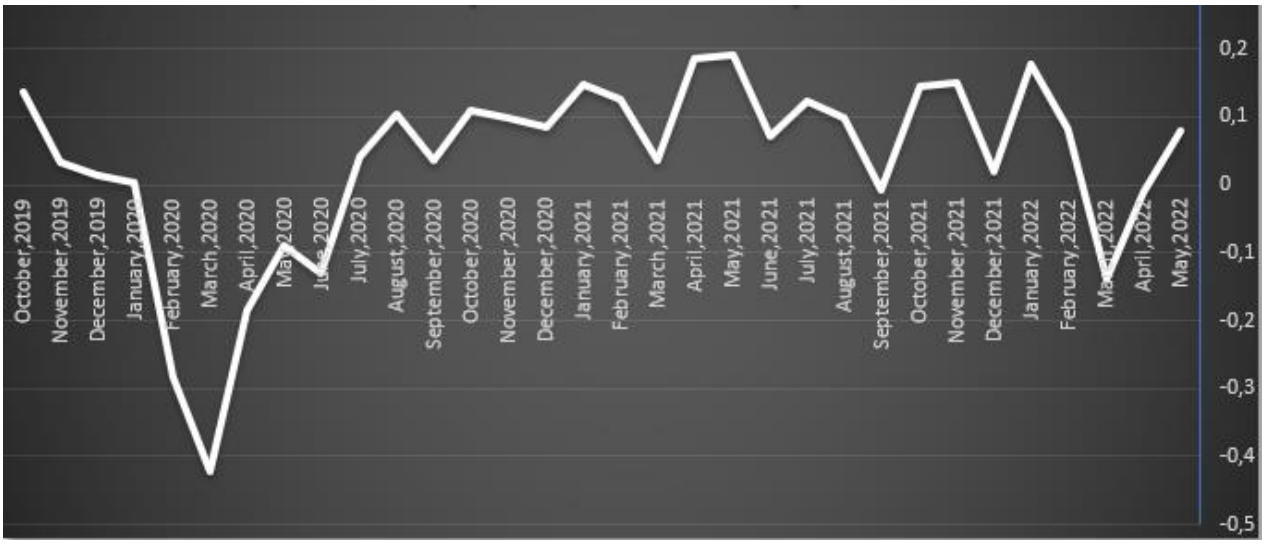
Στις παρακάτω γραφικές παραστάσεις αναπαρίστανται τα μοντέλα μηχανικής μάθησης που χρησιμοποιήθηκαν για την πρόβλεψη των καινούργιων δεδομένων. Συγκεκριμένα, προκύπτει μια μηνιαία αναπαράσταση για ένα διάστημα δυο ετών που αφορά την επίδραση συναισθήματος του Covid-19 και άλλων συμβάντων σε διάφορους τομείς της οικονομίας όπως η αγορά μετοχών (Stock Market), τα κρυπτό νομίσματα (Cryptocurrency) και το εμπόριο (Commodities and Future).

5.1 Αγορά μετοχών

Σε αυτό το σημείο εμφανίζονται οι γραφικές παραστάσεις των μοντέλων μηχανικής μάθησης κάθε κατηγορίας. Συγκεκριμένα, φαίνονται τα μοντέλα με τις υψηλότερες τιμές ακρίβειας για τις δύο μεθόδους εξαγωγής χαρακτηριστικών. Οι τιμές που αναγράφονται αφορούν τους δείκτες συναισθήματος των άρθρων που έχουν προβλεφθεί σε μια διάρκεια δυο ετών με στόχο την παρατήρηση της αλλαγής συναισθήματος της αγοράς με βάση την πανδημία του Covid-19 και άλλων γεγονότων που συμβάλλουν στην αυξομείωση του δείκτη συναισθήματος.

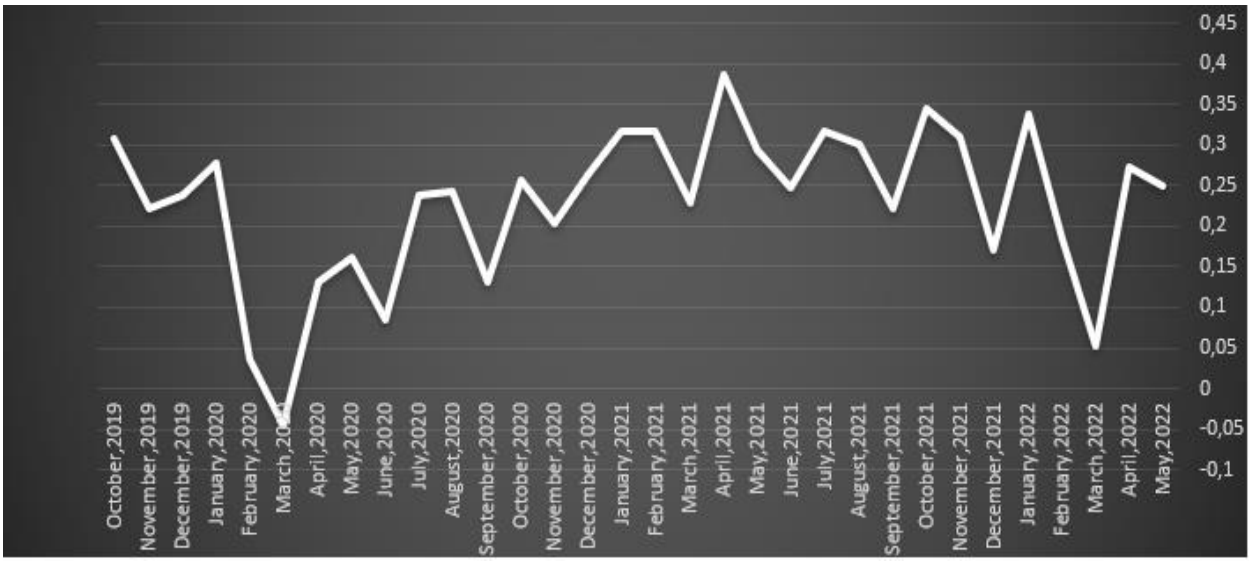


Εικόνα 13 – Ανάλυση συναισθήματος με βάση τον πολύ στρωματικό ταξινομητή perceptron (μέθοδος: Count Vectorizer) στην αγορά μετοχών

