ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

UNIVERSITY OF PIRAEUS

Δημιουργία μοντέλου για την αξιολόγηση αποφάσεων επενδύσεων στηριζόμενοι σε τεχνικές sentiment analysis και στατιστικές μεθόδους.

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Τσόπελας Θωμάς

Επιβλέπων Καθηγητής : Μιχαήλ Φιλιππάκης

Τμήμα Ψηφιακών Συστημάτων Πανεπιστημίου Πειραιώς.

Π.Μ.Σ «Ανάπτυξη Προηγμένων Πληροφοριακών Συστημάτων»

Πειραιάς 2022

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου κύριο Μιχαήλ Φιλιππάκη για τις οδηγίες και κατευθύνσεις που μου παρείχε.

Οφείλω επίσης ένα μεγάλο ευχαριστώ στην οικογένειά μου για την στήριξη καθ' όλη τη διάρκεια των μεταπτυχιακών σπουδών μου.

Σημαντική ήταν η βοήθεια που μου προσέφεραν οι συνάδελφοι μου στην κατανόηση εξειδικευμένων οικονομικών εννοιών και του τρόπου όπου λειτουργούν τα αντισυμβαλλόμενα μέλη στη σύγχρονη οικονομία, ώστε να αποκτήσω μία αναλυτική σκέψη.

Τέλος πρέπει να αναγνωριστεί όλη η πολύτιμη προσπάθεια όλων των ανώνυμων προγραμματιστών οι οποίοι μοιράζουν τις γνώσεις τους αφιλοκερδώς στα επιστημονικά forums στο διαδίκτυο επιτρέποντας σε άτομα άλλων γνωστικών πεδίων να γνωρίζουν την ομορφιά της επιστήμης των μεγάλων δεδομένων .

Τσόπελας Θωμάς

02/07/2022

We need to be super careful with AI. Potentially more dangerous than Nukes.

-Elon Musk

Περιεχόμενα

Περίληψη

Η παρούσα διπλωματική εργασία περιέχει μια μελέτη στην ερευνητική περιοχή των Predictive Analytics ,κάνοντας χρήση των δυνατοτήτων της γλώσσας Python και των διαθέσιμων βιβλιοθηκών καθώς γίνεται μια προσπάθεια πρόβλεψης της τιμής μιας μετοχής ως χρονοσειρά σε αρχικό χρόνο.

Ως προς το περιεχόμενο, αρχικά παρουσιάζονται κάποιες βασικές έννοιες που συνδέονται με τις μετοχές και τις επενδύσεις, όπως είναι η αναμενόμενη απόδοση,efficient market hypothesis, η πραγματική απόδοση μιας επένδυσης, ο κίνδυνος που ενέχει μια επένδυση η ανάλυση φυσικής γλώσσας και συναισθήματος και πώς αυτά τα δύο πεδία έχουν εφαρμογή στην σύγχρονη οικονομία και καθημερινότητα.

Γίνεται ανάλυση της έννοιας της χρονοσειράς και των βασικότερων χαρακτηριστικών της σε θεωρητικό επίπεδο, όπως η στασιμότητα, η αυτοσυσχέτιση , η περιοδικότητα. Παρουσιάζονται, επίσης, τα γραμμικά μοντέλα AR(p), MA (q) και ARIMA (p,d,q), τα οποία δύνανται να «προσαρμοστούν» στα δεδομένα μιας στάσιμης χρονοσειράς και να παρέχουν προβλέψεις με μια σχετική ακρίβεια.

Τέλος, το θεωρητικό υπόβαθρο των χρονοσειρών και των γραμμικών μοντέλων εφαρμόζεται στην πράξη σε μια μελέτη πραγματικής μετοχής για την περίοδο 17/11/2016 – 16/11/2021. Η έρευνα αφορά τις μετοχές των εταιρειών «Home Depot» η οποία είναι εισηγμένη στο Αμερικανικό Χρηματιστήριο Dow Jones.

Η ανάπτυξη του παγκόσμιου ιστού έχει οδηγήσει στην ραγδαία αύξηση των δεδομένων όπου είναι διαθέσιμα προς επεξεργασία και ανάλυση για την εξαγωγή συμπερασμάτων ,για το λόγο αυτό έχουν κατασκευαστεί εργαλεία για την άντληση και επεξεργασία των δεδομένων αυτών. Η ανάλυση της προσωπικής άποψης και πεποίθησης αποτελεί ένα βασικό σημείο όπου μπορεί να πραγματοποιηθεί μέσω των παραπάνω δεδομένων.

Στη συνέχεια γίνεται χρήση ενός ήδη εκπαιδευμένου νευρωνικού δίκτυο για την συσχέτιση δημοσιεύσεων έγκυρων οικονομικών πηγών με την πορεία της μετοχής μέσω ανάλυσης φυσικής γλώσσας και στη προσπάθεια εξαγωγής αποτελεσμάτων στον τομέα της ανάλυσης συναισθήματος του οικονομικού site ενημέρωσης finviz,με τη χρήση της τεχνικής ενός διαθέσιμου λεξικού.

Για την ανάπτυξη των μοντέλων ανάλυσης χρνοσειρών και ανάλυσης συναισθήματος, χρησιμοποιήθηκε η γλώσσα προγραμματισμού Python, η οποία τα τελευταία χρόνια Οι βασικότερες βιβλιοθήκες που χρησιμοποιήθηκαν είναι οι Pandas, NumPy, Seaborn και Vader, οι οποίες προσφέρουν τις απαραίτητες τεχνικές που απαιτούνται για την παρούσα εργασία.

Λέξεις – Κλειδιά: Predictive Analytics, Μετοχές, Χρονοσειρές, Προβλέψεις,efficient market hypothesis

,Μοντέλα ARIMA,Pytohn,NLP,neural network,VADER.

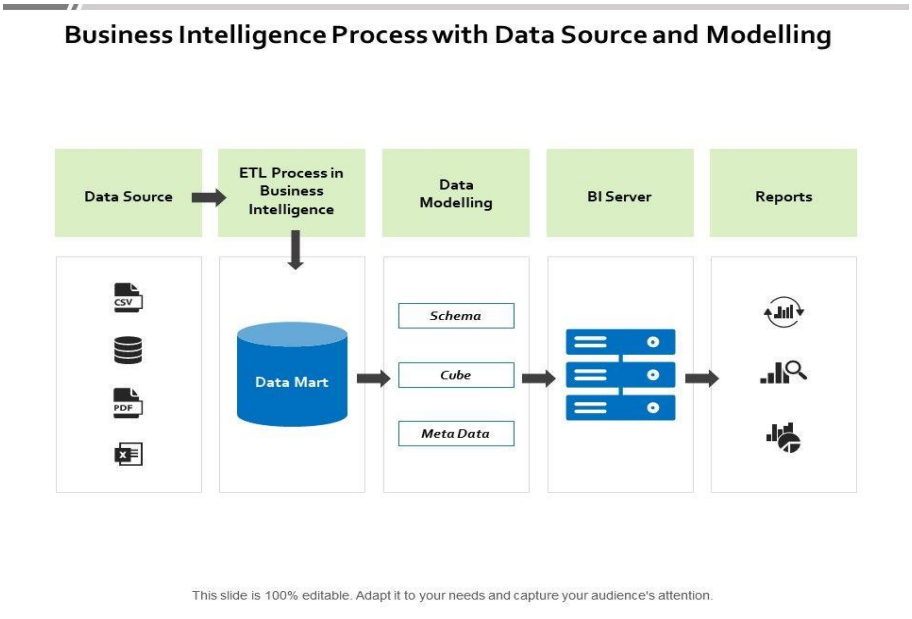
ΚΕΦΑΛΑΙΟ 1

1.1 Εισαγωγή

Στη σύγχρονη οικονομία καθίσταται όχι μόνο προφανές αλλά και αναγκαίο όλα τα συμβαλλόμενα μέρη που συμμετέχουν στην οικονομική ζωή και στην λήψη αποφάσεων να μπορούν να λαμβάνουν ορθολογικές αποφάσεις. Δεν είναι λίγες οι περιπτώσεις που ακόμη και σήμερα πολλές επιχειρήσεις επιλέγουν να σχεδιάσουν και να λαμβάνουν σημαντικές αποφάσεις και διεργασίες όπως στρατηγικός σχεδιασμός,business plan,νέες επενδύσεις και προώθηση νέων προϊόντων με γνώμονα εμπειρικά δεδομένα και παρωχημένες τεχνικές. Χαρακτηριστικά παραδείγματα αποτελούν η οικονομική «φούσκα» του ελληνικού Χρηματιστηρίου τη περίοδο 1997-1999 όσον αφορά την ελληνική οικονομία αλλά και την κατάρρευση της αμερικάνικης αγοράς των ακινήτων του έτους 2008 με την ασύστολη αγοροπωλησία τιτλοποιημένων δανείων η οποία οδήγησε στην παγκόσμια οικονομική κρίση.

Τα παραπάνω παραδείγματα έχουν ένα κοινό χαρακτηριστικό την λανθασμένη διαχείριση και αξιολόγηση των υπαρχόντων πληροφοριών. Στη σύγχρονή εποχή ο κλάδος της τεχνολογίας και της πληροφορικής 'σε σχεδιασμό με την επιστήμη της οικονομίας με ότι αυτό συνεπάγεται - τραπεζική, διαχείρισή επιχειρήσεων- έχουν αναπτύξει εργαλεία που είναι πού χρήσιμα στην καθημερινή χρήση. Όροι όπως **«Επιχειρηματική Ευφυΐα»** (Business Intelligence) και της **«Επιστήμης Δεδομένων»** (Data Science) αποτελούν πλέον αναπόσπαστο τμήμα την ζωής των σύγχρονων επιχειρήσεων και στο σύγχρονο οικονομικό περιβάλλον αβεβαιότητας.

Ως Επιχειρηματική Ευφυΐα ορίζουμε ένα σύνολο από μεθόδους ανάλυσης, τεχνολογίες, ικανότητες και στρατηγικές, οι οποίες στόχο έχουν την επεξεργασία των διαθέσιμων δεδομένων και την εξαγωγή χρήσιμης πληροφορίας από αυτά, για την υποστήριξη της διαδικασίας λήψης επιχειρηματικών αποφάσεων. Η Επιχειρηματική Ευφυΐα επιτρέπει σε έναν οργανισμό να μαθαίνει, να αντιλαμβάνεται καταστάσεις και συμβάντα, να σκέφτεται αφαιρετικά, να προβλέπει τάσεις και μελλοντικά συμβάντα, να σχεδιάζει και να καινοτομεί. Η παραγόμενη πληροφορία μετουσιώνεται σε γνώση που αξιοποιείται από τα διοικητικά στελέχη, ώστε να δρομολογήσουν κατάλληλες δράσεις, που θα οδηγήσουν στον καθορισμό και την επίτευξη επιχειρηματικών στόχων, με τρόπο αποτελεσματικό και αποδοτικό και ελαχιστοποιεί όσο μπορεί το ρίσκο της λήψης.



Εικόνα 1 Διαδικασία Business Inteligence

Οι παραπάνω κλάδοι επηρεάζουν σε μεγάλο βαθμό την διαδικασία της προβλεπτικής αναλυτικής «Predictive Analytics»,δηλαδή την πρόβλεψη ενός μελλοντικού γεγονότος στηριζόμενοι σε ιστορικές παρατηρήσεις του γεγονότος. Τα βασικά στάδια την άνω διαδικασίας είναι τα εξής:

- Η συλλογή σωστών δεδομένων-πληροφοριών και ο κατάλληλος μετασχηματισμός τους σε μορφή χρήσιμη για εμάς
- Την επεξεργασία των άνω δεδομένων για την παραγωγή του κατάλληλου μοντέλου
- Την εξαγωγή των προβλέψεων την αξιολόγηση της απόδοσης και την επανάληψη των άνω σταδίων σε περίπτωση που κρίνεται απαραίτητο για την βελτίωση της απόδοσης.

Στη παρούσα διπλωματική θα ασχοληθούμε με την προσπάθεια πρόβλεψης τις τιμές κλεισίματος των μετοχών της αμερικανικής εταιρείας «Home Depot» με την χρήση θεωρίας χρονοσειρών ,έπειτα θα γίνει η προσπάθεια δημιουργίας ενός νευρωνικού δικτύου για την εξαγωγή συμπερασμάτων για το αν οι δημοσιεύσεις σε οικονομικά άρθρα επηρεάζουν την τιμή κλεισίματος.

Η ανάλυση γίνεται μέσω της γλώσσας προγραμματισμού Python,με IDE Google colab και θα προσπαθήσουμε να απαντήσουμε στα εξής βασικά ερωτήματα:

- Είναι επιστημονικά ορθό να αντιμετωπίσουμε τις τιμές των μετοχών μόνο ως χρονοσειρες αγνοώντας άλλα οικονομικά μεγέθη(λογιστική αξία εταιρείας,future prices κτλ.)
- Πόσο καλό είναι το παραγόμενο μοντέλο
- Μπορεί το μοντέλο να συνδυαστεί με το νευρωνικό δίκτυο sentiment analysis.

Τα επόμενα κεφάλαια έχουν ως στόχο την απάντηση των παραπάνω ερωτημάτων όσο το δυνατόν σε ικανοποιητικό επίπεδο και βαθμό.

ΚΕΦΑΛΑΙΟ 2

2.Χρηματιστήριο και Μετοχές

Ως Χρηματιστήριο ορίζουμε μια επίσημη, οργανωμένη και ελεγχόμενη αγορά κινητών αξιών, οι τιμές των οποίων προσδιορίζονται από τις δυνάμεις προσφοράς και ζήτησης. Το χρηματιστήριο είναι ο χώρος όπου συναντώνται οι αντίθετες προσδοκίες των επενδυτών για τη διαμόρφωση των τιμών των μετοχών μια συγκεκριμένη χρονική στιγμή. Οι πρώτοι, προσπαθώντας να πουλήσουν τις μετοχές που κατέχουν, πιέζουν την τιμή της μετοχής πτωτικά, ενώ οι δεύτεροι προσπαθώντας να αγοράσουν, πιέζουν την τιμή της μετοχής ανοδικά και με αυτόν τον τρόπο διαμορφώνεται μια τιμή όπου η προσφορά και η ζήτηση ισορροπούν μια δεδομένη χρονική στιγμή. Ο επενδυτής βλέπει το χρηματιστήριο ως μια εναλλακτική μορφή τοποθέτησης των χρημάτων που αποταμιεύει, με σκοπό την επιδίωξη ικανοποιητικής απόδοσης, υψηλότερης από αυτήν που προσφέρουν επενδύσεις όπως οι τραπεζικές καταθέσεις και τα κρατικά ομόλογα.

Το προϊόν του χρηματιστήριου που θα μας απασχολήσει στη παρούσα εργασία είναι η μετοχή. Ως μετοχή ορίζουμε ένα από τα ίσα τμήματα στα οποία διαιρείται η κεφαλαιακή αξία μίας εταιρείας. Αποτελεί ένα αξιόγραφο δηλαδή ο κάτοχός της διαθέτει δικαιώματα και υποχρεώσεις που πηγάζουν από τη κατοχή της τα οποία διαφέρουν ανάλογα με τον αριθμό και το είδος τους. Η βασικότερη μορφή μετοχής είναι η κοινή μετοχή η οποία επιτρέπει τη συμμετοχή στα κέρδη, έκδοση νέων μετοχών και προτεραιότητα σε περίπτωση εκκαθάρισης .

2.1 Απόδοση Μετοχής

Η έννοια αυτή είναι αρκετά σημαντική καθώς μετράει την αξία του κατόχου της και διαμορφώνει την ζήτησή της. Η απόδοση μιας μετοχής σε μία περίοδο είναι ίση προς τη ποσοστιαία διαφορά μεταξύ της αρχικής και της τελικής περιουσίας του κατόχου δηλαδή:

$$r = \frac{Τελική\ Τιμή - Αρχική\ Τιμή + Μέρισμα\ Περιόδου}{Αρχική\ Τιμή}$$

Με πιο απλά λόγια εκφράζει το κέρδος ενός επενδυτή για μία συγκεκριμένη χρονική περίοδο, η οποία περιορίζεται στον χρόνο t με t-1.Το μέγεθος αυτό είναι αρκετά σημαντικό για τον επενδυτή ή κάτοχο ή για τον δυνητικό αγοραστή της καθώς αποτελεί ένα μέγεθος μέτρησης της επιτυχίας ή όχι δηλαδή εάν η απόδοση έχει θετικό πρόσημο κρίνεται θετικά από τον επενδυτή ενώ εάν είναι αρνητικό τότε αρνητικά. Με άλλα λόγια μία θετική απόδοση συνδέεται μεγαλύτερη καθαρή παρούσα αξία της αρχικής επένδυσης.

2.2 Κίνδυνος

Ως κίνδυνος ορίζεται η μεταβλητότητα των δυνητικών αποτελεσμάτων από την αναμενόμενη τιμή ή τον αριθμητικό μέσο δηλαδή η πραγματική απόδοση δεν θα είναι ίση με την αναμενόμενη. Το κυριότερο στατιστικό μετρό που χρησιμοποιείται για την μέτρηση της διασποράς γύρω από το αριθμητικό μέσο αποτελεί η τυπική απόκλιση ή η μέση απόκλιση τετραγώνου. Κατά συνέπεια η τυπική απόκλιση αποτελεί μέτρο συνολικού κινδύνου μίας μετοχής ή ενός χαρτοφυλακίου ή γενικότερα μίας επενδυτικής προσπάθειας. Ο κίνδυνος είναι εμφανής σε όλες τις δραστηριότητες, όλων των οργανισμών ανεξάρτητα από το σκοπό και από την

διάρθρωση των λειτουργιών του καθενός. Ένα περιουσιακό στοιχείο έχει υψηλό κίνδυνο, όταν υπάρχει μεγάλη πιθανότητα να απέχει πολύ η πραγματική του απόδοση από την αναμενόμενη. Τα συμβαλλόμενα μέλη θα πρέπει με κατάλληλες μεθόδους και διαδικασίες να εκτιμήσουν και να διαχειριστούν τους κινδύνους που μπορεί να υπάρξουν. Με αυτό τον τρόπο θα μπορούν να εξασφαλίσουν την βιωσιμότητά τους.

Οι κυριότεροι κίνδυνοι στο στην οικονομική ζωή είναι οι εξής :

• Κίνδυνος Αγοράς (Market Risk): αφορά τον κίνδυνο μείωσης του επιπέδου τιμών της αγοράς στο σύνολο της ή σε κάποια από τα στοιχεία του ενεργητικού κάποιου επενδυτικού προϊόντος. Για παράδειγμα η μεταβολή αυτή μπορεί να οφείλεται στην αυξομείωση των επιτοκίων ή των τιμών των επενδυτικών τίτλων.

• Πιστωτικός Κίνδυνος (Credit Risk): είναι ο κίνδυνος που διατρέχει μια επιχείρηση ή ένας οικονομικός οργανισμός να μην εισπράξει τις απαιτήσεις του λόγω αδυναμίας του αντισυμβαλομένου. Χαρακτηριστικές μορφές του είναι :

- Κίνδυνος αθέτησης
- Κίνδυνος έκθεσης
- Κίνδυνος ανάκτησης

• Επιτοκιακός Κίνδυνος (Interest Rate Risk) : είναι ο κίνδυνος να μεταβληθεί η αξία μιας επένδυσης κάτι το οποίο οφείλεται σε μεταβολές στο επίπεδο των επιτοκίων.

• Κίνδυνος Ρευστότητας (Liquidity Risk) : οφείλεται στην αβεβαιότητα που δημιουργείται όταν κάποια επένδυση δεν μπορεί να ρευστοποιηθεί έγκαιρα και σε μία τιμή παραπλήσια της αγοραίας αξίας του ή στην αδυναμία του επενδυτή να αντλήσει κεφάλαια ώστε να ανταποκριθεί στις ταμειακές ροές.

• Κίνδυνος Πληθωρισμού: γνωστός και ως «κίνδυνος αγοραστικής δύναμης» κίνδυνος απώλειας της πραγματικής αξίας των απαιτήσεων ή κάποιας επένδυσης λόγω μεγαλύτερης της αναμενόμενης αύξησης ή μείωσης του πληθωρισμού. Προκαλείται εφόσον ο πληθωρισμός μεταβάλλεται κατά τρόπο διαφορετικό από αυτόν που έχει προβλεφθεί και καθιστά αβέβαιη την μελλοντική πραγματική αξία μιας επένδυσης. Ιδιαίτερα επιρρεπείς είναι οι μετοχές των αναδυόμενων αγορών.

2.3 Μέτρηση Κινδύνου

Για την σωστή και έγκυρη αντιμετώπιση απαιτείται η σωστή μέτρηση των πιθανών κινδύνων. Αυτό επιτυγχάνεται μέσω ποσοτικών δεδομένων (ιστορικά στοιχεία),ποιοτικά στοιχεία (consulting) Για το λόγο αυτό απαιτείται η εκτίμηση της πιθανότητας εμφάνισης μίας κατηγορίας κινδύνου καθώς και οι επιπτώσεις που θα επιφέρει στη δομή του οργανισμού ή στην αξίας ενός χαρτοφυλακίου στη περίπτωση ενός επενδυτή.

2.4 Διαχείριση Κινδύνου

Είναι προφανές ότι ο κίνδυνος στην σύγχρονη οικονομική ζωή είναι αδύνατον να εξαλειφθεί για το λόγω αυτό απαιτείται η αποτελεσματική ,έγκαιρη διαχείριση ώστε να περιοριστούν οι δυσμενείς συνέπειες του. Οι οργανισμοί, επενδυτές έχουν την υποχρέωση να προσεγγίζουν τους κινδύνους που σχετίζονται με την δραστηριότητα τους με τέτοιο τρόπο ώστε διαχρονικά να εξασφαλίσουν την ανάπτυξη τους. Μέσω της παραπάνω διαδικασίας οι επενδυτές, οργανισμοί μπορούν να αποφύγουν προβλέψιμούς κινδύνους, να προστατευθούν από λάθος επενδυτικές αποφάσεις και να μειώσουν τις απώλειες στο ενεργητικό τους σε περίπτωση εκδήλωσης.

Βασικές τεχνικές διαχείρισης και αντιστάθμισης κινδύνου είναι οι παρακάτω:

- Μεταφοράς Κινδύνου(tranfer risk): σε άλλα συμβαλλόμενα μέρη (προμηθευτές, ασφαλιστικές εταιρίες, έμποροι ,τράπεζες κτλ).

- Ελέγχου κινδύνου (control risk): εφαρμογή κατάλληλων πολιτικών και διαδικασιών εσωτερικού ελέγχου που έχει θεσμοθετήσει κάθε οργανισμός.

- Απαλλαγή κινδύνου (avoid risk): η εταιρεία-οργανισμός αποφεύγει να επενδύσει σε δραστηριότητές με συγκεκριμένο κίνδυνο.

