ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

Π.Μ.Σ ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ & ΥΠΗΡΕΣΙΕΣ
ΚΑΤΕΥΘΥΝΣΗ: ΜΕΓΑΛΑ ΔΕΔΟΜΕΝΑ & ΑΝΑΛΥΤΙΚΗ

Ανάλυση συναισθήματος κοινωνικών δικτύων για την πρόβλεψη της τιμής του Bitcoin με χρήση LSTM αλγορίθμου.

ΤΖΑΒΑΡΑΣ ΙΩΑΝΝΗΣ ΜΕ1946

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ: ΜΙΧΑΗΛ ΦΙΛΙΠΠΑΚΗΣ

ΙΑΝΟΥΑΡΙΟΣ 2022

Περίληψη

Η Blockchain τεχνολογία αφορά ένα ταχέως αναπτυσσόμενο πεδίο τα τελευταία χρόνια. Με την αγορά των κρυπτονομισμάτων να σημειώνει συνολική αξία 1.03 τρισεκατομμύρια δολάρια και εκατομμύρια χρήστες να την αναφέρουν καθημερινά στα social media αποτελεί ένα χαοτικό πεδίο για έρευνα μιας και η αγορά επηρεάζεται από εκατοντάδες παράγοντες ταυτόχρονα. Στην παρούσα εργασία συλλέχθηκαν δεδομένα από πηγές κοινωνικών δικτύων (Tweeter , ιστοσελίδες και blogs), τα οποία αναλύθηκαν και κατηγοριοποιήθηκαν με διαφορετικές τεχνικές ανάλυσης συναισθήματος. Τα κειμενικά δεδομένα αποθηκευτήκαν σε βάση δεδομένων και εξετάστηκαν για τον βαθμό επιρροής τους χρησιμοποιώντας ένα σύνολο τεχνικών επιλογής χαρακτηριστικών. Παράλληλα μελετήθηκαν τα νευρωνικά δίκτυα RNNs (Recurrent neural networks) και αναπτύχθηκε μοντέλο μηχανικής μάθησης LSTM (Long-short memory). Το μοντέλο LSTM μετά από έρευνα σε επίπεδο αρχιτεκτονικής και υπερπαραμέτρων αξιοποιήθηκε για την πρόβλεψη της τιμής του κρυπτονομίσματος Bitcoin για χρονικό πλαίσιο εφτά ημερών και περιόδους έξι ωρών. Το νευρωνικό δίκτυο που αναπτύχθηκε σημείωσε αποτελεσματικότητα στην πρόβλεψη της τάσης για κάθε επόμενη περίοδο της χρονοσειράς του κρυπτονομίσματος 71.4%. Οι προτεινόμενες μέθοδοι μπορούν να χρησιμοποιηθούν για πρόβλεψη μεγαλύτερου μελλοντικού παραθύρου.

Λέξεις Κλειδιά: μηχανική μάθηση, RNN, LSTM, ανάλυση συναισθήματος, VADER, WordNet, SVM, κρυπτονόμισμα, πρόβλεψη τιμής, Bitcoin, Tweeter, κοινωνικά δεδομένα, επιλογή χαρακτηριστικών, SQL Server.

Abstract

The blockchain technology is a fast-growing field over the last years. Having a combined market value of totaled 1.03 trillion dollars and be mentioned by millions of users on the social media, the cryptocurrency stock market is a shambolic place for study as is affected by hundreds of factors simultaneously. On the current study, data from social media (Tweeter, sites and blogs) have been gathered, analyzed and classified with many techniques of sentiment analysis. The textual data have been hosted on a database and their effect examined with different ways of feature selection. Additionally, the RNNs networks studied and a LSTM neural network have been implemented. The LSTM model after iterations and experiments on the architecture and the hyper-parameters, have been used for the prediction of Bitcoin's price for the next seven days every six hours. The implemented model achieving forecasts with an accuracy of 71.4% on the trend of the timeseries. The proposed method can be used for a long time frame prediction.

Keywords: machine learning, RNN, LSTM, sentiment analysis, VADER, WordNet, SVM, cryptocurrency, price prediction, Bitcoin, Tweeter, social media, feature selection, SQL Server.

Περιεχόμενα

Κατάλογος Εικόνων

Κατάλογος Πινάκων

1.Εισαγωγή στις Χρονοσειρές

Με το όρο χρονοσειρά (*time series*) ονομάζουμε μια ακολουθία παρατηρήσεων που λαμβάνεται ανά τακτά χρονικά διαστήματα. Μια χρονοσειρά μπορεί να είναι συνεχής ή διακριτή. Στις συνεχείς χρονοσειρές η καταγραφή των δεδομένων είναι συνεχής σε κάποιο χρονικό διάστημα (παραδείγματος χάριν η τιμή της θερμοκρασίας στη διάρκεια ενός έτους). Αντίθετα, ονομάζεται διακριτή όταν οι παρατηρήσεις λαμβάνονται καθορισμένες χρονικές στιγμές, όπως για παράδειγμα η τιμή «κλεισίματος» μιας μετοχής στο τέλος της ημέρας.

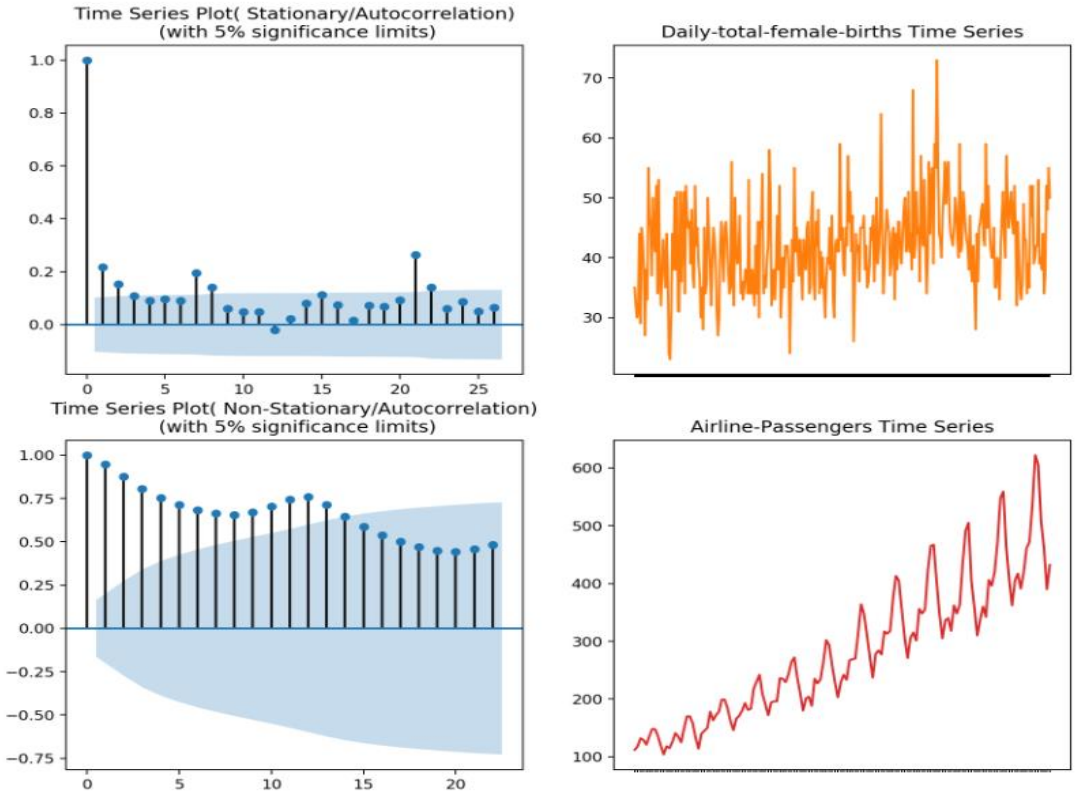
1.1 Στάσιμες Χρονοσειρές

Ένας από τους σημαντικότερους τύπους χρωνοσειρών είναι οι στάσιμες χρονοσειρές. Μια χρονοσειρά ονομάζεται στάσιμη εάν η κοινή κατανομή πιθανότητας των παρατηρήσεων $y_t, y_{t+1}, \dots, y_{t+n}$ είναι η ίδια με την κοινή κατανομή των παρατηρήσεων $y_{t+k}, y_{t+k+1}, \dots, y_{t+k+n}$ [1].

Με άλλα λόγια, οι στάσιμες χρονοσειρές χαρακτηρίζονται από παρόμοιες στατιστικές ιδιότητες (παραδείγματος χάριν σταθερή κατανομή πιθανότητας) στην διάρκεια του χρόνου. Γενικά, είναι συχνά αποδεκτό να χαρακτηρίζουμε μια χρονοσειρά με στασιμότητα (ή ασθενής στασιμότητα) εάν στις δύο πρώτες περιόδους ικανοποιούνται οι συνθήκες:

a) η αναμενόμενη τιμή δεν εξαρτάται από τον χρόνο

b) η συνάρτηση αυτόσυσχέτισης για υστέρηση ***k***, εξαρτάται μόνο από την ***k*** παράμετρο και όχι από τον χρόνο. Επιπλέον, ένα διάγραμμα αυτόσυσχέτισης με ισχυρές φθίνουσες παρατηρήσεις μπορεί να προϋποθέτει στασιμότητα [1].

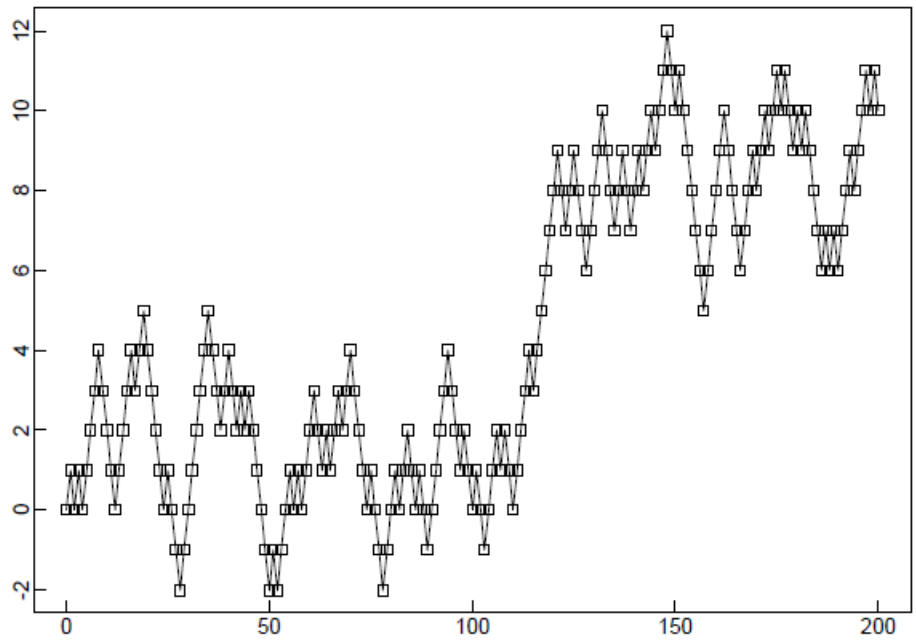


Εικόνα 1: Στάσιμη χρονοσειρά.

Στην παραπάνω εικόνα παρατηρούμε το διάγραμμα αυτοδιακύμανσης μίας στάσιμης και μιας χρονοσειράς με τάση αντίστοιχα. Η μη μεταβολή της δομής της αυτοδιακύμανσης σε σχέση με τον χρόνο (κάτω μέρος εικόνας) υποδηλώνει στασιμότητα [2].

1.2 Τυχαίος Περίπατος (Random Walk)

Ο τυχαίος περίπατος αναφέρετε σε οποιαδήποτε διεργασία για την οποία δεν μπορεί να αναγνωριστεί κάποιο πρότυπο ή τάση. Οι τιμές μιας μεταβλητής δηλαδή ή η κίνηση ενός αντικειμένου είναι εντελώς τυχαίες.



Εικόνα 2: Τυχαίος Περίπατος.

Έχει παρατηρηθεί ότι χρονολογικές σειρές οικονομικών δεδομένων ακολουθούν το υπόδειγμα του τυχαίου περιπάτου με περιπλάνηση. Η επιλογή τέτοιον μεταβλητών μπορεί να οδηγήσει σε λανθασμένα συμπεράσματα συσχέτισης μεταξύ μεταβλητών. Τέτοιες παλινδρομήσεις πρέπει να γίνονται αποδεκτές μόνο εάν οι μεταβλητές συνολοκληρώνονται [3].

Για την μελέτη των στάσιμων χρωνοσειρών χρησιμοποιείται η αυτοσυσχέτιση. Μια χρονοσειρά της μορφής ονομάζεται στάσιμη $\{ X_t, t = 0, \pm 1 , \dots \}$ ονομάζεται στάσιμη εάν οι στατιστικές τιμές είναι ίδιες με την «μετατοπισμένη» χρονοσειρά $\{X_{t-h}, t = 0, \pm 1 , \dots \}$ οπού *h* φυσικός αριθμός [5].

Η στασιμότητα μίας χρονοσειρά διακρίνεται σε ασθενή στασιμότητα και σε αυστηρή στασιμότητα. Η αυστηρή στάσιμη χρονοσειρά είναι ανεξάρτητη τη μεταβλητής του χρόνου σε αντίθεση με την ασθενώς στάσιμη.

Ασθενώς στάσιμη ονομάζεται η χρονοσειρά με σταθερό μέσο όρο [6]:

$$\mu_x(t) = E(x_t) \; \forall \; t$$

και σταθερή πεπερασμένη διακύμανση :

$$Var(Y_t) = \; \sigma_Y^2 < \; \infty \; \forall \; t$$

Η αυτοδιακύμανση είναι ανεξάρτητη του t και ορίζεται:

$$\gamma_\kappa = Cov(Y_t, Y_{t-k}) = Cov(Y_t, Y_{t+k}).$$

Αυστηρώς στάσιμη χρονοσειρά ονομάζεται η χρονοσειρά της οποίας η από κοινού κατανομή της είναι ανεξάρτητη του χρόνου. Δηλαδή ισχύει:

$$f(Y_{t1} , Y_{t2}, \dots , Y_{tn}) = f(Y_{t1+k}, Y_{t2+k}, \dots , Y_{tn+k})$$

Όπου k θετικός ακέραιος.

Μια αυστηρώς στάσιμη χρονοσειρά είναι απαραίτητα και ασθενώς στάσιμη. Το αντίθετο ισχύει μόνο στην περίπτωση της κανονικής κατανομής.

1.3 Ποιοτικά Χαρακτηριστικά Χρωνοσειρών

Η ανάλυση χρωνοσειρών μεταφράζεται στην μελέτη βασικών χαρακτηριστικών τους. Αυτά είναι η εποχικότητα, η τάση , η κυκλικότητα και οι ακραίες τιμές (outliers).

1.3.1 Τάση

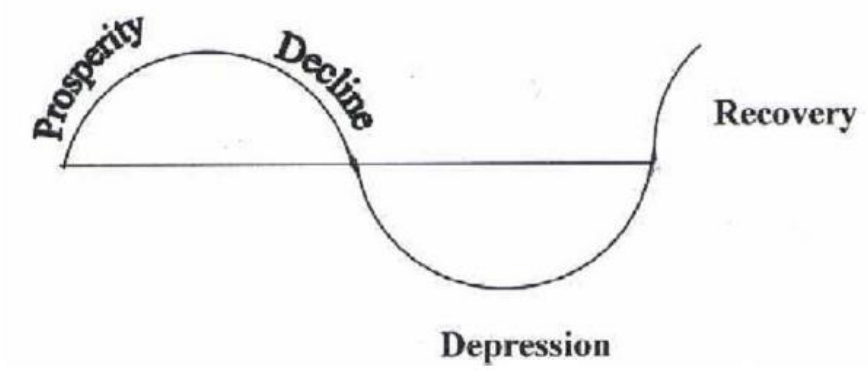
Με τον όρο Τάση εννοούμε το πρότυπο εκείνο των τιμών μιας χρονοσειράς το οποία παρουσιάζει κινητικότητα προς υψηλότερες ή χαμηλότερες τιμές σε ένα χρονικό διάστημα. Συνήθως η τάση είναι μακροχρόνια (περίπου 10 χρόνια) και μπορεί να θεωρηθεί ανοδική ή καθοδική. Η τάση θεωρείτε ανύπαρκτη, όταν η κεντρική κίνηση της χρονοσειράς είναι παράλληλη προς τον άξονα του χρόνου, χωρίς να παρουσιάζει τάση προς αύξηση ή μείωση.

Κάποιες από τις μεθόδους προσδιορισμού της τάσης είναι η μέθοδος ελαχίστων τετραγώνων, η μέθοδος κινητών μέσων, καθώς και τα test στασιμότητας όπως Dickey-Fuller test και Zivot-Andrews test [7].

1.3.2 Κυκλικότητα

Ως κυκλικότητα ορίζεται η περιοδικότητα με περίοδο μεγαλύτερη του έτους. Οι περίοδοι αυτοί είναι μεγαλύτερες του έτους και συνήθως της τάξεως της πενταετίας, χωρίς όμως να είναι σταθερού μήκους κατ' ανάγκη. Η κυκλικότητα εντοπίζεται κυρίως σε οικονομικές Χρονοσειρές. Ένας εύκολος τρόπος για να διακρίνουμε την κυκλικότητα είναι η εύρεση των περιοδικοτήτων με την βοήθεια των κορυφών του διαγράμματος [7].

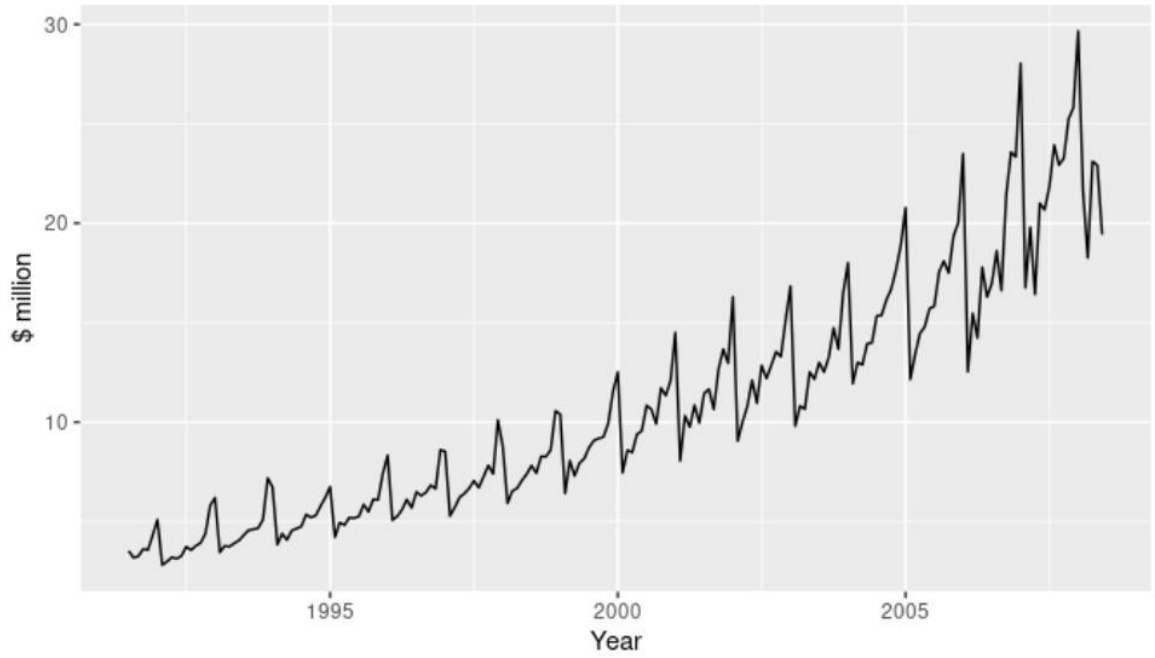
Οι περισσότερες χρηματοοικονομικές χρονοσειρές παρουσιάζουν μία μορφή κυκλικότητας. Για παράδειγμα ο κύκλος μίας εταιρίας αποτελείται από τέσσερίς φάσεις. Αυτές είναι οι: i) Ευημερία(Prosperity) ii) Πτώση(Decline) iii)Ύφεση (Depression) iv) Ανάκαμψη(Recovery)



Εικόνα 3: Τέσσερείς φάσεις χρηματοοικονομικών σειρών.

1.3.3 Εποχικότητα

Η εποχικότητα εντοπίζεται όταν η χρονοσειρά επηρεάζεται από εποχικές συνιστώσες όπως η μέρα της εβδομάδας ή κάποιος μήνας του χρόνου. Η εποχικότητα έχει καθορισμένη περιοδικότητα. Η παρακάτω εικόνα δείχνει τις μηνιαίες πωλήσεις αντιδιαβητικών φαρμάκων στην Αυστραλία. Παρατηρείται εποχικότητα και αύξηση των πωλήσεων κάθε τέλος του χρόνου [1].



Εικόνα 4: Μηνιαίες πωλήσεις *φαρμάκων στην Αυστραλία (Hyndman, R.J., & Athanasopoulos, G. (2018))*

1.3.4 Ακραίες τιμές

Οι ακραίες τιμές είναι παρατηρήσεις οι οποίες διαφέρουν σημαντικά από την μέση τιμή των υπόλοιπων παρατηρήσεων της χρονοσειράς. Οι ακραίες τιμές χαρακτηρίζονται από τυχαιότητα αλλά σε ορισμένες περιπτώσεις εξωτερικοί παράγοντες συμβάλουν καθοριστικά στις ασυνήθιστες αυτές παρατηρήσεις. Για παράδειγμα, ο αριθμός ετήσιων θανάτων μιας περιοχής μετά από κάποια φυσική καταστροφή στην περιοχή εκείνη. Στο παρακάτω διάγραμμα παρατηρούμε τον μέσο χρόνο εργασίας των εργαζομένων μιας εταιρίας και την ακραία τιμή μιας ημέρας απεργίας.