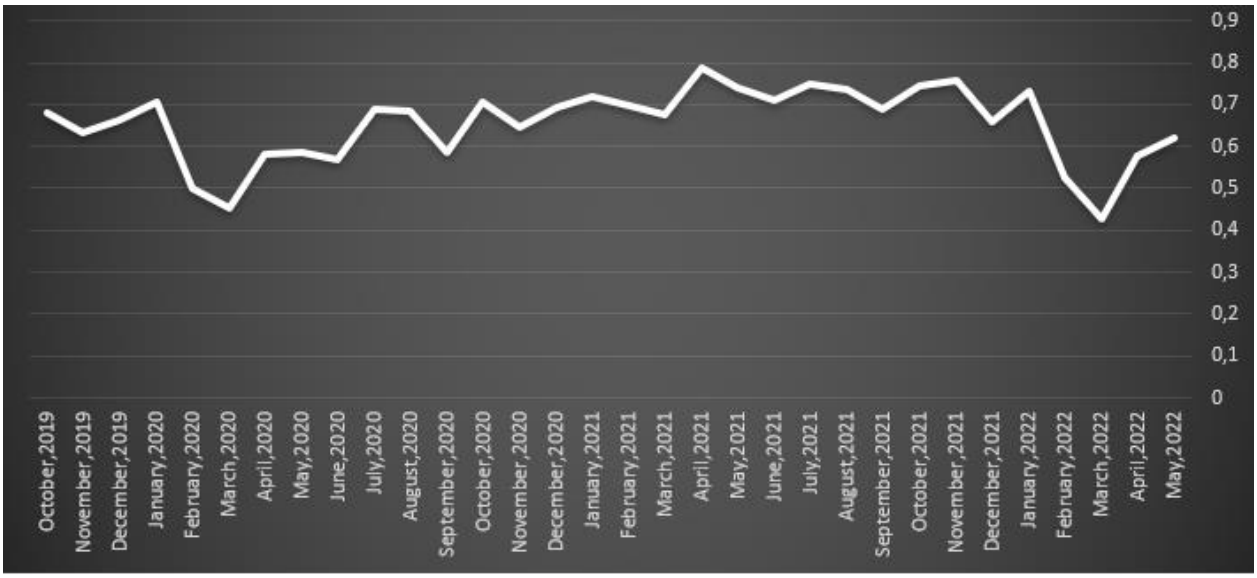


Εικόνα 14 – Ανάλυση συναισθήματος με βάση την Στοχαστική κλίση καθόδου (μέθοδος: Count Vectorizer) στην αγορά μετοχών

Οι παραπάνω γραφικές παραστάσεις υποδεικνύουν την αυξομείωση του δείκτη συναισθήματος της αγοράς μετοχών. Μπορούμε να παρατηρήσουμε και στα δυο γραφήματα υπάρχουν αυξημένες αλλά και μειωμένες τιμές. Ο Μάρτιος του 2020 αποτελεί ένα μήνα αρνητικού συναισθήματος καθώς σηματοδοτεί την αφετηρία της πανδημίας. Ακόμα ,αποτελεί μια ημερομηνία με χαμηλό δείκτη εξαιτίας του πληθωρισμού, των συνεχώς αυξημένων επιτοκίων, της εισβολής στην Ουκρανία καθώς και πιθανής ύφεσης.



Εικόνα 15 - Ανάλυση συναισθήματος με βάση την Στοχαστική κλίση καθόδου (μέθοδος: TF-IDF) στην αγορά μετοχών

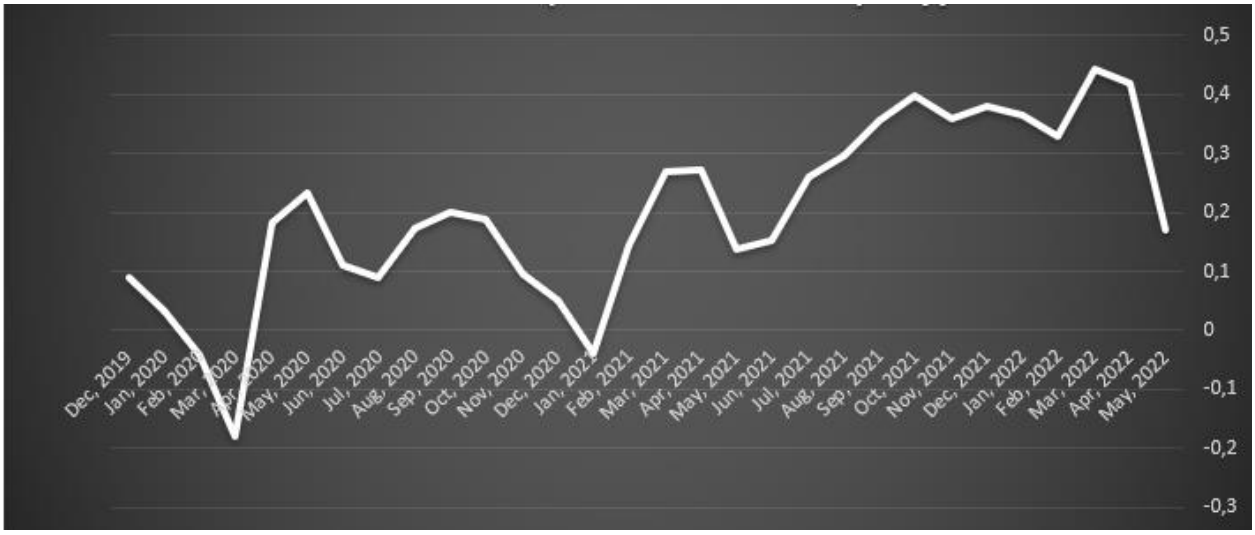


Εικόνα 16- Ανάλυση συναισθήματος με βάση τα Διανύσματα μηχανών υποστήριξης (μέθοδος: TF-IDF) στην αγορά μετοχών

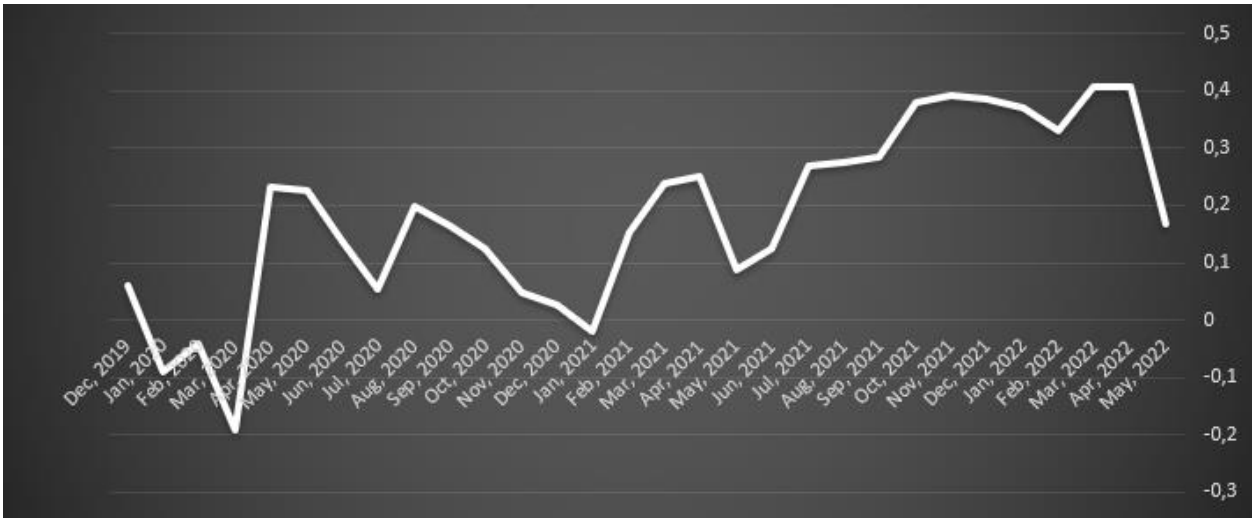
Τα παραπάνω γραφικά σχήματα υποδεικνύουν τις ίδιες μειωμένες τιμές των δεικτών για τους ίδιους λόγους που αναφέρθηκαν παραπάνω. Μπορούμε να παρατηρήσουμε ότι στο τελευταίο σχήμα οι διαφορές των συναισθημάτων δεν είναι ιδιαίτερα έντονες γεγονός που οφείλεται κυρίως στο τρόπο με τον οποίο ταξινομούνται τα δεδομένα σε συνδυασμό με την μέθοδο TF-IDF.