- Διαφοροποίηση Χαρτοφυλακίου: με την προσθήκη επενδυτικών στοιχείων σε ένα χαρτοφυλάκιο ο συνολικός του κίνδυνος μειώνεται. Καθώς προσθέτουμε επενδυτικά προϊόντα το χαρτοφυλάκιο μας τείνει να πλησιάζει το κίνδυνο του χαρτοφυλακίου της αγοράς δηλαδή εκείνο το οποίο περιέχει τα στοιχεία που προσφέρονται για επενδύσεις κεφαλαίου σε μια δεδομένη περίοδο. Ο κίνδυνος του αυτού του χαρτοφυλακίου εξαρτάται από τις συνθήκες όπου επικρατούν στην εγχώρια και παγκόσμια οικονομία την τρέχουσα περίοδο. Καλή οικονομική πρακτική αποτελεί η προσθήκη οικονομικών στοιχείων με αντίθετη συσχέτιση ,σε περίπτωση όπου ο συντελεστής συσχέτισης ρ ισούται με +1 δεν προκύπτει οικονομικό όφελος καθώς έχουν τέλεια θετική ενώ το όφελος μεγιστοποιείται με συντελεστή ρ=-1.Τα στελέχη μπορούν να αντιμετωπίσουν τον κίνδυνο του χαρτοφυλακίου επιλέγοντας συνδυασμούς στοιχείων που δεν είναι τέλεια θετικά συσχετισμένα δηλαδή $-1<\rho<1$.

ΚΕΦΑΛΑΙΟ 3

3.Θεωρία Χρονοσειρων(time -series)

Με τον όρο χρονολογική σειρά ή χρονοσειρά ορίζουμε μια σειρά δεδομένων με κύριο χαρακτηριστικό τη διατεταγμένη χρονική διάταξη των παρατηρήσεων της σειράς με συγκεκριμένο βήμα. Με άλλα λόγια, χρονοσειρά είναι ένα σύνολο από δεδομένα που δίνονται με συγκεκριμένη χρονική διάταξη. Μερικά παραδείγματα χρονοσειρών που χρήζουν μελέτης είναι το ετήσιο μέσο ύψος της στάθμης του νερού μιας λίμνης ή ενός ποταμιού, ο ετήσιος αριθμός ηλιακών κηλίδων, η ετήσια μέση τιμή της θερμοκρασίας του πλανήτη, το Α.Ε.Π μιας χώρας στην πάροδο των χρόνων,η ετήσια ποσότητα βροχής ή τιμή της μετοχής μιας εταιρίας στο χρηματιστήριο που αποτελεί και το ερευνητικό αντικείμενο της εργασίας.

Τα δεδομένα της χρονοσειράς επηρεάζονται σε μεγάλο βαθμό από το σταθερό βήμα της δειγματοληψίας .Η πλειοψηφία των χρονοσειρών έχει μίκρο και σταθερό βήμα όμως αρκετές φορές αυτό δεν είναι εφικτό. Χαρακτηριστικό παράδειγμα αποτελεί ο χρηματιστηριακό δείκτης μίας χώρας καθώς λόγω εορτών ή Σαββατοκύριακων είναι κλειστός.Αυτές οι περιπτώσεις απαιτούν ειδικές παραδοχές ή τεχνικές συμπλήρωσης των κενών τιμών με τον μέσο όρο της προηγουμένης και επόμενης παρατήρησης κτλ.

Οι χρονολογικές σειρές χωρίζονται σε κατηγορίες ως προς τη μέτρηση του χρόνου και ως προς το πλήθος των μεταβλητών που επηρεάζεται από την χρονική εξέλιξη. Έτσι, υπάρχουν χρονοσειρές διακριτού χρόνου (ημερήσιες, μηνιαίες, ετήσιες) και χρονοσειρές συνεχούς χρόνου (που αναφέρονται σε κάθε σημείο του χρόνου), όπως επίσης μονομεταβλητές χρονοσειρές (univariate time series) και πολυμεταβλητές χρονοσειρές (multivariate time series). Στην προσπάθεια μελέτης των ημερήσιων τιμών της μετοχής θα χρησιμοποιηθούν μονομεταβλητές χρονοσειρές διακριτού χρόνου, αφού το μοναδικό μέγεθος που επηρεάζεται είναι η τιμή κλεισίματος της μετοχής με χρονική μεταβολή μιας ημέρας.

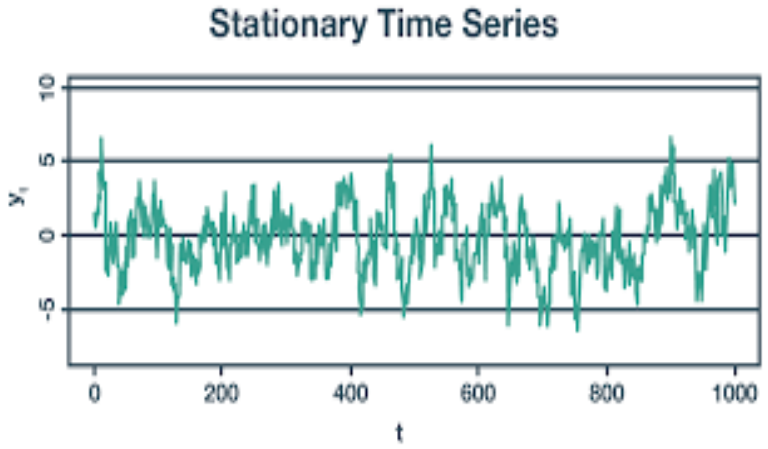
Τα τρία κυριότερα στοιχεία της ανάλυσης χρονοσειρών είναι η περιγραφή, η επεξήγηση και η πρόβλεψη των εξαρτημένων δεδομένων. Η περιγραφή επιτυγχάνεται με τη βοήθεια διαφόρων γραφημάτων, η επεξήγηση χρησιμοποιώντας κάποιας μορφής μοντέλα για να εξερευνηθούν οι μηχανισμοί δημιουργίας της χρονοσειράς, και η πρόβλεψη περιλαμβάνει τη χρησιμοποίηση ενός μμοντέλου για να προβλεφθούν μμελλοντικές τιμές της σειράς.

3.1Χαρακτηριστικά Χρονοσειρών(θεωρητική προσέγγιση)

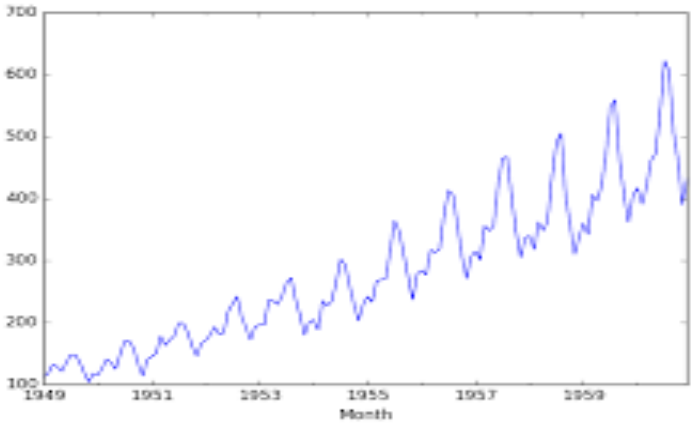
Όλες οι χρονοσειρές ανεξαρτήτως το μέγεθος που περιγράφουν (τιμή κλεισίματος μετοχής ,πωλήσεις παγωτών κτλ.) παρουσιάζουν ορισμένα βασικά χαρακτηριστικά τα οποία είναι τα εξής:

- Στασιμότητα(stationary)
- Τάση(trend)
- Περιοδικότητα(seasonality)
- Αυτοσυσχέτιση(auto-correlation)

Η στασιμότητα μίας χρονοσειράς επηρεάζεται από τη μέση τιμή και την διακύμανση - συνδιακύμανση στη πάροδο του χρόνου δηλαδή , εάν σε δύο διαδοχικές χρονικές στιγμές t_1 και t_2 τα στατιστικά μεγέθη δεν έχουν μεταβληθεί τότε είναι στατική. Για λόγους ευκολίας έχει επικρατήσει ο έλεγχος της μη μεταβολής της μέσης τιμής και της διακύμανσης ,και έχει γίνεται διάκριση σε αυστηρή και ασθενή στασιμότητα. Στη περίπτωση μας παρακάτω κάνουμε και τον στατιστικό έλεγχο dickey-fuller test.Από την γραφική απεικόνιση τους. Ακολουθούν παραδείγματα



Εικόνα 2 Στάσιμη Χρονοσειρά



Εικόνα 3 Μη στάσιμη Χρονοσειρά

Πριν την έναρξη της μελέτης και επεξεργασίας μίας χρονοσειράς είναι αναγκαίο ο έλεγχος για το εάν η χρονοσειρά είναι στατική ή όχι καθώς αποτελεί σημαντικό παράγοντα στις τεχνικές ανάλυσης και πρόβλεψης καθώς τα περισσότερα μοντέλα υποθέτουν ότι οι παρατηρήσεις της χρονοσειράς είναι ανεξάρτητες μεταξύ τους. Η ύπαρξη στασιμότητας σημαίνει ότι τα βασικά στατιστικά μεγέθη δεν μεταβάλλονται σημαντικά δηλαδή ότι οι παρατηρήσεις μεταβάλλονται με ομοιόμορφο τρόπο .Στη συνεχεία θα εξετάσουμε βασικές τεχνικές μετατροπής σε στατική χρονοσειρά.

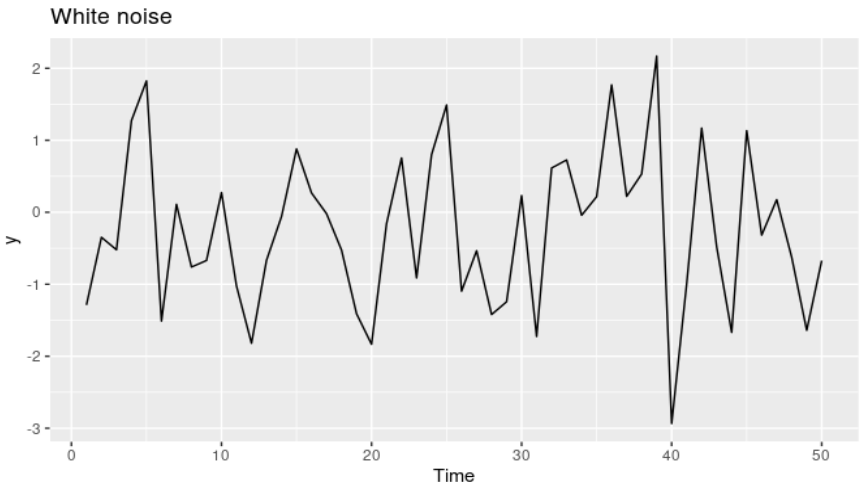
Ως τάση ορίζεται η μεταβολή των τιμών της χρονοσειράς στη διάσταση του χρόνου. Εξαρτάται από την κατεύθυνση που ακολουθεί η γραφική παράσταση των δεδομένων, καθώς μπορεί να είναι σταθερή, ανοδική ή να εναλλάσσεται κατά περιόδους. Η τάση μπορεί να είναι μία ευθεία γραμμή ή μία καμπύλη. Η μαθηματική της έκφραση είναι ο κινητός μέσος όρος .

Ως περιοδικότητα ορίζεται η μεταβολή των τιμών της χρονοσειράς κατά την διάρκεια συγκεκριμένων χρονικών διαστημάτων. Τα επαναλαμβανόμενα γεγονότα αποτελούν την εποχικότητα και μπορούν να βοηθήσουν στην άντληση χρήσιμων πληροφοριών. Πολύ συχνό είναι το φαινόμενο στο χώρο των πωλήσεων δηλαδή παρατηρείται αύξηση των πωλήσεων παγωτών κατά τους καλοκαιρινούς μήνες.

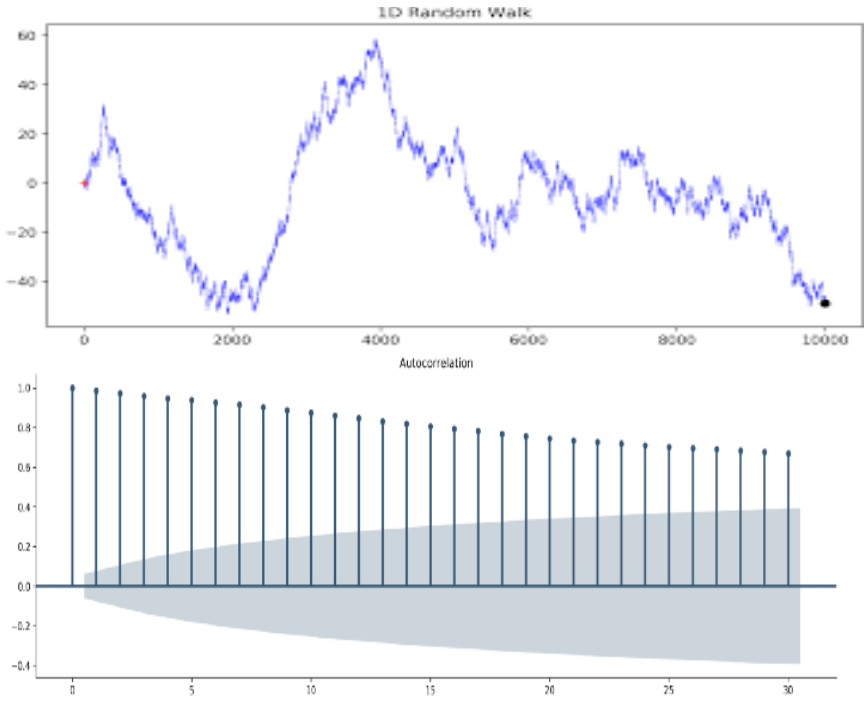
3.2 White Noise(Λευκός Θόρυβος) -Random Walk (Τυχαίος Περίπατος)

Στο σημείο αυτό αξίζει να γίνει ειδική μνεία για δύο χαρακτηριστικές περιπτώσεις χρονοσειράς , τον λευκό θόρυβο και τον τυχαίο περίπατο.

1. Ο λευκός θόρυβος αποτελεί μία από τις πιο απλές μορφές χρονοσειράς χωρίς τάση ή περιοδικότητα. Μια τέτοια χρονοσειρά είναι εντελώς τυχαία και δεν περιέχει αυτοσυσχετίσεις μεταξύ των παρατηρήσεων. Μια χρονοσειρά λέγεται λευκός θόρυβος (white noise) και θα συμβολίζουμε την κατανομή της ως WN(0, σ2 β), με μέση τιμή 0 και διασπορά σ 2 β . Αν επιπλέον τα στοιχεία της χρονοσειράς λευκού θορύβου ακολουθούν κανονική (Γκαουσιανή) κατανομή, τότε η χρονοσειρά λέγεται Γκαουσιανός λευκός θόρυβος (Gaussian white noise).
2. Τυχαίος περίπατος ορίζεται ως μία μη στάσιμη χρονοσειρά στην οποία κάθε τυχαία μεταβλητή στη χρονική στιγμή t προκύπτει από την προηγούμενη τυχαία παρατήρηση στη χρονική στιγμή t_{t-1}. Οι φαινομενικές συσχετίσεις που φαίνονται από ένα διάγραμμα ιστορίας τυχαίου περιπάτου οφείλονται στο ότι το τυχαίο βήμα σε κάθε χρονική στιγμή, το οποίο έχει γνωστό σημείο εκκίνησης.



Εικόνα 4 Γραφική Παράσταση Λευκού Θορύβου



Εικόνα 5 Γραφική Παράσταση Τυχαίου Περιπάτου(επάνω)-Διάγραμμα αυτοσυσχέτισης

Για να επιτευχθεί η ανάπτυξη ενός οικονομετρικού μοντέλου, ικανού να προβλέψει την εξέλιξη μιας των παρατηρήσεων μίας χρονοσειράς είναι απαραίτητο να ελεγχθεί το ενδεχόμενο, η χρονοσειρά να αποτελεί λευκό θόρυβο ή τυχαίο περίπατο. Στις δύο περιπτώσεις η πρόβλεψη της χρονοσειράς θα είναι αδύνατη, αφού για τον μεν λευκό θόρυβο θα πρόκειται για μια εντελώς τυχαία σειρά γεγονότων, αδύνατων να προσεγγιστούν από κάποιο μαθηματικό μοτίβο, για τον δε τυχαίο περίπατο θα πρόκειται για μια μη στάσιμη χρονοσειρά, για την οποία είναι εξ αρχής σαφές πως είναι αδύνατο να προβλεφθεί καθώς οι παρατηρήσεις δεν είναι απλώς τυχαίες αλλά συσχετίζονται με όλες τις προηγούμενες παρατηρήσεις . Στην πορεία του κεφαλαίου θα παρουσιαστεί αναλυτικά τόσο ο τρόπος με τον οποίο ελέγχεται η στασιμότητα μιας χρονοσειράς.

3.3 Βασικοί Στατιστικοί έλεγχοι (ανεξαρτησίας-στασιμότητας)

Ο πρώτος έλεγχος για τη στασιμότητα ή τη μη στασιμότητα μιας χρονοσειράς γίνεται οπτικά, παρατηρώντας προσεκτικά το γράφημά των παρατηρήσεών της. Όπως έχει ήδη αναφερθεί, μια στάσιμη χρονοσειρά χαρακτηρίζεται από σταθερή μέση τιμή και διακύμανση κατά την πάροδο του χρόνου.

Εκτός του παραπάνω οπτικού-εμπειρικού κανόνα υπάρχουν και στατιστικοί έλεγχοι για την επιβεβαίωση της στασιμότητας ή μη μίας χρονοσειράς. Τα πιο γνωστά είναι Augmented Dickey-Fuller Test (ADF test, 1996) και το Kwiatkowski Phillips Schmidt Shin Test (KPSS test, 1992).Στη συγκεκριμένη μελέτη θα γίνει χρήση του ADF test.

Τα δύο αυτά τεστ εξετάζουν την ύπαρξη ή την μη ύπαρξη μοναδιαίας ρίζας που συνεπάγεται τη μη στασιμότητα ή τη στασιμότητα της χρονοσειράς αντίστοιχα .Πιο συγκεκριμένα:

- Το ADF test έχει ως :

  Μηδενική υπόθεση H_0: την ύπαρξη μοναδιαίας ρίζας, δηλαδή μη στάσιμη χρονοσειρά.

  Εναλλακτική υπόθεση H_1: την μη ύπαρξη μοναδιαίας ρίζας, δηλαδή στάσιμη χρονοσειρά.

- Το KPSS test έχει ως:
  Μηδενική υπόθεση H_0: την μη ύπαρξη μοναδιαίας ρίζας, δηλαδή στάσιμη χρονοσειρά.

  Εναλλακτική υπόθεση H_1: την ύπαρξη μοναδιαίας ρίζας, δηλαδή μη στάσιμη χρονοσειρά.

Γίνεται αντιληπτό πως για να θεωρηθεί μια χρονοσειρά στάσιμη πρέπει η μηδενική υπόθεση του ADF test να απορρίπτεται, ενώ η μηδενική υπόθεση του KPSS test να επιβεβαιώνεται στο διάστημα εμπιστοσύνης που θα ορίζεται. Εφόσον η στασιμότητα επιβεβαιωθεί, τότε η διαδικασία μελέτης της χρονοσειράς μπορεί να συνεχιστεί κανονικά. Αν όμως η χρονοσειρά θεωρείται μη στάσιμη, τότε θα πρέπει να μετατραπεί σε στάσιμη, ώστε να μπορεί να αναπτυχθεί ένα μαθηματικό μοντέλο, ικανό να προβλέπει τις μελλοντικές της τιμές με όσο το δυνατόν μεγαλύτερή ακρίβεια.

Στο σημείο αυτό θα αναλύσουμε μερικούς βασικούς τρόπους μετατροπής μίας μη στατικής χρονοσειράς σε στατική.

3.4 Μετατροπή χρονοσειράς σε στάσιμη

Η έλλειψη στασιμότητας στις παρατήσεις μίας χρονοσειράς οφείλεται στο γεγονός ότι η μέση τιμή και η διακύμανση των παρατηρήσεων μεταβάλλεται με τη πάροδο του χρόνου. Οι παρατηρήσεις δηλαδή εμφανίζουν τάση ή εποχικότητα. Είναι αρκετά συνηθισμένο το φαινόμενο της εποχικότητας σε χρονοσειρές που σχετίζονται με πωλήσεις προϊόντων ή υπηρεσιών όπως αναψυκτικά τους καλοκαιρινούς μήνες, αυξήσεις πωλήσεων εισιτηρίων σε περιόδους διακοπών κτλ .

Για να μετατραπεί μια μη στάσιμη χρονοσειρά σε στάσιμη, πρέπει να γίνει απαλοιφή της τάσης και της εποχικότητας. Η απαλοιφή αυτή γίνεται για να «απομονωθεί» το σύστημα που θέλουμε να μελετήσουμε από εξωγενείς παράγοντες. Στο στάδιο της πρόβλεψης της χρονοσειράς, η τάση και η περιοδικότητα προστίθενται στο μοντέλο που έχει εκτιμηθεί, ώστε οι να επιτευχθεί η βέλτιστη προσαρμογή του μοντέλου στα δεδομένα της χρονοσειράς.

Γενικά μια χρονοσειρά μπορούμε να τη χωρίσουμε σε τρεις συνιστώσες ως εξής :

$$X_t = \mu_t + s_t + y_t$$

όπου μ_t είναι ή συνιστώσα της τάσης, s_t η συνιστώσα της περιοδικότητας για κάποια περίοδο d ($s_t - d = s_t$) και y_t είναι η χρονοσειρά των υπολοίπων αν αφαιρέσουμε από την παρατηρούμενη χρονοσειρά την τάση και την περιοδικότητα. Η τάση και η περιοδικότητα είναι και οι δυό συναρτήσεις του χρόνου και δεν περιέχουν πληροφορία για τη δυναμική του συστήματος, δηλαδή την εξάρτηση της παρατήρησης X_t από τις προηγούμενες παρατηρήσεις. Σε κάποια προβλήματα όλη η πληροφορία που μας ενδιαφέρει μπορεί να εντοπίζεται στη συνιστώσα της τάσης και της περιοδικότητας, οπότε το πρόβλημα περιορίζεται στην εκτίμηση των μ_t και s_t. Σε άλλες περιπτώσεις θέλουμε να εξαλείψουμε την τάση και την περιοδικότητα για να διερευνήσουμε τη δυναμική του συστήματος και θα πρέπει αφού εκτιμήσουμε τις συναρτήσεις μt και s_t να τις αφαιρέσουμε από της X_t για να πάρουμε τη χρονοσειρά των υπολοίπων y_t.

3.4.1 Διαφορά τάξεων

Στη βιβλιογραφία προτείνεται ο μετασχηματισμός των δεδομένων μιας χρονοσειράς ώστε να εξασφαλιστεί η στασιμότητα, όπως για παράδειγμα οι διαφορές 1ης ή 2ης τάξης κτλ .