Εικόνα 5: Διάγραμμα *ωρών εργασίας ανά ημέρα.*

1.4 Διάσπαση χρονοσειράς

Μια χρονοσειρά επηρεάζεται από τέσσερις βασικές συνιστώσες, τις οποίες μπορούμε να τις διαχωρίσουμε από τις παρατηρήσεις τις χρονοσειράς μας. Αυτές είναι οι: *Τάση, Εποχικότητα, Κυκλικότητα, Σφάλματα*. Με βάση τα τέσσερα χαρακτηριστικά παράγονται δύο μοντέλα, το προσθετικό και το πολλαπλασιαστικό μοντέλο. Εάν το Y_t είναι η πραγματική παρατήρηση σε χρόνο t, εκφράζουμε την τάση της σειράς ως T_t , τον εποχικό παράγοντα ως S_t , τον κυκλικό παράγοντα ως C_t , και τις ακραίες τιμές ως I_t .

Οι σχέση των παραπάνω μπορεί να πάρει τις μορφές:

$$Y_t = T_t * S_t * C_t * I_t$$

$$Y_t = T_t + S_t + C_t + I_t$$

Στο πολλαπλασιαστικό μοντέλο, τα τέσσερα αυτά χαρακτηριστικά δεν είναι αναγκαστικά ανεξάρτητα και μπορούν να επηρεάσουν το ένα το άλλο. Αντιθέτως στο προσθετικό μοντέλο οι συνιστώσες είναι ανεξάρτητες.

Το προσθετικό μοντέλο εφαρμόζεται καλύτερα σε μια χρονοσειρά της οποίας η κυκλικότητα, ή η εποχικότητα, δεν μεταβάλετε με τις αυξομειώσεις των τιμών των παρατηρήσεων. Όταν το μέγεθος της κυκλικότητας εξελίσσεται ανάλογα με τις τιμές των παρατηρήσεων της χρονοσειράς, τότε το πολλαπλασιαστικό μοντέλο εφαρμόζει καλύτερα [1].

1.5 Αυτοσυσχέτιση (ACF)

Η αυτοσυσχέτιση είναι μία περίπτωση συσχέτισης δύο μεταβλητών, αλλά αναφέρεται στην συσχέτιση δύο διαδοχικών τιμών τις ίδιας μεταβλητής. Πιο συγκεκριμένα, το στατιστικό αυτό μέτρο αναπαριστά πως συσχετίζονται οι παρατηρήσεις μιας χρονοσειράς μεταξύ τους [1].

Στην περίπτωση τυχαίας χρονοσειράς οι αυτόσυσχέτισης Y_t, Y_{t-k} για κάθε k είναι κοντά στο 0 άρα οι διαδοχικές τιμές της χρονοσειράς για k χρονική υστέρηση (lag) είναι ασυσχέτιστες. Όταν η χρονοσειρά έχει τάση τότε οι διαδοχικές παρατηρήσεις είναι ισχυρά συσχετισμένες και οι συντελεστές αυτόσυσχέτισης είναι διάφοροι του 0. Για μία χρονοσειρά n παρατηρήσεων η k δειγματική αυτοσυσχέτιση r_k υπολογίζεται από την σχέση:

$$r_k = \frac{\sum_{t=k+1}^{n}(Y_t - \bar{Y})(Y_{t-k} - \bar{Y})}{\sum_{t=1}^{n}(Y_t - \bar{Y})^2}$$

Για τον υπολογισμό τις αυτόσυσχέτισης χάνεται μία παρατήρηση από το δείγμα των παρατηρήσεων της χρονοσειράς. Για παράδειγμα, για τον υπολογισμό της k αυτόσυσχέτισης θα χάνονται από το δείγμα k παρατηρήσεις, άρα θα έχουμε συνολικό δείγμα *n-k*. Το διάγραμμα αυτόσυσχέτισης μας βοηθάει στον καθορισμό των παραμέτρων των μοντέλων *MA* και *AR*. Αντίστοιχα, η μερική αυτοσυσχέτιση ***(PACF),*** η οποία δείχνει το βαθμό συσχέτισης μίας παρατήρησης *k* χρονικών περιόδων στο παρελθόν και της τωρινής παρατήρησης, είναι κατάλληλη για τον ορισμό των μέγιστων τιμών των παραμέτρων των μοντέλων *MA* και *AR* [1].

1.6 Αυτοδιακύμανση

Η αυτοδιακύμανση είναι ένα βασικό εργαλείο ανάλυσης, περιγραφής και πρόβλεψης των χρωνοσειρών. Η συνάρτηση της αυτοδιακύμανσης περιγράφει την σειριακή εξάρτηση μιας μονομεταβλητής, στάσιμης χρονοσειράς. Για μία ασθενώς στάσιμη χρονοσειρά *Y* η συνδιακύμανση για χρονική υστέρηση *k* συμβολίζεται με γ_k και ισούται με:

$$\gamma_\kappa = E(X_t - \mu)(X_{t+k} - \mu)\,, k = 0, \pm 1, \ldots$$

Για *k=0* προκύπτει η διακύμανση της χρονοσειράς.

Η επιλογή σωστού μοντέλου χρωνοσειρών είναι εξαιρετικά σημαντική για την μελλοντική πρόβλεψη. Οι ορισμοί της αυτοδιακύμανσης όπως και της αυτόσυσχέτισης έχουν νόημα κυρίως όταν αναφερόμαστε σε στάσιμες χρονοσειρές [49].

2. Μετρά αξιολόγησης της Πρόβλεψης και Δείκτες Σφαλμάτων

Για την επιλογή της καλύτερης μεθόδου πρόβλεψης μιας πραγματικής ή προσομοιωμένης χρονοσειράς χρησιμοποιούμε συναρτήσεις μεταξύ των πραγματικών και τον τιμών που έχουν προβλεφθεί, τα λεγόμενα μέτρα. Για την εφαρμογή μιας μεθόδου–μοντέλου, η χρονοσειρά χωρίζεται σε τρία set δεδομένων.

Το set εκπαίδευσης, το set επικύρωσης και set δοκιμών. Το set δοκιμών χρησιμοποιείται για να υπολογιστεί πόσο ακριβής είναι η πρόβλεψη με την μέθοδο που χρησιμοποιούμε. Σημαντικά χαρακτηριστικά των χρωνοσειρών που θα παρατηρηθούν στις παρακάτω συναρτήσεις είναι η τιμή

που αναπαριστά την πραγματική τιμή y_t, f_t η τιμή που αναπαριστά την πρόβλεψη, $e_t = y_t - f_t$ το σφάλμα τις πρόβλεψης και *n* το μέγεθος του set δοκιμών.

Επιπλέον, ως μέση τιμή ορίζεται η τιμή $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_t$ και η τιμή $\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_t - \bar{y})^2$ διακύμανση.

Κάποια από τα ποιο σημαντικά μέτρα τα οποία χρησιμοποιούνται ευρύτατα στην βιβλιογραφία είναι τα παρακάτω:

2.1 MFE

Ως μέσο σφάλμα πρόβλεψης MFE (Mean Forecast Error) ορίζεται η μέση απόκλιση των τιμών που έχουν εκτιμηθεί από τις πραγματικές. Ορίζεται ως $MFE = \frac{1}{n}\sum_{i=1}^{n} e_t$. Το μέσο σφάλμα πρόβλεψης δεν λαμβάνει ή επηρεάζεται διαφορετικά από τα ακραία σφάλματα. Με αλλά λόγια τα αρνητικά σφάλμα και τα θετικά σφάλματα αληλοεξουδετερώνονται. Όσο ποιο κοντά στο 0 είναι η τιμή του συγκεκριμένου μέτρου τόσο ποιο μεροληπτική και η πρόβλεψη.

2.2 MAE

Ως απόλυτο μέσο σφάλμα πρόβλεψης MAE (Mean Absolute Error) ορίζεται η μέση απόλυτη απόκλιση των τιμών που έχουν εκτιμηθεί από τις πραγματικές. Στην Βιβλιογραφία εμφανίζεται και ως *Mean Absolute Deviation (MAD)*. Μιας και οι τιμές των σφαλμάτων είναι απόλυτες το μέτρο δεν μας δίνει πληροφορία για την κατεύθυνση των σφαλμάτων. Ορίζεται ως $MFE = \frac{1}{n}\sum_{i=1}^{n} |e_t|$. Όσο ποιο κοντά στο 0 η τιμή του μέτρου τόσο καλύτερη και η πρόβλεψη.

2.3 MAPE

Το μέσο ποσοστό σφάλματος (*Mean Absolute Percentage Error*) είναι ένα από τα ποιο γνωστά μέτρα αποτελεσματικότητας στην βιβλιογραφία. Αποτελεί ποσοστό αποτελεσματικότητας και ορίζεται ως το απόλυτο πηλίκο του σφάλματος με τις πραγματικές τιμές. Το μετρό δίνεται από την συνάρτηση: $MAPE = \frac{1}{n}\sum_{i=1}^{n} \left|\frac{e_t}{y_t}\right| x100$.

2.4 MSE

Το μέσο τετραγωνικό σφάλμα (*Mean Squared Error*) δείχνει πόσο κοντά είναι η γραμμή παλινδρόμησης σε σχέση με ένα set σημείων. Αυτό επιτυγχάνεται με τον τετραγωνισμό των αποστάσεων των σημείων από την γραμμή (σφάλματα). Με τον τετραγωνισμό των σφαλμάτων εξαλείφονται οι αρνητικές τιμές και «τιμωρούνται περισσότερο» τα ακραία σφάλματα. Υπολογίζεται μέσο της σχέσης: $MSE = \frac{1}{n}\sum_{i=1}^{n} {e_t}^2$.

2.5 RMSE

Το μέτρο *RMSE* είναι η ρίζα του μέσου τετραγωνικού σφάλματος(*MSE*). Μοιράζεται τις ίδιες ιδιότητες με το MSE και ορίζεται ως $RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} {e_t}^2}$. Χρησιμοποιείται εκτεταμένα στην μετεωρολογία και την ανάλυση παλινδρόμησης για την επαλήθευση πειραμάτων.

2.6 NMSE

Το κανονικοποιημένο μέσο τετραγωνικό σφάλμα θεωρείτε ένα αντικειμενικό και αποτελεσματικό μέτρο. Ορίζεται ως $\frac{MSE}{\sigma^2} = \frac{1}{n\,\sigma^2}\sum_{i=1}^{n} {e_t}^2$ και όσο μικρότερη η τιμή του τόσο καλύτερη η πρόβλεψη.

2.7 Teil's U-statistics

Ο Henri (Hans) Theil πρότεινε δύο μέτρα που έχουν εφαρμογή σε οικονομικά προβλήματα. Το ένα μέτρο αφορά την αποτελεσματικότητα μιας πρόβλεψης U_1 και το άλλο την ποιότητα της πρόβλεψης U_2. Ορίζονται ως [10]:

$$U_1 = \frac{\left(\sum_{i=1}^{n}(\overline{y_t}-y_t)^2\right)^{1/2}}{\left(\sum_{i=1}^{n} {y_t}^2\right)^{1/2}}$$

$$U_2 = \frac{\left(\frac{1}{n}\sum_{i=1}^{n}(\overline{y_t}-y_t)^2\right)^{\frac{1}{2}}}{\left(\frac{1}{n}\sum_{i=1}^{n} {y_t}^2\right)^{\frac{1}{2}} + \left(\frac{1}{n}\sum_{i=1}^{n} {\overline{y_t}}^2\right)^{\frac{1}{2}}}$$

3. Μέθοδοι εξομάλυνσης

Αποτελούν μεθόδους προβλέψεις χρωνοσειρών. Κάποιοι από τις μεθόδους εξομάλυνσης είναι ο κινητός μέσος, η εκθετική εξομάλυνση, η μέθοδος Holt και η μέθοδος Winters.

3.1 Κινητός Μέσος (*Moving Average*)

Ως κινητός μέσος τάξης ***q*** είναι η μέση τιμή των ***q*** διαδοχικών παρατηρήσεων. Ο κινητός (κυλιόμενος) μέσος σε ένα χρονικό σημείο είναι ο αριθμητικός μέσος ενός χρονικού διαστήματος που έχει ως κέντρο το συγκεκριμένο σημείο. Ορίζεται ως [7]:

$$\hat{Y} = \frac{Y_t + Y_{t-1} + \ldots + Y_{t-k+1}}{q}$$

Όπου, Y_t η πραγματική τιμή , q ο αριθμός των παραγόντων του κινητού μέσου και Y_{t-1} η πρόβλεψη για την επόμενη περίοδο.

Η παραπάνω σχέση προκύπτει από την ελαχιστοποίηση του αθροίσματος των τετράγωνων των σφαλμάτων. $SSE = \sum_{i=t-k+1}^{t}(Y - \hat{Y})^2$

Στην παρακάτω εικόνα παρατηρούμε τους κυλιόμενους μέσους *MA(10) (*μπλε χρώμα*)* και *MA(500) (*κίτρινο χρώμα*)* για την συναλλαγματική αξία του εικονικού νομίσματος *WIN* σε σχέση με το εικονικό νόμισμα *USDT (Tether).* Παρατηρούμε ότι ο κυλιόμενος μέσος 50 περιόδων προσαρμόζεται πιο «*δύσκολα*» στις αλλαγές τις τιμής.

Εικόνα 6: Ενναλακτήρια τιμή USDT-WIN με εφαρμογή Κινητού Μέσου. Πηγή CoinMarket.com

Η μέθοδος αυτή χρησιμοποιείτε με την παρουσία τάσης, για να έχουμε καλύτερα αποτελέσματα από τον απλό μέσο όρο. Στις περιπτώσεις που ο κυλιόμενος μέσος χρησιμοποιεί άρτιο παράθυρο, για παράδειγμα *MA(4),* δεν υπάρχει κεντρική περίοδος στην οποία θα αντιστοιχηθεί ο κυλιόμενος μέσος. Σε αυτές τις περίπτωσης κάνουμε χρήση του επικεντρωμένου κυλιόμενου μέσου [12].

3.2 Κινητός μέσος με τάξης 1

Ο ποιο απλός παράγοντας που συναντάμε στο μοντέλο του κινητού μέσου είναι όταν ***q=1***. Η εξίσωση γίνεται:

$$Y_t = \varepsilon_t - \theta\varepsilon_{t-1}$$

Όπου, ε_t είναι η διαδικασία του λευκού θορύβου με μέσο όρο 0 και διακύμανση σ^2. Η διαδικασία του κινητού μέσου με βήμα 1 εκφράζει τις τιμές τις χρονοσειράς με τις τιμές του λευκού θορύβου της ποιο πρόσφατης περιόδου και τις προηγούμενης της πολλαπλασιαζόμενης με συντελεστή $\boldsymbol{\vartheta}$. Όταν η χρονοσειρά έχει μέση τιμή $\mu \neq 0$ τότε η *MA(1)* ορίζεται σε αποκλίσεις από την μέση τιμή τη χρονοσειράς, δηλαδή [13]:

$$Y_t - \mu = \varepsilon_t - \theta\varepsilon_{t-1}$$

3.3 Εκθετική Εξομάλυνση

Η μέθοδος της εκθετικής εξομάλυνσης (exponential smoothing) λαμβάνει υπόψιν όλες τις τιμές τις χρονοσειράς δίνοντας όμως διαφορετικό βάρος στην κάθε παρατήρηση.

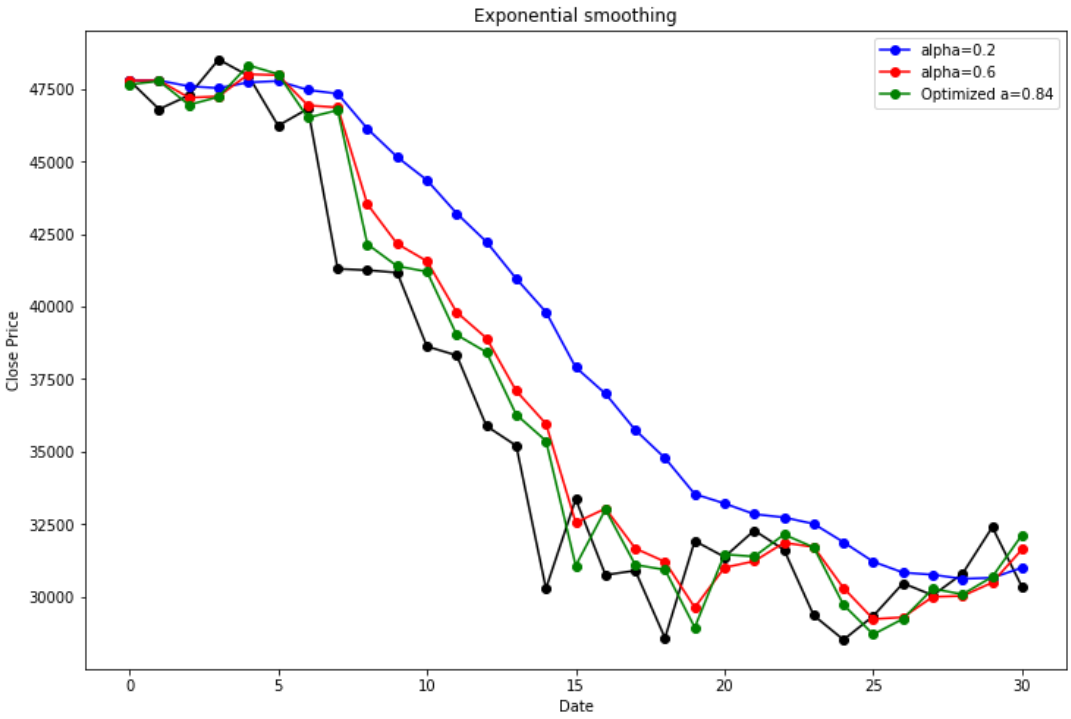
Η εξίσωση είναι

$$\widehat{Y_{t+1}} = y_t a + (1-a)\widehat{y_t}$$

Οπού t η τρέχουσα περίοδος, το α είναι η σταθερά της εξομάλυνσης (παίρνει τιμές 0 έως 1) και οι $\widehat{y_t}$, $\widehat{Y_{t+1}}$ τιμές αναφέρονται στην πρόβλεψη της μελλοντικής τιμής καθώς και στην τιμή της παρατήρησης αυτήν την περίοδο.

Όσο ποιο μεγάλη η μεταβλητή α τόσο και μεγαλύτερη η βαρύτητα που δίνετε στην πιο πρόσφατη παρατήρηση. Με άλλα λόγια μόνο οι ποιο πρόσφατες παρατηρήσεις επηρεάζουν την πρόβλεψη.

Στο παρακάτω παράδειγμα παρατηρούμε την τιμή της συναλλαγματικής αξίας του BTC/EUR για τον μήνα Ιούνιο και την εκθετική εξομάλυνση για τα δεδομένα αυτά.



Εικόνα 7:Εκθετική εξομάλυνση χρονοσειράς BTC/EUR.

Στην παραπάνω περίπτωση προβλέψαμε την μελλοντική πορεία της τιμής με τρείς διαφορετικές τιμές της παραμέτρου a. Για a=0.2, για a=0.6 και το καλύτερο αποτέλεσμα το πήραμε για a=0.84.

Η εκθετική εξομάλυνση εφαρμόζεται σε χρονοσειρές χωρίς τάση [13].

3.4 Εκθετική Εξομάλυνση με προσαρμογή τάσης (Holt Method)

Η μέθοδος αυτή προτάθηκε το *1957* και επιτρέπει την πρόβλεψη δεδομένων με τάση. Η εξίσωση πρόβλεψης περιλαμβάνει δύο υπό-εξισώσεις. Η μία αφορά την εξομάλυνση τη τάσης και η άλλη την εξομάλυνση της χρονοσειράς. Ποιο συγκεκριμένα,

$$A_t = ay_t + (1-a)(A_{t-1} + T_{t-1}) \quad (1)$$

$$T_t = \beta(A_t - A_{t-1}) + (1-\beta)T_{t-1} \quad (2)$$

$$\widehat{y_{t+h}} = A_t + hT_t$$

Η εξίσωση νούμερο (1) αφορά την εξομάλυνση της χρονοσειράς, η εξίσωση (2) δίνει την εκτίμηση της τάσης για την χρονική στιγμή *t* και η εξίσωση (3) δίνει την πρόβλεψη την χρονική στιγμή *t+h.* Η παράμετρος *h* αφορά την περίοδο μετά την τελευταία παρατήρηση για την οποία κάνουμε πρόβλεψη. Παρατηρούμε ότι η παράμετρος *a* είναι η σταθερά εξομάλυνσης της χρονοσειράς και η *β* η σταθερά εξομάλυνσης της τάσης. Ισχύει ότι $0 \leq \alpha, \beta \leq 1$ [1].

3.5 Πρόβλεψη χρωνοσειρών με την χρήση στοχαστικών Μοντέλων

Η επιλογή του κατάλληλου μοντέλου πρόβλεψης είναι ιδιαίτερα σημαντική. Ένα μοντέλο χρονοσειράς ονομάζεται γραμμικό ή μη-γραμμικό χρονοσειρά εάν μια παρατήρηση έχει γραμμική ή μη-γραμμική συσχέτιση με τις προηγούμενες παρατηρήσεις. Στην βιβλιογραφία, παρουσιάζονται συχνά τα γραμμικά μοντέλα ***AR (Autoregressive model)*** και ***MA (Moving Average model).*** Ο συνδυασμός αυτών των δύο παράγει τα μοντέλα ***ARMA (Autoregressive Integrated Moving Average)*** και ***ARMA (Autoregressive Moving Average).*** Στην περίπτωση χρωνοσειρών που παρουσιάζουν κυκλικότητά ή εποχικότητα χρησιμοποιείται μια παραλλαγή των ***ARIMA*** μοντέλων, τα ***SARIMA (Seasonal Autoregressive Integrated Moving Average).*** Τα παραπάνω στοχαστικά μοντέλα έχουν είναι γνωστά και ως ***Box-Jenkins*** μοντέλα.