5.2 Κρυπτό νομίσματα

Στο σημείο αυτό, παρουσιάζεται η δεύτερη κατηγορία της διπλωματικής εργασίας τα κρυπτό νομίσματα. Όπως και στην προηγούμενη περίπτωση φαίνονται οι αριθμημένοι δείκτες συναισθήματος που στόχο έχουν να προσδώσουν συναίσθημα στα νέα άρθρα με βάση τους αλγόριθμους μηχανικής μάθησης που παρουσιάζονται. Ο διαχωρισμός των γραφικών παραστάσεων αφορά την χρήση δύο τεχνικών εξαγωγής χαρακτηριστικών τον TF

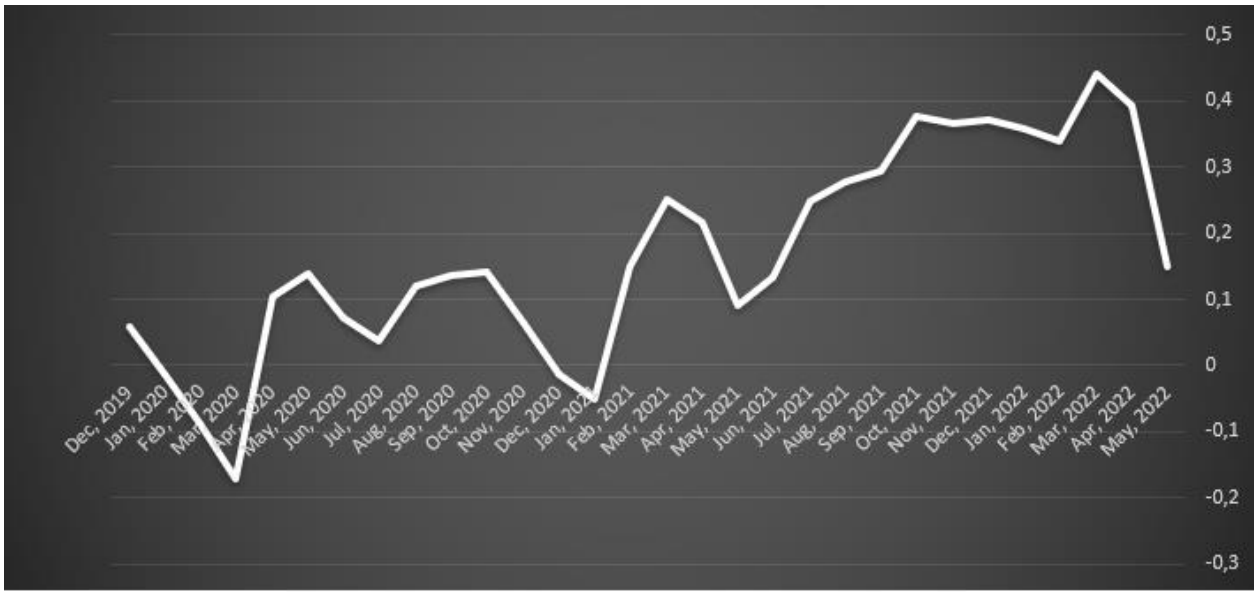


Εικόνα 17- Ανάλυση συναισθήματος με βάση τον πολύ στρωματικό ταξινομητή perceptron (μέθοδος: Count Vectorizer) στα κρυπτονομίσματα

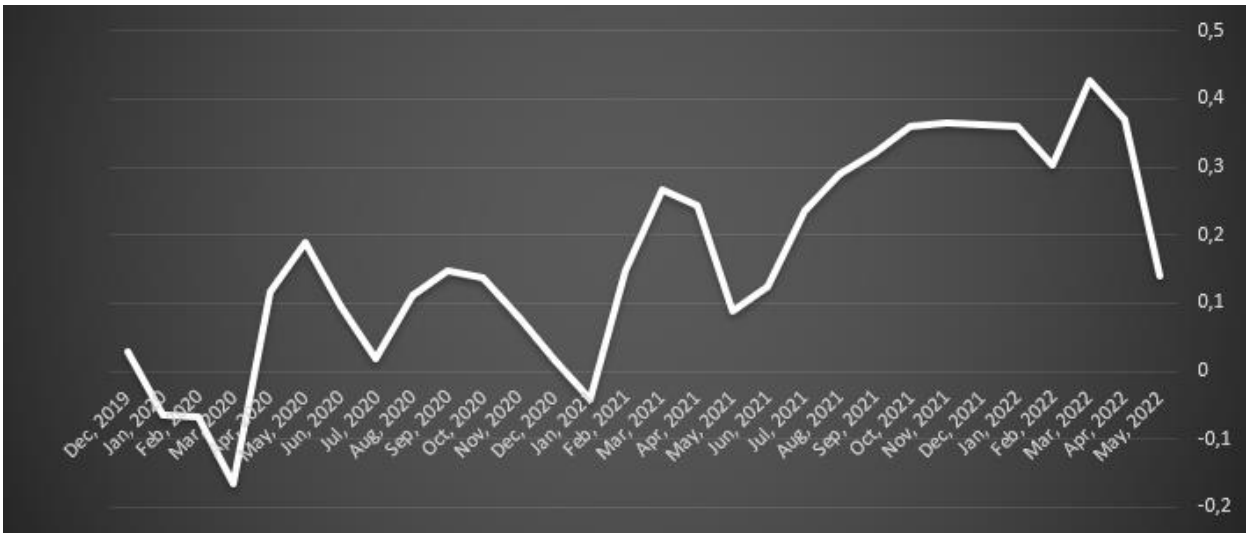


IDF και της μέτρησης διανύσματος. Όπως και προηγουμένως παρατηρείται αυξομείωση του δείκτη συναισθήματος με βάση τα γεγονότα που εκτυλίχθηκαν στα έτη 2019-2022 και σηματοδοτούν την αφετηρία της πανδημίας και άλλων γεγονότων.