Μπορούμε να δημιουργήσουμε μια χρονοσειρά η οποία προκύπτει από την διαφορά των παρατηρήσεων της δηλαδή:

Διαφορά $1_{ης}$ τάξης: $Zt = Yt - Yt-1$

Διαφορά $2_{ης}$ τάξης: $Zt = Yt - Yt-2$

Διαφορά $2_{ης}$ τάξης: $Zt = Yt - Yt-3$

Αφού αφαιρεθούν τα στοιχεία της τάσης ή της εποχικότητας ή και τα δύο μαζί, γίνεται επανέλεγχος για τη στασιμότητα της χρονοσειράς που έχει προκύψει από τους μετασχηματισμούς που χρησιμοποιήθηκαν.

Ο έλεγχος γίνεται παρατηρώντας ξανά τόσο το γράφημα της νέας χρονοσειράς, όσο και της συνάρτησης αυτοσυσχέτισής της, καθώς επίσης και τα ADF και KPSS tests.

Αν η χρονοσειρά εξακολουθεί να είναι μη στάσιμη, τότε μπορούν να χρησιμοποιηθούν οι διαφορές 1ης ή 2ης τάξης, μέχρις ότου να οδηγηθούμε σε στασιμότητα της υπό εξέταση χρονοσειράς.

Έχοντας λύσει το ζήτημα της στασιμότητας, μπορεί να συνεχιστεί η ανάλυση των δεδομένων των χρονοσειρών, εκτιμώντας ένα μοντέλο που να προσαρμόζεται σε αυτά και να μπορεί να οδηγήσει σε μια πρόβλεψη.

3.4.2 (Seasonal Differencing)

Στη μέθοδο αυτή ,αντί να υπολογίζουμε την διαφορά μεταξύ δύο συνεχόμενων παρατηρήσεων επιλέγουμε να υπολογίσουμε τη διαφορά μεταξύ δύο συνεχόμενων παρατηρήσεων της ίδιας περιόδου. Δηλαδή στις περιπτώσεις που έχουμε εβδομαδιαίες παρατηρήσεις μια μέτρηση της ημέρας Δευτέρας θα αφαιρεθεί από την παρατήρηση της προηγούμενης Δευτέρας δηλαδή περίοδο 7 ημερών.

Έτσι για ετήσια δεδομένα έχουμε $Y_t - Y_{t-12,}$για εξαμηνία δεδομένα $Y_t - Y_{t-6,}$ για εβδομαδιαία $Y_t - Y_{t-7}$.

ΚΕΦΑΛΑΙΟ 4

4. Γραμμικά Μοντέλα Ανάλυσης Χρονοσειρών και Πρόβλεψης

Κύριο στόχο στην μελέτη και ανάλυση χρονοσειρών αποτελεί η διερεύνηση των μελλοντικών τιμών της εκάστοτε χρονοσειράς δηλαδή η πρόβλεψη. Η πρόβλεψη χρονοσειρών, δηλαδή η ανάπτυξη μοντέλων που επιτρέπουν την περιγραφή και την πρόβλεψη της εξέλιξης των τιμών της χρονοσειράς, αποτελεί σημαντική πηγή πληροφόρησης για την υποστήριξη λήψης αποφάσεων ιδιαίτερα στο κλάδο της οικονομίας. Σε αυτό το κεφάλαιο παρουσιάζονται οι πιο γνωστές μέθοδοι πρόβλεψης χρονοσειρών, όπως είναι διάφορα μοντέλα γραμμικών στοχαστικών διαδικασιών (AR, MA, ARMA), η ανάλυση ARIMA.

Για να μπορέσουν οι μεθοδολογίες που ακολουθούν να εφαρμοστούν με την μέγιστη αποτελεσματικότητα θωρούμε ότι εφαρμόζονται σε στάσιμες χρονοσειρές.

4.1 Γραμμική παλινδρόμηση (Linear Regression)

Η πρώτη προσέγγιση απλών γραμμικών στοχαστικών διαδικασιών, αφορά το μοντέλο της γραμμικής παλινδρόμησης (linear regression) και είναι το πιο απλό και κατανοητό. Γενικά, το μοντέλο εκφράζει τις σχέσεις μεταξύ ανεξάρτητων και εξαρτημένων μεταβλητών, διερευνώντας το βαθμό συσχέτισής τους μέσω μίας απλής συνάρτησης.

4.1.1 Απλή γραμμική παλινδρόμηση (Simple Linear Regression)

Στη περίπτωση της απλής γραμμικής παλινδρόμησης (Simple Linear Regression), εκφράζεται μέσω της γραμμική σχέσης ανάμεσα σε μια εξαρτημένη μεταβλητή Y (μεταβλητή πρόβλεψης) και σε μια ανεξάρτητη μεταβλητή X Ο μαθηματικός ορισμός του μοντέλου πρόβλεψης της απλής γραμμικής παλινδρόμησης για τις χρονοσειρές με μία ανεξάρτητη μεταβλητή:

$$\boldsymbol{Y}_t = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \mathbf{x}_t + \boldsymbol{\varepsilon}_t$$

όπου $\boldsymbol{\beta}0$, $\boldsymbol{\beta}1$ οι άγνωστοι συντελεστές του μοντέλου και $\boldsymbol{\varepsilon}t$ το τυχαίο σφάλμα .

4.2 Αυτοπαλινδρούμενα Μοντέλα (Auto Regressive Models)

Στις γραμμικές στοχαστικές διαδικασίες ανήκουν και τα αυτοπαλινδρούμενα μοντέλα. Πιο συγκεκριμένα, είναι μοντέλα γραμμικής παλινδρόμησης όπου η εξαρτημένη μεταβλητή είναι η τυχαία μεταβλητή της χρονοσειράς X σε μία χρονική στιγμή t, δηλαδή τη Xt και ως ανεξάρτητη μεταβλητή ορίζεται η τυχαία μεταβλητή X της χρονοσειράς αλλά σε προηγούμενους χρόνους, δηλαδή $Xt-1$, … , $Xt-p$. Ο αριθμός p, αφορά την τάξη (order) του μοντέλου και αναφέρεται στο μήκος της υστερήσεως . Δηλαδή πρόκειται για μία παλινδρόμηση της ίδιας μεταβλητής μεταξύ παροντικών και παρελθοντικών χρονικών στιγμών t .

Απαιτείται η εφαρμογή μεθόδων εξομάλυνσης δηλαδή η μείωση ή η εξάλειψη του θορύβου στα δεδομένα της χρονοσειράς ώστε να εμφανιστούν τα βασικά χαρακτηριστικά της εν λόγω χρονοσειράς όπως στασιμότητα ή μη ,κυκλικότητα και εποχικότητα και να εφαρμοστεί το κατάλληλο μοντέλο.

4.3 Μοντέλα Κινητού Μέσου (Moving Average)

Τα μοντέλα κινητού μέσου (MA) ανήκουν στις γραμμικές στοχαστικές διαδικασίες όπου αντί για την ανάλυση της μεταβλητής πρόβλεψης στο μοντέλο της παλινδρόμησης αναλύονται οι παρελθοντικές τιμές του τυχαίου σφάλματος. Πιο ειδικά, για κάθε παρατήρηση μιας χρονοσειράς Xt , προκύπτει ως ένα σταθμισμένο άθροισμα μίας σταθεράς c, μιας χρονοσειράς λευκού θορύβου και q καθυστερημένων εκδοχών της χρονοσειράς λευκού θορύβου .

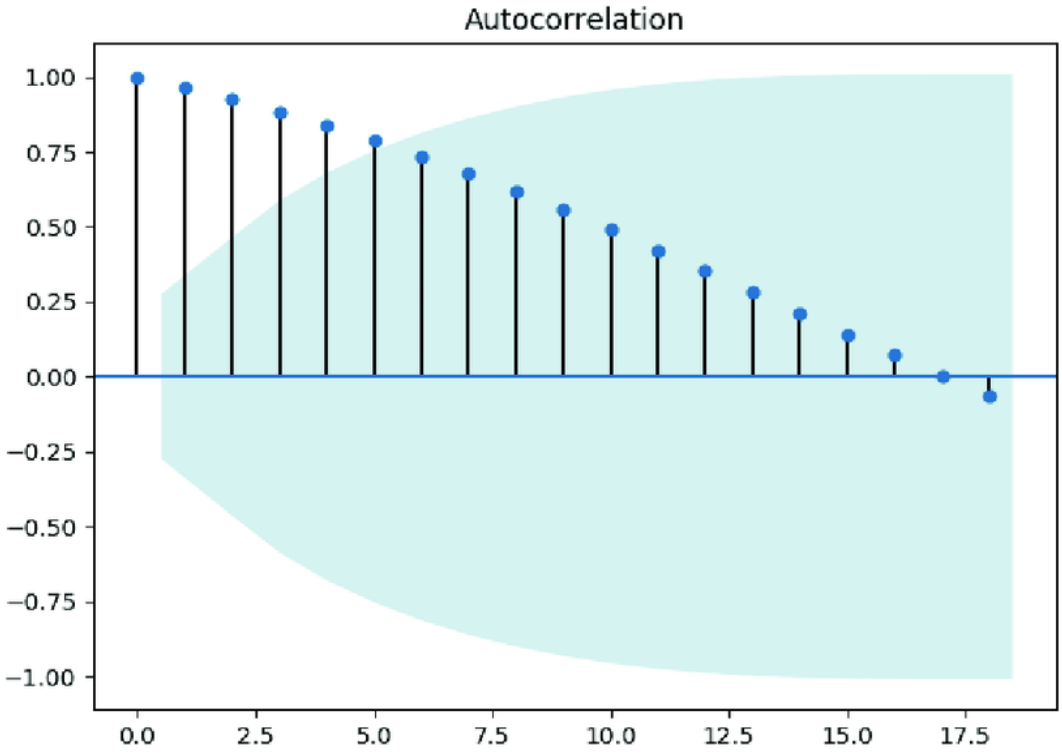
Τα μοντέλα κινούμενου μέσου θεωρούνται τα πλέον κατάλληλα για την περιγραφή μιας οικονομικής χρονοσειράς, αφού έχουν τη δυνατότητα να περιγράφουν φαινόμενα στα οποία τα γεγονότα έχουν μια άμεση επίδραση βραχυχρόνιας διάρκειας .Αυτό πρακτικά σημαίνει ότι μια πληροφορία ή ένα γεγονός επηρεάζει την επόμενη τιμή της χρονοσειράς, ωστόσο η επιρροή αυτή συνεχίζει να υφίσταται και στις αμέσως επόμενες χρονικά τιμές.Υπάρχουν δύο βασικές κατηγορίες ο απλός και κινητός μέσος όρος.

Απλός Μέσος Όρος: υπολογίζουμε τον μέσο όρο των παρατηρήσεων του δείγματος για να προβλέψουμε τη μελλοντική τιμή. Η διαδικασία της πρόβλεψης είναι αρκετά απλή και δεν χρησιμοποιείται στην βιβλιογραφία.

Κινητός Μέσος Όρος: Στη συγκεκριμένη μέθοδο, υπολογίζουμε τον μέσο όρο των n πρόσφατων παρατηρήσεων του δείγματος και τον χρησιμοποιούμε για να προβλέψουμε την επόμενη περίοδο. Για κάθε νέα παρατήρηση που εισέρχεται στο δείγμα, καταργούμε την παλαιότερη παρατήρηση και υπολογίζουμε εκ νέου μια νέα μέση τιμή, η οποία αποτελεί την πρόβλεψη της επόμενης περιόδου

Επειδή πρόκειται για μια χρονοσειρά λευκού θορύβου, η διαδικασία κινούμενου μέσου αποτελεί μια στάσιμη χρονοσειρά με μηδενική μέση τιμή. Επομένως, για ένα μοντέλο κινούμενου μέσου τάξης q MA(q).

Για την επιλογή του ακεραίου αριθμού της τάξης του μοντέλου γίνεται χρήση της συνάρτησης αυτοσυσχέτισης (auto correlation function ή acf), αφού στο γράφημά της (κορελόγραμμα) εύκολα διαπιστώνει κανείς πως γίνεται μηδέν μετά από q χρονικές υστερήσεις.



Εικόνα 6 Διάγραμμα αυτοσυσχέτισης

4.4 Μεικτά μοντέλα ή αυτοπαλινδρομούμενα κινητού μέσου ARIMA (p,q)

Η ιδιότητα της στασιμότητας είναι ιδιαίτερα σημαντική για τις χρονοσειρές και δεν πρέπει να παραλείπεται. Στις περισσότερες περιπτώσεις όμως, οι χρονοσειρές δε διέπονται από αυτήν την ιδιότητα και χρειάζεται η μετατροπή τους σε στάσιμες χρονοσειρές. Αν η μη-στασιμότητα οφείλεται στην ύπαρξη τάσης), υπάρχουν διάφοροι μέθοδοι εξομάλυνσης της τάσης οι οποίες αναφέρθηκαν στην ενότητα 3.4 .

Μία από αυτές, είναι η μέθοδος της διαφοράς η οποία είναι και η πιο απλή μέθοδος κατά την οποία μια μη-στάσιμη χρονοσειρά μετατρέπεται σε στάσιμη, χρησιμοποιώντας τον μετασχηματισμό των διαφορών με προγενέστερες παρατηρήσεις. Παίρνοντας τις πρώτες d διαφορές η χρονοσειρά ονομάζεται ολοκληρωμένη πρώτης τάξης και συμβολίζεται με I(d). Εάν και πάλι, η χρονοσειρά είναι μη-στάσιμη, τότε η διαδικασία διαφοράς επαναλαμβάνεται παίρνοντας τις διαφορές δεύτερης ή και μεγαλύτερης τάξης μέχρι να μετατραπεί σε στάσιμη χρονοσειρά. Ο αριθμός των διαφορών που μετατρέπει μία χρονοσειρά σε στάσιμη συμβολίζεται με d και η χρονοσειρά ονομάζεται ολοκληρωμένη d τάξεως και συμβολίζεται με I(d), αντίστοιχα.

Τα αυτοπαλίνδρομα μοντέλα κινητού μέσου ARMA(p,q), εφαρμόζονται μόνο σε στάσιμες χρονοσειρές, μπορούν να χρησιμοποιηθούν και σε μη στάσιμες χρονοσειρές οι οποίες μετατρέπονται σε. Συνδυάζοντας την μέθοδο της διαφόρισης για d επαναλήψεις μίας μη-στάσιμης χρονοσειράς με το μοντέλο ARMA(p,q), προκύπτει το ολοκληρωμένο μικτό μοντέλο ή ολοκληρωμένο αυτοπαλινδρομούμενο μοντέλο κινούμενου μέσου (Auto Regressive Integrated Moving Average model, ARIMA) που συμβολίζεται με ARIMA(p, d, q) όπου:

p:ο αριθμός τάξης του μοντέλου αυτοσισχέτισης

d:ο αριθμός τάξης διαφορών για την μετατροπή σε στάσιμη χρονοσειρά

q:η τάξη του κινούμενου μέσου όρου.

Στο σημείο αυτό να τονιστεί ότι η τιμή της παραμέτρου d ορίζεται από τον στατιστικό Dickey Fuller test είτε κατασκευάζοντας τη συνάρτηση αυτοσυσχέτισης και την γραφική της παράσταση. Ο καθορισμός των παραμέτρων p και q υπολογίζονται με τη χρήση διαγραμμάτων της αυτοσυσχέτισης (ACF) και της μη αυτοσυσχέτισης (PACF).

4.4.2 Μοντέλα Χρονοσειράς με Εποχικότητα

Μία ακόμα κατηγορία μοντέλων αφορά στις χρονοσειρές που χαρακτηρίζονται από εποχικότητα (περιοδικότητα). Σε αυτήν την κατηγορία εντάσσονται λοιπόν μοντέλα τα οποία ουσιαστικά αποτελούν υποκατηγορία των μοντέλων ARIMA, δηλαδή αποτελούνται κι αυτά από ένα αυτοπαλίνδρομο μέρος (AR), ένα μέρος διαφορών (I) κι ένα μέρος κινούμενου μέσου (MA). Τα εποχικά μοντέλα, καλούνται κατά αντιστοιχία Seasonal AutoRegressive Integrated Moving Average models (SARIMA) .

Το εποχικό υπόδειγμα ARIMA είναι το παρακάτω

$$ARIMA(p,d,q) \times (P,D,Q)$$

Όπου:

p:ο αριθμός τάξης του μοντέλου αυτοσισχέτισης

d: ο αριθμός τάξης διαφορών για την μετατροπή σε στάσιμη χρονοσειρά

q:η τάξη του κινούμενου μέσου όρου.

P : το πλήθος των εποχικών αυτοπαλίνδρομων όρων(SAR)

D : το πλήθος των εποχικών διαφορών

Q: το πλήθος των εποχικών όρων κινητού μέσου(SMA)

Όπως και στο μοντέλο ARIMA έτσι και στο εποχικό μοντέλο απαιτείται ο καθορισμός της τιμής της εποχικής διαφοράς εφόσον υπάρχει.

4.5 Κριτήρια επιλογής μοντέλου

Η επιλογή του καταλληλότερου μοντέλου ARIMA για την ερμηνεία και πρόβλεψή μίας χρονοσειράς δεν είναι πάντα προφανής και εύκολη. Συχνά περισσότερα από ένα μοντέλα μπορούν σχεδόν να ταυτίζονται. Επιπλέον, μπορεί κάποιο μοντέλο να προσαρμόζεται αποτελεσματικότερα με καλύτερα αποτελέσματα σε μία χρονοσειρά αλλά η πολυπλοκότητά του να είναι σημαντικά μεγαλύτερη γεγονός που καθιστά την επιλογή του απαγορευτική είτε για λόγους πόρων είτε για λόγους αποδοτικότητας.

Τα κυριότερα κριτήρια επιλογής του αποδοτικότερου μοντέλου ARIMA είναι το Bayesian Information Criterion (BIC) και το ο Akaike's Information Criterion (AIC).Και τα δύο αυτά κριτήρια μας δείχνουν κατά πόσοι αξίζει τα υπό εξέταση μοντέλα να γίνουν πιο πολύπλοκα ώστε να αυξηθεί η πιθανότητα οι παραγόμενες τιμές να είναι όσο το δυνατόν πλησιέστερες με τις πραγματικές.

Στα μοντέλα πρόβλεψης ισχύει το γεγονός ότι όσο αυξάνονται οι παρατηρήσεις του train-set τόσο αυξάνεται και η πολυπλοκότητα αλλά μειώνεται η προκατάληψή του. Στο σημείο αυτό ελλοχεύει ο κίνδυνος του over-fitting στα δεδομένα του train-set και να μειωθεί σοβαρά η απόδοση του στο validation-set λόγω της αύξηση της διακύμανσης των σφαλμάτων των παρατηρήσεων. Επομένως το ιδανικό σημείο είναι το σημείο στο οποίο καθορίζουμε την πολυπλοκότητα του μοντέλου ώστε να έχουμε μικρό τιμή προκατάληψης και υψηλή απόδοση.

Τα κριτήρια αυτά δεν έχουν κάποια σταθερή τιμή σύγκρισης αλλά μόνο με τις τιμές των υπό εξέταση μοντέλων. Για το λόγω αυτό επειδή υπολογίζονται με τη μέθοδο μέγιστης πιθανοφάνειας επιλέγονται το μοντέλα με την ελάχιστη τιμή.

- **Akaike's Information Criterion (AIC):**

$$AIC = -2\log L + 2(p+q+k+1)$$

,όπου κ=0 εάν η σταθερά του μοντέλου c ισούται με 0 και κ=1 σε αντίθετο σενάριο.

- **Bayesian Information Criterion (BIC)**

$$BIC = AIC + \log(n)(p+q+k+1)$$

Το κριτήριο BIC δίνει έμφαση στην πολυπλοκότητα του μοντέλου για την αποφυγή του φαινομένου υπέρ-προσαρμογής(over-fitting)

Εφόσον έχουν υπολογισθεί οι βέλτιστες τιμές των παραμέτρων ARIMA για όλα τα μοντέλα, συγκρίνονται οι τιμές των κριτηρίων και επιλέγονται αυτά με το ελάχιστο.

ΚΕΦΑΛΑΙΟ 5

5. Ανάλυση τιμών μετοχών στο περιβάλλον Google Colab.

Στο σημείο αυτό θα παρουσιαστεί βήμα προς βήμα η διαδικασία ARIMA για τις τιμές κλεισίματος των μετοχών της αμερικανικής εταιρείας HOME DEPOT.Η συλλογή των δεδομένων έγινε από τον ιστοσελίδα https://finance.yahoo.com/quote/HD?p=HD&.tsrc=fin-srch.

Η περίοδος μελέτης αφορά τη χρονική περίοδο 2016-11-17 έως 2021-11-16. Όπως έχει ήδη αναφερθεί, θα χρησιμοποιηθούν οι τιμές κλεισίματος (close prices) των μετοχών ως χρονοσειρές διακριτού χρόνου, οι οποίες έχουν ακανόνιστη συμπεριφορά ως προς την περιοδικότητα, αφού δεν υπάρχουν τιμές στις αργίες, στις εθνικές εορτές και τα σαββατοκύριακα. Η όλη διαδικασία θα υλοποιηθεί στο περιβάλλον του Google Colab , εκμεταλλευόμενοι τα διάφορα πακέτα βιβλιοθηκών της γλώσσας προγραμματισμού Python. Αρχικά, λοιπόν, παρουσιάζονται οι βιβλιοθήκες (libraries) και οι εντολές – συναρτήσεις που θα χρησιμοποιηθούν στην τρέχουσα ανάλυση.

5.1 Γλώσσα Python και Google Colab

Δημιουργήθηκε από τον Ολλανδό Κίντο βαν Ρόσσουμ (Guido van Rossum) στο ερευνητικό κέντρο Centrum Wiskunde & Informatica (CWI) το 1989. Αρχικά, η Python ήταν γλώσσα σεναρίων .Είναι μία γλώσσα αντικειμενοστραφούς προγραμματισμού και παρέχει ένα μεγάλο φάσμα στατιστικών τεχνικών, όπως η γραμμική και μη γραμμική παλινδρόμηση, η ομαδοποίηση, η ανάλυση χρονοσειρών ανάπτυξη νευρωνικών δικτύων και άλλα.

Το πιο βασικό πλεονέκτημα της Python είναι ότι πρόκειται για μια γλώσσα ανοιχτού κώδικα (open source) και είναι ελεύθερη για τον καθένα να τη χρησιμοποιήσει. Προσφέρει μεγάλη ευκολία στη κατανόηση καθώς είναι αρκετά κοντά στην φυσική γλώσσα, ο διερμηνευτής της παρουσιάζει άμεσα τα λάθη που εμφανίζονται και το moto της σύμφωνα με τον ιδρυτή της είναι το **"there should be one and preferably only one obvious way to do instead of There's more than one way to do it"**

Ο κώδικας θα γραφεί στο google colad ένα web based python ide.Οι βασικές βιβλιοθήκες που κλήθηκαν στην εργασία είναι η pandas,numpy βασικές διότι κυρίως για τον διανυσματικό προγραμματισμό και για την ευκολία επεξεργασίας μεγάλου όγκου δεδομένων και για την ταχύτητα επεξεργασίας.