3.5.2 AR (Αυτοπαλινδρρούμενα Μοντέλα)

Στα αυτοαναδρομικά μοντέλα (*Autoregressive models)* κάνουμε πρόβλεψη της μεταβλητής κάνοντας χρήση ενός γραμμικού συνδυασμού των *προηγούμενων* τιμών της μεταβλητής. Ο όρος αυτοαναδρομικό – αυτοπαλίνδρομο μοντέλο δηλώνει ότι πρόκειται για παλινδρόμηση της μεταβλητής ως προς της παλαιότερες τιμές τις.

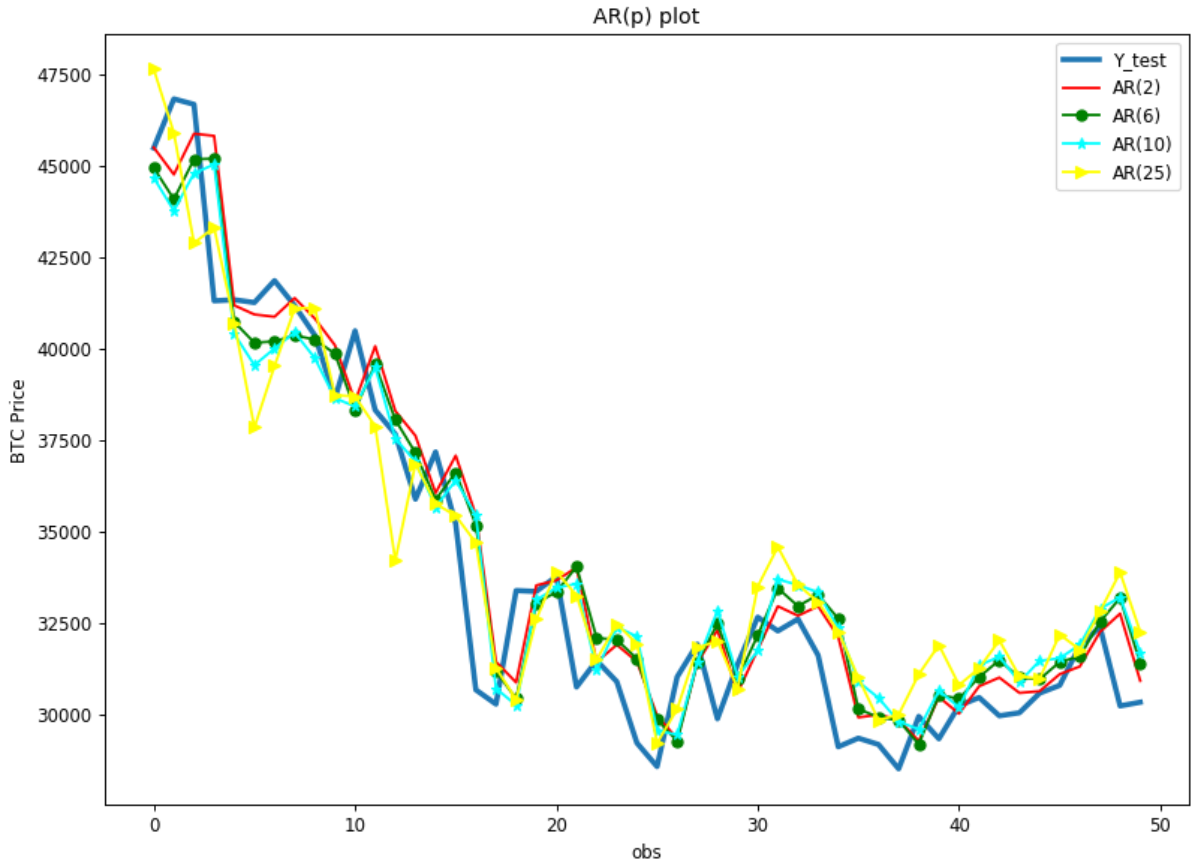
Ένα αυτοπαλίνδρομο υπόδειγμα *p* τάξεως, ή AR(*p*) έχει μορφή:

$$\boldsymbol{y_t = c + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \ldots + \varphi_p y_{t-p} + e_t},$$

Όπου ως τάξη *p* ορίζεται η χρονική υστέρηση, η μεταβλητή θεωρείται ότι είναι λευκός θόρυβος και ως Y_{t-p} ορίζουμε τις *παλαιότερες* τιμές της εξαρτημένης μεταβλητής $\boldsymbol{y_t}$.

Για τις διάφορες τιμές του φ καταλήγουμε σε διαφορετικές χρονοσειρές. Συγκεκριμένα, για ένα μοντέλο AR(1) :

- Όταν $\boldsymbol{\varphi_1 = 0}$ τότε, η εξαρτημένη μεταβλητή $\boldsymbol{y_t}$ είναι παρόμοια του λευκού θορύβου,
- Όταν $\boldsymbol{\varphi_1 = 1}$ ***και*** $\boldsymbol{c = 0}$ τότε, η εξαρτημένη μεταβλητή $\boldsymbol{y_t}$ συμπεριφέρεται όπως ο τυχαίος περίπατος.
- Όταν $\boldsymbol{\varphi_1 = 1}$ ***και*** $\boldsymbol{c \neq 0}$ τότε, η εξαρτημένη μεταβλητή $\boldsymbol{y_t}$ συμπεριφέρεται όπως τυχαίος περίπατος με μετατόπιση
- Όταν $\boldsymbol{\varphi_1 < 0}$ τότε, η εξαρτημένη μεταβλητή $\boldsymbol{y_t}$ τείνει να *ταλαντώνεται* γύρω από την μέση τιμή.



Εικόνα 8: Εφαρμογή AR μοντέλου διαφορετικών παραθύρων (lags) στην τιμή του BTC.

3.5.3 Πρώτες Διαφορές (differencing)

Η διαδικασία της διαφόρισης περιορίζει της διακυμάνσεις επιπέδου μιας χρονοσειράς αφαιρώντας την τάση και την εποχικότητα. Ο υπολογισμός των διαφορών μεταξύ διαδοχικών παρατηρήσεων αποτελεί μία τεχνική μετατροπής μίας μη-στάσιμης χρονοσειράς σε στάσιμη, στην απαλοιφή με άλλα λόγια της τάσης. Η πρώτες διαφορές μπορούν να εκφραστούν μέσο της σχέσης [12] :

$$y_t' = y_t - y_{t-1}$$

Όπως είναι κατανοητό η χρονοσειρά των διαφορών θα έχει Τ-1 παρατηρήσεις, αφού η τιμή της πρώτης y_1' παρατήρησης δεν μπορεί να υπολογιστεί.

Σε περιπτώσεις που οι χρονοσειρά δεν μετατραπεί σε στάσιμη μπορεί να χρειαστεί να υπολογίσουμε τις διαφορές δεύτερης τάξης, οι οποίες δίνονται από την σχέση:

$$y''_t = y'_t - y'_{t-1} = (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}) = y_t - 2y_{t-1} + y_{t-2}$$

3.5.4 Εποχική Διαφορά

Η εποχική διαφορά, αφορά στην διαφορά μεταξύ μιας παρατήρησης και της προηγούμενη της από την ίδια περίοδο. Άρα, έχουμε:

$$y'_t = y_t - y_{t-m}$$

όπου m ο αριθμός τω περιόδων. Αν τα δεδομένα μας μετά την διαφόριση είναι λευκός θόρυβος τότε ένα μοντέλο πρόβλεψης της αρχικής χρονοσειράς που προτείνει η βιβλιογραφία είναι το

$$y'_t = y_{t-m} + e_t$$

Οι προβλέψεις του μοντέλου αυτού είναι ίσες με την τιμή της τελευταίας παρατήρησης.
Αποτελεσματικά, αυτή η μέθοδος δίνει αποτελέσματα παρόμοια με την Naïve μέθοδο, σε επίπεδο εποχών.

3.6 ARIMA μοντέλα

Ο συνδυασμός της αυτοπαλινδρόμησης, του μοντέλου του κινητού μέσου και της διαδικασίας των διαφορών(διαφόριση) προκύπτει το αυτοπαλινδρονούμένο μοντέλο κινητού μέσου. Το μοντέλο χαρακτηρίζεται από την σχέση:[1]

$$y'_t = c + \varphi_1 y'_{t-1} + \ldots + \varphi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q} + \varepsilon_t$$

και περιγράφεται ως ***ARIMA(p,d,q)*** με,

- **p =** παράγοντας της αυτοπαλινδρόμησης
- **d =** βαθμός διαφορών
- **q =** παράγοντας του κινητού μέσου

Πολλά από τα μοντέλα της βιβλιογραφίας μπορούν να περιγραφούν από ένα μοντέλο ARIMA.

Λευκός Θόρυβος	*ARIMA(0,0,0)*
Τυχαίος Περίπατος	*ARIMA(0,1,0)*
Τυχαίος Περίπατος με διαφορές	*ARIMA(0,1,0) with constant*
Αυτό παλινδρόμηση	*ARIMA(p,0,0)*
Κινητός μέσος	*ARIMA(0,0,q)*

Εικόνα 9:Περιγραφή μοντέλων ARIMA (Πηγή: [1], Κεφάλαιο: 8.5).

Η επιλογή των παραγόντων πολλές φορές αποδεικνύεται δύσκολη. Για την επιλογή των παραγόντων μπορεί να χρησιμοποιηθεί το μέτρο μέγιστης πιθανοφάνιας (MLE - maximum likelihood estimation).

Αυτή η τεχνική είναι παρόμοια με την μέθοδο ελαχίστων τετραγώνων για τα μοντέλα ARIMA, το αποτέλεσμα δηλαδή αν ελαχιστοποιήσουμε τον παράγοντα*:* $\sum_{t=1}^{T} \varepsilon_t^{\,2}$

Επιπλέον, όπως για τα άλλα αυτοπαλινδρρούμενα μοντέλα, μπορούν να χρησιμοποιηθούν τα κριτήρια πληροφορίας AIC και BIC για τον καθορισμό των παραμέτρων.[1]

Ποιο συγκεκριμένα,

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

Για το αντίστοιχο Μπεϋζιανό μοντέλο,

$$BIC = AIC + [log(T) - 2](p + q + k + 1)$$

Όσο ποιο μικρή η τιμή των παραπάνω κριτηρίων τόσο καλύτερο το μοντέλο μας. *Αξίζει να σημειωθεί ότι τα κριτήρια πληροφορίας αποτελούνε οδηγό κυρίως για τις παραμέτρους p και q και όχι τόσο για τον παράγοντα διαφόρισης d.*

4. Επεξεργασία Φυσικής Γλώσσας

Παρά το γεγονός ότι στη σημερινή εποχή πολλές φορές οι συναλλαγές του κοινού ή των πελατών γίνονται χωρίς να απαιτείται φυσική παρουσία, είναι εξαιρετικά σημαντικό για τις επιχειρήσεις και τους οργανισμούς να γνωρίζουν, να αντιλαμβάνονται και να δύνανται να αναλύσουν τα συναισθήματα των συναλλασσόμενων. Μάλιστα, όσο περισσότερο περιορίζεται η φυσική παρουσία και η προσωπική επαφή τόσο σημαντικότερη καθίσταται η ανάγκη αναγνώρισης των συναισθημάτων με τρόπους που αξιοποιούν την σύγχρονη τεχνολογία και τις δυνατότητές της. Από την άλλη πλευρά, η ραγδαία ανάπτυξη των ψηφιακών τεχνολογιών και του διαδικτύου οδήγησε σε μια άνευ προηγουμένου αύξηση των δεδομένων κειμένου, δημιουργώντας νέες ευκαιρίες και προκλήσεις. Οι ερευνητές και επιστήμονες του κλάδου αυτού, όπως γλωσσολόγοι, έχουν πλέον πρόσβαση στα δεδομένα, που μπορεί να διαφέρουν δραστικά ανάλογα με το αντικείμενο ή το θεματικό πεδίο όσον αφορά στο περιεχόμενο και την ευρύτητα [15].

Δεδομένης της έκρηξης του όγκου αλλά και του βαθμού διακίνησης των δεδομένων, σήμερα υπάρχουν διαθέσιμα από πολλαπλές πηγές και σε πολλαπλές μορφές δεδομένα, τα οποία παρέχουν τεράστιες δυνατότητες κατανόησης των συναισθημάτων, αρκεί κανείς να επιτύχει την αποτελεσματική επεξεργασία τους με μαζικό ή επιλεκτικό τρόπο, φιλτράροντας τον όποιο θόρυβο τα περιβάλλει. Ως εκ τούτου, ένα από τα σύγχρονα ερευνητικά προβλήματα που έχει αναδυθεί τα τελευταία χρόνια είναι η προσπάθεια αυτοματοποιημένης ανάλυση συναισθημάτων (sentiment analysis), όρος που συναντάται και ως εξόρυξη γνώμης (opinion mining), και έχει ως στόχο την ανίχνευση και ανάλυση κειμένων και ταξινόμησή των γραφόμενων ανάλογα με τα συναισθήματα που εμπεριέχουν [16].

Ο θεμελιώδης στόχος επιχειρήσεων και οργανισμών είναι να αξιοποιήσουν τα δεδομένα αυτά με όσο γίνεται πιο αποδοτικό και αποτελεσματικά για αυτούς τρόπο. Εάν προσπαθήσει κάποιος να δώσει έναν τεχνικό ορισμό στην ανάλυση συναισθημάτων ή εξόρυξη γνώμης, θα μπορούσε να πει ότι πρόκειται για μια ανάλυση δεδομένων κειμένου, που μπορεί να δείξει το είδος των συναισθημάτων που εμπεριέχονται, δηλαδή θετικά, ουδέτερα ή αρνητικά συναισθήματα, που μπορούν επιπλέον να ποσοτικοποιηθούν με κάποιο score. Δεδομένου ότι η ανάλυση συναισθημάτων προκύπτει από την αξιολόγηση των απόψεων που εμπεριέχουν τα δεδομένα κειμένου, είναι επίσης γνωστή και ως εξόρυξη γνώμης. Με άλλα λόγια, θα μπορούσε κανείς να ορίσει την ανάλυση συναισθημάτων ως μια

διαδικασία εντοπισμού και ανάλυσης απόψεων και γνωμών που διατυπώνονται σε τμήματα κειμένου, βάσει της οποίας προκύπτει αν τα υπό εξέταση τμήματα κειμένου εκφράζουν θετικές, ουδέτερες, ή αρνητικές απόψεις [17].

Αν και η ανάλυση συναισθημάτων έχει γίνει πολύ γνωστή τα τελευταία χρόνια, τα πρώτα βήματα θεμελίωσης της ξεκίνησαν από τις αρχές του 21ου αιώνα. Κατά τη διάρκεια αυτής της περιόδου, υπήρξαν πολλές προσπάθειες, οι οποίες ασχολήθηκαν κατά βάση με τον τρόπο συλλογής και ανάλυσης δεδομένων. Ένας από τους σπουδαιότερους λόγους, για τον οποίο η ανάλυση συναισθημάτων είναι σημαντική, είναι ότι δίνει τη δυνατότητα σε όποιον επιθυμεί να το κάνει, να αντιληφθεί και να αναλύσει τη γενική γνώμη του κοινού για ένα ζήτημα, ένα προϊόν, μια υπηρεσία [18].

Η μεγαλύτερη δυσχέρεια που παρουσιάζει η ανάλυση συναισθημάτων πηγάζει από τις ιδιαιτερότητες της εκάστοτε φυσικής γλώσσας, οι οποίες δύνανται να διαφοροποιήσουν το νόημα και την ερμηνεία κατά περίπτωση και ανάλογα με τα συμφραζόμενα, κυρίες σε πιο δυσδιάκριτες περιπτώσεις όπου συναντούμε ειρωνεία ή σαρκασμό [19].

4.1 Τύποι ανάλυσης συναισθημάτων

Ο τρόπος με τον οποίο χρησιμοποιείται κάθε φορά η φυσική γλώσσα διαφοροποιείται ανάλογα με τους συμμετέχοντες, την εκάστοτε οπτική τους και τον τρόπο με τον οποίο αλληλοεπιδρούν. Ο όρος ανάλυση συναισθημάτων ορίζεται συνήθως ως η διαδικασία με την οποία τμηματοποιούνται κομμάτια κειμένου, που εκτείνονται από μεμονωμένες λέξεις και φράσεις ή προτάσεις έως πλήρη έγγραφα, σε ένα μικρό αριθμό κατηγοριών που αντιπροσωπεύουν διαφορετικά είδη συναισθημάτων. Ο όρος συναίσθημα στη συγκεκριμένη γνωστική περιοχή αποτελεί συνήθως συνώνυμο όρων όπως συγκίνηση, επίδραση κ.λπ. Στην απλούστερη διατύπωσή της, η ανάλυση συναισθημάτων θεωρείται δυαδικό πρόβλημα, κατά την επίλυση του οποίου εντοπίζονται και διαχωρίζονται συναισθηματικά τα θετικά από το αρνητικά συναισθήματα, διαδικασία γνωστή και ως ανίχνευση πολικότητας (Turney, 2002). Σε μεταγενέστερες ερευνητικές προσπάθειες, κυριάρχησε η προσέγγιση κατά την οποία στην ανάλυση συναισθημάτων πραγματοποιείται κατηγοριοποίησή τους σε τρεις κατηγορίες, αρνητικά, ουδέτερα και θετικά, ενώ υπάρχει και βιβλιογραφία που χρησιμοποιεί κλιμακωτή διαβάθμιση σε περισσότερες από τρεις κατηγορίες, για παράδειγμα ασθενώς έως έντονα αρνητική, ουδέτερη, ασθενώς ως έντονα θετική [20].

Λόγω της πολύ ευρείας έρευνας που βρίσκεται σε εξέλιξη στην εν λόγω περιοχή, υπάρχουν και πολλά είδη διαχωρισμού της ανάλυσης συναισθημάτων. Θα μπορούσε κανείς να συνοψίσει τους τύπους ανάλυσης συναισθημάτων στις εξής τέσσερις κατηγορίες:

▪ Λεπτομερής Ανάλυση Συναισθήματος (Fine-grained Sentiment Analysis): Πρόκειται για ανάλυση με την οποία μπορούν να γίνουν κατανοητά το ύφος και η πολικότητα των σχολίων επί κάποιου θέματος. Ο συγκεκριμένος τύπος ανάλυσης είναι εν γένει απαιτητικός και υψηλού κόστους σε σύγκριση με άλλα είδη, διότι λαμβάνει χώρα σε μεγάλο επίπεδο λεπτομέρειας.

▪ Ανάλυση Ανίχνευσης Συναισθήματος (Emotion Detection Sentiment Analysis): Είναι προχωρημένος τύπος ανάλυσης και αναγνώρισης συναισθήματος ο οποίος αξιοποιεί τεχνικές όπως χρήση λεξικών και μηχανική μάθηση (machine learning). Ένα από τα πιο γνωστά μοντέλα αυτού του τύπου είναι το μοντέλο «Big Six» του Ekman, το οποίο διαβαθμίζει τα αισθήματα σε θυμό, φόβο, ευτυχία, έκπληξη, αηδία και θλίψη [21].

Επίσης, στην κατηγορία αυτή κατατάσσονται τα διαστατικά (dimensional) μοντέλα, τα οποία κατηγοριοποιούν τα συναισθήματα σε περιοχές όπως ευχαρίστηση, διέγερση, κυριαρχία, κ.λπ. Ειδικότερα, τα λεξικά περιλαμβάνουν λίστες λέξεων με θετική ή αρνητική ερμηνεία, ώστε να καθίσταται εφικτή η κατανόηση του συναισθήματος, ενώ για την συνολική αξιολόγηση ενός κειμένου χρησιμοποιούνται αλγόριθμοι τεχνητής νοημοσύνης και μηχανικής μάθησης [23].

▪ Μονοδιάστατη Ανάλυση Συναισθήματος (Aspect-based Sentiment Analysis): Πρόκειται για τύπο ανάλυσης συναισθήματος που επικεντρώνεται σε μία πτυχή μιας υπηρεσίας ή ενός προϊόντος

▪ Ανάλυση Προθέσεων (Intent Analysis): Ο συγκεκριμένος τύπος ανάλυσης συναισθήματος εμβαθύνει την ανάλυση σε τέτοιο βαθμό ώστε να μπορεί να δώσει και προβλέψεις της πρόθεσης ενός ατόμου με βάση τα χαρακτηριστικά του συναισθήματός του. Για παράδειγμα, η αγορά ή όχι ενός προϊόντος ή μιας υπηρεσίας από πλευράς ενός πελάτη.

4.2 Μέθοδοι και τεχνικές ανάλυσης συναισθημάτων

Η ανάλυση συναισθημάτων επιτυγχάνεται τεχνικά όχι μόνο με την εξέταση κάθε δεδομένου/ λέξης μεμονωμένα, αλλά επίσης και συνδυαστικά ώστε να συνδεθούν και να ερμηνευτούν με ολιστικό τρόπο όλες οι έννοιες που έχουν καταγραφεί και η ερμηνεία τους. Τα δεδομένα κειμένου συνήθως προέρχονται από διάφορες πηγές όπως μέσα κοινωνικής δικτύωσης, μηνύματα (email, sms, chat, κ.λπ.), ιστότοπους, μεταδεδομένα (εικόνων, βίντεο, κ.λπ.) και άλλα.