Εικόνα 18- Ανάλυση συναισθήματος με βάση την Στοχαστική κλίση καθόδου (μέθοδος: Count Vectorizer) στα κρυπτονομίσματα



Εικόνα 19- Ανάλυση συναισθήματος με βάση την Στοχαστική κλίση καθόδου (μέθοδος: TF-IDF) στα κρυπτονομίσματα



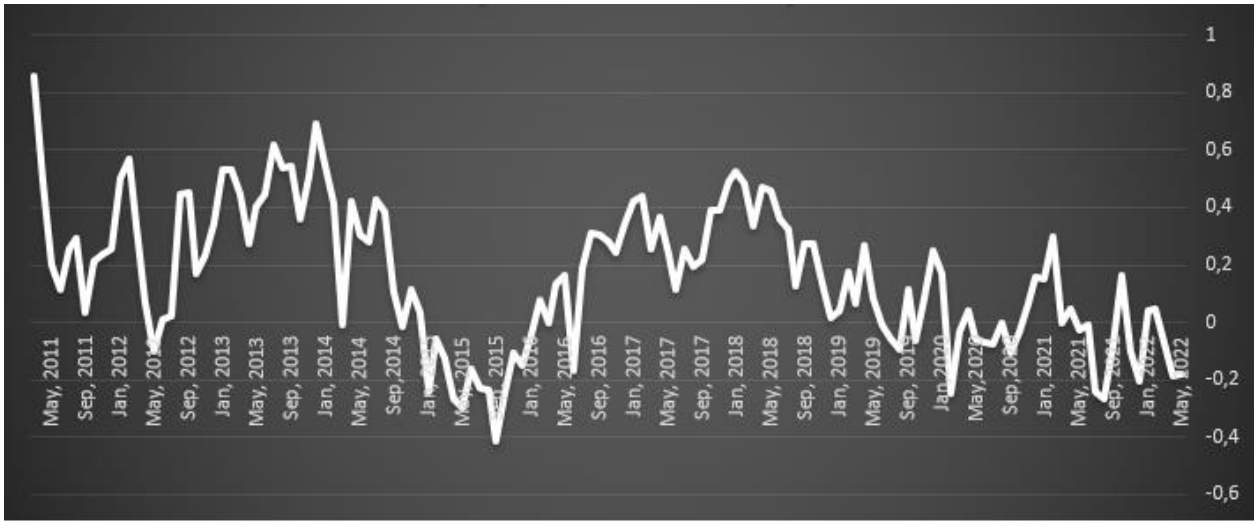
Εικόνα 20- Ανάλυση συναισθήματος με βάση τα Διανύσματα μηχανών υποστήριξης (μέθοδος: TF-IDF) στα κρυπτονομίσματα

Όπως και στην κατηγορία της αγοράς μετοχών οι παραπάνω γραφικές παραστάσεις υποδεικνύουν την αντίδραση των συναισθημάτων της αγοράς σε σχέση με τα γεγονότα που συμβαίνουν. Είναι σημαντικό να παρατηρήσουμε ότι και σε αυτή την κατηγορία κατά τη περίοδο του Μαρτίου 2020, δηλαδή κατά την αφετηρία της πανδημίας, ο δείκτης παρουσιάζεται αρκετά μειωμένος. Κατά την περίοδο του Φεβρουαρίου 2021 η επιθετική άνοδος των επιτοκίων οδήγησε στην μείωση του δείκτη συναισθήματος. Συγκριτικά με την κατηγορία της αγοράς μετοχών η περίοδος του Μαρτίου 2022 δεν θεωρείται μειωμένη συναισθηματικά αφού η εισβολή στην Ουκρανία καθώς και η ύφεση φαίνεται ότι δεν επηρέασαν την αγορά των κρυπτό νομισμάτων.

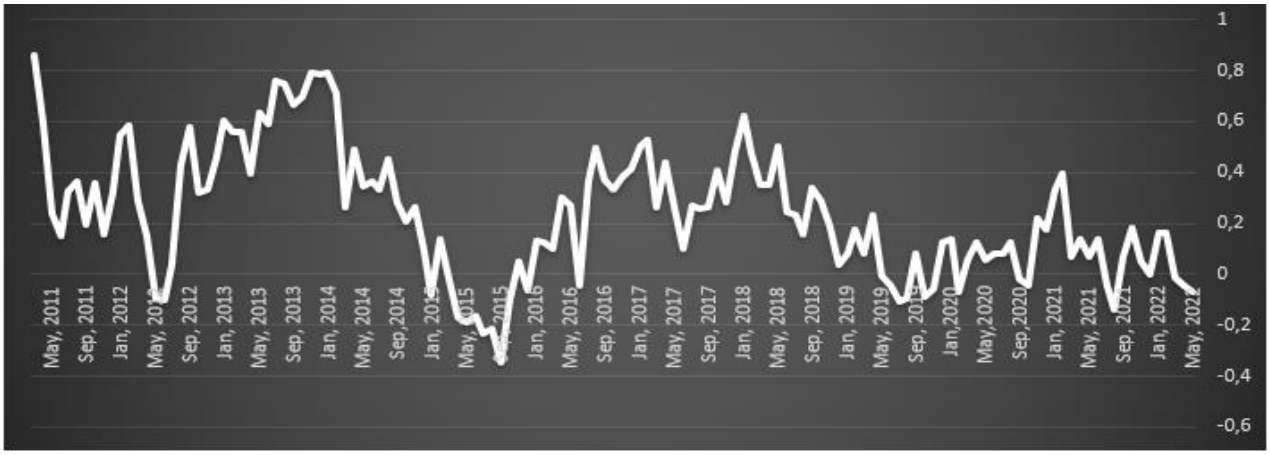
Στις παραπάνω γραφικές παραστάσεις παρουσιάστηκαν οι λόγοι κατά τους οποίους οι δείκτες άλλοτε παρουσιάζουν αυξημένο και άλλοτε μειωμένο δείκτη συναισθήματος. Οι ειδήσεις και τα γεγονότα που καθημερινά συμβαίνουν και δημοσιεύονται είναι ο κύριος λόγος αυξομείωσης των διαγραμμάτων που αναλύονται. Παρακάτω αναλύεται η κατηγορία του παγκόσμιου εμπορίου όπου για διαφορετικούς ή κοινούς λόγους παρουσιάζεται αυξομείωση.