Πιο συγκεκριμένα για την ανάλυση χρονοσειράς έγινε κλήση της **statsmodels.graphics.tsaplots** για τον υπολογισμό και οπτικοποίηση των γραφικών παραστάσεων συσχέτισης και αυτοσυσχέτισης.

Για τον υπολογισμό των παραμέτρων του μοντέλου ARIMA έγινε χρήση της **pmdarima** .

5.2 Ανάλυση των τιμών μετοχής HOME DEPOT-Πείραμα.

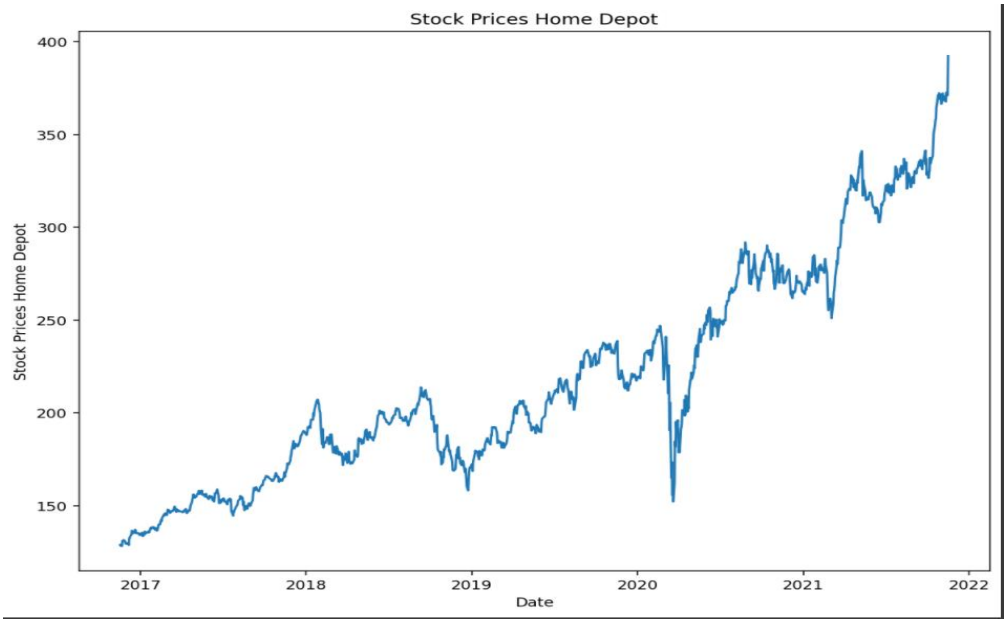
Αρχικά εισάγουμε στο ide τα δεδομένα με τις τιμές σε μορφή .csv αρχείου ,και με τη την βιβλιοθήκη pandas το μετατρέπουμε σε dataframe .

Date	Open	High	Low	Close	Adj Close	Volume
2016-11-17	125.870003	129.139999	125.699997	128.929993	114.675789	6801200
2016-11-18	128.929993	129.360001	127.720001	128.330002	114.142120	4428900
2016-11-21	128.279999	128.899994	127.410004	128.220001	114.044281	4072100
2016-11-22	128.380005	131.300003	128.380005	130.979996	116.499130	5535800
2016-11-23	131.360001	131.990005	130.860001	131.210007	116.703728	3582700
...	...	...	...	...	...	...
2021-11-10	368.250000	371.209991	367.720001	368.579987	368.579987	2161400
2021-11-11	371.000000	372.000000	365.829987	367.630005	367.630005	2387300
2021-11-12	369.119995	373.500000	366.700012	372.630005	372.630005	2792600
2021-11-15	374.390015	374.869995	369.290009	371.079987	371.079987	3650600
2021-11-16	382.000000	394.380005	379.000000	392.329987	392.329987	8648200

Εικόνα 7 Πίνακας τιμών μετοχών HM

Όπως φαίνεται στην άνω εικόνα στον πίνακα υπάρχουν διαθέσιμες προς επεξεργασία η ημερομηνία, η μέγιστη αξία της μετοχής ανά ημέρα, η ελάχιστη αξία, η τιμή ανοίγματος, η τιμή κλεισίματος, η προσαρμοσμένη τιμή κλεισίματος και ο αριθμός συναλλαγών.

Δημιουργούμε την χρονοσειρά για την περίοδο 17/11/2016-16/11/2021 με παρατηρήσεις τις τιμές κλεισίματος και μετατρέπουμε τη στήλη Date σε δείκτη(index) του dataframe και έπειτα οπτικοποιούμε τη χρονοσειρά.



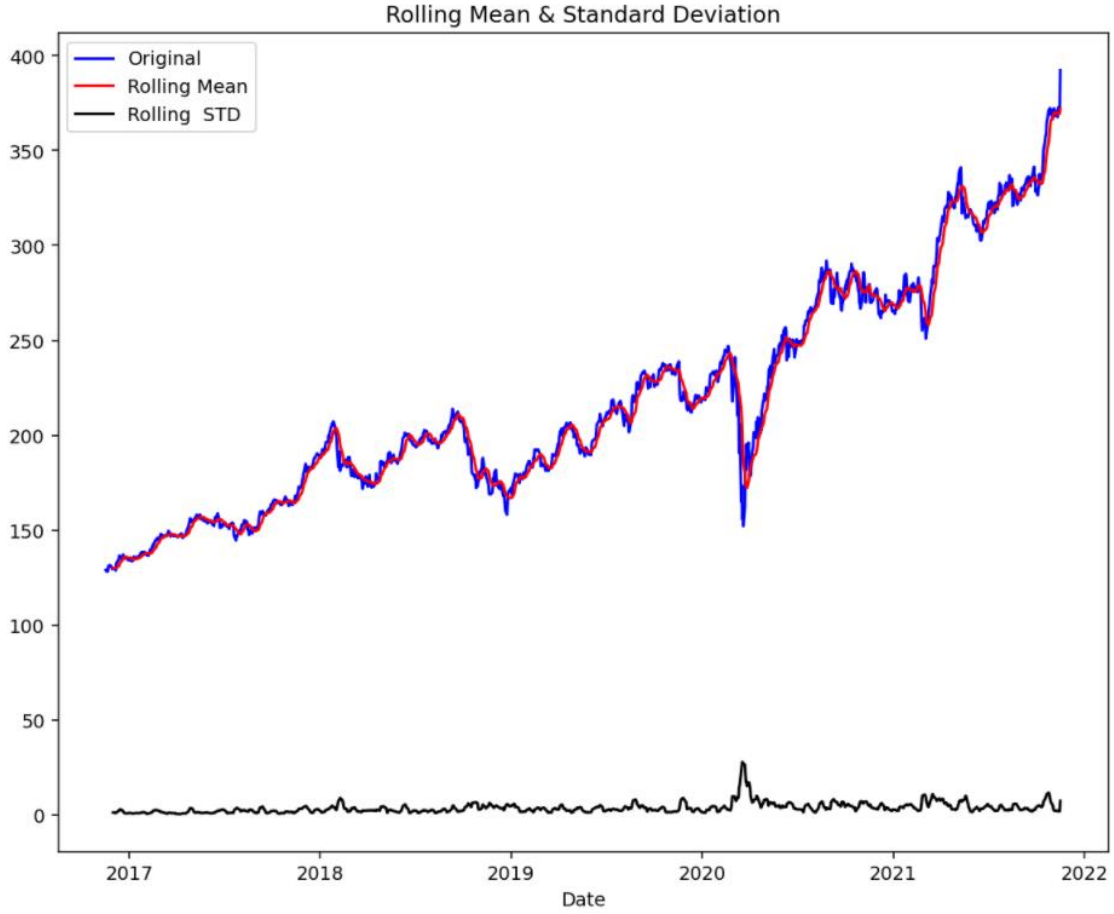
Εικόνα 7 Γραφική παράσταση τιμής κλεισίματος HOME DEPOT

Στο σημείο αυτό και πριν συνεχίσουμε στα επόμενα στάδια πρέπει να ελέγξουμε την στασιμότητας η την μη της χρονοσειράς. Αρχικά από την γραφική παράσταση των παρατηρήσεων είναι προφανές

ότι πρόκειται για μία μη στάσιμη χρονοσειρά ,καθώς η μέση τιμή μεταβάλλεται με τη πάροδο του χρόνου.Θα γίνει ο έλεγχος με δύο τρόπους

- Με την οπτικοποίση του κινητού μέσου
- Και του στατιστικού ελέγχου ADF

Κινητός μέσος όρος τάξης 10 παρατηρήσεων.



Εικόνα 7 Γραφική παράσταση κινητού μέσου όρου

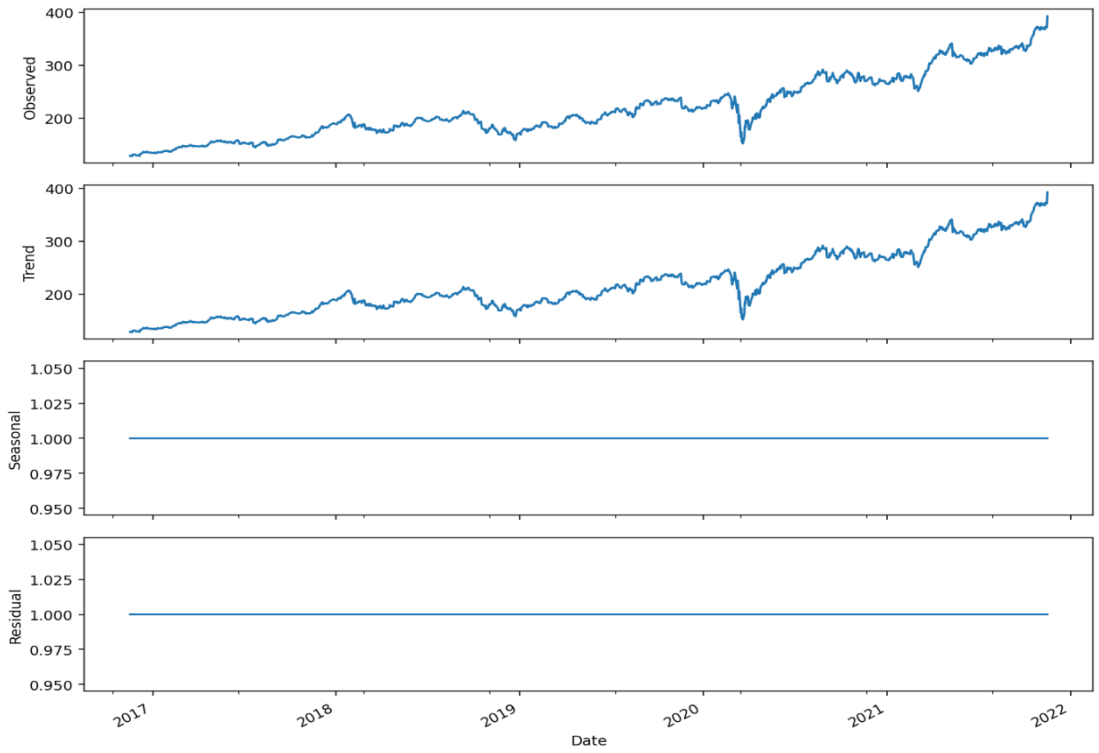
Είναι οπτικά ξεκάθαρο ότι ο μέση τιμή μεταβάλλεται με τη πάροδο του χρόνου.

```
ad_test(data['Close'])

1. ADF :  0.4744873861859998
2. P-Value :  0.9840766819064242
3. Num Of Lags :  9
4. Num Of Observations Used For ADF Regression: 1248
5. Critical Values :
        1% :  -3.4356006420838963
        5% :  -2.8638586845641063
        10% :  -2.5680044958343604
```

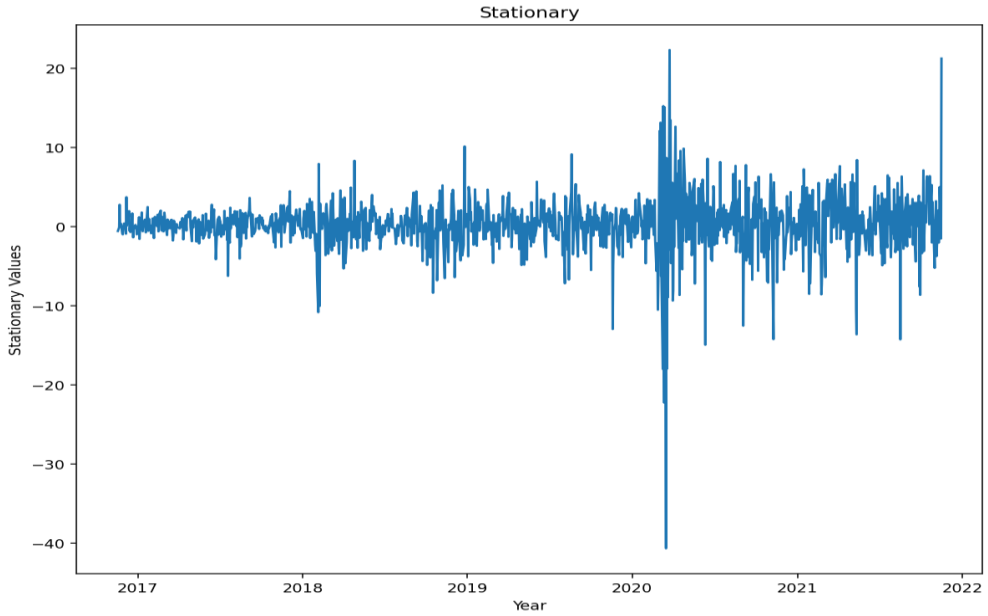
Για το ADF test η μηδενική υπόθεση που συνεπάγεται μη στασιμότητας επαληθεύεται, αφού το p – value ισούται με 0,98407 > 0,05.Επομένως για να γίνει χρήση του μοντέλου ARIMA πρέπει να μετατραπεί σε στάσιμη πρώτα.

Επιπλέον μπορούμε να οπτικοποιήσουμε με τη βοήθεια της βιβλιοθήκης **statsmodels.tsa.seasonal** να διασπάσουμε την χρονοσοσειρά σε στοιχεία της όπως η τάση και η εποχικότητα και τα κατάλοιπα. Στις παραμέτρους χρησιμοποιούμε το πολλαπλασιαστικό μοντέλο διότι η γραφική παράσταση έχει ανοδική τάση.



Εικόνα 8 Γραφική παράσταση τάσης,εποχικότητας και καταλοίπων

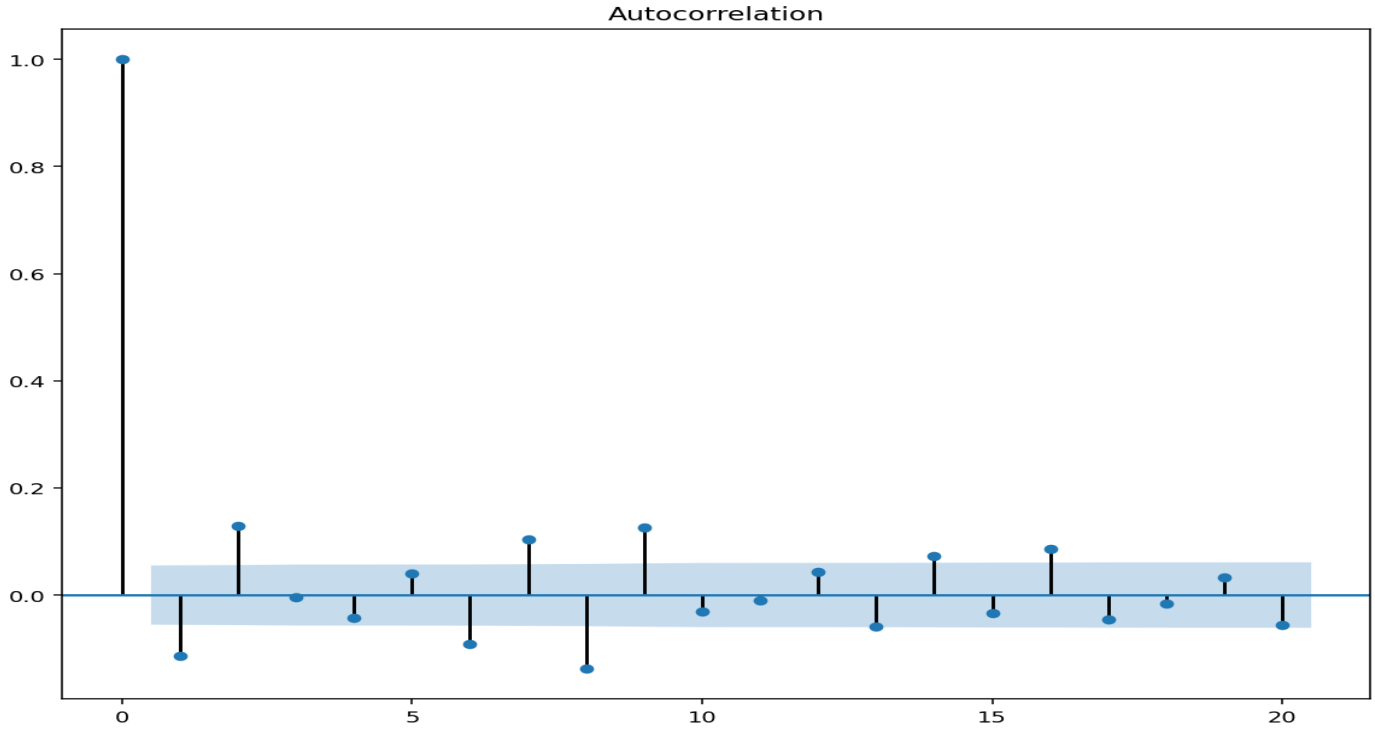
Η αρχική χρονοσειρά μετατρέπεται σε στάσιμη μέσω της μεθόδου των διαφορών. Αυτό έχει ώς αποτέλεσμα τη δημιουργία μίας νέας στάσιμης χρονοσειράς το γράφημα της οποίας είναι το εξής.



Εικόνα 8 Γραφική παράσταση στάσιμης χρονοσειράς

'Αν εξαιρεθεί η ακραία πτώση της τιμής κατά την έναρξη του έτους 2020 παρατηρείται μία σχετική σταθερότητα στη μέση τιμή της νέας χρονοσειράς στο πέρασμα του χρόνου.

Το γεγονός αυτό επιβεβαιώνεται με το στατιστικό έλεγχο ADF αλλά και από τη γραφική παράσταση των διαγραμμάτων αυτοσυσχέτισης.



Εικόνα 9 Γραφική παράσταση αυτοσυσχέτισης στάσιμης χρονοσειράς

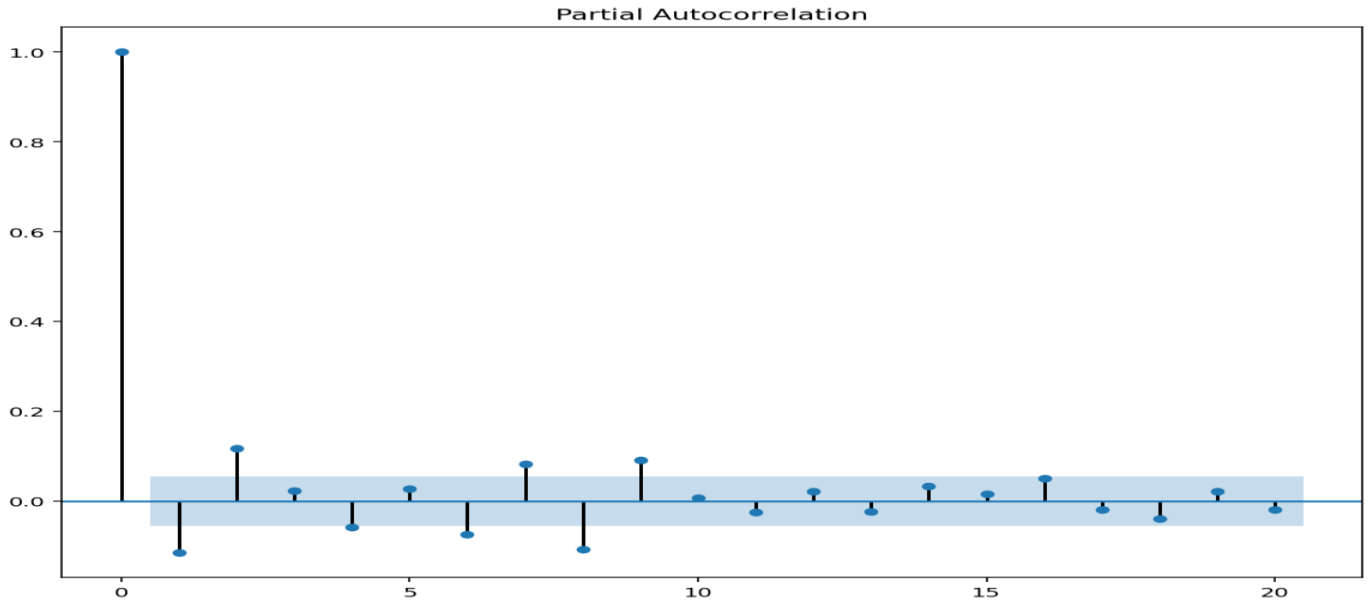
```
ad_test(df['Close'])

1. ADF :  -11.245858524357729
2. P-Value :  1.7604023661008886e-20
```

Εικόνα 10 ADF test

Το γράφημα της συνάρτησης αυτοσυσχέτισης «σβήνει» γρήγορα, γεγονός που οδηγεί σε ένδειξη στασιμότητας. Επίσης, η μηδενική υπόθεση του ADF test απορρίπτεται, αφού το p – value ισούται με τιμή πολύ μικρότερη του 0,05 threshold της μηδενικής υπόθεσης.

Στο σημείο αυτό προχωράμε στην αναζήτηση των παραμέτρων του μοντέλου ARIMA.Γνωρίζουμε ότι η τάξη για την μετατροπή σε στάσιμη χρονοσειρά είναι 1 άρα d=1.Αρχικά κατασκευάζουμε τα διαγράμματα αυτοσυσχέτισης και μερικής αυτοσυσχέτικης,επομένως έχουμε :



Εικόνα 11 Γραφική παράσταση μερικής αυτοσυσχέτισης στάσιμης χρονοσειράς

Σε ότι αφορά το γράφημα αυτοσυσχέτισης (ACF), υπάρχουν τρείς (3) στατιστικά σημαντικές τιμές, ενώ μετά από τέσσερις (4) χρονικές υστερήσεις (lags) η τιμή γίνεται μηδέν. Δίνεται μια πρώτη ένδειξη για ένα μοντέλο κινούμενου μέσου MA (3) ή MA (4).