Για την υλοποίηση των προαναφερθέντων τύπων ανάλυσης συναισθήματος χρησιμοποιούνται διαφορές τεχνικές. Στην απλή περίπτωση οι τεχνικές αυτές στηρίζονται σε κανόνες (rule-based), ενώ στους πλέον προηγμένους τύπους ανάλυσης, όπως ήδη αναφέρθηκε, χρησιμοποιούνται αυτοματοποιημένες τεχνικές που στηρίζονται σε αλγόριθμους τεχνικής νοημοσύνης (artificial intelligence) και μηχανικής μάθησης (machine learning). Ακολούθως παρατίθενται οι βασικές τεχνικές ανάλυσης συναισθημάτων [17]:

▪ Τεχνικές Ανάλυσης Συναισθήματος που Στηρίζονται σε Κανόνες (Rule-based): Καθορίζονται από τον προγραμματιστή με ντετερμινιστικό τρόπο. Η ανάλυση στηρίζεται σε απλούς ή σύνθετους κανόνες, η εκτέλεση των οποίων καθορίζει και το αποτέλεσμα. Παράδειγμα τέτοιου κανόνα θα μπορούσε να είναι ο υπολογισμός ενός score με βάση τον συνολικό αριθμό των θετικών και αρνητικών λέξεων που εμφανίζονται σε ένα κείμενο. Το ύψος του score που προκύπτει με τον τρόπο αυτό καθορίζει και το αν τελικά το υπό εξέταση κείμενο θα χαρακτηριστεί «θετικό», «ουδέτερο» ή «αρνητικό». Οι rule-based τεχνικές δεν ενδείκνυνται για περιπτώσεις κατά τις οποίες τα δεδομένα μεταβάλλονται δυναμικά ή έχουν μεγάλο βαθμό πολυπλοκότητας.

▪ Τεχνικές Μηχανικής Μάθησης: Εφαρμόζουν αλγόριθμους που δεν απαιτούν χειροκίνητο (manual) χειρισμό. Κατά την εφαρμογή αυτών των τεχνικών οι αλγόριθμοι που χρησιμοποιούνται εκπαιδεύουν το σύστημα να αναγνωρίζει με τρόπο δυναμικό τα δεδομένα και να τα διαβαθμίζει ως θετικά, ουδέτερα ή αρνητικά. Επομένως, οι εν λόγω τεχνικές ενδείκνυνται για αναλύσεις μεγαλύτερης πολυπλοκότητας. Χαρακτηριστικά παραδείγματα αλγορίθμων μηχανικής μάθησης που χρησιμοποιούνται για ανάλυση συναισθημάτων είναι η γραμμική παλινδρόμηση (linear regression), ο αλγόριθμος Naïve Bayes, οι τεχνικές διανυσματικής μηχανικής (vector machines) και αλγόριθμοι βαθιάς μάθησης (deep learning).

▪ Υβριδικές Τεχνικές: Όταν οι rule-based τεχνικές συνδυάζονται με τεχνικές μηχανικής μάθησης, προκύπτουν τεχνικές που συχνά αποκαλούνται υβριδικές (hybrid). Το πλεονέκτημα αυτών των τεχνικών είναι ότι ισορροπούν τον απαιτούμενο αυτοματισμό με την πολυπλοκότητα υλοποίησης και ενδείκνυνται στην περίπτωση που αφενός οι rulebased τεχνικές δε δουλεύουν καλά και αφετέρου οι τεχνικές μηχανικής μάθησης είναι ιδιαίτερα πολύπλοκες χωρίς να φέρνουν ανάλογη της πολυπλοκότητας αυτής προστιθέμενη αξία.

4.3 Ανάλυση συναισθήματος (Sentiment analysis)

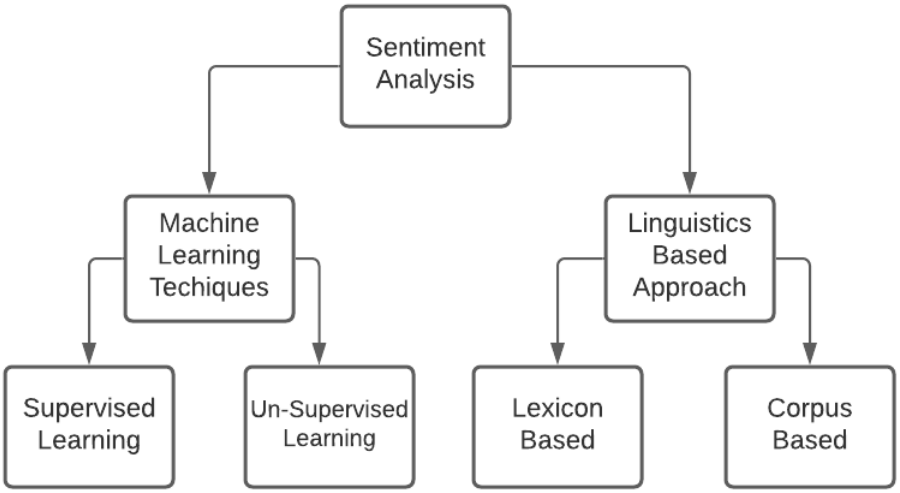
Οι τεχνικές εξόρυξης γνώμης από κείμενα-γραπτό λόγο είναι αποτελεί νέα τάση στον χώρο των επιχειρήσεων και της αναλυτικής. Σήμερα, οι έξυπνες υπολογιστικές μέθοδοι, η αναγνώριση προτύπων καθώς και η μηχανική μάθηση αποτελούν ισχυρά εργαλεία ανταγωνισμού στην αγορά στα χέρια των επιχειρήσεων. Ένα γνωστό παράδειγμα, είναι εκείνο των επιχειρήσεων που προσπαθούν ,μέσω της ανάλυσης των δεδομένων και της ανακάλυψης προτύπων μέσα από τα δεδομένα, να βελτιώσουν και να κατανοήσουν την εμπειρία του καταναλωτή – πελάτη. Ένα μεγάλο μέρος των σημερινών δεδομένων του διαδικτύου στις μέρες μας αποτελεί γραπτό λόγο (παραδείγματος χάρη κριτικές, απόψεις, περιγραφές, blogposts, social media και άλλα). Χαρακτηριστικό είναι ότι το Twitter ανακοίνωσε ότι μέσος αριθμός tweets ανά λεπτό για το 2020 ήταν 350000 . Μέσα από αυτόν των τεράστιο όγκο δεδομένων οι επιχειρήσεις μπορούν να αντλήσουν πλούσια πληροφορία σχετικά με την εμπειρία των καταναλωτών, την βελτίωση των προϊόντων και υπηρεσιών καθώς και για μελλοντικές προβλέψεις. Οι παραπάνω λόγοι είναι που κάνουν τις τεχνικές sentiment analysis (ανάλυση συναισθήματος) τόσο διαδεδομένοι στην επιστήμη των υπολογιστών και σε κοινωνικές επιστήμες [24].

Η εξόρυξη γνώμης (sentiment analysis) αποτελεί ένα νέο πεδίο της επιστήμης των δεδομένων, και έχει ως στόχο την ανακάλυψη της υποκειμενικότητας (προσωπική άποψη) σε κείμενα και/ή την εξαγωγή και κατηγοριοποίηση των συναισθημάτων και τον απόψεων. Η επιστήμη αυτή έχει τις ρίζες της στην Επεξεργασία Φυσικής Γλώσσας (Natural Language Processing (NLP)) και μπορεί να εφαρμογή [26].

4.3.1 Sentiment Analysis Methods

Οι μέθοδοι της ανάλυσης συναισθήματος χωρίζονται σε τεχνικές μηχανικής μάθησης (Machine Learning), τεχνικές βασισμένες σε στην γλώσσα (Linguistic Approach) και τον συνδυασμός τους. Η διαδικασία μπορεί να εφαρμοστεί σε επίπεδο κείμενου, σε επίπεδο προτάσεων ή και σε υπό-μέρη προτάσεων (λέξεις) .

- Κείμενο: Σε αυτό το επίπεδο, ο τελικός χαρακτηρισμός αποδίδεται σε ολόκληρο το κείμενο.
- Πρόταση: Σε αυτό το επίπεδο θεωρείτε ότι κάθε πρόταση περιγράφει ένα συναίσθημα. Ο τελικός χαρακτηρισμός του κειμένου αποτελείτε από το άθροισμα των χαρακτηρισμών των επιμέρους προτάσεων.
- Επίπεδο λέξης-Σύνολο λέξεων: Μια πρόταση μπορεί να περιλαμβάνει σύνολο λέξεων ή λέξεις η οποίες μπορούν να έχουν τον δικό τους χαρακτηρισμό ως προς το συναίσθημα που δηλώνουν [25].



Εικόνα 10: Μέθοδοι ανάλυσης συναισθήματος.

Στις τεχνικές μηχανικής μάθησης ένα σύνολο δεδομένων με εκ των προτέρων κατηγοριοποίηση του συναισθήματος δίνεται με σκοπό την εκπαίδευση του αλγορίθμου στο σύνολο αυτό. Οι τεχνικές βασισμένες σε χρήση λεξικού (lexicon-based) περιλαμβάνουν λίστες λέξεων η ακόμη και φράσεων η οποίες έχουν αντιστοιχηθεί σε κάποιο συναίσθημα.

a. Μέθοδοι μηχανικής μάθησης
   i. Επιβλεπόμενη Μάθηση

      Τέτοιες μέθοδοι χρησιμοποιούνται όταν υπάρχει σύνολο δεδομένων για εκπαίδευση. Η διαδικασία περιέχει την εκπαίδευση κάποιου αλγορίθμου μηχανικής μάθησης με ένα υποσύνολο του συνόλου εκπαίδευσης και την εφαρμογή, πρόβλεψή και επικύρωση του στο άλλο υποσύνολο. Στην συνέχεια ο εκπαιδευμένος αλγόριθμος μπορεί να εφαρμοστεί σε άγνωστο σύνολο δεδομένων. Τέτοιοί αλγόριθμοι μηχανικής μάθησης είναι οι: SVM (Support Vector Machines) και τα δέντρα απόφασης (Decision Trees).
   ii. Μη-επιβλεπόμενη

      Στην περίπτωση της μη-επιβλεπόμενης μάθησης, ο αλγόριθμος δημιουργεί ένα μοντέλο βασισμένο στις παρατηρήσεις, χωρίς να γνωρίζει τις επιθυμητές εξόδους. Σε αυτήν την κατηγορία ανήκουν οι αλγόριθμοι ομαδοποίησης, τα νευρωνικά δίκτυα και οι αλγόριθμοι συσχετίσεων [29].

a) Βασισμένες σε Λεξικό (Lexicon based approach)

Με την ίδια λογική όπου τα κείμενα, οι παράγραφοι και οι προτάσεις μπορούν να χαρακτηριστούν ως προς το συναίσθημα που αντικατοπτρίζουν (θετικό, αρνητικό ή ουδέτερο) το ίδιο ισχύει και με τις λέξεις. Για την ακρίβεια οι λέξεις είναι εκείνες οι οποίες προσδίδουν συναίσθημα στις προτάσεις. Επομένως, θα μπορούσαμε να μετρήσουμε την θετικότητα ή το αρνητικό συναίσθημα που αντανακλά ένα κείμενο, υπολογίζοντας της αρνητικές και τις θετικές λέξεις του κειμένου. Το λεξικό

είναι μία συλλογή λέξεων η οποίες έχουν κατηγοριοποιηθεί ως προς το συναίσθημα με το οποίο σχετίζονται. Όπως είναι κατανοητό το πλήθος των λέξεων καθώς και η ειδίκευσή που περιλαμβάνει κάθε λεξικό (οικονομικά δεδομένα κα.) παίζουν καθοριστικό ρόλο στην αποτελεσματικότητα. Οι τεχνικές αυτές χωρίζονται σε δύο υποκατηγορίες.

b) Corpus-Based

Η τεχνική Corpus-Based αφορά τον χαρακτηρισμό λέξεων βασισμένη σε μια λίστα λέξεων που έχουν χαρακτηριστεί εκ των προτέρων. Λέξεις οι οποίες δεν υπάρχουν στην κατηγοριοποιημένη λίστα, αντιστοιχίζονται σε αυτές που έχουν το πιο σχετικό περιεχόμενο. Ο τελικός χαρακτηρισμός του κειμένου επηρεάζεται από την συχνότητα των λέξεων που εκφράζουν «χαρά» ή «λύπη».

i) VADER

Το VADER (Valence Aware Dictionary and Sentiment Reasoner) είναι ένας συνδυασμός λεξικού και κανόνων βασισμένων στην ανάλυση συναισθήματος που αναπτύχθηκε από τους Hutto και Gilbert. Το VADER έχει την δυνατότητα αναγνώριση της πολικότητας των απόψεων, την κατηγοριοποίηση δηλαδή της γνώμης σε ‘Θετική’, ‘Αρνητική’ και ‘Ουδέτερη’ , καθώς και την ένταση-δυναμικότητα της γνώμης. Ποσοτικοποίησή με αλλά λόγια την εξορυγμένη γνώμη. Το VADER έχει αναπτυχθεί κυρίως για την ανάλυση της γλώσσας, των συμβόλων και των διαφορετικών τρόπων γραφής κειμένων στα μέσα μαζικής ενημέρωσης [38].

Το VADER περιλαμβάνει συνοπτικές λέξεις, αργκό και φάτσες “emoticons”. Το VADER περιλαμβάνει 7500 χαρακτηριστικά καθώς και κανόνες οι οποίοι έχουν εξαχθεί από την ανάλυση tweets. Η ανάλυση των tweets από τους δημιουργούς του λεξικού (Hutto και Gilbert) κατέληξε σε πέντε βασικά χαρακτηριστικά- κανόνες τα οποία επηρεάζουν την ένταση της γνώμης. Ο συνδυασμός των χαρακτηριστικών και τον κανόνων συνιστούν το λεξικού [37].

4.4 Τεχνικές Καθαρισμού δεδομένων

Όπως στα περισσότερα προβλήματα μηχανικής μάθησης, έτσι και στην εξαγωγή συναισθήματος από κείμενο, ο καθορισμός των δεδομένων παίζει ουσιαστικό ρόλο. Ο καθορισμός κείμενο. Κύριες μέθοδοι αποτελούν:

- Η διαγραφή σημείων στίξης
- Η διαγραφή των URL
- Η διαγραφή των *stopwords* (παραδείγματος χάριν «ΤΟ» , «ΑΥΤΟ»)
- Διαίρεση σε σύμβολα
- Προσαρμογή γραμμάτων σε Κεφαλαία ή Πεζά
- Stemming
- Lemmatization

4.4.1 Stemming

Για λόγους γραμματικής, τα κείμενα χρησιμοποιούν διάφορες μορφές της ίδιας λέξης, παραδείγματος χάριν *'παίζω', 'παίζοντας'*. Επιπλέον, υπάρχουν λέξεις οι οποίες ανήκουν στην ίδια οικογένεια λέξεων και έχουν παρόμοια έννοια, όπως *'δημοκρατία'* και *'δημοκρατικός'*. Με την αποκοπή των καταλήξεων *(stemming)* πολλές φορές καταφέρνουμε να επιτύχουμε τον στόχο, την απλοποίηση και διατήρηση του κορμού της λέξης. Ως αποτέλεσμα όλες οι λέξεις που ανήκουν στην ίδια οικογένεια λέξεων θα κατηγοριοποιηθούν και θα ληφθούν το ίδιο υπόψιν σε ένα πρόβλημα εξαγωγής άποψης και συναισθήματος από κείμενο [40].

4.4.2 Lemmatization

Η απλοποίηση μιας λέξης στο αρχικό της λήμμα είναι μία ακόμη τεχνική προεπεξεργασίας δεδομένων σε προβλήματα ανάλυσης κειμένου και εξαγωγής συναισθήματος. Η λημματοποίηση στηρίζεται σε λεξιλογικούς και μορφολογικούς κανόνες για την ένταξη ενός λήμματος στην κατάλληλη ρίζα της λέξης [40]. Κύρια χαρακτηριστικά αυτής της τεχνικής είναι η απαλοιφή των κλίσεων μιας λέξης και η αφαίρεση καταλήξεων [39] .

4.5 Support Vector Machine (SVM)

Σκοπός των SVM's είναι η ανακάλυψη – δημιουργία ενός υπερεπιπέδου, το οποίο διαχωρίζει με τον βέλτιστο τρόπο (η απόσταση του κοντινότερου σημείου της μίας κατηγορίας με το υπερεπίπεδο να είναι ίση με την απόσταση του κοντινότερου σημείου της άλλης κατηγορίας με το υπερεπίπεδο) δύο κατηγορίες. Η εξίσωση του υπερεπιπέδου μπορεί να γραφεί ως: $\mathrm{H}\colon \mathrm{w}^{\mathrm{T}}\mathrm{x} + \mathrm{b} = 0$

Όπως είναι κατανοητό το υπερεπίπεδο σε έναν δισδιάστατο χώρο θα είναι μία ευθεία, σε τρισδιάστατο θα αναπαριστά επίπεδο και ούτω καθεξής.

Οι SVM έχουν την δυνατότητα επίλυσής γραμμικών και μη γραμμικών προβλημάτων. Έχει αποδειχθεί ότι σε προβλήματα ανάλυσης και εξαγωγής συναισθήματος από κείμενο με χρήση SVM, η χρήση της μεθόδου χ-τετραγώνου για την επιλογή των χαρακτηριστικών και την μείωση των διαστάσεων αυξάνει την αποτελεσματικότητα [27].

4.6 Naïve Bayes Classifier

Οι Bayesian κατηγοριοποιητές είναι βασισμένοι στον νόμο του Bayes, ο οποίος μας δίνει την πιθανότητα να πραγματοποιηθεί ένα γεγονός X δεδομένης μιας κατάστασης Y.

$$P(X|Y) = \frac{P(X)P(Y|X)}{P(Y)}$$

Στα προβλήματα εξαγωγής συναισθήματος ο νόμος προσαρμόζεται σε «πιθανότητα είναι συναίσθημα να είναι θετικό/αρνητικό δεδομένου του περιεχομένου» και η εξίσωσης σε:

$$\mathrm{P}(\text{sentiment} \mid \text{sentence}) = \frac{\mathrm{P}(\text{sentiment})\ \mathrm{P}(\text{sentence} \mid \text{sentiment})}{\mathrm{P}(\text{sentence})}$$

Οι Μπεϋζιανοί κατηγοριοποιητές υποθέτουν ότι η επίδραση ενός γνωρίσματος σε μια δεδομένη κατηγορία είναι ανεξάρτητη από τις τιμές των άλλων γνωρισμάτων. Συμπερασματικά στην περίπτωση της εξόρυξης συναισθήματος από κείμενο, δεν υπάρχει σύνδεση-εξάρτηση μεταξύ των λέξεων.

4.7 DECISION TREES

Τα δέντρα απόφασης (Decision Trees) έχουν μία δεντρική μορφή οπού τα φύλλα του αναπαριστούν της κατηγορίες και οι κόμβοι τα χαρακτηριστικά (λέξεις). Όσο ποιο ψηλά (κοντά στην ρίζα) στην δομή βρίσκετε ένα χαρακτηριστικό τόσο μεγαλύτερη η πληροφορία του. Η σημαντικότητα του χαρακτηριστικού μπορεί να μετρηθεί με στατιστικές μεθόδους όπως ο συντελεστής τζινί (Gini Index) και το κέρδος πληροφορίας (Information Gain).

Information Gain

Το κέρδος πληροφορίας χαρακτηρίζει το πόση πληροφορία «*φέρει*» ένα χαρακτηριστικό. Για παράδειγμα εάν το *A* είναι ένα χαρακτηριστικό σε ένα σύνολο *S:*

$$G(A,S) = E(S) - \sum_{i=1}^{m} f_s(A_i) * E(S_{Ai})$$

- ***E(...)*** η συνάρτηση εντροπίας
- **m** το πλήθος των τιμών A_i που παίρνει το *A* στο *S*
- $\boldsymbol{f_s(A_i)}$ το ποσοστό των δειγμάτων στο *S* που παίρνουν την τιμή A_i
- $\boldsymbol{S_{Ai}}$ το υποσύνολο του *S* όπου η τιμή του *A* είναι A_i

Ως εντροπία ορίζεται ο βαθμός αβεβαιότητας ενός συνόλου δεδομένων *S.*

$$Entropy\ (S) = \sum_{n=1}^{n} -p_i\ log p_i$$

Οπού $p_1, p_2, \dots, p_i$ οι πιθανότητες των ενδεχομένων που περιλαμβάνονται στο σύνολο. Η εντροπία είναι ένα μέγεθος που παίρνει τιμές από το 0 έως 1. Αν όλα τα δείγματα είναι ομοιογενή ως προς μια κατηγορία, τότε η Εντροπία είναι 0.

Gini Index

Ο συντελεστής Τζίνι υπολογίζει την πιθανότητα ενός τυχαία επιλεγμένου χαρακτηριστικού να έχει κατηγοριοποιηθεί λανθασμένα. Ο συντελεστής Gini είναι ένα μέγεθος που παίρνει τιμές από 0 έως

1. Το 1 συμβολίζει την τυχαία κατανομή των δεδομένων σε σχέση με τις διαφορετικές κλάσης. Η τιμή 0,5 δείχνει την ίση διασπορά των δεδομένων ανάμεσα στις κλάσης.