5.3 Παγκόσμιο εμπόριο

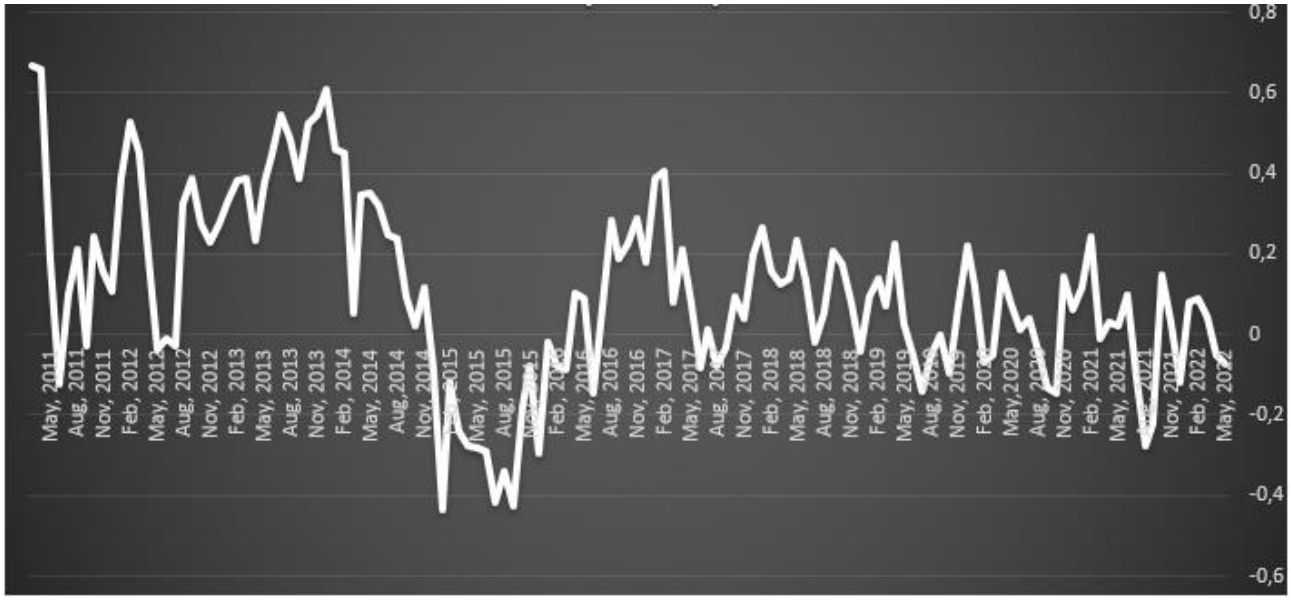
Από την εξίσωση δεν θα μπορούσε να λείπει η κατηγορία του παγκόσμιου εμπορίου. Όπως και παραπάνω παρατηρείται ο διαχωρισμός των γραφικών παραστάσεων με βάση τα μοντέλα μηχανικής μάθησης που παρουσίασαν τα καλύτερα ποσοστά αναλογικά με τις τεχνικές εξαγωγής χαρακτηριστικών. Στην συγκεκριμένη κατηγορία, χρησιμοποιήθηκε δείγμα περισσότερων ετών καθώς τα άρθρα για τα έτη 2019-2022 δεν είναι αρκετά για την εξαγωγή αξιόπιστων δεικτών.



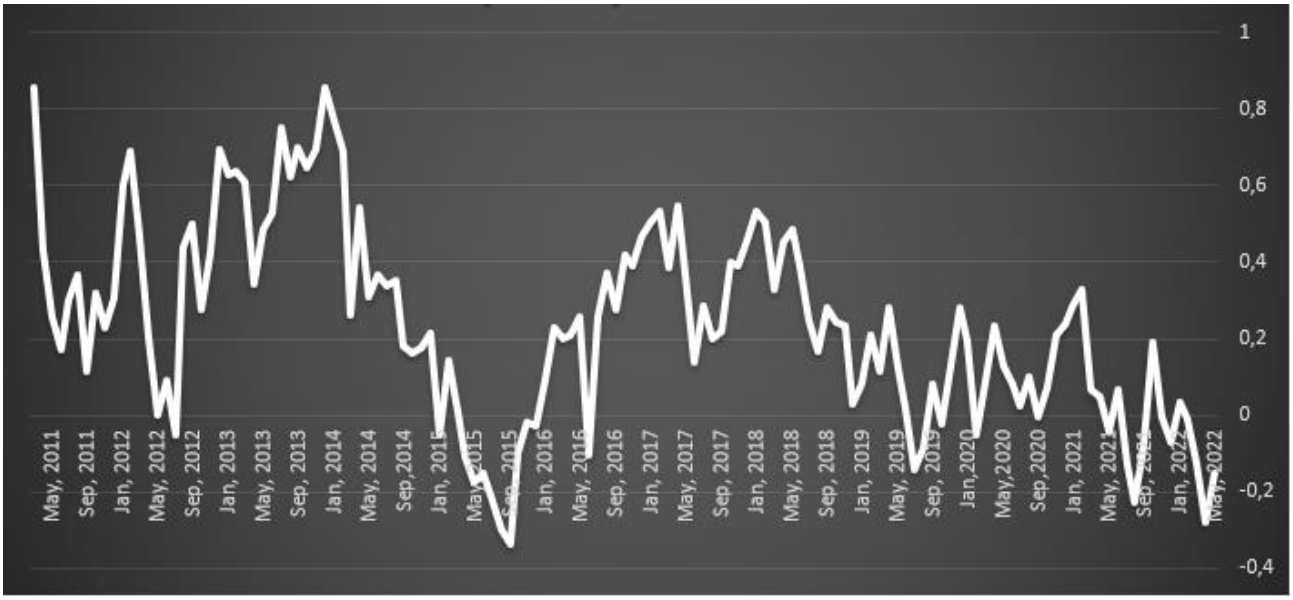
Εικόνα 22 – Ανάλυση συναισθήματος με βάση τον πολύ στρωματικό ταξινομητή perceptron (μέθοδος: Count Vectorizer) στο Commodities and Future



Εικόνα 21 - Ανάλυση συναισθήματος με βάση την Στοχαστική κλίση καθόδου (μέθοδος: Count Vectorizer) στο Commodities and Future



Εικόνα 23- Ανάλυση συναισθήματος με βάση την Στοχαστική κλίση καθόδου (μέθοδος: TF-IDF) στο Commodities and Future



Εικόνα 24 – Ανάλυση συναισθήματος με βάση τα βάση τα Διανύσματα μηχανών υποστήριξης (μέθοδος: TF-IDF) στο Commodities and Future

Στην κατηγορία του παγκόσμιου εμπορίου το εύρος των άρθρων είναι αρκετά μεγαλύτερο εξαιτίας των παραπάνω άρθρων που χρησιμοποιήθηκαν για την σύνθεση την βάσης δεδομένων. Μπορούμε να παρατηρήσουμε ότι και στα 4 σχήματα η περίοδος των δυο ετών 2019- 2022 είναι μειωμένη σε σχέση με παλαιότερα έτη λόγω της αφετηρίας της πανδημίας.