Σε ότι αφορά το γράφημα μερικής αυτοσυσχέτισης, υπάρχει μόνο τρείς (3) στατιστικά σημαντική τιμή, ενώ στη συνέχεια φθίνει στο μηδέν, υποδεικνύοντας ως πιθανό μοντέλο το αυτοπαλίνδρομο AR (3) ή AR(4).

Στο σημείο αυτό μπορούμε να κάνουμε συνδιασμους και να ελέγχουμε την απόδοση του μοντέλου μας αλλα θα κάνουμε χρήση της συνάρτησης auto_arima() της βιβλιοθήκης pmdarima της γλώσσας python.

Χρησιμοποιώντας την παραμετροποίηση stepwise = FALSE, εκτελούμαι την προσομοίωση σε πιο αργό ρυθμό, με στόχο την καλύτερη δυνατή προσέγγιση, ενώ δίνοντας στην παράμετρο trace την τιμή TRUE ο χρήστης μπορεί να δει αναλυτικά καθένα από τα πιθανά μοντέλα που εξετάζονται ως προς την καταλληλότητα. Επιπλέον στη τιμή m επιλέγουμε την ποσότητα 52 διότι 52 είναι οι εβδομάδες του έτους και έχουμε ημερήσια δεδομένα .

Η επιλογή του καλύτερου μοντέλου γίνεται με βάση το κριτήριο πληροφορίας AIC (Akaike criterion).Σύμφωνα με το οποίο δεν προτιμάται ένα μοντέλο το οποίο είναι πιο περίπλοκο σε σχέση με ένα άλλο όταν προσφέρει ελάχιστα μεγαλύτερη ακρίβεια. Εκτελώντας έτσι την συνάρτηση εξάγεται ότι το μοντέλο ARIMA(3,1,2) είναι το καλύτερο με AIC=6674.287:

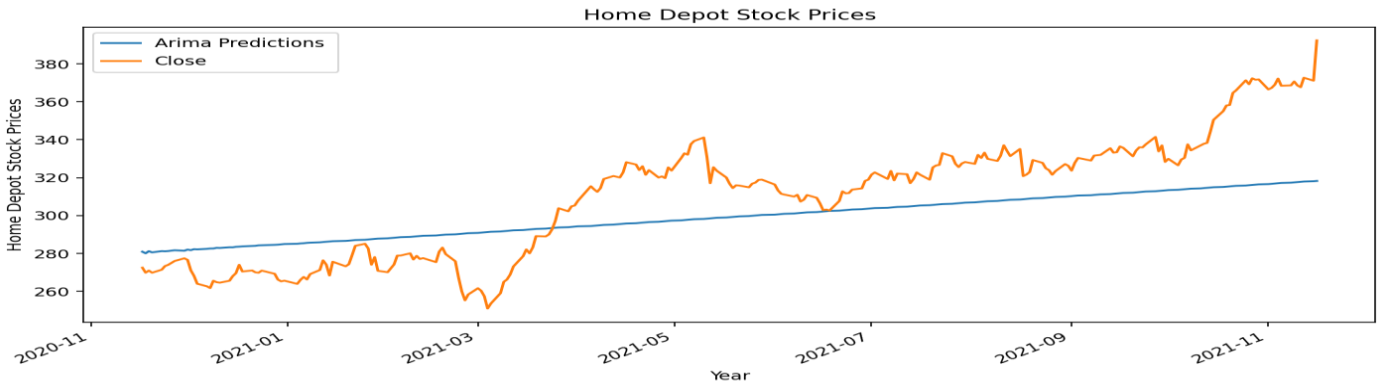
```
step=auto_arima(data,start_p=0,start_q=0,start_d=0,max_p=9,max_q=9,seasonal=False,trace=True,m=52)

Performing stepwise search to minimize aic
 ARIMA(0,1,0)(0,0,0)[0] intercept   : AIC=6732.592, Time=0.05 sec
 ARIMA(1,1,0)(0,0,0)[0] intercept   : AIC=6717.555, Time=0.08 sec
 ARIMA(0,1,1)(0,0,0)[0] intercept   : AIC=6721.026, Time=0.47 sec
 ARIMA(0,1,0)(0,0,0)[0]             : AIC=6735.047, Time=0.04 sec
 ARIMA(2,1,0)(0,0,0)[0] intercept   : AIC=6701.491, Time=0.26 sec
 ARIMA(3,1,0)(0,0,0)[0] intercept   : AIC=6702.807, Time=0.42 sec
 ARIMA(2,1,1)(0,0,0)[0] intercept   : AIC=6703.138, Time=0.65 sec
 ARIMA(1,1,1)(0,0,0)[0] intercept   : AIC=6711.130, Time=0.39 sec
 ARIMA(3,1,1)(0,0,0)[0] intercept   : AIC=6686.027, Time=2.06 sec
 ARIMA(4,1,1)(0,0,0)[0] intercept   : AIC=6683.787, Time=3.78 sec
 ARIMA(4,1,0)(0,0,0)[0] intercept   : AIC=6700.603, Time=0.74 sec
 ARIMA(5,1,1)(0,0,0)[0] intercept   : AIC=6684.467, Time=4.13 sec
 ARIMA(4,1,2)(0,0,0)[0] intercept   : AIC=6672.008, Time=8.36 sec
 ARIMA(3,1,2)(0,0,0)[0] intercept   : AIC=6671.495, Time=2.98 sec
 ARIMA(2,1,2)(0,0,0)[0] intercept   : AIC=6702.698, Time=1.87 sec
 ARIMA(3,1,3)(0,0,0)[0] intercept   : AIC=6706.698, Time=4.11 sec
 ARIMA(2,1,3)(0,0,0)[0] intercept   : AIC=6704.736, Time=2.19 sec
 ARIMA(4,1,3)(0,0,0)[0] intercept   : AIC=inf, Time=4.62 sec
 ARIMA(3,1,2)(0,0,0)[0]             : AIC=6674.287, Time=0.59 sec

Best model:  ARIMA(3,1,2)(0,0,0)[0] intercept
Total fit time: 37.836 seconds
```

Εικόνα 12 Auto Arima

Η στάσιμη χρονοσειρά έχει συνολικά 1825 παρατηρήσεις. Έτσι διασπάται το σύνολο σε train και test set με ποσοστά 0,8 και 0,2 αντίστοιχα. Αφού ολοκληρώσουμε την εκπαίδευση κάνουμε χρήση της συνάρτησης model.predict() και κατασκευάζουμε σε κοινό διάγραμμα τις προγραμματικές τιμές του test set και τις υπολογισμένες τιμές.



Εικόνα 13 Προβλεπόμενες και πραγματικές τιμες κλεισίματος

Ο έλεγχος ακρίβειας του μοντέλου γίνεται με την μέθοδο των ελαχίστων τετραγώνων ανάμεσα στις πραγματικές τιμές και παραγόμενες τιμές του μοντέλου της εικόνας 13 . Οι τιμές των σφαλμάτων είναι ιδιαίτερα χαμηλές, γεγονός που ενισχύει την καταλληλόλητα του μοντέλου καθώς έχουμε μέσο τετραγωνικό σφάλμα 24,26.

Εικόνα 14 Μέσο τετραγωνικό σφάλμα μοντέλου

```
from sklearn.metrics import mean_squared_error
from math import sqrt
test['Close'].mean()
rmse=sqrt(mean_squared_error(pred,test['Close']))
print(rmse)
```

```
24.262542120751682
```

Στο σημείο αυτό επιβάλλεται να αναρωτηθούμε ποια η σημασία καθώς και ποια η ερμηνεία του άνωθεν αποτελέσματος, δηλαδή είναι το μοντέλο μας είναι αποτελεσματικό η όχι;. Η απάντηση αυτή δεν είναι καθόλου εύκολη καθώς εξαρτάται από πολλούς παράγοντες και μερικοί από αυτούς δεν είναι μετρήσιμοι όπως η αποστροφή του επενδυτή στον κίνδυνο η όχι. Οι άνθρωποί είναι όντα τα οποία ανέκαθεν ενδιαφέρονταν να μάθουν τι τους επιφυλάσσει το μέλλον. Κύριος στόχος του αποτελεί να μειώσει την απόκλιση της πρόβλεψης με την πραγματική αλήθεια. Το αίσθημα αυτό είναι κοινό από τα ανώτατα στελέχη έως και απλών μικροεπενδυτών, γεγονός που είναι λογικό γιατί στην διαμόρφωση της πραγματικής αλήθειας ενέχονται και απρόβλεπτοι παράγοντες .Στο σημείο αυτό έχει απόλυτη εφαρμογή η βασική αρχή της οικονομίας δηλαδή η σχετικότητα αναφοράς και του περιβάλλοντος, δηλαδή αξίζει να δαπανηθεί χρόνος και πόροι για μία μικρή βελτίωση; Το αποτέλεσμα αυτό είναι μέσα στο επιτρεπόμενο budget.Όπως ο βραβευμένος οικονομολόγος Paul Krugman έχει αναφέρει ότι στην οικονομία τα γεγονότα πρέπει να περνάνε «το τεστ του καθρέφτη» δηλαδή να έχουν απλούς ελέγχους κατανοητούς από όλους. Εξηγώντας αν κάναμε πρόβλεψη για την τιμή ενός πακέτου τσίχλες το μοντέλο είναι λάθος καθώς σε περιόδους μη ακραίων καταστάσεων(υπερ-πληθωρισμού κτλ) δεν μπορεί να είναι τόσο ακριβό το προϊόν.

Πολλές διαφορετικές μέθοδοι προβλέψεων έχουν προταθεί και προταχθεί από τους επιστήμονες και κυρίως τους ακαδημαϊκούς εκ των οποίων μερικές βασίζονται μόνο σε θεωρητικό υπόβαθρο, ενώ άλλες απαιτούν και την συμβολή της τεχνολογίας και μάλιστα με μεγάλη υπολογιστική ισχύ.

ΚΕΦΑΛΑΙΟ 6

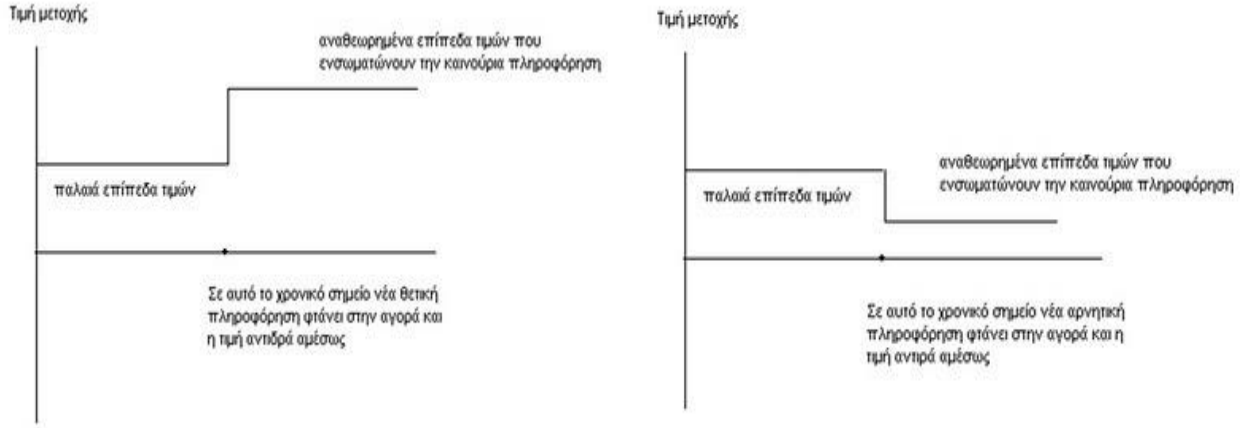
6.Ανάλυση συναισθήματος άρθρων οικονομικού περιεχομένου

Μέχρι το σημείο αυτό η προσπάθεια ανάλυσης και πρόβλεψης των τιμών κλεισίματος στο χρηματιστήριο αντιμετωπίζετε μόνο με στατιστικό τρόπο, δηλαδή οι τιμές κλεισίματος αποτελούν εξέλιξη αριθμητικών δεδομένων κατά τη πάροδο του χρόνου χωρίς να ενσωματώνουν κάποια λογική πίσω από την αλλαγή. Στη σύγχρονη όμως οικονομία τα πράγματα δεν είναι τόσο απλά, καθώς χιλιάδες παράγοντες επηρεάζουν τις αποφάσεις όλων των εμπλεκόμενων μελών για την λήψη μίας επιχειρηματικής απόφασης. Ωστόσο πολλοί παράγοντες δεν είναι απόλυτα μετρήσιμοι όπως ΑΕΠ, εξωτερικό χρέος, ισοζύγιο τρεχουσών συναλλαγών κτλ. Αλλά προσδοκίες σχετικά με το μέλλον ή η ικανοποίηση η μη απέναντι σε μία απόφαση ειλημμένη ή μελλοντική. Θα μπορούσαμε να κάνουμε την ανάλυση και την προσπάθεια πρόβλεψης κάνοντας χρήση στοιχείων από την οικονομική επιστήμη όπως με τις μελλοντικές τιμές των future χρεογράφων.

6.1 Θεωρία αποτελεσματικών Αγορών(Efficient Market Hypothesis)

Θεωρία της αποτελεσματικής αγοράς είναι μια θεμελιώδης οικονομική θεωρία που ορίζει ότι οι χρηματαγορες είναι διαρκώς και πλήρως ενημερωμένες, δηλαδή, οι παρούσες τιμές των χρεογράφων αντικατοπτρίζουν πλήρως κάθε σχετική και διαθέσιμη πληροφορία κατά τρόπο αποτελεσματικό και αλλάζουν συνεχώς προκειμένου να ενσωματώσουν οποιαδήποτε νέα πληροφορία προκύψει.

Γι' αυτό το λόγο είναι αδύνατο να νικήσει κάποιος την αγορά χρησιμοποιώντας οποιαδήποτε πληροφορία αφού αυτή, σύμφωνα με τη θεωρία, έχει ήδη προεξοφληθεί και ενσωματωθεί στην τιμή του χρεογράφου.Για το λόγο αυτό σε μία αποτελεσματική αγορά πρέπει να ισχύει η αύξηση της τιμής σε θετικά νέα και μείωση σε αρνητικά νέα δηλαδή όπως στις παρακάτω εικόνες.



Εικόνα 15 Αποτελεσματικές αντιδράσεις σε θετική (αριστερά) και αρνητική (δεξιά) πληροφόρηση

Σύμφωνα με την οικονομική επιστήμη και την συγκεκριμένη θεωρία διακρίνονται τρία επίπεδα πληροφοριακής αποτελεσματικότητας:

- **Μορφή Ασθενούς Αποτελεσματικότητας (weak-form efficiency)**

Οι ιστορικές πληροφορίες(παρελθοντικές τιμές, αποδόσεις και όγκος συναλλαγών)είναι ήδη προσαρμοσμένες στις παρούσες τιμές και δεν μπορούν να χρησιμοποιηθούν από τους παίκτες της αγοράς για την επίτευξη υπερ-απόδοσης(arbitrage)

- **Μορφή Ημι-Ισχυρής Αποτελεσματικότητας (semi-strong form efficiency)**

Οι δημοσιευμένες πληροφορίες των (ισολογισμοί,μερίσματα,κέρδη) είναι ήδη προσαρμοσμένες στις παρούσες τιμές και δεν μπορούν να χρησιμοποιηθούν από τους παίκτες της αγοράς για την επίτευξη υπερ-απόδοσης(arbitrage).

- **Μορφή Ισχυρής Αποτελεσματικότητας (strong form efficiency)**

Οι μη δημοσιευμένες πληροφορίες(εσωτερική πληροφόριση) είναι ήδη προσαρμοσμένες στις παρούσες τιμές και δεν μπορούν να χρησιμοποιηθούν από τους παίκτες της αγοράς για την επίτευξη υπερ-απόδοσης(arbitrage).

Παρόλα αυτά ακόμα και στις αποτελεσματικές αγορές οι ορθολογικοί παράγοντες όπως οι επενδυτές κάνουν επενδύσεις και οι καταναλωτές λαμβάνουν αποφάσεις κατανάλωσης οδηγούμενοι εν μέρει από το συναίσθημα και τη συγκίνηση, επηρεάζονται από γνωστικές προκαταλήψεις, και συνήθως στηρίζονται σε ελλιπείς, ανακριβείς και «θορυβώδεις» πληροφορίες για την απόφασή τους και στις διαδικασίες λήψης αποφάσεων.

Η πρόσφατη παγκόσμια χρηματοπιστωτική κρίση ανέδειξε το ρόλο του συναισθήματος και της συμπεριφοράς του επενδυτή στη διαμόρφωση των κρίσεων, σε περιόδους ακραίων χρηματιστηριακών μεταβολών, όπου επικρατεί φόβος στην αγορά. Υπό αυτές τις συνθήκες οι συμμετέχοντες στις αγορές είναι δύσκολο να δράσουν πλήρως ορθολογικά και να εξαλείψουν την επίδραση του συναισθήματος. Η πολυπλοκότητα της λειτουργίας του παγκοσμιοποιημένου χρηματοοικονομικού συστήματος και του συνόλου των παραγόντων που επηρεάζουν τις αγορές, αλλά και η ίδια η ανθρώπινη φύση, οδηγούν τους επενδυτές σε αλλεπάλληλα σφάλματα, τα οποία πολλές φορές οδηγούν σε σημαντικές απώλειες.

Ως οικονομικό συναίσθημα εννοούμε την ένδειξή για το πόσο θετικά ή αρνητικά αντιλαμβάνεται μία μεμονωμένη μονάδα της οικονομίας την παρούσα οικονομική κατάσταση. Η συνολική τάση όλων των επιμέρους μελών μίας οικονομίας αποτελεί το συναίσθημα της αγοράς(market sentiment).Το συναίσθημα των συμβαλλόμενων μελών επηρεάζεται από τις προσωπικές αξίες, πεποιθήσεις η ακόμα και από τις πεποιθήσεις μιας ομάδας στην οποία είτε ανήκει είτε πιστεύει ότι αποτελεί μέλος της(herding sentiment).Η προσπάθεια κατανόησης του συναισθήματος είναι ιδιαίτερα σημαντική διότι το συναίσθημα καθορίζει την προσφορά, ζήτηση επομένως και τη τιμή κάθε προσφερόμενου προϊόντος.

Οικονομικές ιστοσελίδες, περιοδικά, εφημερίδες επενδύσεων και δημοσιεύσεις κρατούν ενήμερο έναν επενδυτή και τον βοηθούν να πληροφορηθεί σχετικά με μία επένδυση, για το τι συμβαίνει στην οικονομία, τα νέα που επηρεάζουν τα χρήματά του, το που πρέπει να επενδύσει τα χρήματά του και που μπορεί να τοποθετήσει τα χρήματά του έτσι ώστε να έχει την υψηλότερη δυνατή απόδοση. Οι πηγές αυτές περιέχουν πολύτιμες πληροφορίες σχετικά με τις επιχειρήσεις εν γένει, καθώς και τις τρέχουσες οικονομικές και χρηματοπιστωτικές εξελίξεις, τις ειδήσεις της χρηματιστηριακής αγοράς και τις σχετικές ειδήσεις, όλες αυτές οι πληροφορίες που επηρεάζουν το επενδυτικό κοινό και το πιο σημαντικό, που επηρεάζουν τις επενδύσεις και τις επενδυτικές αποφάσεις εν γένει. Στο σημείο αυτό πρέπει να σημειωθεί ότι υπάρχει και ο παράγοντας των στρεβλών κινήτρων. Επειδή η πληροφόρηση είναι πολύ σημαντική και επηρεάζει όλες τις μονάδες μίας οικονομίας οι παραλήπτες αυτών των πληροφορίων οφείλουν να είναι προσεκτικοί καθώς και να φιλτράρουν τις διαθέσιμες πληροφορίες να αξιολογούν ποιες θα τους φανούν χρήσιμες και ποιες έχουν στόχο να παραπλανήσουν, καθώς πολλές φορές δίδονται στρατηγικά ψευδείς ή αναληθείς δεδομένα για την εξυπηρέτηση οικονομικών συμφερόντων κινούμενα με γνώμονα τον ανταγωνισμό.

6.2 Πηγές πληροφοριών/Αρθρογραφία

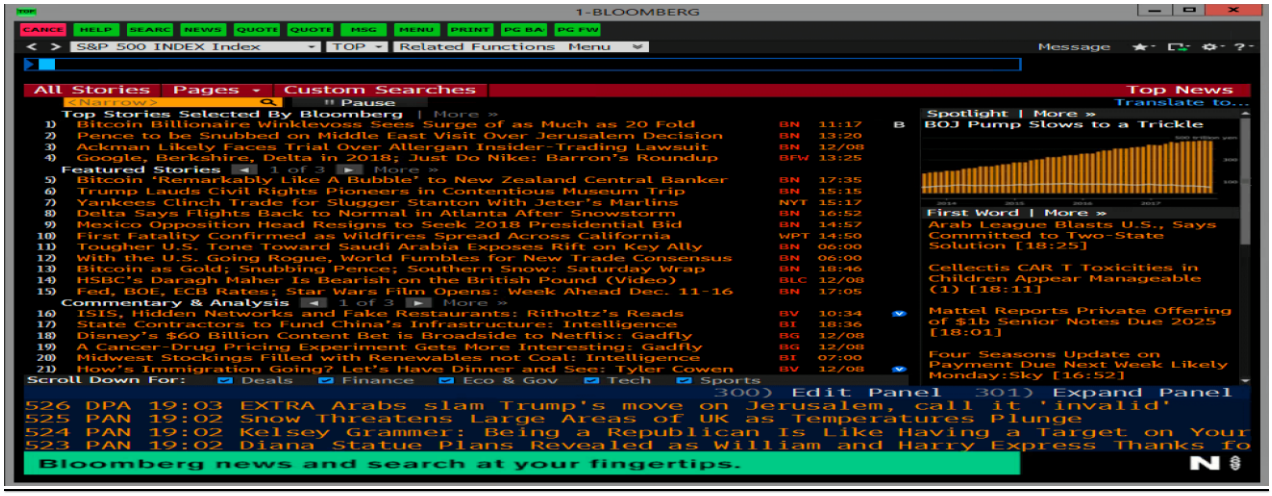
Στο σημείο αυτό θα μελετήσουμε τις διαθέσιμες έγκυρες πηγές από τις οποίες μπορούμε να αντλήσουμε έγκυρα άρθρα ώστε να αναλύσουμε το συναίσθημα των επενδυτών ή τις σκέψεις κατά συνέπεια και τα συναισθήματα μεγάλων ειδησεογραφικών ιστοσελίδων.