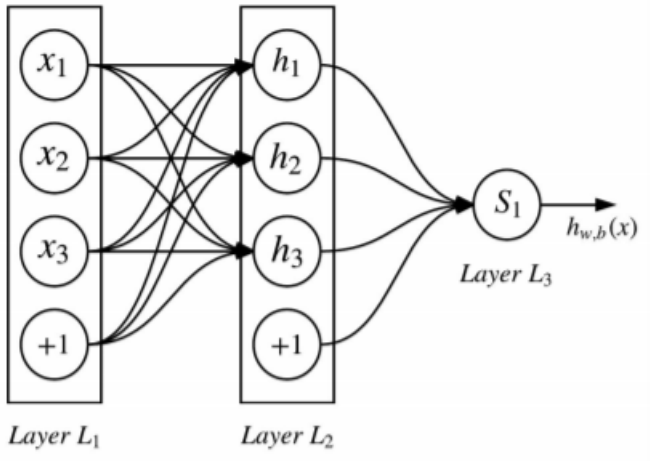
$$Gini\ Index = 1 - \sum_{i=1}^{n} (p_i)^2$$

Τα δέντρα απόφασης αποτελούν καλή τεχνική στην εξόρυξη συναισθήματος με υψηλή ακρίβεια σε μεγάλα σύνολα δεδομένων. Οι ποιο γνωστοί αλγόριθμοι δέντρων απόφασης είναι οι CART, CHAID, ID3 και C5.0 .

4.8 Νευρωνικά Δίκτυα

Τα νευρωνικά δίκτυα (Neural Networks, Connectionist Networks, Parallel Distributed Processing Models) είναι επί της ουσίας μια προσπάθεια συσχετισμού ορισμένων μαθηματικών μοντέλων με ορισμένα βιολογικά μοντέλα που έχουν να κάνουν με τον τρόπο που συνδέονται οι νευρώνες του ανθρώπινου εγκεφάλου. Εμπνευσμένα από την δομή των ανθρώπινων νευρώνων, τα νευρωνικά δίκτυα αποτελούνται από νευρώνες οργανωμένους σε επίπεδα, οι οποίοι ενώνονται μεταξύ τους με συνδέσεις που περιέχουν βάρη. Ανάλογα με την τοπολογία του δικτύου τα νευρωνικά δίκτυα μπορούν να κατηγοριοποιηθούν σε ανατροφοδοτούμενα νευρωνικά δίκτυα (RNN) και σε πρόσθιας τροφοδότησης (feed forward) [28].

Στην περίπτωση των νευρικών δικτύων πρόσθιας τροφοδότησης, οι μονάδες είναι οργανωμένες σε διαφορετικά επίπεδα ώστε οι μονάδες του ενός να τροφοδοτούν τις μονάδες του επόμενου μέχρις ότου τροφοδοτηθούν και οι μονάδες του τελευταίου δικτύου. Αυτό, πρακτικά, σημαίνει ότι δεν υπάρχει έξοδος κάποιας μονάδας που να λειτουργεί ως είσοδος για κάποια άλλη μονάδα.

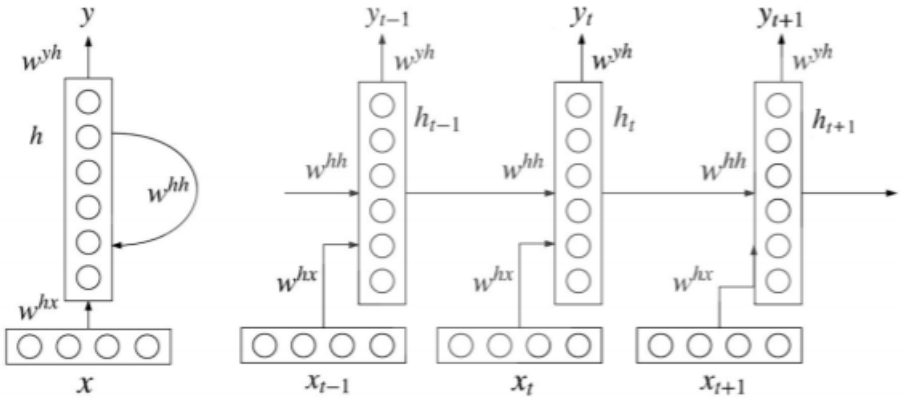


Εικόνα 11:Δίκτυο πρόσθιας τροφοδότησης.

Η παραπάνω εικόνα συνιστά ένα παράδειγμα δικτύου πρόσθιας τροφοδότησης αποτελούμενο από τρία επίπεδα L1,L2,L3. Το επίπεδο L1 είναι το επίπεδο εισόδου με έναν όρο ανακοπής (intercept) +1. L3 είναι το επίπεδο εξόδου το οποίο ανταποκρίνεται στο διάνυσμα S1. L2 είναι το κρυφό επίπεδο. Οι γραμμές ανάμεσα στα επίπεδα αναπαριστούν τις συνδέσεις και την πορεία της πληροφορίας. Κάθε σύνδεση σχετίζεται με ένα βάρος. Τα βάρη καθορίζουν τις συνδέσεις μεταξύ των νευρώνων και

παίρνουν τιμές συνήθων από την διαδικασία της εκπαίδευσης. Στο κρυφό επίπεδο, κάθε νευρώνας δέχεται ως είσοδο x_1, x_2, x_3 και ένα όρο ανακοπής από το προηγούμενο επίπεδο και έχει ως απόκρισή μια τιμή $f(W^t x) = f(\sum_{i=1}^{n} W_i + b)$,όπου *f* η συνάρτηση ενεργοποίησης, W_i τα βάρη των συνδέσεων, *b* ο *intercept* όρος και *n* ο αριθμός των επιπέδων. Οι ποιο γνωστές συναρτήσεις ενεργοποίησης είναι η σιγμοειδής συνάρτηση (*tamh*), συνάρτηση υπερβολής και *RELU* (γραμμικές μονάδες ανόρθωσης).

Αντίθετα με τα ανωτέρω, στα οπισθίως τροφοδοτούμενα δίκτυα, τα οποία καλούνται και ανατροφοδοτούμενα ΤΝΔ, επιτρέπεται στις μονάδες να τροφοδοτούν άλλες μονάδες του επιπέδου ή και προηγούμενων επιπέδων(κίνηση προς τα πίσω). Στην περίπτωση που η ανατροφοδότηση αφορά κόμβους του ίδιου επιπέδου, τότε τα δίκτυα ονομάζονται αυτοσυσχετιζόμενες μνήμες(autoassociated memories, διαφορετικά καλούνται ετεροσυσχετιζόμενες μνήμες (heteroassocieated memories) [29].



Εικόνα 12:Δίκτυο με ανατροφοδότηση.

Σε αντίθεση με τα δίκτυα πρόσθιας τροφοδότησης, τα δίκτυα RNN χρησιμοποιούν την «εσωτερική» μνήμη τους για την επεξεργασία συνεχόμενων ροών δεδομένων. Πιο συγκεκριμένα τα νευρωνικά δίκτυα αυτού του τύπου "θυμούνται" τους υπολογισμούς και την πληροφορία που έχουν επεξεργαστεί και έτσι κάθε καινούρια ροή πληροφορίας συνδέεται κατά κάποιον τρόπο με τους προηγούμενους υπολογισμούς.

Η παραπάνω εικόνα χωρίζεται σε δύο μέρη. Στο αριστερό μέρος είναι ένα μη εμφωλευμένο κυκλικό δίκτυο, ενώ στην δεξιά ένα εμφωλευμένο δίκτυο τριών (Χρονικών) σταδίων-βημάτων. Το μέγεθος των βημάτων καθορίζεται από το μέγεθος της εισόδου. Στην εικόνα χ_t είναι το διάνυσμα εισόδου την χρονική στιγμή *t*. h_t είναι ένα κρυφό στάδιο την χρονική στιγμή *t* το οποίο υπολογίζεται με βάση την τους υπολογισμούς τις προηγούμενης κατάστασης h_{t-1}. Η κατάσταση αυτή χαρακτηρίζεται από την εξίσωση,

$$h_t = f(w^{hh} h_{t-1} + w^{hx} x_t)$$

Οπού *f* η συνάρτηση ενεργοποίησης. Ως w^{hh} ορίζουμε τον πίνακα με τα βάρη της προηγούμενης κατάστασης h_{t-1}. Ως y_t ορίζεται η απόκριση και περιλαμβάνει ένα σύνολο πιθανοτήτων σχετικά με την πρόβλεψη. Για παράδειγμα, εάν το πρόβλημα ήτανε η πρόβλεψη της επόμενης λέξης μιας πρότασης, τότε η απόκριση θα ήτανε ένα διάνυσμα με μία πιθανότητα για κάθε λέξη του λεξικού.

4.8.1 Βασικές αρχιτεκτονικές Νευρωνικών

Η πολυπλοκότητα ενός νευρωνικού δικτύου εξαρτάται από το πλήθος των νευρώνων. Πιο περίπλοκα νευρωνικά δίκτυα δημιουργούνται από πολλούς νευρώνες οι οποίοι συνδέονται μεταξύ τους με συγκεκριμένη δομή. Η δομή αυτή αποτελεί την αρχιτεκτονική του δικτύου [30].

Οι βασικές κατηγορίες είναι:

1. Δίκτυα απλής τροφοδότησης ενός επιπέδου.
   Δίκτυο με μία απλή επίπεδη είσοδο.
2. Πολυεπίπεδο δίκτυο πρόσθιας τροφοδότησης
   Δίκτυα απλής τροφοδότησης που περιλαμβάνει ένα η περισσότερα κρυμμένα επίπεδα και νευρώνες.
3. Δίκτυα με ανατροφοδότηση.
   Δίκτυα στα οποία ένα επίπεδο μπορεί να συνδεθεί ως είσοδος σε προηγούμενο επίπεδο με την ύπαρξη βρόχων ανάδρασης [31].

4.9 Recurrent Neural Networks (RNNs)

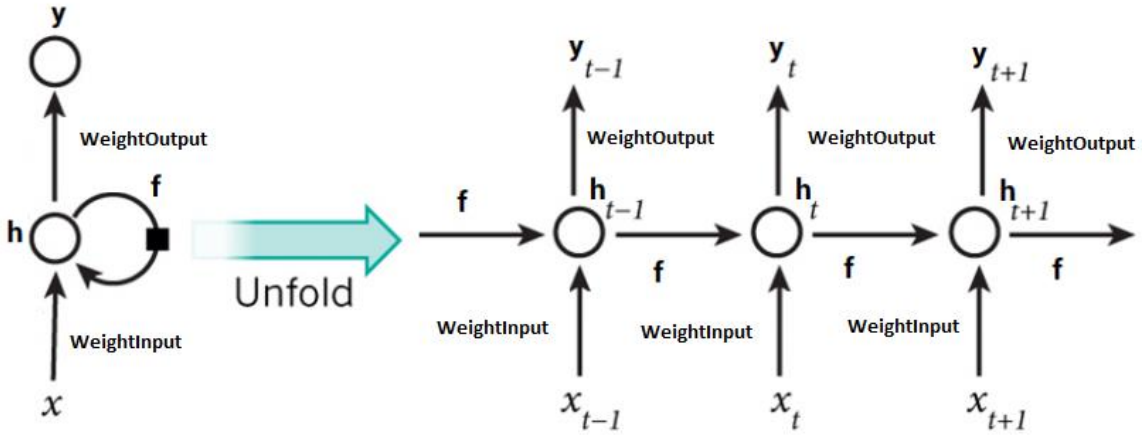
Τα RNNs νευρωνικά δίκτυα αποτελούνται από κόμβους οι οποίοι αλληλοεπιδρούν σε διακριτές χρονικές στιγμές, μεταξύ συνδέσεων με βάρη w_{lm} (από τον κόμβο *l* στον κόμβο *m*). Κάθε κόμβος έχει μία συνάρτηση ενεργοποίησης *y(t)* όπου *t* κάθε διακριτή χρονική στιγμή. Με την ενεργοποίηση κάθε κόμβου, τροφοδοτούνται οι επόμενοι κόμβοι συνθέτοντας έτσι την δομή του δικτύου. Μια συνάρτηση ενεργοποίησης ενός κόμβου *l* υπολογίζεται με βάση από την είσοδο του νευρωνικού.

Συναρτήσεις όπως η Λογιστική συνάρτηση, η σιγμοειδείς συνάρτηση, η εφαπτομένη και η τετραγωνική συνάρτηση αναφέρονται στην βιβλιογραφία ως συναρτήσεις ενεργοποίησης.

Η συνδεσμολογία μεταξύ των κόμβων του νευρωνικού ονομάζεται αρχιτεκτονική ή τοπολογία. Τα RNN δίκτυα μερικώς επαναληπτικά, ή επαναληπτικά νευρωνικά δίκτυα [32].

Η Εκπαίδευση ενός RNN δικτύου με την μέθοδο *gradient descent* (κάθοδος βασισμένη στην κλίση), για παράδειγμα ένα LSTM δίκτυο, μπορεί να αναχθεί σε πρόβλημα ελαχιστοποίησης μιας συνάρτησης σφάλματος E(w) ως προς το διάνυσμα των βαρών με βήμα καθόδου (ρυθμός μάθησης) *a.*

Στην παρακάτω εικόνα παρατηρούμε την πορεία ενός απλού RNN δικτύου ξεδιπλωμένου, ανά χρονική στιγμή [33].



Εικόνα 13: RNN δίκτυο.

4.9.1 Αδυναμίες δικτύων RNN

Τα δίκτυα με ανατροφοδότηση έχουν την δυνατότητα να συνδυάζουν την προγενέστερη πληροφορία για την επίλυση ενός προβλήματος. Η πληροφορία αυτή μπορεί να αποτελεί μια χρήσιμη παράμετρο σε κάποια προβλήματα, αλλά και αντίστροφα.

Για παράδειγμα, σε ένα πρόβλημα μηχανικής μάθησης για την «εύρεση» της επόμενης λέξης (σύννεφα) στην πρόταση «…ηλιόλουστη ημέρα καθώς σήμερα ο ουρανός δεν έχει σύννεφα.» ενός κειμένου, δεν χρειάζεται επιμέρους πληροφορία από προηγούμενες προτάσεις του κειμένου ώστε να αποφασίσουμε ότι η επόμενη λέξη είναι η λέξη «σύννεφα». Σε ένα άλλο παρόμοιο πρόβλημα όμως, για την εύρεση της επόμενης λέξης (Ελληνικά) στην πρόταση «… γιαυτό και η μητρική μου γλώσσα είναι τα Ελληνικά.», η πληροφορία του προηγούμενου κειμένου είναι ουσιαστική για την επίλυση.

Αυτές οι εξαρτήσεις του νευρωνικού δικτύου με τις εισόδους-πληροφορίες του παρελθόντος μελετήθηκαν σε βάθος από τους Hochreiter and Bengio, οι οποίοι κατέληξαν σε κάποια χρήσιμα συμπεράσματα για τα αίτια των αδυναμιών αυτών του νευρωνικού [34].

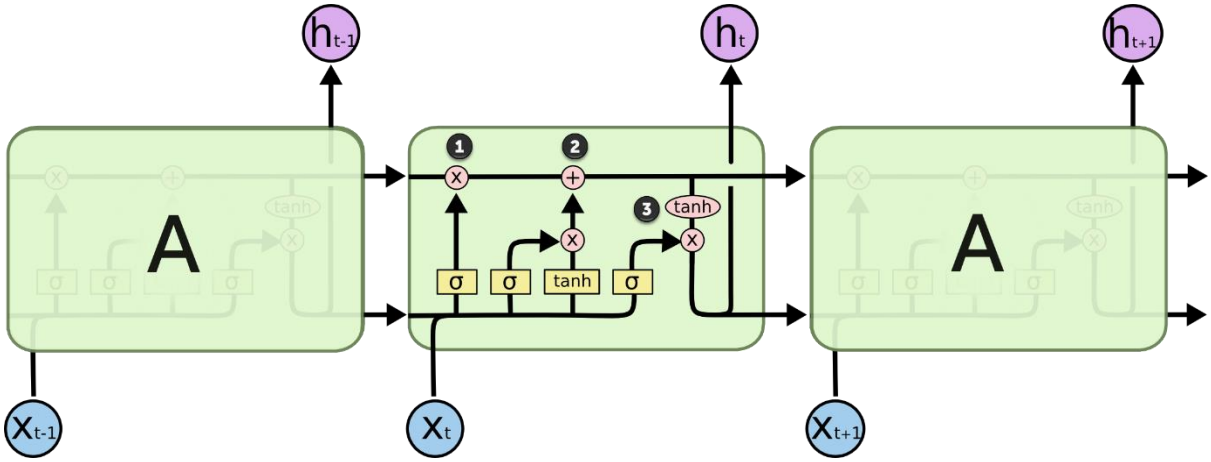
4.10 LSTM

Πέραν των αποδεδειγμένων αποτελεσματικών εφαρμογών των παραπάνω ανατροφοδοτούμενων δικτύων (recurrent networks), παρουσιάζουν και κάποιες αδυναμίες. Οι αδυναμίες αυτές εμφανίζονται στην περίπτωση όπου το χρονικό διάστημα μεταξύ εισόδου και εξόδου είναι υπολογίσιμο.

Τα δίκτυα LSTM (Long Short-Term Memory) έρχονται για να λύσουν αυτό το πρόβλημα των χρονικών κενών. Τα LSTM δίκτυα παρουσιάστηκαν από τον Schmidhuber το 1997 και ακολουθούν την αρχιτεκτονική των RNN δικτύων. Η αρχιτεκτονική τους περιλαμβάνει πύλες για τον έλεγχο της πληροφορίας μεταξύ των κελιών.

Κύριο ρόλο στην λειτουργεία του νευρωνικού παίζει το κελί της μνήμης. Κάθε επαναλαμβανόμενη μονάδα της αλυσίδας του νευρωνικού δικτύου αποτελείται από πύλες και κελιά. Η κατάσταση του κελιού της μνήμης μεταβάλετε από τις πύλες οι οποίες προσθέτουν ή απομακρύνουν πληροφορία από το κελί. Ποιο συγκεκριμένα, οι πύλες που συνθέτουν την δομή είναι η πύλη επιλεκτικής συγκράτησης (forget gate), η πύλη εισόδου και η πύλη εξόδου. Η κάθε πύλη εξαρτάτε από μία συνάρτηση ενεργοποίησης.

Στην παρακάτω εικόνα παρατηρούμε την αλυσίδα ενός απλού LSTM νευρωνικού. Οι πύλες 1-(forget gate),2-Input Gate και 3-Output Gate που παρουσιάζονται με συνδυασμούς της σιγμοειδείς και μίας εφαπτομένης συνάρτησης [35].



Εικόνα 14: Πύλες ενός LSTM δικτύου (Πηγή [34]).

Η παραπάνω αλυσίδα περιγράφει την αρχιτεκτονική ενός απλού LSTM νευρωνικού δικτύου.

5.Εισαγωγή στα Κρυπτονομίσματα

Η έννοια των κρυπτονομισμάτων (cryptocurrencies) είναι σχετικά πρόσφατη καθώς το Bitcoin εμφανίστηκε για πρώτη φορά το 2009 ως το πρώτο γνωστό κρυπτονόμισμα. Ως κρυπτονόμισμα ορίζεται ένα αποκεντρωμένο σύστημα χωρίς κεντρική εξουσία, το οποίο χρησιμοποιεί κρυπτογραφία για τον έλεγχο των συναλλαγών, την αύξηση της προσφοράς και την πρόληψη της απάτης. Οι επιβεβαιωμένες συναλλαγές αποθηκεύονται ψηφιακά και καταγράφονται στο λογιστικό σύστημα των κρυπτονομισμάτων, το οποίο είναι γνωστό ως Blockchain. Όπως είναι φυσικό, για την επεξεργασία των κρυπτονομισμάτων και τη διενέργεια των συναλλαγών είναι απαραίτητη η χρήση υπολογιστών με ισχυρές προδιαγραφές και μνήμη [42]. Το πιο δημοφιλές κρυπτονόμισμα είναι το Bitcoin, το οποίο εμφανίστηκε το 2009 και δημιουργήθηκε από έναν επιστήμονα ηλεκτρονικών υπολογιστών με το ψευδώνυμο "Satoshi Nakamoto". Σύμφωνα με τον Nakamoto (2008), το κόστος της διαμεσολάβησης στην πραγματική αγορά νομισμάτων αυξάνει το κόστος συναλλαγής, περιορίζοντας το ελάχιστο μέγεθος της συναλλαγής και μειώνοντας τη δυνατότητα για απλές, καθημερινές συναλλαγές. Παράλληλα, ένα ευρύτερο κόστος αναφορικά με την απώλεια της

ικανότητας να καταβάλλονται μη αναστρέψιμες πληρωμές για μη αναστρέψιμες υπηρεσίες. Ως αποτέλεσμα αυτών των παραγόντων, οι έμποροι ζητούν όλο και περισσότερες πληροφορίες για να ελέγξουν τη φερεγγυότητα των πελατών τους, ενώ ένα ορισμένο ποσοστό απάτης γίνεται αποδεκτό ως αναπόφευκτο. Αντιθέτως, η εισαγωγή ενός ηλεκτρονικού συστήματος πληρωμών βασισμένο στην κρυπτογραφική απόδειξη επιτρέπει σε δύο μέρη να πραγματοποιούν συναλλαγές απευθείας μεταξύ τους χωρίς την ανάγκη διαμεσολάβησης τρίτου. Επιπλέον, συναλλαγές που δεν είναι πρακτικά εφαρμόσιμες για αντιστροφή προστατεύουν τους πωλητές από την απάτη και οι συνήθεις μηχανισμοί μεσεγγύησης εφαρμόζονται για την προστασία των αγοραστών. Το δίκτυο Bitcoin είναι ένα δίκτυο peer-to-peer που ελέγχει και παρακολουθεί τη δημιουργία νέων Bitcoin (mining - εξόρυξη) και τις συναλλαγές με κρυπτονομίσματα. Το δίκτυο περιλαμβάνει ένα μεγάλο αριθμό υπολογιστών που συνδέονται μεταξύ τους μέσω του διαδικτύου και εκτελεί πολύπλοκες μαθηματικές διαδικασίες με στόχο την εξόρυξη νέων κρυπτονομισμάτων αλλά και την επαλήθευση της ορθότητας των συναλλαγών Bitcoin.