Μία σημαντική πτώση παρουσιάζεται κατά την περίοδο Μαΐου 2015 που οφείλεται στην μείωση των τιμών του πετρελαίου κυρίως λόγω παραγόντων προσφοράς, όπως η άνθηση της παραγωγής πετρελαίου στις ΗΠΑ καθώς και υποχώρηση γεωπολιτικών ανησυχιών.

6 - Αποτελέσματα

Στο κεφάλαιο αυτό, έχοντας ολοκληρώσει την διαδικασία ανάλυσης συναισθήματος για τα άρθρα οικονομικού χαρακτήρα σχετικά με την αγορά μετοχών, τα κρυπτό νομίσματα, και το εμπόριο ,πραγματοποιείται σύγκριση των αποτελεσμάτων που προέκυψαν. Για την κάθε κατηγορία παρουσιάζονται δυο πίνακες με αποτελέσματα με δυο τεχνικές της εξαγωγής χαρακτηριστικών. Στην πρώτη κατηγορία της αγοράς μετοχών και στην τεχνική του TF-IDF καλύτερα ποσοστά για την ακρίβεια και άλλες μετρικές παρέχονται μέσω των αλγορίθμων Στοχαστική κλίση καθόδου (Stochastic Gradient Descent) με ποσοστό ακρίβειας 71,42%, Μηχανές διανυσμάτων υποστήριξης με ποσοστό ακρίβειας 71,42% και πολύ στρωματικού ταξινομητή perceptron με ποσοστό 67,85%. Στις χειρότερες μεθόδους ακολουθούν τα Δένδρα απόφασης (decision tree) με ακρίβεια 65,71 % , Αφελής ταξινομητής Bayes με ακρίβεια 64,28%, k-κοντινότερων γειτόνων και Γραμμική Διακριτική Ανάλυση με ποσοστά 62,14% και 57,14% αντίστοιχα. Αναφερόμενοι στην μέθοδο της μέτρησης διανύσματος καλύτερα αποτελέσματα παρουσιάζουν οι ίδιες τεχνικές δηλαδή η Στοχαστική κλίση καθόδου (Stochastic Gradient Descent) με ακρίβεια 70% ο SVC με ακρίβεια 71,42% και ο πολύ στρωματικός ταξινομητής perceptron με ακρίβεια 72,85%. Στους χειρότερους αλγορίθμους μηχανικής μάθησης της κατηγορίας ανήκουν τα Δένδρα απόφασης (Decision trees) με ακρίβεια 65,71%, ο Αφελής ταξινομητής Bayes με ακρίβεια 62,66%, ο αλγόριθμος k-κοντινότερων γειτόνων με ακρίβεια 55,71% και η Γραμμική Διακριτική Ανάλυση με ποσοστό 57,14%. Επίσης, κρίνεται αναγκαίο να παρουσιαστούν τα ποσοστά της επόμενης κατηγορίας που αφορά τα κρυπτό νομίσματα (cryptocurrency). Στην μέθοδο TF-IDF της εξαγωγής χαρακτηριστικών για τα κρυπτό νομίσματα καλύτερα ποσοστά παρουσιάζονται μέσω των αλγορίθμων Στοχαστική κλίση καθόδου (Stochastic Gradient Descent) με ακρίβεια 67%, Μηχανές διανυσμάτων υποστήριξης με ακρίβεια 68% και πολύ στρωματικός ταξινομητής perceptron με ακρίβεια 66,69%. Τα χειρότερα ποσοστά ακρίβειας παρουσιάζονται μέσω των αλγορίθμων δένδρου απόφασης (Decision trees) με 57%, Αφελής ταξινομητής Bayes με ακρίβεια 44%, k-κοντινότερων γειτόνων και Γραμμική Διακριτική Ανάλυση με ποσοστά 50,54% και 53% αντίστοιχα. Ακόμα, για την μέθοδο της μέτρησης διανύσματος οι ίδιοι αλγόριθμοι παρουσιάζουν τα καλύτερα ποσοστά με ακρίβεια 63,49% για την Στοχαστική κλίση καθόδου (Stochastic Gradient Descent), 60,87% για τις Μηχανές διανυσμάτων υποστήριξης και 66,7% για τον πολύ στρωματικό ταξινομητής perceptron. Τα χειρότερα ποσοστά παρέχονται μέσω των αλγορίθμων δένδρου απόφασης με ποσοστό 56%, Αφελής ταξινομητής Bayes με ποσοστό 59,5%, k-κοντινότερων γειτόνων με ποσοστό 50,54% και Γραμμική Διακριτική Ανάλυση με ποσοστό 51%. Στην κατηγορία του παγκόσμιου εμπορίου καλύτερα ποσοστά παρέχονται στην TF-IDF μέθοδο μέσω της Στοχαστικής κλίσης καθόδου (Stochastic Gradient Descent) με ποσοστό 63%, Μηχανές διανυσμάτων υποστήριξης με 62,8% και πολύ στρωματικό ταξινομητή perceptron με ποσοστό 61%. Τα χειρότερα ποσοστά της

κατηγορίας παρουσιάζονται μέσω των αλγορίθμων δένδρου απόφασης με 59,4%, Αφελής ταξινομητής Bayes με ποσοστό 59,1%, k-κοντινότερων γειτόνων με ακρίβεια 60% και Γραμμική Διακριτική Ανάλυση με ακρίβεια 50%. Αναφορικά της μεθόδου της μέτρησης διανύσματος καλύτερα ποσοστά παρουσιάζουν οι αλγόριθμοι Στοχαστικής κλίσης καθόδου (Stochastic Gradient Descent) με ακρίβεια 60%, Μηχανές διανυσμάτων υποστήριξης με ακρίβεια 60% και πολύ στρωματικό ταξινομητή perceptron με ποσοστό 62,3%. Στους χειρότερους αλγορίθμους, ανήκουν τα δένδρα απόφασης με ποσοστό 54,85%, Αφελής ταξινομητής Bayes με ακρίβεια 54,57%, k-κοντινότερων γειτόνων με 42% και Γραμμική Διακριτική Ανάλυση με ποσοστό 49%. Είναι σημαντικό να αναφερθεί ότι τα μειωμένα ποσοστά των αλγορίθμων που παρουσιάστηκαν οφείλονται στην αδυναμία των αλγορίθμων να επιλύσουν μη γραμμικά προβλήματα όπως είναι η ανάλυση συναισθήματος. Τα ποσοστά με την υψηλότερη ακρίβεια χρησιμοποιούνται ώστε να δημιουργηθούν οι γραφικές παραστάσεις του κεφαλαίου 6. Σύμφωνα με τους δείκτες που παρουσιάζονται η πτώση της αγοράς σε ημερομηνίες κλειδιά δικαιολογείται χάρη στην αφετηρία της πανδημίας του Covid-19 και άλλων σημαντικών γεγονότων. Οι συνεχείς αυξομειώσεις των δεικτών συναισθήματος στις γραφικές παραστάσεις που εξήχθησαν και αναλύθηκαν παραπάνω αφορούν πληθώρα γεγονότων που κάθε ένα επηρεάζει με διαφορετικό τρόπο την κάθε κατηγορία από αυτές που παρουσιάστηκαν. Ειδικότερα, είναι ξεκάθαρο ότι η πανδημία κατά τα έτη 2019-2022 οδήγησε σε μία ραγδαία μείωση του συναισθήματος της αγοράς. Συμπληρωματικά, η εισβολή στην Ουκρανία καθώς άλλα γεγονότα δικαιολογούν τους ιδιαίτερα μειωμένους δείκτες. Τέλος, οι δείκτες των άρθρων που συνθέτουν τα γραφήματα παρουσιάζονται μηνιαία με στόχο την λεπτομερή παρατήρηση των δεικτών συναισθήματος της αγοράς. Επομένως, με όλη αυτή την διαδικασία μπορούμε να αντιληφθούμε την αξία της ανάλυσης συναισθήματος σε τομείς κλειδιά. Μέσω των γραφικών παραστάσεων μπορούμε να αδράξουμε την γνώμη της αγοράς ώστε επενδύσεις, marketing καθώς και διαφημίσεις προς το κοινό να έχουν μεγαλύτερο αντίκτυπο.