6.2.1 Bloomberg Terminal

Ιδρύθηκε από τον επιχειρηματία και επενδυτή Michael Bloomberg και αποτελεί μία από τις μεγαλύτερες επιχειρήσεις που συνδυάζει παγκόσμια οικονομική και χρηματιστηριακή ειδησεογραφία καθώς και χρηματοοικονομικές υπηρεσίες .Ο τερματικός σταθμός της Bloomberg terminal επιτρέπει στους χρήστες του κυρίως επενδυτές και traders να έχουν πρόσβαση σε ζωντανά δεδομένα και σε πραγματικό χρόνο από όλο το πλανήτη. Περιέχει πληροφορίες και ιστορικά στοιχεία για όλες τις κατηγορίες περιουσιακών στοιχείων(μετοχές,cds,ισοτιμίες κτλ),σχολιασμούς και αναλύσεις όχι μόνο των δημοσιογράφων αλλά και οικονομολόγων καθώς και την ακρίβεια τους στη πάροδο του χρόνου .Διαθέτει 192 γραφεία σε 73 διαφορετικές χώρες επιτρέποντας την σωστή οικονομική ενημέρωση.

Απευθύνεται σε πελάτες που ψάχνουν για την πιο ολοκληρωμένη, έγκαιρη και ακριβή υπηρεσία δεδομένων των επιχειρήσεων για να τροφοδοτήσουν κρίσιμες εφαρμογές γραφείου και / ή βάσεις δεδομένων. Η Bloomberg αποτελεί έναν ειδησεογραφικό κολοσσό, και μεταδίδει οικονομικά νέα τη στιγμή που συμβαίνουν. Επίσης η πλατφόρμα συνδέει επενδυτές από όλα τα μήκη και πλάτη του πλανήτη οι οποίοι διατυπώνουν γνώμες και μεταδίδουν ειδήσεις. Ο συνδυασμός των παραπάνω, με την σύγχρονη ενσωμάτωση των ειδήσεων που μεταδίδονται από τα υπόλοιπα παγκόσμια και τοπικά μέσα ενημέρωσης στην ίδια πλατφόρμα, τοποθετεί το Bloomberg Terminal στην κορυφαία θέση, ίσως και πάνω από το ίδιο το αχανές Διαδίκτυο, για την άντληση δημοσιευμένων νέων και κριτικών προβλέψεων σε όλοτον κόσμο.

Εικόνα 16 Bloomberg terminal



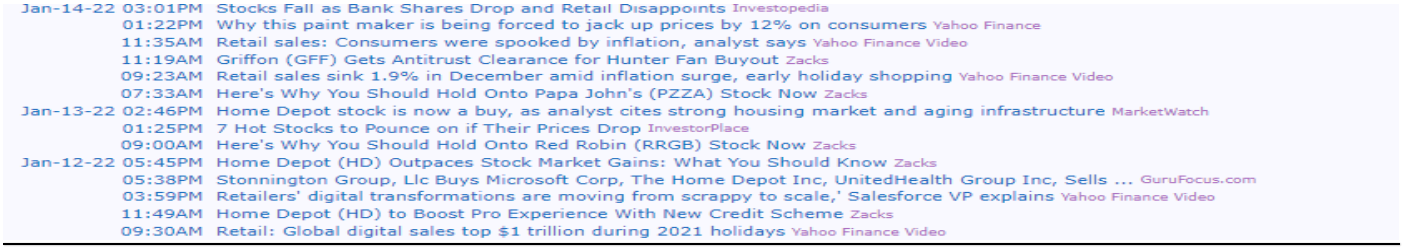
Εικόνα 17 Bloomberg terminal News

Η αρχιτεκτονική της πλατφόρμας επιτρέπει την αναζήτηση και εύρεση συγκεκριμένων και έγκυρων άρθρων καθώς και σχολιασμό της από έμπειρους και εξιδεικευμένους οικονομολόγους και δημοσιογράφους. Αποτελεί ίσως την καλύτερη πλατφόρμα έγκυρης άντλησης άρθρων οικονομικού περιεχομένου για την ανάλυση φυσικής γλώσσας και οικονομικού συναισθήματος.

6.2.2 Finviz (financial visualizations)

Αποτελεί μία ιστοσελίδα η οποία είναι ειδικευμένη στην παροχή ολόκληρου του φάσματος στην οικονομική πληροφορία. Ο χρήστης επισκεπτόμενος τη συγκεκριμένη σελίδα μπορεί να έχει πρόσβαση στα σημαντικότερες εξελίξεις σε μετοχές σε πραγματικό χρόνο όπως αυτό ορίζεται από το newsfeed.Ο χρήστης μπορεί να επιλέξει το όνομα της εταιρείας που επιθυμεί να δεί την πορεία της μετοχής. Συγχρόνως έχει πρόσβαση σε ιστορικά δεδομένα όπως τιμές κλεισίματος παρελθουσών ετών, σε μια τεράστια ποικιλία από γραφήματα, αλλά και σε πραγματικό χρόνο τις κινήσεις όλων των οικονομικών προϊόντων. Συν τοις άλλοις, παρέχεται η δυνατότητα στο χρήστη να δημιουργήσει το δικό του «διαδικτυακό» χαρτοφυλάκιο στο οποίο να τοποθετήσει τους τίτλους της επιλογής του και κατόπιν να ενημερώνεται για τις επιλογές του κατόπιν πληρωμής της συνδρομής. Επιπλέον με την επιλογή μίας συγκεκριμένης εταιρείας πέρα των ιστορικών στοιχείων εμφανίζονται και οι τίτλοι άρθρων που σχετίζονται με την εταιρία. Το κυριότερο από όλα όμως είναι το γεγονός ότι όλο αυτό το πλήθος των πληροφοριών των άρθρων παρέχεται δωρεάν και μπορούμε να αναλύσουμε την φυσική γλώσσα τους διαθέσιμους τίτλους.

Εικόνα 18 Home Depot finviz



Εικόνα 19 Άρθρα σχετιζόμενα με την αναζητούμενη μετοχή

6.2.3 Social Networks

Καθημερινά εκατομμύρια χρήστες των μέσων κοινωνικής δικτύωσης κοινοποιούν φωτογραφίες, δημοσιεύσεις με την προσωπική τους γνώμη για ένα οικονομικό γεγονός ή αντιδρούν σε δημοσιεύσεις άλλων χρηστών για τα τρέχοντα οικονομικά τεκταινόμενα . Οι πλατφόρμες αυτές αποτελούν μία τεράστια βάση για την άντληση δεδομένων ώστε να χρησιμοποιηθούν για ανάλυση συναισθήματος. Το γεγονός αυτό είναι δίσημο καθώς οι επιχειρήσεις μπορούν να κατανοήσουν τις αντιδράσεις του κοινού με γρήγορο τρόπο ώστε να βελτιώσουν τις υπηρεσίες τους ή αντίθετα να δημιουργηθούν στρεβλά κίνητρα για έλεγχο και καθοδήγησή της άποψης του κοινού για μία μετοχή Η δυσκολία σε αυτή την κατάσταση έγγυται με ποιόν τρόπο μπορούν να αντληθούν και κατά συνέπεια να αναλυθούν αποτελεσματικά τα δεδομένα αυτά για την ανάλυση συναισθήματος.

Η πλατφόρμα Facebook μπορεί εύκολα να μετρήσει την επισκεψιμότητα ενός οικονομικού άρθρου, τις αντιδράσεις των χρηστών κατά συνέπεια να εξάγει αποτελέσματα για την προσωπική άποψη του καθενός. Οι αναρτήσεις και τα σχόλια που γράφονται στην πλατφόρμα αυτή μπορούν εύκολα να κατηγοριοποιηθούν σε θετικά ,αρνητικά ή ουδέτερα και να αναλυθούν και να εξαχθούν με αυτόν τον τρόπο συμπεράσματα της αυθόρμητης κρίσης της κοινής γνώμης για την αγορά ή για το συγκεκριμένο οικονομικό γεγονός.

Η πλατφόρμα του tweeter όπως και το Facebook αποτελεί και αυτό μία πλατφόρμα καταγραφής προσωπικών γνωμών και σκέψεων σχετικά με τα οικονομικά γεγονότα. Για το λόγω αυτό πλέον υπάρχουν πολλά τμήματα marketing που ασχολούνται αποκλειστικά με δημοσιεύσεις από της συγκεκριμένης πλατφόρμας.Ωστόσο κάθε δημοσίευση έχει συγκεκριμένο αριθμό χαρακτήρων .Με αυτό το τρόπο ο χρήστης πρέπει να αποτυπώσει την άποψη του ή τη στάση του με σαφή και περιεκτικό τρόπο ,αποφεύγοντας άσκοπες επεκτάσεις κάνοντας την ανάλυση πιο εύκολη και με μεγαλύτερη ακρίβειά . Επιπλέον με την χρήση του ειδικού χαρακτήρα της δίεσης (hashtag) ο χρήστης μπορεί να σηματοδοτήσει την τιμή της μετοχής με στενευμένο τρόπο. Ωστόσο η ανάλυση των δημοσιεύσεων από τη πλατφόρμα του tweeter αποτελεί αρκετά μεγάλη πρόκληση λόγω του αδόμητου περιεχόμενό του και της μικρής αξιοπιστίας των χρηστών του πχ. Fake news για εσωτερική πληροφόρηση και διαρροή πληροφοριών.

6.2.4 Δελτία επίσημων χρηματοοικονομικών και πιστωτικών ιδρυμάτων.

Όλοι σχεδόν οι χρηματοπιστωτικοί οίκοι στην κάθε χώρα του πλανήτη διαθέτουν στο οργανόγραμμά τους, τμήματα οικονομικών μελετών. Οι διευθύνσεις αυτές εκδίδουν ανά τακτά χρονικά διαστήματα ενημερωτικά δελτία για τα χρηματοπιστωτικά προϊόντα της αγοράς και τις οικονομικές εξελίξεις εν γένει. Τα δελτία αυτά, παρατηρείται ανάλογα και με τον εκδότη, μπορούν να είναι ημερήσια, εβδομαδιαία, μηνιαία, τριμηνιαία, εξαμηνιαία ή και ετήσια. Οι εκθέσεις αυτές αντικατοπτρίζουν τη γνώμη κορυφαίων τραπεζικών κολοσσών όπως για παράδειγμα η Goldman Sachs με το εβδομαδιαίο "Weekly Monitor" ή η τεράστια επενδυτική τράπεζα VANGUARD η οποία διαχειρίζεται κεφάλαια ύψους 3 τρισεκατομμυρίων δολαρίων με το ετήσιο "Vanguard's economic and investment

outlook''. Προσφέρουν μία πυκνογραμμένη άποψη για την αγορά και την κατεύθυνσή της, επισημαίνουν τους κινδύνους και τις ευκαιρίες για επενδύσεις σε μεμονωμένα χρεόγραφα ή και ολόκληρους κλάδους. Χαρακτηριστικό των εγγράφων αυτών είναι η εξαιρετικά συμπυκνωμένη πληροφορία και άποψη που εμπεριέχουν καθώς και το γεγονός ότι το περιεχόμενό τους είναι αποκλειστικά περί των χρηματοοικονομικών και χρηματοπιστωτικών αγορών. Επιπλέον η νομοθεσία υποχρεώνει μεγάλα πιστωτικά ιδρύματα να δημοσιεύουν πολλές συμφωνίες που κλείνουν ,μεταξύ τους για λόγους διαφάνειας και ελέγχου.

Συμπερασματικά στη παρούσα διπλωματική δεν θα ασχοληθούμε με τη δημιουργία νευρωνικού δικτύου για την κατηγοριοποίηση και αξιολόγηση των ειδησεογραφικών πηγών σε αξιόπιστες και μη αλλά μόνο στην ανάλυση και εξαγωγή αποτελεσμάτων των τίτλων των σημαντικότερων άρθρων για την μετοχή της Home Depot. Οι τίτλοι αυτοί θα εξαχθούν από την προαναφερθείσα πλατφόρμα Finviz.Ο λόγος που θα ασχοληθούμε με τους τίτλους και όχι με ολόκληρα άρθρα δεν είναι τυχαίος. Τα συμβαλλόμενα μέλη μίας οικονομίας καλούνται να λαμβάνουν τις αποφάσεις τους σε πολύ μικρό χρονικό διάστημα και οι τίτλοι τον άρθρων έχουν ως στόχο να διαμορφώσουν ή να μαγνητίσουν το ενδιαφέρον των επενδυτών με σαφή και γρήγορο τρόπο. Επιπλέον τις περισσότερες φορές διαβάζοντας μόνο το τίτλο ενός άρθρου-κειμένου μπορούμε να αντιληφθούμε σε μεγάλο βαθμό την βασική ιδέα το αρθρογράφου.

ΚΕΦΑΛΑΙΟ 7

7. Επεξεργασία Φυσικής Γλώσσας(NLP)

Το συναίσθημα δεν αναλύεται μέσω τεχνητής αναφοράς, όπως μερικοί άνθρωποι ενδεχομένως να υποθέσουν. Αυτό επιτυγχάνεται μέσω μιας συστηματικής διαδικασίας που περιλαμβάνει τη χρήση ενός λεξικού συναισθημάτων. Το λεξικό αυτό εκχωρεί ένα βαθμό θετικότητας ή αρνητικότητας σε μία λέξη από μόνο του το οποίο στη συνέχεια χρησιμοποιείται για να δώσει νόημα στο σύνολο του κειμένου. Αυτός είναι ένας τρόπος ανάλυσης του συναισθήματος (sentiment). Στη συνέχεια, λαμβάνοντας υπόψη ένα είδος εγγενούς θετικότητας ή αρνητικότητας κάθε λέξης που θα μπορούσε να χρησιμοποιηθεί από κάποιον για να μιλήσει για μια επιχείρηση ή ένα προϊόν. Για παράδειγμα, η λέξη "excellent" θα πρέπει να θεωρείται μια θετική λέξη, καθώς και το "prefer" και το "love". Στο αντίθετο άκρο του φάσματος μπορούμε να δούμε λέξεις όπως "hate", "dislike", κλπ. Υπάρχουν δύο προβλήματα με αυτή τη μεθοδολογία, ωστόσο. Το πρώτο πρόβλημα είναι ότι αυτή η εκχώρηση των θετικών και αρνητικών συναισθημάτων αξιολογεί μια λέξη χωρίς το πλαίσιο στο οποίο είναι γραμμένη. Με άλλα λόγια χωρίς να λαμβάνει υπόψη τις υπόλοιπες λέξεις ή φράσεις του κειμένου ή της πρότασης, καθώς μία λέξη άρνησης πριν από μία αρνητική λέξη προσδίδει θετική χροιά πχ. « **δεν επιβεβαιώθηκαν οι δυσμενείς συνέπειες του πληθωρισμού** ». Το λεξικό είναι εξαιρετικά περιορισμένο στον αριθμό των λέξεων που θα αποδίδουν πάντα ένα θετικό ή αρνητικό συναίσθημα σε μια έκφραση. Το δεύτερο πρόβλημα είναι ότι οι ερευνητές μπορεί να αναθέτουν διαφορετικούς βαθμούς θετικότητας ή αρνητικότητας σε μία λέξη. Ιδιαίτερα στην περίπτωση των διφορούμενων εκφράσεων, ένας ερευνητής μπορεί να είναι περισσότερο διατεθειμένος να σημειώσει μια λέξη ως περισσότερο ή λιγότερο θετική.

Η κατηγοριοποίηση ενός κειμένου δεν φαίνεται πάντα στα διάφορα χαρακτηριστικά που αναφέρονται μέσα σε ένα άρθρο. Η ανάλυση συναισθήματος παραδοσιακά πραγματοποιείται με τη χρήση της τεχνολογίας που αξιολογεί ένα άρθρο σε συνολικό επίπεδο. Μέσα σε ένα κείμενο, ωστόσο, το θέμα δεν μπορεί να συνδέεται με τις περιγραφές. Για παράδειγμα, αν ληφθεί υπόψη η πρόταση : «Το φετινό roland Garros είναι αρκετά εντυπωσιακό παρόλο που πολλοί αστέρες του αθλήματος δεν θα παρενρεθούν» Η πρόταση θα πρέπει είναι θετική, δεδομένου του αριθμού των θετικών λέξεων που διαθέτει. Μόνο στο τέλος μπορεί κανείς να προσδιορίσει την τελεσιδικία της αποφάσεως που είναι συνολικά αρνητική. Τα λεξικά που χρησιμοποιούνται αναπτύσσονται μέσω της ανάλυσης διαφόρων παραγόντων, συμπεριλαμβανομένων της πόλωσης του συναισθήματος και των βαθμών θετικότητας (Όπως το «μου αρέσει» έναντι του «δεν μου αρέσει»), προσδιορίζουν ποια μέρη ενός εγγράφου περιέχουν υποκειμενικό περιεχόμενο (ανίχνευση της υποκειμενικότητας και της αναγνώρισης της γνώμης), προσδιορίζουν τα μέρη ενός εγγράφου που υπόκεινται στο ίδιο αντικείμενο ανάλυσης (από κοινού το θέμα της ανάλυσης συναισθήματος), καθώς και τον καθορισμό του «πολιτικού» προσανατολισμού του κειμένου (απόψεις και προοπτικές). Άλλες μη τεκμηριωμένες πληροφορίες στο κείμενο μπορούν επίσης να ληφθούν υπόψη. Για παράδειγμα, υπάρχουν έξι "καθολικά" αισθήματα: θυμός, αηδία, φόβος, χαρά, λύπη και έκπληξη που μπορούν να αναλυθούν, καθώς και η παρουσία ενός όρου, η συχνότητα ενός όρου και η σύνταξη.

Η μεγαλύτερη δυσκολία της ανάλυσης φυσικής γλώσσας εντοπίζεται στη διφορούμενη ερμηνεία που προκαλεί η ασάφεια στη γλώσσα σε πολλά επίπεδα όπως

- ασάφεια σε επίπεδο σύνταξης (ambiguity at syntactic level).

Χτύπησα το ληστή με το τσεκούρι. Δηλαδή το τσεκούρι ήταν του ληστή ή εγώ τον χτύπησα με τσεκούρι.

- ασάφεια σε επίπεδο λεξιλογικό (ambiguity at lexical level), όταν το νόημα μιας λέξης είναι διφορούμενο. Για παράδειγμα:

*Το πρώτο γράμμα του μήνα .*Εννοούμε το πρώτο γράμμα που έλαβα το μήνα ή το πρώτο γράμμα της λέξης μήνα;

- ασάφεια σε αναφορικό επίπεδο (ambiguity at referential level), όταν δεν είναι ευκρινές το σε ποιον, πού ή σε τι η πρόταση αναφέρεται. Για παράδειγμα:

Ο σκύλος δάγκωσε τη γάτα γιατί νιαούρισε. Η γάτα νιαούρισε στο σκύλο ή η γάτα απλά νιαούρισε;

- ασάφεια σε σημασιολογικό επίπεδο (ambiguity at semantic level), όταν, με διατήρηση της ίδιας συντακτικής ανάλυσης, η πρόταση επιδέχεται τουλάχιστον δυο διαφορετικές ερμηνείες. Για παράδειγμα:

*Ψίλοι στα αυτιά μου μπήκαν .*Η φράση είναι μεταφορική ή κυριολεκτική;

Η επεξεργασία φυσικής γλώσσας πραγματοποιείται σε τρία στάδια ανάλυσης την συντακτική, σημασιολογική και πραγματολογική. Για να είναι η ανάλυση αποτελεσματική οι προτάσεις πρέπει να είναι καλοσχηματισμένες στα τρία προαναφερθέντα επίπεδα.

Η βασική ιδέα για την κατανόηση της ανάλυσης φυσικής γλώσσας και ανάλυσης συναισθήματος βασίζεται στην κατανόηση της βασικής ιδέας ,δηλαδή τί μπορεί να κάνει ο εκάστοτέ ομιλητής με τις κατάλληλες λέξεις και πως οι λέξεις αλληλοεπιδρούν μεταξύ τους δίνοντας το επιθυμητό νόημα.

Η επεξεργασία φυσικής γλώσσας είναι τεράστιας σημασίας στη σύγχρονη εποχή, και ειδικά στο κόσμο των επιχειρήσεων αρκεί να κατανοήσει κανείς τα παρακάτω:

- Ανάλυση δεδομένων μεγάλης κλίμακας. Η επεξεργασία φυσικής γλώσσας επιτρέπει σε υπολογιστές ή πληροφοριακά συστήματα την επεξεργασία αδόμητου λόγου όπως αξιολογήσεις πελατών, παράπονα ή δημοσίευσες σε κοινωνικά μέσα και την εξαγωγή πολύτιμων αναδράσεων για την συσχετιζόμενη εταιρεία.
- Αυτοματοποίηση διαδικασιών σε πραγματικό χρόνο. Εργαλεία επεξεργασίας φυσικής γλώσσας μπορούν να κατανοήσουν, να κατευθύνουν και να αξιολογήσουν τα γραπτά κείμενα ανθρώπων αποτελεσματικότερα και αποδοτικότερα με ελάχιστη ανθρώπινη παρέμβαση πχ. Chatbots εξυπηρέτησης πελατών.
- Παραμετροποίηση λογισμικών και αλγορίθμων ανάλυσης φυσικής γλώσσας ανάλογα με τις ανάγκες κάθε επιχειρησιακού τομέα.
- Επιπλέον η ανάλυση φυσικής γλώσσας αποτελεί το πρώτο στάδιο για τομείς όπως Text-to-Speech.Τα πεδία αυτά μπορούν να εφαρμοστούν σε κοινωνικές ομάδες με προβλήματα ακοής ή όραση διευκολύνοντας σε πολύ μεγάλο βαθμό την καθημερινότητά τους.

Παρόλο που τα τελευταία χρόνια έχουν γίνει μεγάλα άλματα στο τομέα της επεξεργασίας της ανθρώπινης γλώσσας υπάρχουν πολλά υπάρχουν ακόμα σημεία που πρέπει να λυθούν. Οι περισσότερες δυσκολίες εμφανίζονται διότι η ανθρώπινη γλώσσα είναι ασαφής και δίσημη. Χαρακτηριστικό παράδειγμα αποτελεί η έννοια του σαρκασμού. Για να συμβεί αυτό οι επιστήμονες θα πρέπει να εκπαιδεύσουν τα εργαλεία και τους διαθέσιμους αλγορίθμους εκτός των λέξεων ή την σειρά των συντακτικών κανόνων να κατανοούν και το νόημα και το περιεχόμενο των προτάσεων. Αυτό όμως προϋποθέτει ότι λαμβάνονται υπόψιν και εξωτερικοί

παράγοντες όπως ηλικία, καταγωγή φύλο ή άλλοι παράγοντες που πιθανόν δεν έχουν εντοπιστεί ακόμη .