5.1 Πώς αναπτύσσεται ένα κρυπτονόμισμα

Τα στάδια ανάπτυξης ενός κρυπτονομίσματος περιλαμβάνουν την κωδικοποίηση, την εύρεση χρηστών και την αύξηση της δημοτικότητας.

1) Κωδικοποίηση

Για τους χρήστες που διαθέτουν μία στοιχειώδη κατανόηση της κωδικοποίησης, είναι εξαιρετικά εύκολο να δημιουργήσουν τη δική τους μορφή κρυπτογράφησης. Σε πιο προχωρημένο επίπεδο, οι προγραμματιστές μπορούν να κάνουν τροποποιήσεις στον κώδικα που αλλάζει τη λειτουργία του νομίσματος όπως επιθυμούν. Ωστόσο, ακόμα και για τους χρήστες που δεν γνωρίζουν ή δεν κατανοούν σε βάθος την κωδικοποίηση, υπάρχουν ορισμένες διαθέσιμες υπηρεσίες που επιτρέπουν τη δημιουργία κρυπτονομισμάτων έναντι αμοιβής.

2) Εύρεση χρηστών

Η ανάπτυξη ενός κρυπτονομίσματος είναι εύκολη, αλλά η εξεύρεση δικτύου χρηστών είναι μία σαφώς δυσκολότερη διαδικασία. Το κρυπτονόμισμα βασίζεται σε μια αποκεντρωμένη βάση δεδομένων, η οποία με τη σειρά της βασίζεται σε έλεγχο και εγγύηση της ακεραιότητας των δεδομένων. Επομένως, όχι μόνο απαιτείται ένα μεγάλο δίκτυο χρηστών, αλλά θα πρέπει το δίκτυο αυτό να είναι αξιόπιστο και διαθέτει τα κατάλληλα μέσα επεξεργασίας για να διασφαλιστεί ότι όλες οι συναλλαγές γίνονται σωστά. Οι περισσότεροι νέοι τύποι κρυπτονομισμάτων δημιουργούνται στη βάση της λύσης ενός προβλήματος που σχετίζεται με τα παραδοσιακά νομίσματα. Ωστόσο, εάν υπάρχει ανεπαρκής ικανότητα επεξεργασίας για να διευκολυνθεί το σύνολο των συναλλαγών του δικτύου, η κρυπτογράφηση μπορεί να μην γίνει ποτέ επιτυχημένη. Στην ουσία, προκειμένου ένας τύπος νομίσματος να κερδίσει την ευρύτερη προσοχή των χρηστών του οικοσυστήματος Blockchain, πρέπει οι μεταφορές νομισμάτων να γίνονται γρήγορα και με ασφαλή τρόπο. Διαφορετικά, οι χρήστες μπορούν απλά να βρουν έναν άλλο τύπο ψηφιακού νομίσματος προς χρήση.

3) Αύξηση δημοτικότητας

Το τελευταίο βήμα για την ανάπτυξη και την έναρξη κυκλοφορίας ενός κρυπτονομίσματος είναι να κερδίσει δημοτικότητα. Η προσοχή των μέσων μαζικής ενημέρωσης (ΜΜΕ) μπορεί να αποβεί εξαιρετικά χρήσιμη, αλλά και εξαιρετικά μοιραία, εάν το κρυπτονόμισμα δεν θεωρείται γενικότερα φερέγγυο. Όμως, στην περίπτωση που το κρυπτονόμισμα είναι ασφαλές και δεν είναι επιρρεπές σε

παραμορφώσεις ή παραβιάσεις δεδομένων, η προσοχή των ΜΜΕ μπορεί να αποδειχθεί πολύ χρήσιμη. Οι προγραμματιστές κρυπτονομισμάτων θέλουν τα νομίσματά τους να φτάσουν σε όσο το δυνατόν περισσότερους ανθρώπους διότι με τον τρόπο αυτόν αυξάνει και η αξία τους.

5.2 Η τεχνολογία Blockchain και η ανάπτυξη της FinTech

Στην πιο απλή του μορφή, το Blockchain είναι μια κατανεμημένη και αποκεντρωμένη βάση δεδομένων στην οποία κάθε μέλος του δικτύου διατηρεί ένα πλήρες, επαληθευμένο και συγχρονισμένο αντίγραφο όλων των συναλλαγών. Η αρχιτεκτονική του συνδυάζει την προηγμένη κρυπτογραφία, ένα πολύπλοκο σύστημα κινήτρου-ανταμοιβής και ένα μοντέλο που εξασφαλίζει την ακεραιότητα των δεδομένων υπό την απουσία μεσολάβησης ενός τρίτου μέρους, όπως, για παράδειγμα, μία τράπεζα ή ένα χρηματοπιστωτικό ίδρυμα. Ο αγοραστής και ο πωλητής αλληλοεπιδρούν άμεσα χωρίς να απαιτείται επαλήθευση από αξιόπιστο τρίτο μεσάζοντα. Οι συναλλαγές δεν είναι ανώνυμες, αλλά γίνονται με ψευδώνυμα: δημιουργείται μια εγγραφή συναλλαγής, αλλά οι πληροφορίες αναγνώρισης κρυπτογραφούνται και δεν μοιράζονται προσωπικές πληροφορίες. Συνεπώς, το Blockchain είναι ένας αποκεντρωμένος ημερολογιακός κατάλογος όλων των συναλλαγών σε ένα δίκτυο χρήστη προς χρήστη (peer-to-peer) και αποτελεί την τεχνολογία στην οποία βασίζεται το Bitcoin και άλλα κρυπτονομίσματα. Το οικοσύστημα Bitcoin περιλαμβάνει τέσσερα βασικά μέρη: (1) τους χρήστες που στέλνουν και λαμβάνουν πληρωμές, του εξορύκτες (miners) που παράγουν την κρυπτογράφηση, τους επενδυτές που αγοράζουν τα κρυπτονομίσματα, και τους προγραμματιστές που παρακολουθούν και συντηρούν τις διαδικασίες. Κανένα τμήμα της εξίσωσης δεν λειτουργεί χωρίς να υπάρχουν και τα υπόλοιπα μέρη. Ειδικότερα, οι miners διαδραματίζουν πολύ κρίσιμο ρόλο στη διατήρηση του οικοσυστήματος Blockchain καθώς είναι εκείνοι που ουσιαστικά διατηρούν το σύστημα εξόρυξης και προωθούν νέα κρυπτονομίσματα στην αγορά. Οι miners στηρίζονται στην ύπαρξη μίας συνεχούς ροής συναλλαγών που μεταδίδονται στο δίκτυο με στόχο να δώσουν στους καταναλωτές/ επενδυτές τη δυνατότητα να συμμετέχουν στο οικοσύστημα. Δεδομένου ότι τα εμπόδια εισόδου και εξόδου από τη βιομηχανία κρυπτονομισμάτων είναι πολύ χαμηλά, είναι σαφές ότι οι miners συνεχίζουν την εξόρυξη μόνο εάν υπάρχει οικονομικό όφελος [41].

Τα προϊόντα και οι υπηρεσίες ενός οικοσυστήματος Blockchain περιλαμβάνουν ένα πορτοφόλι πολλών νομισμάτων, την ικανότητα άντλησης κεφαλαίων για την ανάπτυξη έργων και μία ολοκληρωμένης αποκεντρωμένης ανταλλαγής. Κανένα από τα τρία αυτά στοιχεία δεν απαιτεί την χρήση τεχνογνωσίας της τεχνολογία Blockchain. Για τον λόγο αυτόν, είναι σχετικά εύκολο μία οποιαδήποτε επιχείρηση να συμμετέχει στο οικοσύστημα και να συμβάλλει στο άνοιγμα της οικονομίας για οποιονδήποτε οργανισμό οποιουδήποτε μεγέθους, σε οποιοδήποτε τομέα.

6. Σκοπός Εργασίας

Σκοπός της παρούσας εργασίας η ανάλυση της αγοράς του ψηφιακού νομίσματος καθώς και οι παράγοντες που την επηρεάζουν, η εξαγωγή χαρακτηριστικών από τα μέσα κοινωνικής δικτύωσης καθώς και ο βαθμός στον οποίο αυτά επηρεάζουν την τιμή του κρυπτονομίσματος Bitcoin (σε σχέση με την αξία του σε ευρώ). Τα μέσα κοινωνικής δικτύωσης που χρησιμοποιήθηκαν για την συλλογή πληροφοριών είναι το Tweeter καθώς και οι ιστοσελίδες BBC NEWS, REDDIT και MoneyControl. Η συλλογή των δεδομένων από το Tweeter έγινε σε πραγματικό χρόνο και η αποθήκευση των δεδομένων πραγματοποιήθηκε με την βοήθεια του SQL Server management studio. Η εξαγωγή της πληροφορίας (web-scraping) από τις Ιστοσελίδες BBC NEWS και MoneyControl πραγματοποιούνταν σε περίοδο έξι ωρών κάθε ημέρα και η αποθήκευση των δεδομένων πραγματοποιήθηκε με την βοήθεια του SQL Server management studio. Παράλληλα, πραγματοποιήθηκε έρευνα σε επίπεδο αρχιτεκτονικής και υπερπαραμέτρων και αναπτύχθηκε αλγόριθμος LSTM για την πρόβλεψη της τιμής, καθώς και της τάξης της τιμής του κρυπτονομίσματος Bitcoin.

6.1 Συλλογή Δεδομένων

Για την συλλογή των δεδομένων χρησιμοποιήθηκαν:

- Χρήση του tweeter API (tweepy) για την συλλογή δεδομένων.

Συγκεκριμένα, με την βοήθεια της γλώσσας προγραμματισμού Python και του API του tweeter, κατέβηκαν δεδομένα από 50 tweeter προφίλ τα οποία σχετίζονται άμεσα ή έμμεσα με την αγορά χρηματιστηριακή αγορά και την αγορά κρυπτονομισμάτων την χρονική περίοδο μεταξύ ’01-01-2020’ έως 08 -08-2021’. Ενδεικτικά κάποια από τα προφίλ αυτά είναι τα [“Ben Armstrong”, “Bitcoin News”, “Bloomberg Quicktake”, “Cardano Community”, “Coinbase Pro”, “Crypto Trading”,” Blockchain.com”, “cz_binance”]. Στην χρονική περίοδο μεταξύ ’08-08-2021’ έως ’15-08-2021’ πραγματοποιήθηκε συλλογή δεδομένων –με αυτοματοποιημένο τρόπο- βασισμένο σε tweets τα οποία περιέχουν hashtags λέξεις κλειδιά όπως #BTC, #CRYPTO, #COINS, καθώς και στα προαναφερθέντα προφίλ.

- Χρήση του API της ιστοσελίδας Binance για την συλλογή δεδομένων σχετικά με το κρυπτονόμισμα Bitcoin.

Με την χρήση του API της ιστοσελίδας Binance, έγινε εξαγωγή δεδομένων τιμών για της συναλλαγματικές αξίας των ζευγών BTC/EUR.

- Χρήση βιβλιοθηκών της Python για Web Scrapping.

Με της βιβλιοθήκης BeutifullSoup έγινε η εξαγωγή των δεδομένων τίτλων και κειμένου από τις ιστοσελίδες BBC NEWS (https://www.bbc.com/news/topics/c734j90em14t/bitcoin), REDDIT (https://www.reddit.com/) και (https://www.moneycontrol.com/cryptocurrency/).

6.2 Προβλήματα και Αντιμετώπιση.

Ένα από τα προβλήματα που αντιμετωπίστηκαν είναι ότι από ένα μέρος των δεδομένων που προήλθαν από τις ιστοσελίδες απουσίαζε ο παράγοντας ώρα. Τα δεδομένα αυτά επιλέχθηκαν να τοποθετηθούν αναλογικά σε σχέση με τις χρονικές στιγμές των υπολοίπων παρατηρήσεων. Δηλαδή το 26% των παρατηρήσεων καταχωρήθηκε στο χρονικό διάστημα 00:00:00-06:00:01, το 20% στο διάστημα 06:00:01-12:00:01, το 28% στο διάστημα 12:00:01-18:00:01 και το 26% στο διάστημα 18:00:01-23:59:59.

Ένα ακόμη προβλημα που αντιμετωπίστηκε σχετικά με το σύνολο των δεδομένων ήτανε η χρήση μίας κοινής ζώνης ώρας (timezone). Τα δεδομένα των τιμών που προέκυπταν από το API τις ιστοσελίδας *REDDIT* από οπου προήλθαν οι ιστορικές τιμές του κρυπτονομίσματος ήτανε στην «Συντονισμένη Παγκόσμια Ζώνη Ώρας» UTC (Coordinated Universal Time). Τα δεδομένα του προήλθαν από το twitter ητανε και αυτά σε UTC ζώνη ώρας. Σε άλλες περιπτώσεις όμως, όπως στα δεδομένα τις ιστοσελίδας *moneycontrol.com* τα δεδομένα αφορούσαν IST (*Indian Stadard Timezone*) δηλαδή *UTC+5:30.* Για τους σκοπούς της εργασίας όλα τα δεδομένα εκπαίδεσης τοποθετήθηκαν σε ζώνη ώρας *UTC.* Στο κομμάτι της εφαρμογής των αποτελεσμάτων της έρευνας δεν αντιμετωπίστηκε τετοιο πρόβλημα μιάς και τα δεδομένα ήταν σε ζωντανό χρόνο.

Επιπλέον, τα δεδομένα μας καθαρίστηκαν από διπλότυπες τιμές.

6.3 Ανάλυση Tweets

Για την ανάλυση και την εξαγωγή συμπεράσματος σχετικά με συναισθήματα που προκαλούν τα tweets τις πλατφόρμας Tweeter χρησιμοποιήθηκαν βιβλιοθήκης και τεχνικές τις γλώσσας προγραμματισμού Python. Ποιο συγκεκριμένα οι βιβλιοθήκη που χρησιμοποιήθηκε είναι η *nltk*.

Το σύνολο των δεδομένων σχηματίστηκε με την συλλογή δεδομένων -tweets- από συγκεκριμένα προφίλ, τα οποία είναι διαπιστευμένα από την πλατφόρμα για το είδος και την εγκυρότητα της πληροφορίας που αναπαράγουν, είτε έχουν ένα σημαντικό πλήθος από followers. Όλα τα προφίλ προέρχονται από τον χώρο της ενημέρωσης και της αγοράς κρυπτονομισμάτων.

Για την διαχείριση των δεδομένων ακολουθήθηκαν τα παρακάτω βήματα:

- ✓ Αναγνώριση των δεδομένων
- ✓ Εξαγωγή συμπερασμάτων και γραφήματα
- ✓ Καθαρισμός δεδομένων
- ✓ Διαίρεση σε σύμβολα
- ✓ Κανονικοποίηση
- ✓ Προσαρμογή και μοντελοποίηση

Το σύνολο των δεδομένων μας είναι 160811 εγγραφές από τις οποίες οι 84880 αποτελούν retweets (53%) άρα είναι και duplicate records. Το χρονικό φάσμα των δεδομένων είναι από ‘2020-08-01’ έως ‘2021-08-08’. Το σύνολο των δεδομένων περιλαμβάνει τη χαρακτηριστικά:

Table 1: TWEEPY API feature extraction.

Feature	Description	Data Type
followers	*Αριθμός ακόλουθων*	*integer*
text	*Κείμενο*	*String*
favourite_count	*Αριθμό like*	*integer*
retweet_count	*Αριθμός retweet*	*integer*
created_at	*Ημερομηνία και ώρα κειμένου*	*Datetime*

Ο καθαρισμός των δεδομένων αφορά την αφαίρεση των *stopwords,* την αφαίρεση συμβόλων και ιστοσελίδων. Σε προηγούμενο στάδιο πραγματοποιήθηκε ο καθαρισμός των δεδομένων. Ποιο συγκεκριμένα, αφαιρέθηκαν οι διπλότυπες εγγραφές, σηματοδοτήθηκαν τα “retweets” και κατηγοριοποιήθηκαν τα δεδομένα χρονικά σε περιόδους των 6 ωρών.

Ο χειρισμός των *emoji* είχε ως εξής. Διαχωρίσαμε το κείμενο από τα *emoji* και εφαρμόσαμε ξεχωριστή ανάλυση και εξαγωγή συμπεράσματος.

Για να έχουμε μία καλύτερη οπτική των δεδομένων μας δημιουργήθηκαν τα παρακάτω διαγράμματα.

Εικόνα 15: Λέξεις με μέγεθος ανάλογο της συχνότητας εμφάνισης τους στο σύνολο δεδομένων.

Για την εύρεση της καλύτερης μεθόδου εξαγωγής του συναισθήματος επιλέχθηκαν δύο μέθοδοί. Αρχικά έγινε χρήση μεθόδων βασισμένων σε κανόνες- με άλλα λόγια χρησιμοποιήθηκαν τα λεξικά **VADER** και **SentiWordNet** για την εξαγωγή συμπεράσματος και στην συνέχεια οι τεχνικές αυτές εμπλουτίστηκαν με επιπλέον κανόνες. Κανόνες οι οποίοι αφορούν την αύξηση ή μείωση της τιμής αξιολόγησης μιας λέξης σε αρνητικό, θετικό και ανάμεικτο συναίσθημα. Σαν δεύτερη προσέγγιση έγινε χρήση αλγορίθμων μηχανικής μάθησης. Συγκεκριμένα, έγινε εκπαίδευσή του μοντέλου SVM σε σύνολο δεδομένων από την ιστοσελίδα Kaggle (https://www.kaggle.com/gauravduttakiit/twitter-sentiment-analysis-nlp/data).

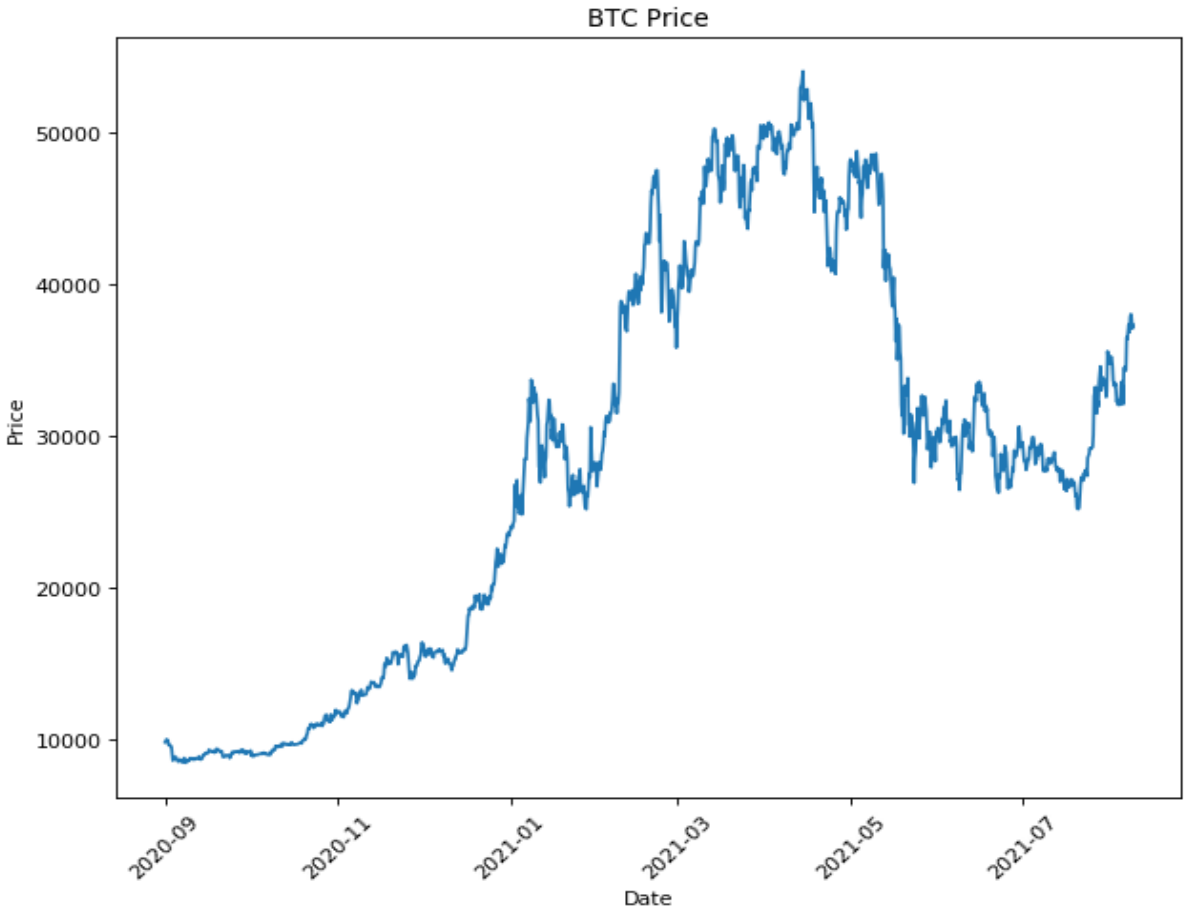
6.4 Rule Based Techniques

Η πρώτη τεχνική αφορά τον συνδυασμό του λεξικού VADER και κάποιον κανόνων που ορίστηκαν τα οποία αφορούν την τιμωρία ή επιβράβευσή του τελικού αποτελέσματος πρόβλεψης του συναισθήματος. Οι λίστες οι οποίες δημιουργήθηκαν περιλαμβάνουν λέξεις οι οποίες παρατηρήθηκαν αρκετά στα δεδομένα (tweets) και για το λεξικό VADER αξιολογούνται με διαφορετικό τρόπο από ότι θα είχαν αξιολογηθεί στον χώρο του χρηματιστηρίου. Παραδείγματος χάρη το emoji μεταφράστηκε με την βοήθεια της βιβλιοθήκης της python *emoji* ως *[‘rocket’].* Η λέξη rocket από μόνη της έχει *polarity*. Στον κόσμο του χρηματιστήριου όμως και ειδικά των κρυπτονομισμάτων αυτό το emoji συμβολίζει την ανοδική πορεία της τιμής ενός κρυπτονομίσματος. Στο παρακάτω πίνακα παρουσιάζεται ένα δείγμα λέξεων– σε κάποιες περιπτώσεις οι λέξεις αποτελούν μετάφραση ενός *emoji*– και το βάρος που θέσαμε καθώς και πως αξιολογούνται αυτές οι λέξεις από τα λεξικά που χρησιμοποιήθηκαν.