7 - Μελλοντικές επεκτάσεις

Η μέθοδος της ανάλυσης συναισθήματος μπορεί να βελτιωθεί με ποικίλους τρόπους. Αρχικά, η βελτίωση των συνόλων δεδομένων μέσω της προσκόμισής άρθρων οικονομικού χαρακτήρα αποτελεί σημαντικό παράγοντα στην αύξηση των μετρικών. Επίσης, η υιοθέτηση υβριδικών μεθόδων μέσω του συνδυασμού lexicons και αλγορίθμων μηχανικής μάθησης αποτελεί σημαντικό παράγοντα στην εξέλιξη της μεθόδου. Η επέκταση της μεθόδου σε περισσότερες γλώσσες αποτελεί ένα επιπλέον σημαντικό βήμα αφού επιτρέπει την εξαγωγή συναισθημάτων από διάφορες γλώσσες χωρίς να είναι απαραίτητη η μετάφραση. Η μέθοδος της ανάλυσης συναισθήματος μπορεί να φανεί ιδιαίτερα χρήσιμη και σε τομείς που δεν αφορούν μόνο την οικονομία όπως τα social media αλλά και άλλους τομείς όπου μέσω της γνώμης του κόσμου δίνεται μια νέα ματιά σε διάφορα θέματα και γεγονότα.

Εκτός από την ήδη υπάρχουσα επεξεργασία δεδομένων θα έπρεπε να υπάρχουν μέθοδοι που να εξετάζουν την αξιοπιστία των δεδομένων που χρησιμοποιούνται. Ειδικότερα,

περισσότεροι κανόνες και λογικοί συλλογισμοί πρέπει να ληφθούν υπόψη όπως το ιστορικό των άρθρων που έχουν δημοσιευθεί από μία εφημερίδα ή έναν αρθρογράφο.

Εξίσου σημαντική είναι η ανάπτυξη μεθόδων που συλλογικά θα παρέχουν καλύτερες μετρικές από τους ήδη υπάρχοντες αλγόριθμους και γραφικές παραστάσεις με λεπτομερέστερους δείκτες.

Τέλος, περαιτέρω επέκταση του τομέα μπορεί να αφορά πειραματισμό και χρήση της ανάλυσης συναισθήματος σε παραπάνω τομείς. Αυτή η ενέργεια μπορεί να αφορά οτιδήποτε έχει να κάνει με την εμπλοκή του κόσμου με διάφορες υπηρεσίες ακόμα και σε κυβερνητικό επίπεδο με στόχο την κριτική για την σωστή εξυπηρέτηση των πελατών μέσα από την έκφραση των συναισθημάτων του κόσμου.

8 - Βιβλιογραφία

[1] I. C. Education, «IBM,» 2 July 2020. [Ηλεκτρονικό]. Available: https://www.ibm.com/cloud/learn/natural-language-processing/.

[2] I. C. Education, «Machine Learning,» *IBM,* 15 July 2020.

[3] Γ. Αικατερίνη, «Repository Kallipos,» 2015. [Ηλεκτρονικό]. Available: https://repository.kallipos.gr/bitstream/11419/3381/3/%CE%A4%CE%B5%CF%87%CE%BD%CE%B7%CF%84%CE%AE%20%CE%9D%CE%BF%CE%B7%CE%BC%CE%BF%CF%83%CF%8D%CE%BD%CE%B7.pdf/.

[4] D. Xiaowen, L. Bing και Y. Philip S., «A holistic lexicon-based approach to opinion mining,» *ACM Digital Library,* p. 231–240, 11 Febrary 2008.

[5] M. KOSTADIN, G. ANA, V. IRENA, C. LUBOMIR και T. DIMITAR, «Evaluation of Sentiment Analysis in Finance: From Lexicons to Transformers,» *Ieeexplore,* pp. 131662 - 131682, 13 June 2020.

[6] A. Khin Zezawar και M. Nyein Nyein, «Sentiment analysis of students' comment using lexicon based approach,» *Ieeexplore,* 24-26 May 2017.

[7] A. Amit και T. Durga, «Application of Lexicon Based Approach in Sentiment Analysis for short Tweets,» *Ieeexplore,* 22-23 June 2018.

[8] A. Ahmed και O. Nazlia, «Integrating a Lexicon based approach and K nearest neighbour for Malay sentiment analysis,» 23 January 2016.

[9] H. Ali, M. Sana, K. Ahmad και S. Shahaboddin, «Machine Learning-Based Sentiment Analysis for Twitter Accounts,» *MDPI,* 16 January 2018.

[10] I. C. Education, «IBM,» 19 August 2020. [Ηλεκτρονικό]. Available: https://www.ibm.com/cloud/learn/supervised-learning/.

[11] A. G. N. C. Palak Bald, «Sentiment Analysis of Movie Reviews using Machine Learning Techniques,» *Research Gate,* pp. 45-49, December 2017.

[12] S. Antonios, K. Paris-Alexandros, S. Panagiotis και P. Ioannis-Chrysostomos, «A Novel Lexicon-based Approach in Determining Sentiment in Financial Data Using Learning Automata,» 13 March 2018.

[13] E. A. Team, «Educative,» 2022. [Ηλεκτρονικό]. Available: https://www.educative.io/answers/what-is-feature-extraction/.

[14] G. F. Geeks, «Geeks For Geeks,» 28 June 2022. [Ηλεκτρονικό]. Available: https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/.

[15] B. Stecanella, «Monkey Learn,» 11 May 2019. [Ηλεκτρονικό]. Available: https://monkeylearn.com/blog/what-is-tf-idf/.