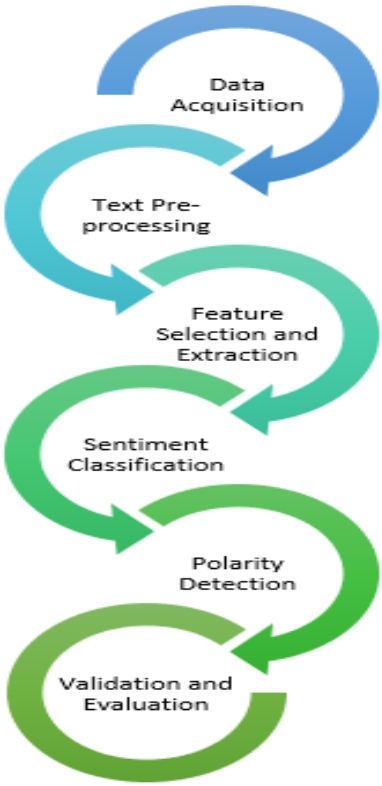
7.2 Ανάλυση Συναισθήματος.

Η ανάλυση των συναισθημάτων είναι ένας από τους ταχύτερα αναπτυσσόμενους τομείς έρευνας στα πεδία της πληροφορικής-τεχνολογίας και των μεγάλων δεδομένων(Big Data) των τελευταίων ετών. Οι ρίζες της εντοπίζονται στις μελέτες για την ανάλυση της κοινής γνώμης καθώς και στην ανάλυση συναισθηματικής ερμηνείας ενός κειμένου, από τον παγκόσμιο ιστό, τα κοινωνικά δίκτυα ή πλατφόρμες εξυπηρέτησες πελατών κτλ. . Τα τελευταία χρόνια εντάθηκε το ενδιαφέρον για την ανάλυση συναισθημάτων ως αποτέλεσμα της υψηλής διαθεσιμότητας υποκειμενικών κειμένων στο διαδίκτυο και της εν δύναμη προσφοράς τους σε πολλούς κλάδους όπως διαφήμιση ,marketing εξυπηρέτηση πελατών. Η ανάλυση και η ερμηνεία καθίσταται ιδιαίτερης υψηλής σημασίας καθώς μέσω των συναισθημάτων οι ενδιαφερόμενοι μπορούν να κρίνουν την ανθρώπινη συμπεριφορά.

Η ανάλυση συναισθήματος είναι ένα σύνολο μεθόδων, τεχνικών και εργαλείων για την ανίχνευση και την εξαγωγή πληροφοριών από την καταγεγραμμένη προσωπική άποψη ή πεποίθηση του ομιλητή. Τέτοιες πηγές είναι προσωπικές γνώμες και απόψεις σχετικά με ένα θέμα. Η ανάλυση δεν αποτελεί κάτι νέο καθώς για χιλιάδες χρόνια οι άνθρωποι προσπαθούσαν να κατανοήσουν την συναισθηματική κατάσταση του ομιλητή ή συγγραφέα. Αρχικά περιορίστηκε στην πολικότητα της κοινής γνώμης και αντικείμενο των σχετικών ερευνών ήταν η θετική, αρνητική ή ουδέτερη στάση απέναντι σε ζητήματα και γεγονότα ,με σκοπό την . Η διάκριση της ανάλυσης συναισθήματος και της εξόρυξης γνώμης δεν είναι ευδιάκριτη και συχνά οι έννοιες τους ταυτίζονται. Στη σύγχρονη εποχή με την ανακάλυψη του παγκόσμιου ιστού όπου καθημερινά παράγονται χιλιάδες δεδομένα από χρήστες καθίσταται επιτακτική η ανάγκη κατασκευής αλγορίθμων για τον υπολογισμό της πολικότητας.

Η παραπάνω διαδικασία είναι αρκετά περίπλοκη και πραγματοποιείται στα παρακάτω στάδια :

- **Συλλογή Δεδομένων(Data Acquisition)**: Η συλλογή δεδομένων είναι μια σημαντική φάση, καθώς πρέπει να καθοριστεί ένα κατάλληλο σύνολο δεδομένων για την ανάλυση και την ταξινόμηση του κειμένου.
- **Προ-επεξεργασία κειμένου** (**Text preprocessing**): Μετά τη συλλογή των δεδομένων, η προ-επεξεργασία επιτρέπει τη μείωση του θορύβου στα δεδομένα. Αυτό γίνεται με την αφαίρεση των περιττών λέξεων διακοπής, επαναλαμβανόμενων λέξεων, απορρίψεων, αφαίρεσης emoticon, αφαίρεσης διευθύνσεων URL κτλ.
- **Επιλογή και εξαγωγή χαρακτηριστικών**(**Feature selection and extraction**): Η σωστή επιλογή και εξαγωγή χαρακτηριστικών παίζει βασικό ρόλο στον προσδιορισμό της ακρίβειας του μοντέλου όπου θα αναπτυχθεί. Ως εκ τούτου, η κατάλληλη τεχνική εξαγωγής χαρακτηριστικών πρέπει να επιλεγεί για την εξαγωγή των χαρακτηριστικών.
- **Ταξινόμηση συναισθημάτω**ν(**Sentiment classification**): Σε αυτή τη φάση, εφαρμόζονται διάφορες τεχνικές ταξινόμησης συναισθημάτων για την ταξινόμηση του κειμένου. Μερικές δημοφιλείς τεχνικές ταξινόμησης συναισθημάτων είναι οι Naïve Bayes (NB) και η Υποστήριξη Διανυσματικές Μηχανές (SVM).
- **Ανίχνευση πολικότητας(Polarity Detection)**: Μετά την ταξινόμηση των συναισθημάτων, προσδιορίζεται η πολικότητα του συναισθήματος. Ο στόχος της ανίχνευσης πολικότητας είναι να αποφασίσει εάν ένα κείμενο εκφράζει θετικό, αρνητικό ή ουδέτερο συναίσθημα.
- **Επικύρωση και αξιολόγηση(Model Validation)**: Τέλος, πραγματοποιείται επικύρωση και αξιολόγηση των ληφθέντων και παραγόμενων αποτελεσμάτων ώστε να προσδιοριστεί η συνολική ακρίβεια των τεχνικών που χρησιμοποιούνται για την ανάλυση συναισθήματος ή απαιτείται η αλλαγή του μοντέλου.



Εικόνα 21 : Στάδια Ανάλυσης Συναισθλήματος

7.2.1 Μεθοδολογίες-Προσεγγίσεις

Οι πιο διαδεδομένες μεθοδολογίες είναι οι παρακάτω:

- **Τεχνικές Μηχανικής Μάθησης(ML):** Πρόκειται για τεχνικές όπου από την επεξεργασία φυσικής γλώσσας αναζητείται να εντοπιστούν τα συναισθήματα που εκφράζουν τα κείμενα. Με τις διαδικασίες αυτές οι υπολογιστές είναι σε θέση να επεξεργάζονται ακολουθίες συμβολοσειρών και να επιστρέφουν τα συναισθήματα που αυτές αντιπροσωπεύουν. Οι τεχνικές αυτές χρησιμοποιούν ισχυρούς αλγορίθμους προκειμένου να εκπαιδεύονται οι υπολογιστές στην εκτίμηση συναισθημάτων. Επιπλέον λαμβάνονται υπόψιν και γλωσσικές ιδιομορφίες.
- **Τεχνικές Λεξικών(Lexicon models):** Γίνεται χρήση έτοιμων γλωσσικών λεξικών συναισθημάτων, τα οποία είναι συλλογές σχολιασμένων και προ-επεξεργασμένων όρων συναισθήματος. Οι συναισθηματικές τιμές αποδίδονται σε λέξεις που περιγράφουν τη θετική, αρνητική και ουδέτερη στάση του ομιλητή. Η τεχνική αυτή διαχωρίζεται περεταίρω στα μοντέλα όπου ελέγχουν λέξεις δηλαδή χρησιμοποιεί ένα μικρό σετ βασικών λέξεων από ένα διαδικτυακό λεξικό. Η στρατηγική εδώ είναι το αρχικό σύνολο λέξεων με τους γνωστούς προσανατολισμούς τους συλλέγονται και στη συνέχεια αναζητούνται διαδικτυακά λεξικά για να βρεθούν τα πιθανά συνώνυμα και ανώνυμα τους. Το δείγμα ταξινομείται με βάση την παρουσία τέτοιων λέξεων σηματοδότησης. Η δεύτερη προσέγγιση είναι αυτή όπου ελέγχει φράσεις και προσπαθεί να αναγνωρίσει το περιεχόμενο τους σε σχέση με όλο το κείμενο.

Οι παραπάνω τεχνικές είναι ευρέως διαδεδομένες στην ανάλυση και κατηγοριοποίηση των συναισθημάτων. Τεχνικές μηχανικής μάθησης είναι ιδιαίτερα χρήσιμες σε εξειδικευμένες θεματικές ενότητές όπου απαιτείται η ανάλυση, καθώς επιτρέπουν πλήρως στον χρήστη την παραμετροποίηση ενός μοντέλου ,ωστόσο η εύρεση κατάλληλου και έγκυρου σύνολο δεδομένων για την εκπαίδευση πολλές φορές είναι δύσκολη έως και αδύνατη. Σε αντίθεση τεχνικές όπου βασίζονται σε λεξικά δεν απαιτούν εκπαίδευση άρα σημαντική εξοικονόμηση χρόνου περίπου την ίδια υπολογιστική ικανότητα

με την μηχανική μάθηση ,η ακρίβεια όμως εξαρτάται από τις λέξεις του λεξικού και πόσο συχνά ενημερώνονται.

7.2.2 Εφαρμογές Ανάλυσης Συναισθήματος

Η ανάλυση συναισθήματος και η προσπάθεια για την εξαγωγή συμπερασμάτων ήδη εφαρμόζεται σε πολλά διαφορετικά πεδία την σύγχρονης ανθρώπινης κοινωνίας .Αυτό συμβαίνει όπως έχουμε αναφέρει και παραπάνω στην υπερπληθώρα των δεδομένων που παράγει κάθε χρήστης καθημερινά μέσω του παγκόσμιου ιστού και κυρίως των Social Media.Παρακάτω γίνεται μία σύντομη ανάλυση επιστημονικών τομέων όπου γίνεται εκτεταμένη χρήση.

- Παρακολούθηση Social Media : Είναι προφανές ότι οι κοινοποιήσεις των χρηστών κατά πλειοψηφία εκφράζουν την ειλικρινή γνώμη και στάση σχετικά με ένα προϊόν ,υπηρεσία ή γεγονός. Με κατάλληλα λογισμικά μπορούμε να επεξεργαστούμε πολύ μεγάλο αριθμό δεδομένων και να κατανοήσουμε την στάση του κοινού.
- Εξυπηρέτηση πελατών : Στη σύγχρονη επιχειρηματική ζωή ο ανταγωνισμός έχει ενταθεί σε πολύ μεγάλο βαθμό .Για το λόγο αυτό η διατήρηση της υπάρχουσα πελατειακής βάσης αποτελεί κύριο στόχο. Για να κατανοήσει κανείς τη σημασία στατιστικές μελέτες τηλεφωνικών εταιρειών δείχνουν ότι η προσέλκυση ενός νέου πελάτη είναι 7 φορές ακριβότερη σε σχέση με τη διατήρηση. Πολλές φορές οι υπάρχουσες δομές και φόρμες συμπλήρωσης παραπόνων ή αξιολόγησης των υπηρεσιών δεν είναι αρκετά ώστε να καρατομήσουν οι οργανισμοί τις σωστές στρατηγικές που πρέπει να ακολουθήσουν .Για το λόγο αυτό αντλούνται στοιχεία από κοινωνικά δίκτυα ή από τηλεφωνικές συνομιλίες εξυπηρέτησης πελατών. Με αυτό το τρόπο υπάρχει αλληλεπίδραση σε πραγματικό χρόνο για την εύρεση της καταλληλότερης λύσης.
- Πολιτική :Στη πολιτική ζωή η γνώση της πραγματικής γνώμης του κοινού είναι ανεκτίμητη .Οι υπεύθυνοι πολιτικών εκστρατειών μπορούν με αυτό το τρόπο να προσαρμόσουν την στρατηγική τους για την εύρεση και προσέλκυση ψηφοφόρων. Τα επιτελεία μπορούν να έχουν σε πραγματικό χρόνο το πώς τάσσεται το εκλογικό σώμα απέναντι τους, τις προσδοκίες του από τους υποψηφίους και τι επηρεάζει θετικά η αρνητικά τους ψηφοφόρους .Στο σημείο αυτό τίθενται και θέματα ηθικής και στρεβλών κινήτρων μέσω της χειραγώγησης της κοινής γνώμης. Χαρακτηριστικό παράδειγμα αποτελεί το σκάνδαλο της Cambridge Analytica το 2018 με την χρήση προσωπικών δεδομένων ψηφοφόρων στις αμερικάνικες εκλογές του 2018 με τις εξατομικευμένες αναρτήσεις στα ιστολογία της πλατφόρμας Facebook με σκοπό τον επηρεασμό της ψήφου τους .
- Προωθητικές ενέργειες : Προωθητικές ενέργειες και εκπόνηση νέων καμπανιών προϋποθέτει την λήψη σωστών και έγκυρων αποφάσεων για την τμηματοποίηση των πελατών για την μεγιστοποίηση των κερδών. Οι νέες καμπάνιες εκτός από τα δημογραφικά χαρακτηριστικά πρέπει να στοχεύουν σε πελάτες οι οποίοι θα ανταποκριθούν θετικά σε μία ενέργεια. Μέχρι στιγμής ο περισσότεροι οργανισμοί κάνουν χρήση ιστορικών δεδομένων σε παρόμοιες ενέργειας με τη χρήση δέντρων απόφασης. Σε αυτό το σημείο μπορούν να προστεθούν τα σχόλια του κοινού και αναρτήσεις είτε σε forums είτε σε πλατφόρμες παραπόνων για την στόχευση του κατάλληλου κοινού.
    - Έγκαιρος εντοπισμών προβλημάτων. Η κριτική των καταναλωτών πολλές φορές αποτελεί την ειλικρινέστερη ανάδραση του συστήματος οργανισμός-κοινό. Με τη έγκαιρη διόρθωση τους οι οργανισμοί μπορούν να προλάβουν ζημιογόνες καταστάσεις.
    - Σύγκριση με ανταγωνισμό και εύρεση νέων τάσεων και ευκαιριών. Με την ανάλυση οι οργανισμοί μπορούν να κατανοήσουν σε τη θέση βρίσκονται οι υπηρεσίες τους σε σχέση με άλλους ανταγωνιστές του ίδιου κλάδου ώστε να βελτιώσουν τυχόν ασθενή στοιχεία τους. Η κριτική των πελάτων φανερώνει τις τάσεις των ανθρώπων στις δραστηριότητες του. Με αυτόν τον τρόπο μπορεί να προωθήσει συγκεκριμένες υπηρεσίες σε συγκεκριμένη ομάδα πελατών ή τη δημιουργία νεών προϊόντων τα οποία

απευθύνονται σε ομάδες με συγκεκριμένα χαρακτηριστικά και ιδιότητες όπως ηλικία ,δραστηριότητες κτλ. .

7.3 Ειδικά Λεξικά Ανάλυσης Συναισθήματος(Sentiment Analysis)-Lexicon Model

Στη παρούσα διπλωματική εργασία θα γίνει χρήση της λεξιλογικής βιβλιοθήκης **VADER(**Valence Aware Dictionary and sEntiment Reasoner**)** της Python. Αποτελεί μία λίστα λέξεων οι οποίες είναι ήδη φορτισμένες συναισθηματικά θετικά ή αρνητικά ώστε να υπολογίσει το συναισθηματικό φορτίο της πρότασης. Στη πραγματικότητα επιστρέφει τη πιθανότητα μία πρόταση να κατηγοριοποιηθεί ως θετική αρνητική ή ουδέτερη δηλαδή όχι μόνο μιλάει για τη βαθμολογία Θετικότητας και Αρνητικότητας αλλά μας λέει επίσης για το πόσο θετικό ή αρνητικό είναι ένα συναίσθημα.. Η εν λόγω βιβλιοθήκη κρίνεται ιδανική για την ανάλυση συναισθήματος για δεδομένα από μέσα κοινωνικής δικτύωσης.

```
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
analyzer = SentimentIntensityAnalyzer()
sentence = "Star wars is the best sci-fi film"
vs = analyzer.polarity_scores(sentence)
print(vs)
```

```
{'neg': 0.281, 'neu': 0.391, 'pos': 0.328, 'compound': 0.1531}
```

Εικόνα 22: Κατηγοριοποίηση πρότασης

7.3.1 Τρόπος λειτουργίας VADER.

Οι δημιουργοί της συγκεκριμένης βιβλιοθήκης έκαναν χρήση της πλατφόρμας αξιολόγησης της Amazon ώστε να αξιολογήσουν τις λέξεις. Αρχικά όταν κάνει parse ένα κείμενο προσπαθεί να εντοπίσει αυτές τις ιδίες λέξεις της λίστας που διαθέτει. Παράγει τέσσερις βασικές μετρικές:

- positive
- negative
- neutral
- compound

Οι πρώτες τρείς μετρικές παρουσιάζουν το ποσοστό των λέξεων που εμπίπτουν σε αυτές τις κατηγορίες, ενώ η τελευταία αποτελεί το άθροισμα των λεξιλογικών βαθμολογιών της Amazon σε κανονικοποιημένη μορφή [-1,1].

Το γεγονός το οποίο καθιστά τη συγκεκριμένη βιβλιοθήκη ιδανική για την ανάλυση συναισθήματος τίτλων άρθρων και γενικά δημοσιεύσεων των social media είναι η συνεχής ενημέρωση από την πλατφόρμα αξιολόγησης της Amazon,και η δυνατότητα να λαμβάνει υπόψιν της τον τρόπο γραφής(κεφάλαια,emojis).Όπως είναι φυσιολογικό η γλώσσα και ο τρόπος έκφρασης συνεχώς μεταβάλλονται γεγονός που καθιστά δύσκολο και χρονοβόρο συνεχώς να ενημερώνονται λεξικά. Αυτό μπορούμε να το δούμε και στο παρακάτω παράδειγμα της ίδιας πρότασης ,αλλά με διαφορετική συναισθηματική ένταση και φόρτιση:

```
sentence = "I like spaghetti"
vs = analyzer.polarity_scores(sentence)
print(vs)
```

```
{'neg': 0.0, 'neu': 0.444, 'pos': 0.556, 'compound': 0.3612}
```

```
sentence = "I LIKE spaghetti"
vs = analyzer.polarity_scores(sentence)
print(vs)
```

```
{'neg': 0.0, 'neu': 0.382, 'pos': 0.618, 'compound': 0.4995}
```

```
sentence = "I LIKE spaghetti <3"
vs = analyzer.polarity_scores(sentence)
print(vs)
```

```
{'neg': 0.0, 'neu': 0.246, 'pos': 0.754, 'compound': 0.7297}
```

Εικόνα 23 : Διαφορετικό συναίσθημα ανάλογα τον τρόπο γραφής

Έχοντας αναφέρει τα παραπάνω είναι προφανές ότι η δυνατότητα της βιβλιοθήκης αυτής δεν είναι μόνο να εντοπίζει τις λέξεις ,αλλά να λαμβάνει υπόψιν τον τρόπο γραφής γεγονός που είναι ιδιαίτερα σημαντικό στην σύγχρονη οικονομική αρθρογραφία ,καθώς δεν είναι λίγες οι φορές που οι δημοσιογράφοι κάνουν χρήση λεξιλογίου slang ή κεφαλαίων χαρακτήρων ώστε να αποδώσουν στα λεγόμενα τους μία νέα διάσταση ερμηνείας. Μία άλλη δυνατότητα της βιβλιοθήκης είναι η δυνατότητα να κατανοεί την αλλαγή του νοήματος μίας πρότασης όταν συναντά λέξεις αντίθεσης όπως but,although κτλ. Γίνετε μία προσπάθεια συμψηφισμού του συναισθήματος πριν και μετά από αυτές τις λέξεις ώστε να εξαχθεί μία συνολική μέτρηση όλης της πρότασης.

Τέλος τα πλεονεκτήματα της χρήσης της συγκεκριμένης βιβλιοθήκης συνοψίζονται στους παρακάτω λόγους:

- δεν απαιτεί εκπαίδευση το νευρωνικό δίκτυο καθώς είναι ήδη πλήρως ενημερωμένο
- σύμφωνα με τη βιβλιογραφία είναι αρκετά αποτελεσματικό στην κατανόηση του συναισθήματος καθώς λαμβάνει υπόψιν πολλές παραμέτρους όπως στίξη, κεφαλαία ιδιωματισμούς.
- Ιδανικό για κοινωνικά δίκτυα.

ΚΕΦΑΛΑΙΟ 8

8.Συσχέτιση δημοσιευμένης Αρθρογραφίας με την Τιμή της μετοχής

Οι επενδυτές της σύγχρονης χρηματοοικονομικής αγοράς βρίσκονται σε συνεχή πίεση, στην αναζήτηση αποδόσεων και σωστών επενδυτικών αποφάσεων τόσο την μεγιστοποίηση των κερδών των χαρτοφυλακίων τους όσο και για την ελαχιστοποίηση των ζημιών τους. Είτε αφορούν ανώτατα στελέχη τραπεζικών ιδρυμάτων, ασφαλιστικών οργανισμών ή μεμονωμένοι επενδυτές , όλοι τους αντιμετωπίζουν τον κοινό κίνδυνο του περιορισμένου χρόνου για την αξιολόγηση και λήψη σωστών αποφάσεων.

8.1 Μεθοδολογία

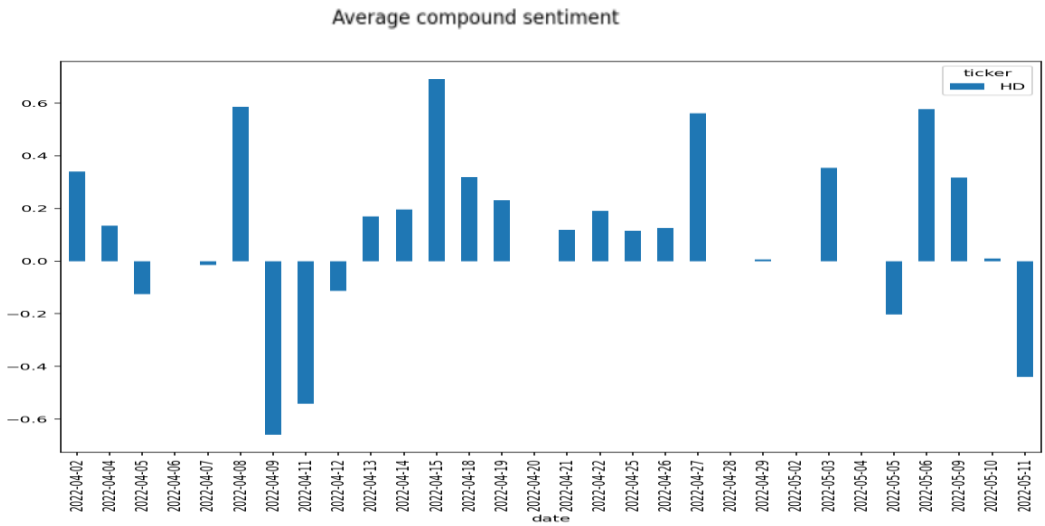
Αρχικά θα αντληθούν οι τίτλοι για την από την πλατφόρμα FINVIZ μέσω της κατασκευής μίας web_scrapping συνάρτησης για την περίοδο(02/04/2022-11/05/2022).Υπάρχει η περίπτωση για κάθε μέρα να υπάρχουν παραπάνω του ενός άρθρου το οποίο να σχετίζεται με την εξεταζόμενη μετοχή. Αρχικά θα υπολογισθεί το συναίσθημα για κάθε άρθρο μεμονωμένα και στη συνέχεια για κάθε μέρα θα υπολογισθεί ο μέσος όρος του συναισθήματος με τιμές [-1,1].Με αυτόν το τρόπο μπορούμε να έχουμε αντικειμενικότητα στη μέτρηση του συναισθήματος κάθε μέρα καθώς αντλούμε πληροφορία από διαφορετικές πήγες. Στην συνέχεια για την ίδια περίοδο θα αντληθούν οι πραγματικές τιμές κλεισίματος της μετοχής. Ωστόσο, οι δύο προαναφερθείσες χρονοσειρές έχουν διαφορετική κλίμακα , για το λόγω αυτό θα κανονικοποιηθούν οι τιμές τους στο διάστημα [0,1] ώστε να είναι εύκολή η εύρεση συσχέτισης και η οπτικοποίηση στο ίδιο διάγραμμα.