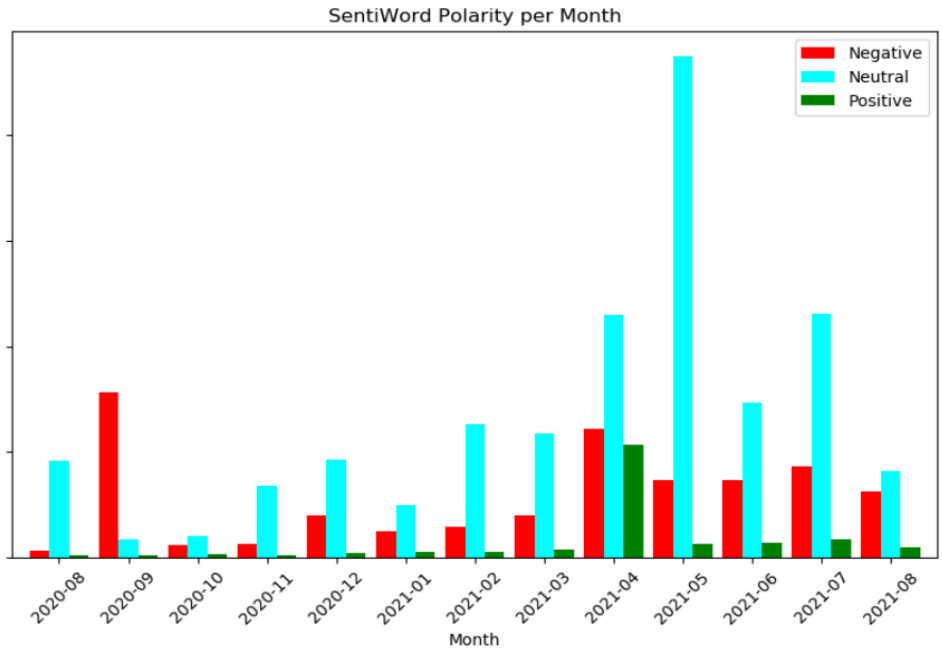
Table 2: Δημιουργία λίστας για την ενίσχυση του αλγορίθμου.

Λέξη – Φράση που Παρατηρήθηκε	**VADER Polarity**	**SentiWordNet Polarity**		**SVM Polarity**	**Ενίσχυση**
		Positive (+) – Negative (-)			
HOLD	0	0	0	0	+0.25
MOON	0	0	0	0	+0.25
BUY	0	0	0	1	+0.25
ROCKET	0	0	0	0	+0.25
KEEP	0	0.625	0	1	+0.25
SURGE	0	0	0.25	1	+0.25
JUMP	0	0	0	1	+0.25
ATH	0	-	-	0	+0.25
SELL	0	0	0	0	-0.25
DOWN	0	0	0.125	0	-0.25
DROP	-0.27	0.125	0	0	-0.25
CRASH	-0.4	0	0	0	-0.25

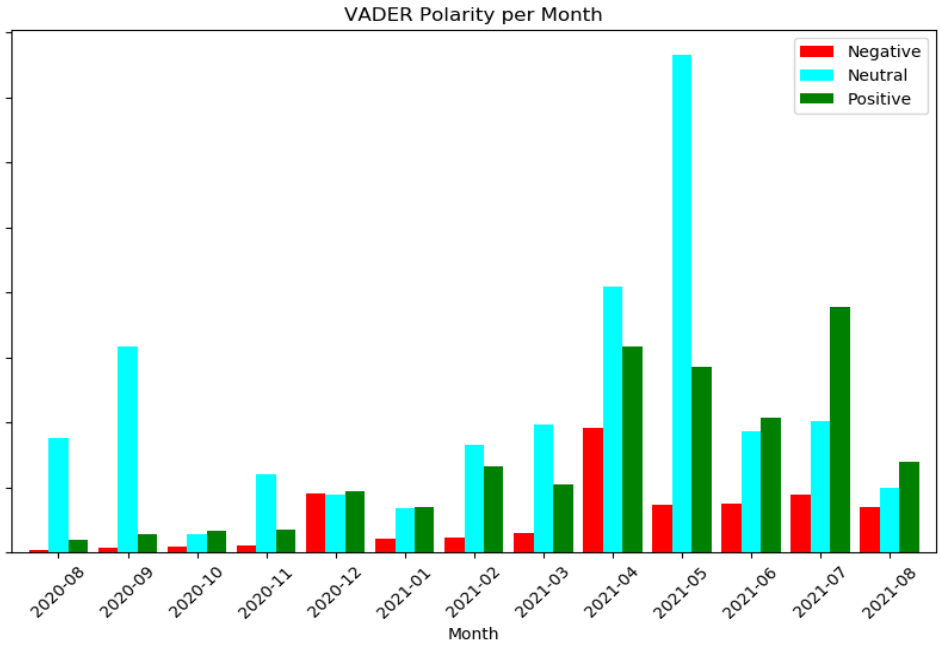
Τα αποτελέσματα της εφαρμογής των δύο rule-based μεθόδων παρουσιάζονται στα παρακάτω γραφήματα.



Εικόνα 16: Χρονοσειρά κρυπτονομίσματος Bitcoin.



Εικόνα 17: Αποτελέσματα Ανάλυσης με τεχνική SentiWord ανά μήνα.



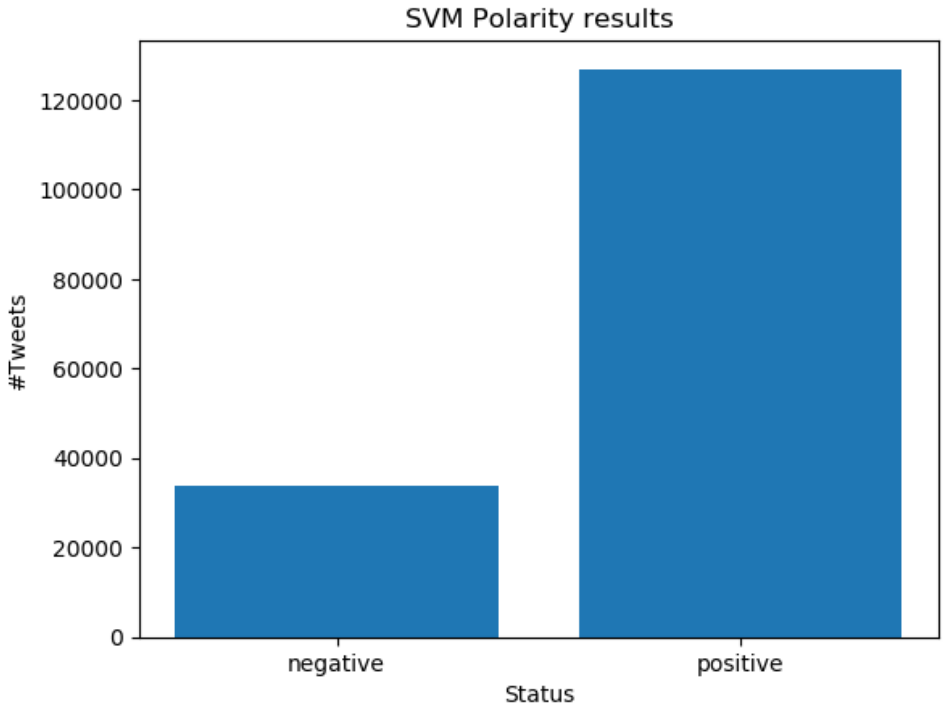
Εικόνα 18: Αποτελέσματα Ανάλυσης με τεχνική VADER ανά μήνα.

6.5 Machine Learning Approach

Η εκπαίδευση του SVM υλοποιήθηκε με την βοήθεια της βιβλιοθήκης *sklearn* της python. Αρχικά έγινε εκπαίδευση αλγορίθμου SVM με πυρήνα Kernel σε σύνολο δεδομένων το οποίο από την ιστοσελίδα Kaggle. Το σύνολο των δεδομένων εκπαίδευσης περιλαμβάνει περισσότερα από ένα εκατομμύριο κατηγοριοποιημένα tweets. Μετά της εκπαίδευση του αλγορίθμου έγινε εφαρμογή στο σύνολο των δεδομένων το οποίο συλλέχθηκε για τους σκοπούς της εργασίας.

Η κύρια διαφορά της τεχνικής αυτής είναι ότι το αποτέλεσμα του αλγορίθμου αφορά κατηγορία και όχι εύρος τιμών όπως η προηγούμενη τεχνική η οποία ήτανε βασισμένη σε λεξικό. Άλλη μια ουσιαστική διαφορά είναι η κατηγοριοποίηση του αποτελέσματος, η οποία αφορά δύο κλάσης και όχι τρείς.

Η αποτελεσματικότητα του αλγορίθμου στο σύνολο πρόβλεψης, σε σύνολο πρόβλεψης 480000 κειμένων είναι 76.9%.



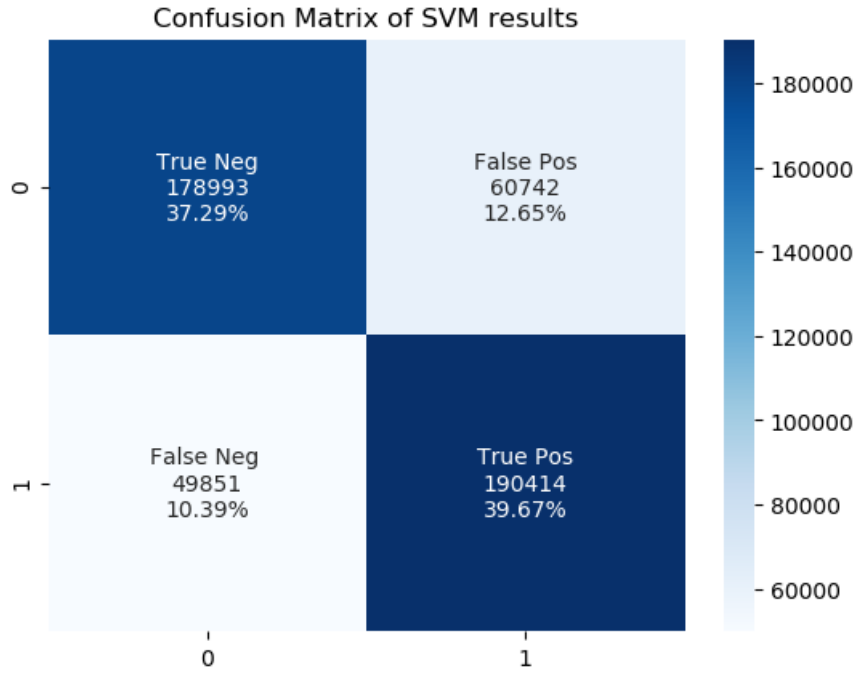
Εικόνα 19: Πλήθος αποτελεσμάτων SVM στο σύνολο των Tweets.

Η κατηγοριοποίηση στο σύνολο των δεδομένων πρόβλεψης έχει ως εξής:

Table 3: Classification Report

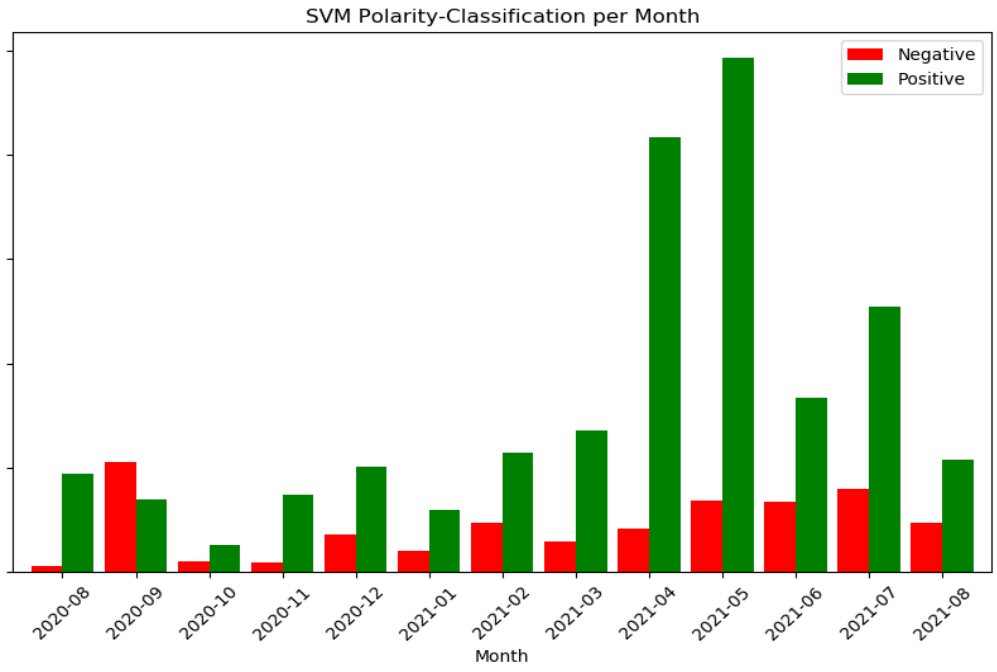
	Precision	Recall	f1-score	Support
Negative	0.78	0.74	0.76	239735
Positive	0.75	0.79	0.77	240265

Στο παραπάνω confusion matrix, παρατηρούμε τα αποτελέσματα του SVM αλγορίθμου.



Εικόνα 20: Confusion Matrix του συνόλου εκπαίδευσης.

Στο παρακάτω διάγραμμα παρατηρούμε τα αποτελέσματα του αλγορίθμου ανά μήνα, ενώ παραπάνω παρουσιάζεται το διάγραμμα της τιμής του κρυπτονομίσματος. Είναι εμφανής ο πανομοιότυπος τρόπος αύξησης της τιμής και ταυτόχρονης αύξησης του θετικού συναισθήματος στα δεδομένα μας.



Εικόνα 21: SVM κατηγοριοποίηση συναισθήματος ανά μήνα.

7. Ανάλυση συναισθήματος σε κειμενικά δεδομένα Blog

Δεύτερη πηγή κείμενωνικων δεδομένων αποτέλεσαν τα δεδομένα από ιστότοπους όπως blogs καθώς και επικεφαλίδες γνωστών ενημερωτικών ιστοσελίδων. Ποιο συγκεκριμένα, συγκεντρώθηκαν οι επικεφαλίδες των ιστοσελίδων:

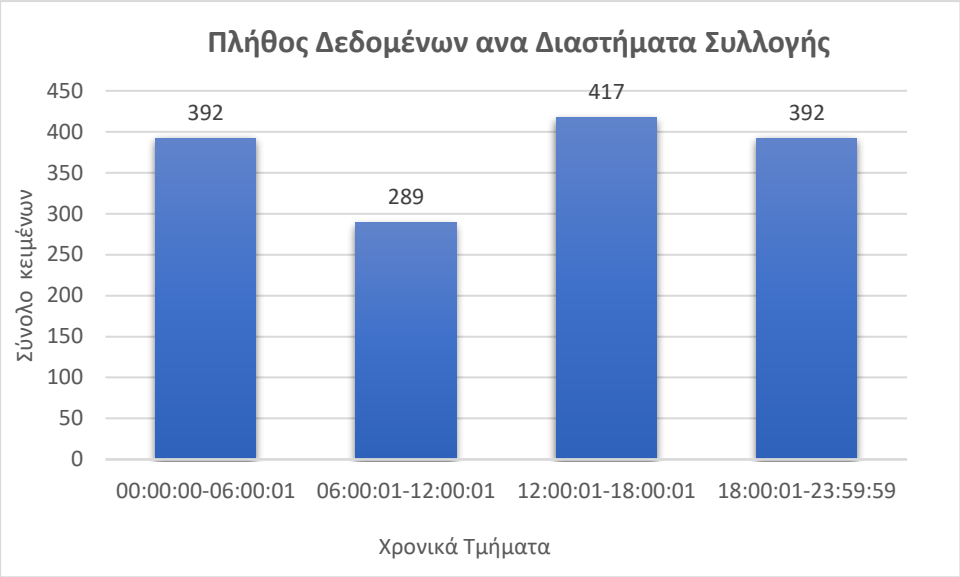
- '*moneycontrol*' (*https://www.moneycontrol.com/news/business/cryptocurrency)*
- 'BBC NEWS' (*https://www.bbc.com/news/topics/*) και
- *'REDDIT' (https://www.reddit.com/)*

για χρονικό διάστημα ενός χρόνου από τον Αύγουστο του 2020. Επιπλέον, με την βοήθεια της γλώσσας προγραμματισμού *Python* και της βιβλιοθήκης *BeautifulSoup*, για τους σκοπούς της εργασίας «κατέβηκαν» δεδομένα των αναρτήσεων περισσότερων από 25 ομάδων «Reddit communities» που αφορούν την αγορά των κρυπτονομισμάτων. Ενδεικτικά, κάποιες από τις κοινότητες των οποίων τα δεδομένα συλλέχθηκαν (web scrapping):

- `Bitcoincash` (Κοινότητα που αφορά το Bitcoin.)
- `Altcoin` (Κοινότητα που αφορά το κρυπτονομίσματα λιγότερο γνωστά στο ευρύ κοινό.)
- `cryptoCurrency` (Κοινότητα που αφορά τα κρυπτονομίσματα.)
- `binance` (Κοινότητα που αφορά την πλατφόρμα BINANCE.)
- `Bitcoin` (Κοινότητα που αφορά το Bitcoin.)
- `cardano` (Κοινότητα που αφορά το κρυπτονόμισμα Cardano.)
- `crypto` (Κοινότητα που αφορά τα κρυπτονομίσματα.)
- `ethereum` (Κοινότητα που αφορά το Ethereum.)
- `ethereumcoin` (Κοινότητα που αφορά το Ethereum.)
- `Money` (Κοινότητα που αφορά την ενότητα "χρήματα".)
- `btc` (Κοινότητα που αφορά το Bitcoin.)
- `ETH` (Κοινότητα που αφορά το Ethereum.)
- `Business` (Κοινότητα που αφορά τον χώρο των επιχειρήσεων.)
- `Coindesk` (Κοινότητα που αφορά την πλατφόρμα CoinDesk.)

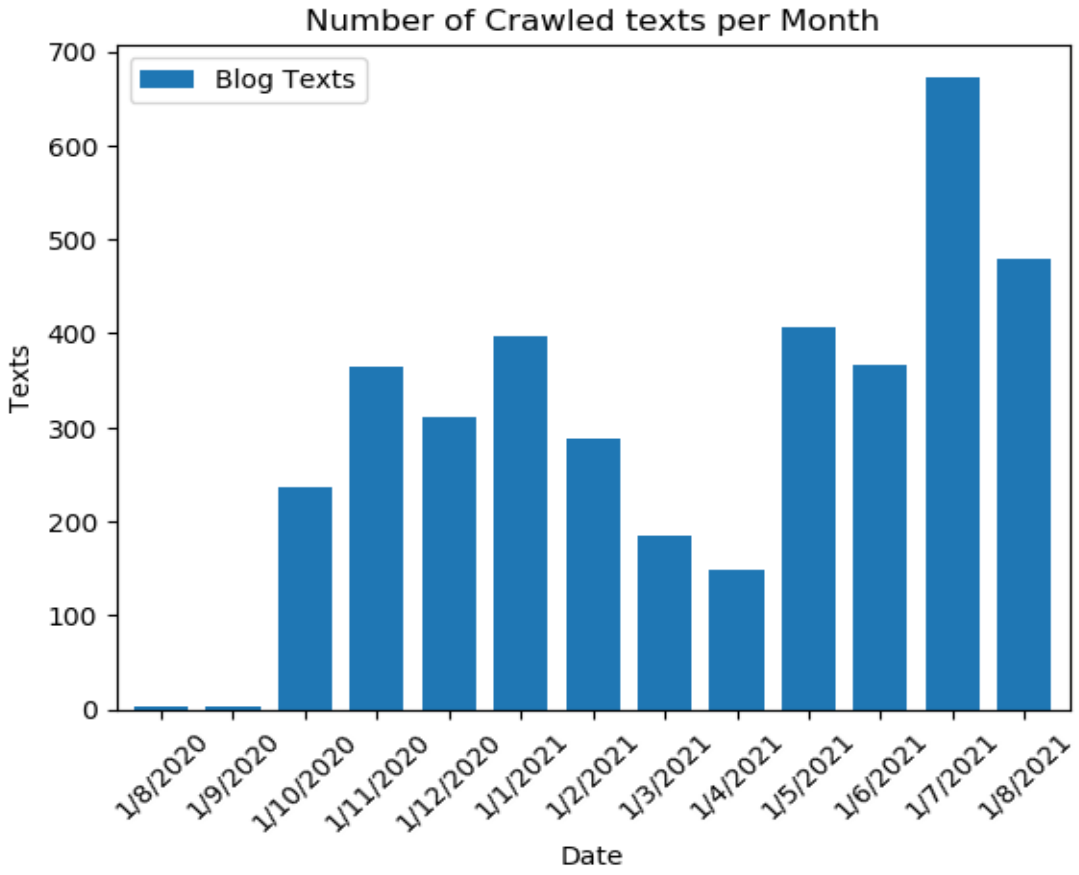
7.1 Καθαρισμός δεδομένων και Κατηγοριοποίηση

Το σύνολο των δεδομένων μας ενοποιήθηκε σε ένα συγκεντρωτικό σετ δεδομένων. Τα δεδομένα καθαρίστηκαν από διπλότυπες εγγραφές και κατηγοριοποιήθηκαν σε παρτίδες των έξι ωρών ώστε να συμβαδίζουν με τα δεδομένα τα χρονικά φάσματα της τιμής του κρυπτονομίσματος.