[16] F. Malik, «Medium,» 19 February 2020. [Ηλεκτρονικό]. Available: https://medium.com/fintechexplained/what-is-grid-search-c01fe886ef0a.

[17] L. G. V, «Analytics Vidhya,» 21 May 2021. [Ηλεκτρονικό]. Available: https://www.analyticsvidhya.com/blog/2021/05/4-ways-to-evaluate-your-machine-learning-model-cross-validation-techniques-with-python-code/.

[18] C3, «C3.ai,» 2022. [Ηλεκτρονικό]. Available: https://c3.ai/glossary/data-science/model-training/#:~:text=Model%20training%20is%20the%20phase,function%20over%20the%20prediction%20range./.

[19] S. Tas, «Medium,» 17 Mar 2021. [Ηλεκτρονικό]. Available: https://seymatas.medium.com/why-do-we-split-datasets-55c46964fd84/.

[20] P. Gupta, «Towards Data Science,» 17 May 2017. [Ηλεκτρονικό]. Available: https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052.

[21] A. Navlani, «Datacamp,» December 2018. [Ηλεκτρονικό]. Available: https://www.datacamp.com/tutorial/decision-tree-classification-python.

[22] P. Sharma, «Analytics Vidhya,» 29 November 2021. [Ηλεκτρονικό]. Available: https://www.analyticsvidhya.com/blog/2021/11/implementation-of-gaussian-naive-bayes-in-python-sklearn/.

[23] J. Brownlee, «Machine learning mastery,» 30 April 2018. [Ηλεκτρονικό]. Available: https://machinelearningmastery.com/a-gentle-introduction-to-calculating-normal-summary-statistics/.

[24] O. Harrison, «Towards Data Science,» 10 September 2018. [Ηλεκτρονικό]. Available: https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761.

[25] H. Mohammad Rezwanul, A. Ahmad και R. Anika, «Sentiment Analysis on Twitter Data using KNN and SVM,» *International Journal of Advanced Computer Science and Applications,* 2017.

[26] A. Kumar, «Vital flux,» 4 August 2022. [Ηλεκτρονικό]. Available: https://vitalflux.com/k-nearest-neighbors-explained-with-python-examples/.

[27] C. Math, «Cue Math,» 2022. [Ηλεκτρονικό]. Available: https://www.cuemath.com/euclidean-distance-formula/.

[28] G. F. Geeks, «Geeks For Geeks,» 10 November 2021. [Ηλεκτρονικό]. Available: https://www.geeksforgeeks.org/ml-linear-discriminant-analysis/.

[29] S. K. Dash, «Analytics Vidhya,» 18 August 2021. [Ηλεκτρονικό]. Available: https://www.analyticsvidhya.com/blog/2021/08/a-brief-introduction-to-linear-discriminant-analysis/.

[30] «Michael, Fuchs,» 11 11 2019. [Ηλεκτρονικό]. Available: https://michael-fuchs-python.netlify.app/2019/11/11/introduction-to-sgd-classifier/.

[31] R. Kwiatkowski, «Towards Data Science,» 22 May 2021. [Ηλεκτρονικό]. Available: https://towardsdatascience.com/gradient-descent-algorithm-a-deep-dive-cf04e8115f21/.

[32] R. Gandhi, «Towards Data Science,» 7 June 2018. [Ηλεκτρονικό]. Available: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47/.

[33] A. Nair, «Analytics India Mag,» 20 June 2019. [Ηλεκτρονικό]. Available: https://analyticsindiamag.com/a-beginners-guide-to-scikit-learns-mlpclassifier/.

[34] M. Fuchs, «Michael Fuchs Python,» 03 02 2021. [Ηλεκτρονικό]. Available: https://michael-fuchs-python.netlify.app/2021/02/03/nn-multi-layer-perceptron-classifier-mlpclassifier/.

[35] A. Mohanty, «Becoming human,» 15 May 2019. [Ηλεκτρονικό]. Available: https://becominghuman.ai/multi-layer-perceptron-mlp-models-on-real-world-banking-data-f6dd3d7e998f.

[36] I. C. Education, «Neural Networks,» *IBM,* 17 August 2020.

[37] F. PEIXOTO, «Analytics Vidhya,» 13 December 2020. [Ηλεκτρονικό]. Available: https://www.analyticsvidhya.com/blog/2020/12/mlp-multilayer-perceptron-simple-overview/.

[38] «Developers.Google.com,» 18 July 2022. [Ηλεκτρονικό]. Available: https://developers.google.com/machine-learning/crash-course/classification/true-false-positive-negative#:~:text=A%20true%20positive%20is%20an,incorrectly%20predicts%20the%20positive%20class./.

[39] J. Brownlee, «How to Calculate Precision, Recall, and F-Measure for Imbalanced Classification,» *Machine Learning Matery,* 25 March 2020.

[40] D. Google, «Developers Google,» 18 July 2022. [Ηλεκτρονικό]. Available: https://developers.google.com/machine-learning/crash-course/classification/accuracy/.

[41] C3.ai, «C3.ai,» 2022. [Ηλεκτρονικό]. Available: https://c3.ai/glossary/machine-learning/precision/#:~:text=Precision%20is%20one%20indicator%20of,the%20number%20of%20false%20positives)./.

[42] T. Wood, «DeepAI,» 2022. [Ηλεκτρονικό]. Available: https://deepai.org/machine-learning-glossary-and-terms/f-score/.

[43] C. Dilmegani, «In-Depth Guide to Web Scraping for Machine Learning in 2022,» *Research Aimutiple,* 13 August 2021.

[44] B. Mahesh, «Machine Learning Algorithms - A Review,» pp. 381-386, January 2020.

[45] R. Susmita, «A Quick Review of Machine Learning Algorithms,» *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon),* pp. 35-39, 2019.

[46] A. Krouska, C. Troussas και M. Virvou, «The effect of preprocessing techniques on Twitter sentiment analysis,» *2016 7th International Conference on Information, Intelligence, Systems & Applications (IISA),* pp. 1-5, 2016.

[47] F. Osisanwo, J. Akinsola, O. Awodele, J. Hinmikaiye, O. Olakanmi και Akinjobi, «Supervised machine learning algorithms: classification and comparison,» *International Journal of Computer Trends and Technology (IJCTT),* pp. 128-138, June 2017.

[48] S. Pradha, M. N. Halgamuge και N. Tran Quoc Vinh, «Effective Text Data Preprocessing Technique for Sentiment Analysis in Social Media Data,» *2019 11th International Conference on Knowledge and Systems Engineering (KSE),* pp. 1-8, 2019.

[49] D. Sharma και N. Kumar, «A review on machine learning algorithms, tasks and applications,» *International Journal of Advanced Research in Computer Engineering \& Technology (IJARCET),* pp. 1548-1552, October 2017.

[50] A. Singh, N. Thakur και A. Sharma, «A review of supervised machine learning algorithms,» *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom),* pp. 1310-1315, 2016.