8.2 Μετρήσεις

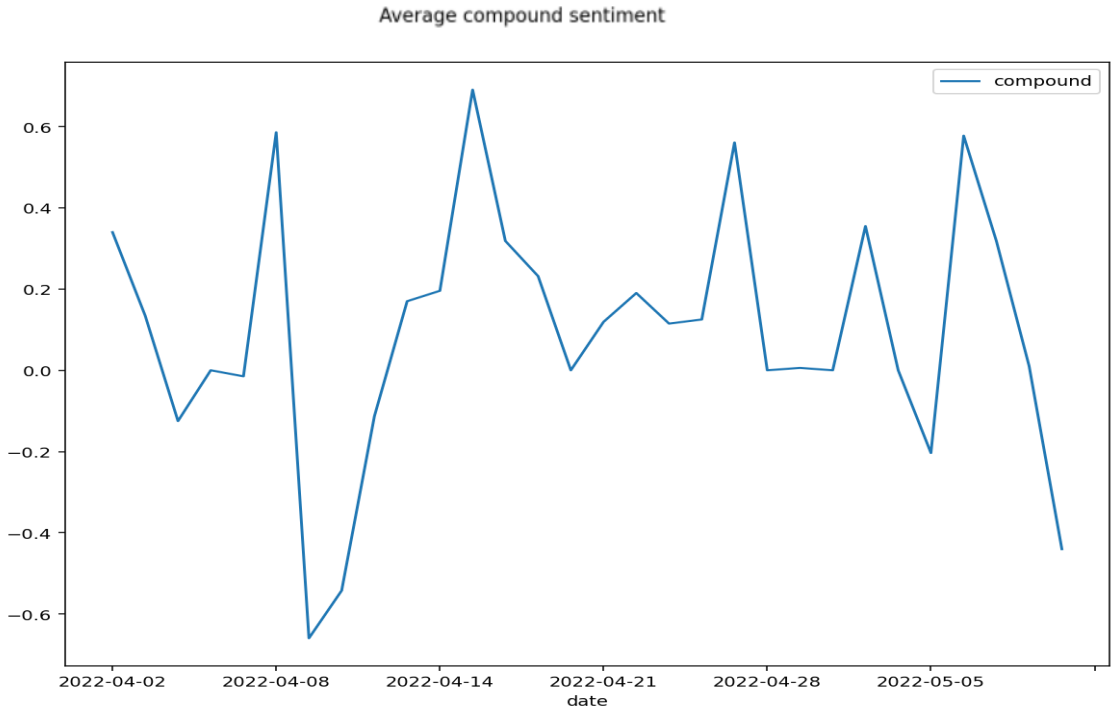
Ακολουθεί το διάγραμμα με το μέσο συναίσθημα ανά μέρα

Α)με τη μορφή bar-chart

Β)με τη μορφή Line-chart

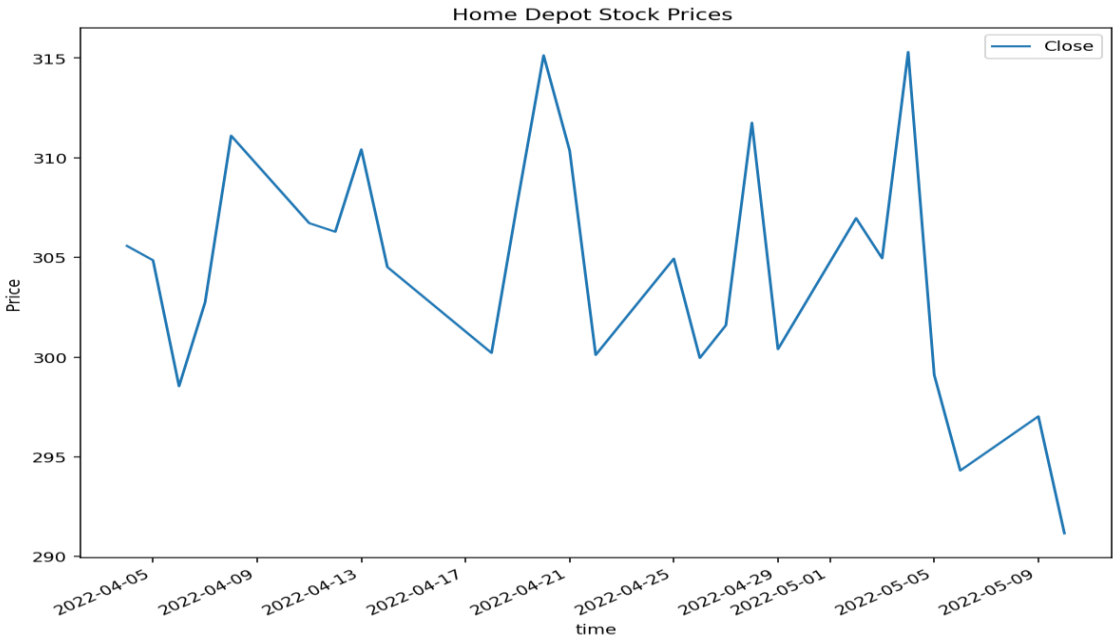


Εικόνα 24 : Μέσο συναίσθημα περιόδου(02/04/2022-11/05/2022) bar-chart



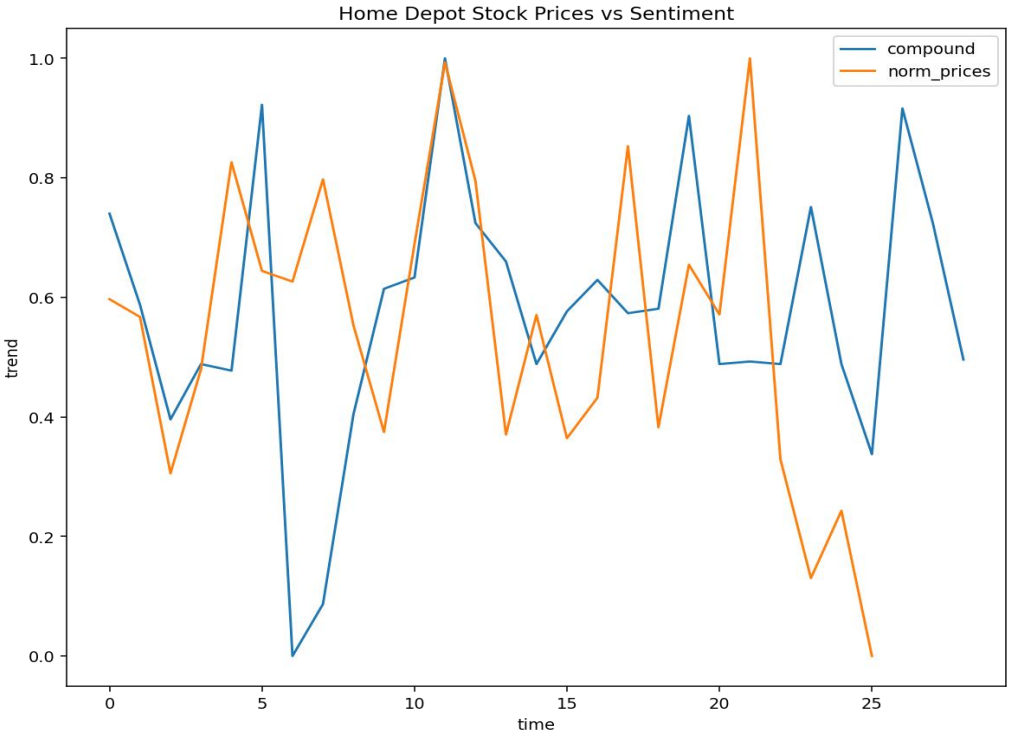
Εικόνα 25 : Μέσο συναίσθημα περιόδου(02/04/2022-11/05/2022) line-chart

Για την ίδια περίοδο οπτικοποιούμε και τις πραγματικές τιμές κλεισίματος στο ακόλουθο διάγραμμα



Εικόνα 26 : Πραγματικές τιμές κλεισίματος περιόδου(02/04/2022-11/05/2022) line-chart

Εφόσον ολοκληρωθεί η διαδικασία κανονικοποίησης στο διάστημα [0,1] αποτυπώνονται στο ίδιο διάγραμμα η γραφική παράσταση της τιμής κλεισίματος και συναισθήματος των δημοσιευμένων άρθρων της ίδιας περιόδου.



Εικόνα 27 : Κοινή απεικόνιση συναισθήματος και τιμής κλεισίματος

Είναι εμφανές ότι παρατηρώντας τις αριθμητικές τιμές και μόνο βλέπουμε μία συσχέτιση ,παρόλα αυτά δεν είναι σαφές εάν υπάρχει κάποιο συσχέτιση θετική ,αρνητική ή αποτελεί ένα τυχαίο γεγονός. Αυτό όμως που καθίσταται σαφές είναι το γεγονός ότι στις περιπτώσεις όπου έχουμε το χειρότερο συναίσθημα, δηλαδή αρνητικά φορτισμένο τίτλο έχουμε και πτώση της της τιμής πχ ημέρα t=4 και αντίστροφα την ημέρα t=12 παρατηρούμε θετικότητα τόσο στο συναίσθημα όσο και στην τιμή της μετοχής. Το γεγονός αυτό επιβεβαιώνεται και από την θετική συσχέτιση των δύο μεταβλητών(0,139).

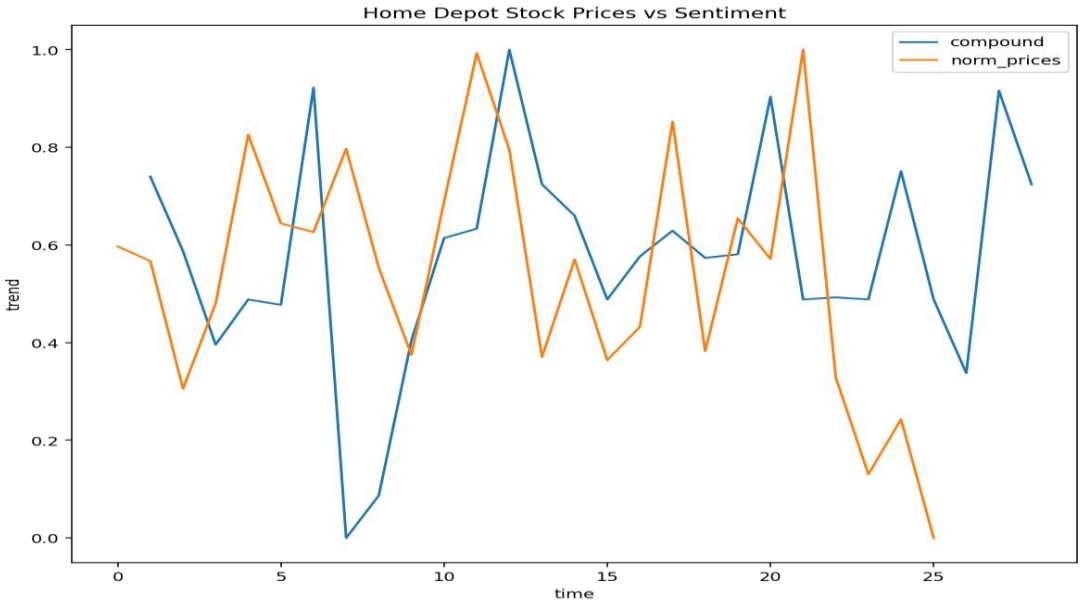
	compound	norm_prices
compound	1.000000	0.139937
norm_prices	0.139937	1.000000

Εικόνα 28 :Θετική συσχέτισή μεταξύ τιμής κλεισίματος και συναισθήματος

Η προαναφερθείσα ανάλυση είναι αρκετά στατική ,με την έννοια ότι αντιμετωπίζει τις δύο μεταβλητές (κανονικοποιημένη τιμή κλεισίματος και συναίσθημα)με αριθμητική προσέγγιση χωρίς να λαμβάνει υπόψιν της την οικονομική ερμηνεία .Αρκετές φορές στην πραγματική οικονομική ζωή τα πράγματα δεν είναι τόσο απλά και μονοδιάστατα. Στη σύγχρονη οικονομία τα περισσότερα γεγονότα προ-εξοφλούνται στην διάσταση του χρόνου αρκεί ένα απλό παράδειγμα για να γίνει κατανοητή η έννοια αυτή. Έστω ότι μία εταιρεία λόγω της καλής της οικονομικής πορείας αναμένεται ότι θα εξαγορασθεί από μία μεγαλύτερη την χρονική στιγμή t . Όλοι οι ενδιαφερόμενοι επενδυτές θα σπεύσουν να αγοράσουν μετοχές αναμένοντας αύξηση της τιμής ώστε να τις πωλήσουν, επειδή οι αγορές ρυθμίζονται μόνες τους και δεν υπάρχει περιθώριο arbitrage ,η ξαφνική αύξηση της ζήτηση της μετοχής θα οδηγήσει σε μεγαλύτερη αύξηση της ζήτησής με αποτέλεσμα η εξαγορά να επισπευσθεί την t-k .Έτσι λειτουργεί και η πληροφορία στη οικονομία δηλαδή ένα θετικό άρθρο μπορεί να

αποτυπωθεί στην τιμή της μετοχής σε μεταγενέστερο χρόνο(αύξηση) ή ένα αρνητικό σε μείωση στο μέλλον.

Για το λόγω αυτό κρίνεται σκόπιμο η κοινή απεικόνιση και η μελέτη τα σχέσης δημοσιευμένων άρθρων με την τιμή κλεισίματος της επόμενης ημέρας.



Εικόνα 29 : Κοινή απεικόνιση συναισθήματος και τιμής κλεισίματος επόμενης ημέρας

Συγκρίνοντας την άνω γραφική παράσταση με το διάγραμμα της εικόνας 28 παρατηρούμε ότι δεν εμφανίζει ιδιαίτερά κοινή πορεία γεγονός που αποτυπώνεται και στον συντελεστή συσχέτισής με τιμή 0,03 .

	compound	norm_prices
compound	1.000000	0.036778
norm_prices	0.036778	1.000000

ΚΕΦΑΛΑΙΟ 9

9.Περιορισμοί-Συμπεράσματα

9.1 Περιορισμοί

Κατά την εκπόνηση της εργασίας οι κυριότεροι περιορισμοί ετέθησαν στο Β μέρος δηλαδή την ανάλυση συναισθήματος. Η βιβλιογραφία των άρθρων περιορίστηκε μόνο στην αγγλική γλώσσα, και η περίοδος των άρθρων που αντλήθηκαν ήταν μόνο για ένα μήνα καθώς υπήρχε περιορισμός στην χρονική διάρκεια της πλατφόρμας. Επιλέχθηκε η ανάλυση μόνο των προτεινόμενων άρθρων της οικονομικής πλατφόρμας finviz τα οποία είχαν επιλεχθεί από τους αρχισυντάκτες του site ως τα πιο αντιπροσωπευτικά.Θα μπορούσε να ισχυριστεί κάνεις ότι περιορισμό αποτελεί και η χρήση λεξικού ωστόσο η Βιβλιοθήκη Vader ενημερώνεται διαρκώς από τις πλατφόρμες της Amazon.

9.2 Συμπεράσματα

Η ανάλυση χρονολογικών σειρών με την τεχνική ARIMA φαίνεται πως μπορεί να ακολουθήσει την τάση της πορείας μίας μετοχής και οι προβλέψεις της να είναι ικανοποιητικές σε μεγάλο βαθμό αν και δεν λαμβάνει παραμέτρους περαιτέρω χαρακτηρίστηκα όπως τιμή ανοίγματος, δείκτες χρηματιστηρίων και μακροοικονομικά μεγέθη. Παρά μόνο τον χρόνο και την επιρροή αυτού σε μία οικονομική μεταβλητή, πχ. Τιμή κλεισίματος. Επίσης οι μετασχηματισμοί οι οποίοι λαμβάνουν χώρα στην εξάλειψη της στασιμότητας επηρεάζουν τα αποτελέσματα.

Παρόλα αυτά η ανάλυση χρονοσειρών, και η γραφική απεικόνισή των βασικών οικονομικών μεγεθών μπορεί να δώσει μία σαφή εικόνα στα έμπειρα στελέχη στη κατανόηση και λήψη σωστών στρατηγικών στη λήψη αποφάσεων για το εκάστοτε χαρτοφυλάκιο ή επιχειρησιακό πρόβλημα που αντιμετωπίζουν.

Υπάρχουν πολλές κατευθύνσεις και τεχνικές όπου μπορεί να ακολουθήσει κανείς στην άντληση των δεδομένων και εξαγωγή πληροφοριών για την Ανάλυση συναισθημάτων. Παρατηρήθηκε ταύτιση της τάσης του συναισθήματος και της τιμής κλεισίματος στις ακραίες περιπτώσεις μεταβολής της τιμής της μετοχής και στις δύο προσεγγίσεις ίδιας και επόμενης ημερομηνίας. Στο σημείο αυτό να σημειωθεί ότι σκοπός της εργασίας δεν ήταν η πρόβλεψη της τιμής κλεισίματος της μετοχής καθώς αυτό είναι αδύνατον(**nobody can beat the market**)εάν δεν υπάρχει εσωτερική πληροφόρηση, αλλά η προσπάθεια εύρεσης μια συσχέτισης ανάμεσα στην οικονομική βιβλιογραφία και τιμή μέσω των λέξεων όπου χρησιμοποιούν οι αναγνώστες ώστε οι επενδυτές και αναλυτές να έχουν μία ένδειξη.

9.3 Μελλοντικές Προεκτάσεις

Όπως κάθε επιστημονικό εγχείρημα έχει τη δυνατότητα επέκτασης και βελτίωσης, έτσι και η παρούσα εργασία μπορεί εξίσου να αναβαθμιστεί και να επεκταθεί σε πιο ευρύ κοινό. Οι μελλοντικές επεκτάσεις αφορούν τόσο στο τρόπο εξαγωγής και συλλογής πληροφορίας και δεδομένων όσο και στις συσχετίσεις που θα ερμηνεύει ή θα εντοπίζει.

Μια μελλοντική προσέγγιση θα ήταν η προσπάθεια συσχέτισης μέσω της τεχνικής μηχανικής μάθησης με τη δημιουργίας ενός train-set και σηματοδότησης με εξιδεικευμένους όρους οικονομίας και χρηματιστηρίου επιλέγοντας τα κατάλληλα άρθρα ή η παρακολούθηση και ερμηνεία του πώς οι αναρτήσεις-άρθρα σημαντικών παικτών της οικονομίας επηρεάζουν τις μετοχές των χρηματιστηρίων π.χ. tweets Elon Musk-κρυπτονομίσματα και να προσπαθήσουμε να ερμηνεύσουμε το market sentiment ενός κλάδου της οικονομίας.

Βιβλιογραφία

[1] Βελαώρα, Χ. (2016), «Η Επίδραση της Ημέρας της Εβδομάδας στις Αποδόσεις των Μετοχών»-Διπλωματική Εργασία

[2] Κύρκος, Ε. (2015), «Επιχειρηματική Ευφυΐα και Εξόρυξη Δεδομένων» (ηλεκτρονικό αποθετήριο Κάλλιπος)

[3] Συμβουλάκης Ν.(2015) «Ανάπτυξη Μεθοδολογίας για την Αξιολόγηση Αποφάσεων και Επεδύσεων στηριζόμενοι σε Τεχνικές sentiment analysis και Στατιστικές Μεθόδους»

[4] Douglas C. Montgomery, Cheryl L Jennings, Kulahci M. (2015). Introduction to Time Series Analysis and Forecasting 2nd Edition, Wiley

[5] Κουγιουμτζής Δ. «Χρονοσειρές» 2021

[6] Allen B. Downey,2nd edition «Think in Python» ISBN : 978-960-645-090-7

[7] Γεωργούλη Κ. «Τεχνητή Νοημοσύνη : Μια Εισαγωγική Προσέγγιση» Ηλεκτρονικό Βιβλίο

[8] A Million News Headlines, News headlines published over a period of 18 Years, License CCO: Public Domain, Kaggle

[9]Laszlo N&Attila Kiss «Prediction of stock values changes using sentiment analysis of stock news headlines» 2020

[10] Αντζουλάτος Α. «Κυβερνήσεις Χρηματαγορές και Μακροοικονομία» ISBN : 978-960-89648-2-2

Χρήσιμοι σύνδεσμοι

[1] https://ieeexplore.ieee.org/document/9298333

[2] http://repfiles.kallipos.gr/html_books/93/07a-main.html

[3]https://medium.com/backyard-programmers/sentiment-analysis-of-stock-news-using-vader-5ba554d7cc19

[4] https://medium.com/swlh/temperature-forecasting-with-arima-model-in-python-427b2d3bcb53

[5] https://www.repustate.com/blog/sentiment-analysis-steps/

[6] https://www.r-bloggers.com/2017/03/forecasting-stock-returns-using-arima-model/

[7] https://www.researchgate.net/publication/261179224_Stock_price_prediction_using_the_ARIMA_model

[8] https://euretirio.com/apotelesmatiki-agora/

[9] https://finviz.com/

[10] https://users.auth.gr/dkugiu/Teach/DataAnalysis/Lecture7NoPause.pdf

[11] http://repfiles.kallipos.gr/html_books/93/index.html

[12] https://en.wikipedia.org/wiki/Facebook%E2%80%93Cambridge_Analytica_data_scandal

[13] https://towardsdatascience.com/time-series-from-scratch-white-noise-and-random-walk-5c96270514d3

[14] https://users.auth.gr/dkugiu/Teach/TimeSeries/TimeSeries.pdf

[15] https://devopedia.org/sentiment-analysis