Εικόνα 22: Πλήθος κειμενικών δεδομένων ανά Διάστημα Συλλογής.

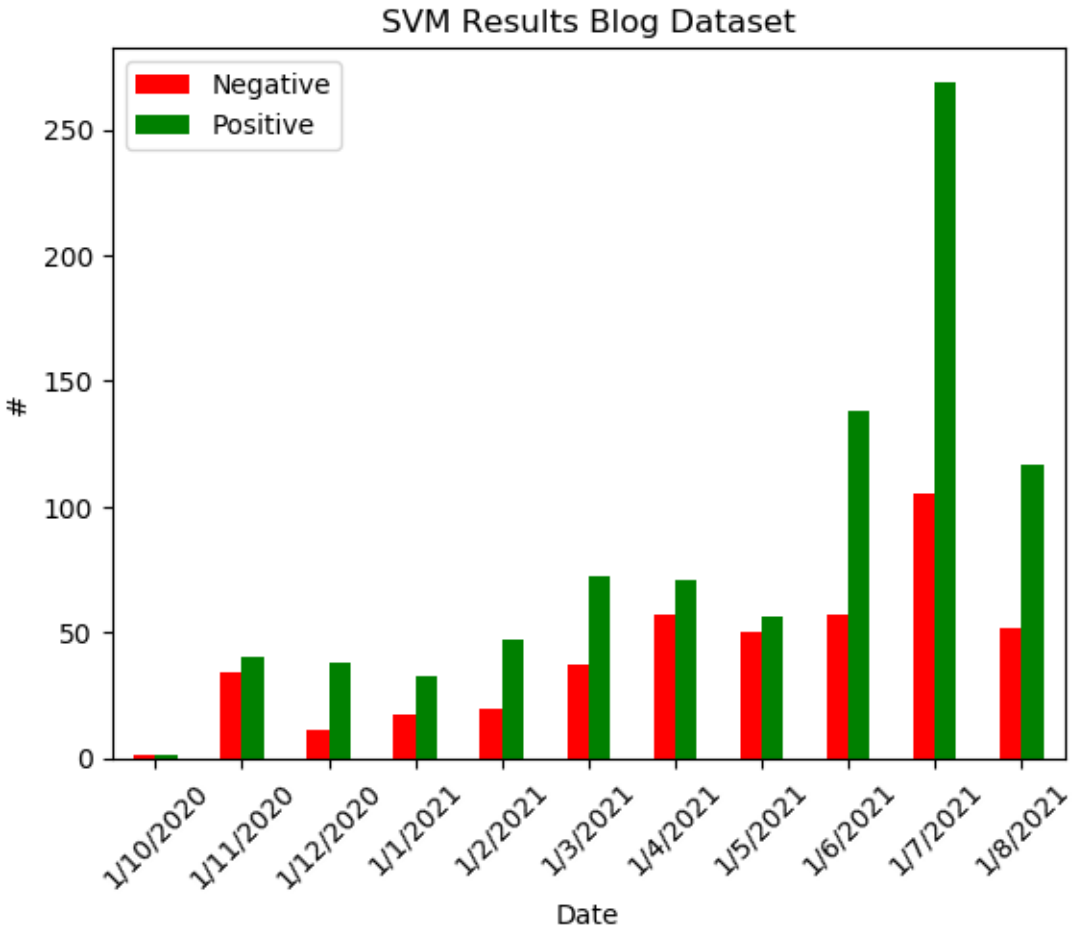
Το σύνολο των δεδομένων μας αποτελείτε από 3921 κείμενα – αναρτήσεις τα οποία κατανέμονται χρονικά όπως φαίνεται στο παραπάνω διάγραμμα.



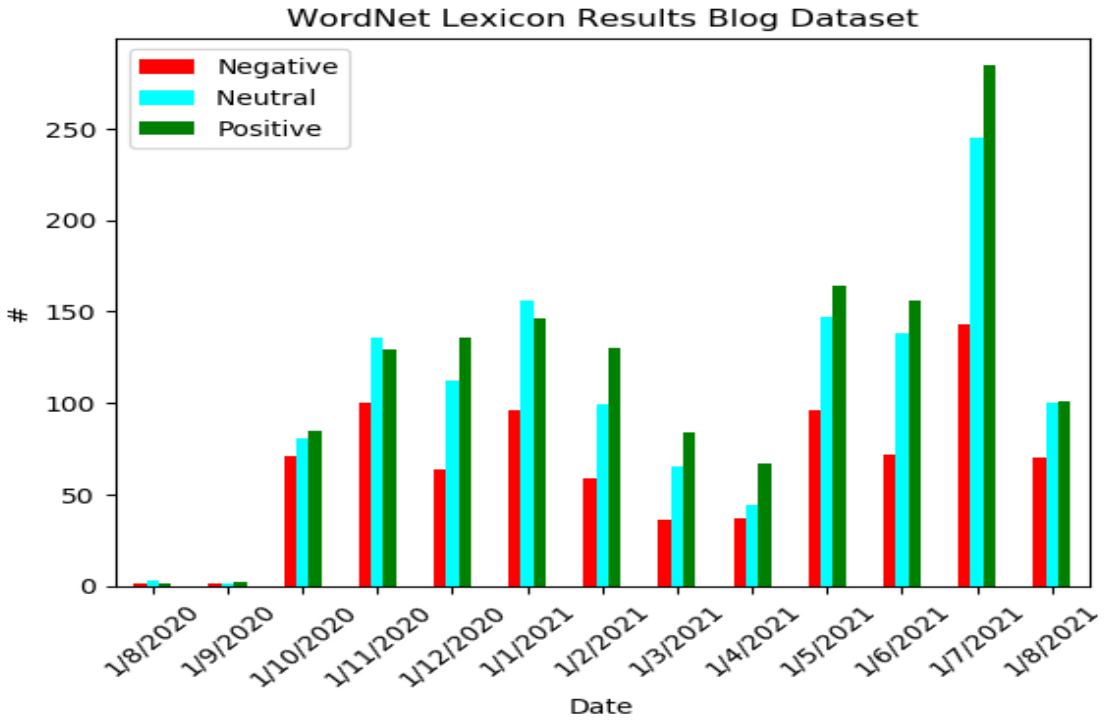
Εικόνα 23: Πλήθος δεδομένων που "κατέβηκαν" ανά μήνα.

Η κατηγοριοποίηση των δεδομένων έγινε πάλι με τρεις προσεγγίσεις. Δυο από αυτές αφορούν τεχνικές βασισμένες σε λεξικό (VADER, SentiWordNet) και η μία αποτελεί μέθοδο μηχανικής μάθησης. Ποιο συγκεκριμένα το ίδιο μοντέλο SVM αλγορίθμου το οποίο εκπαιδεύτηκε με τα δεδομένα του συνόλου (https://www.kaggle.com/gauravduttakiit/twitter-sentiment-analysis-nlp/data) και εφαρμόστηκε στα δεδομένα του Tweeter, χρησιμοποιήθηκε και στην περίπτωση των κειμένων που εξάχθηκαν από τις προαναφερθείσες ιστοσελίδες.

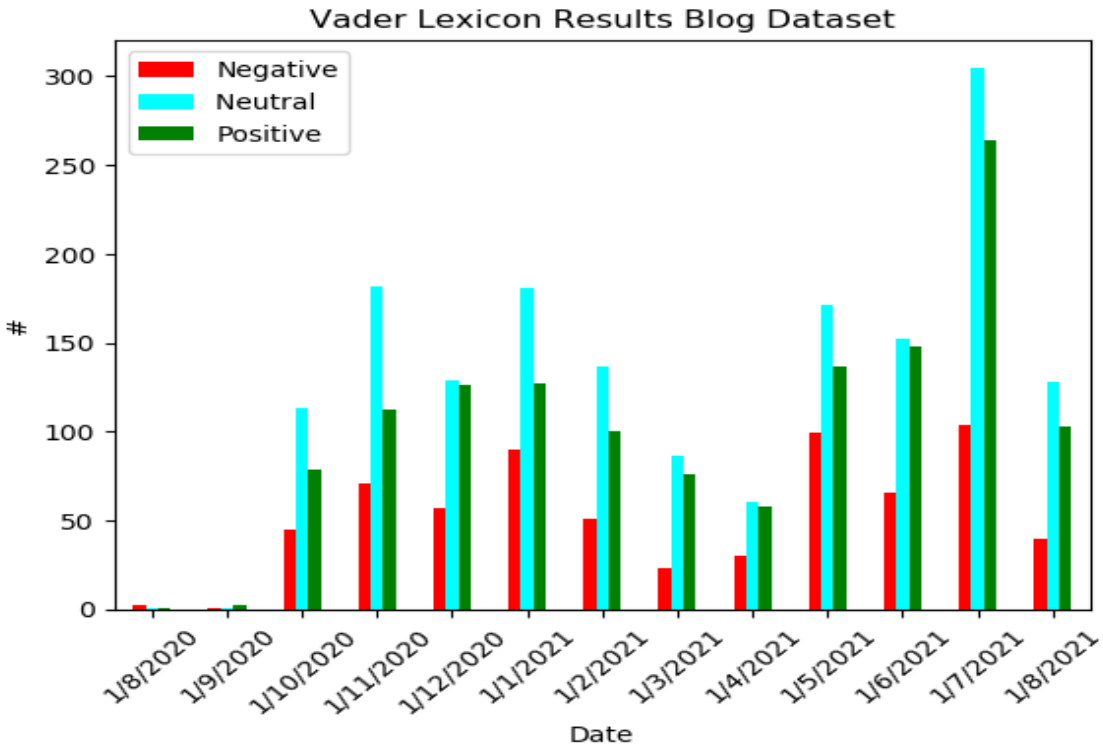
Παρακάτω τα αποτελέσματα,



Εικόνα 24: Κατηγοριοποίηση δεδομένων BLOG με SVM.



Εικόνα 26: Κατηγοριοποίηση BLOG δεδομένων με βάση το WordNet Λεξικό.

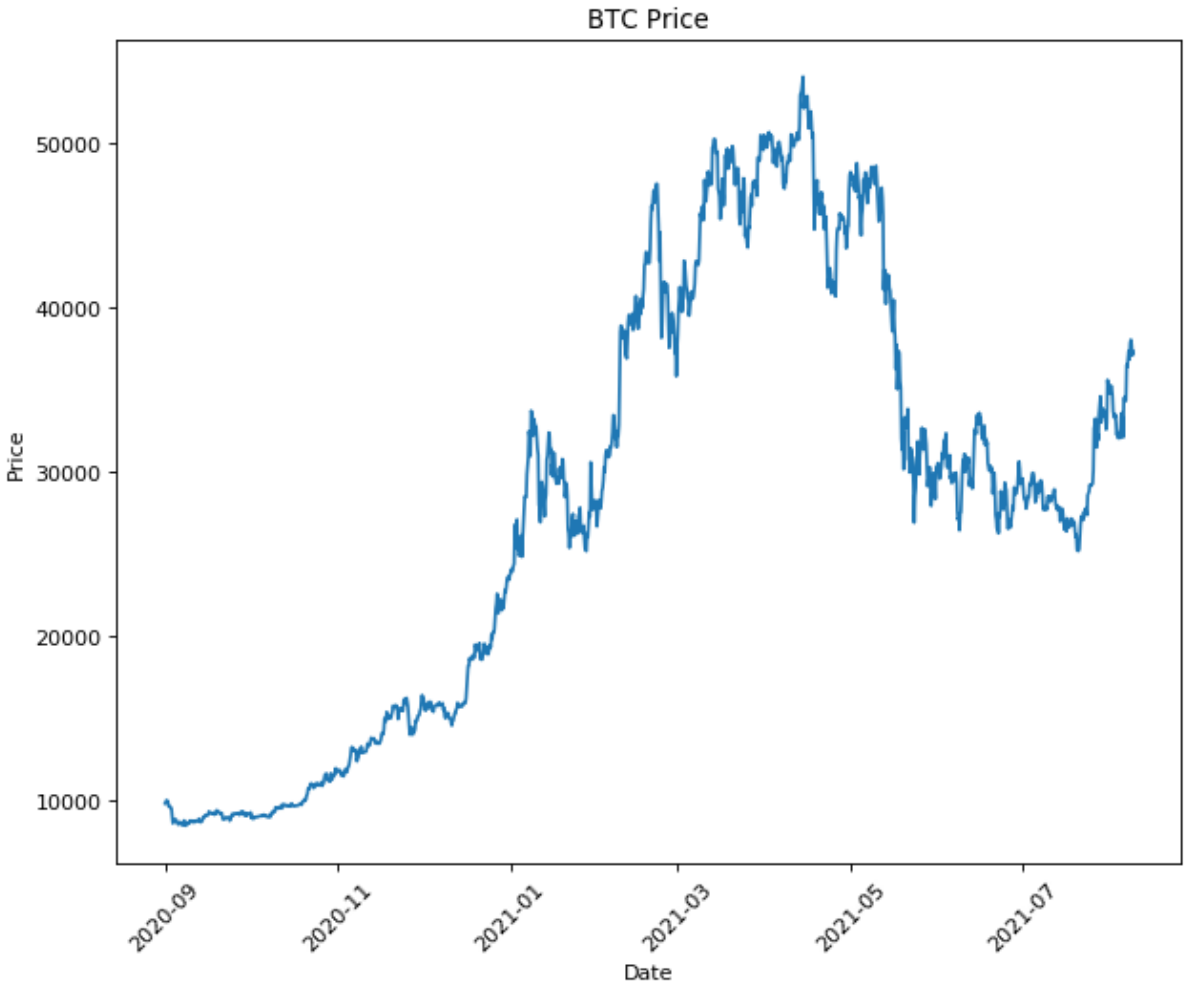


Εικόνα 25: Κατηγοριοποίηση BLOG δεδομένων με βάση το Λεξικό VADER.

8. Ανάλυση Χρονοσειράς Κρυπτονομίσματος

Η ανάλυση μιας χρονοσειράς αποτελεί ένα βήμα της προετοιμασίας ανάπτυξης ενός προβλεπτικού μοντέλου. Η ανάλυση μιας χρονοσειράς περιλαμβάνει την κατανόηση της φύσης της χρονοσειράς καθώς και των χαρακτηριστικών της, με σκοπό την ανάπτυξη ενός ουσιαστικού και αποτελεσματικού μοντέλου πρόβλεψης.

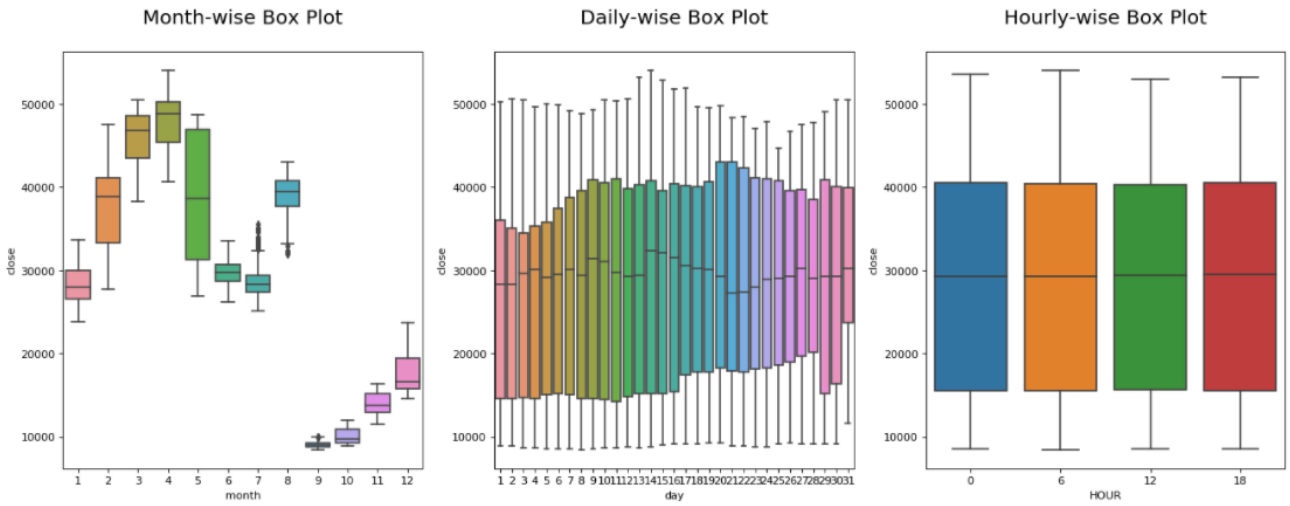
Τα δεδομένα που συνθέτουν την χρονοσειρά είναι η τιμή του κρυπτονομίσματος BITCOIN την χρονική περίοδο 31-08-2020 έως 08-08-2021. Ως «*τιμή*» για χάριν ευκολίας θα ορίσουμε την τιμή κλεισίματος του νομίσματος κάθε έξι ώρες.



Εικόνα 27: Χρονοσειρά κρυπτονομίσματος Bitcoin.

Σε μια πρώτη εικόνα η χρονοσειρά παρουσιάζει κάποια εμφανή ανοδική τάση όχι όμως και κάποια εμφανή εποχικότητα.

Τα γραφήματα Box-Plot μπορούν να μας δώσουνε μία εικόνα της διασποράς των τιμών στο σύνολο των δεδομένων μας. Κάνοντας αναπαράσταση των box-Plots σε επίπεδο ημέρας, σε επίπεδο μήνα και ωρών παρατηρούμε ότι:

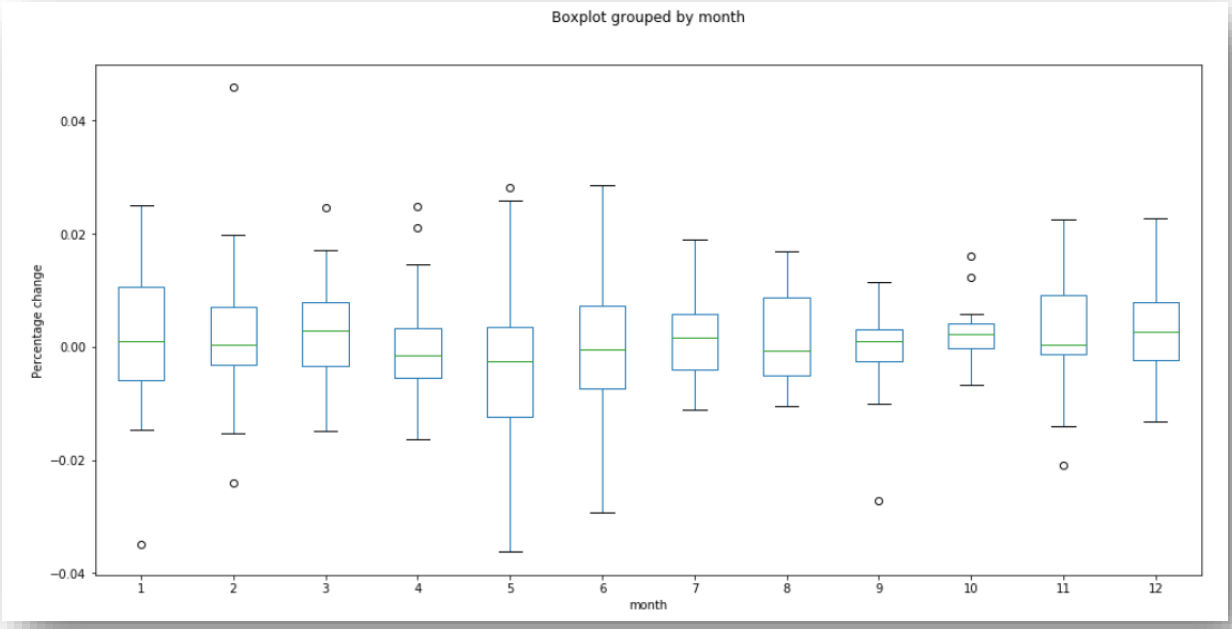


Εικόνα 28: Box-Plot διαγράμματα σε διαφορετικά χρονικά διαστήματα.

Παρατηρούμε της διαφορά του εύρους των τιμών στους τελευταίους μήνες οι οποίοι αφορούν την σεζόν 2020 όπου η τιμή του κρυπτονομίσματος ήτανε αρκετά ποιο χαμηλή. Έτσι εξηγείται και η διαφορά μεταξύ Αυγούστου και Σεπτεμβρίου. Στο διάγραμμα ημερών μπορούμε να διακρίνουμε την αύξηση της αξίας της τιμής στα μέσα του μήνα (13,14,15) και την απότομή πτώση τις επόμενες ημέρες σε συνδυασμό με την αστάθεια που παρουσιάζεται.

Το παραπάνω διάγραμμα όμως εξαιτίας του μικρού αριθμού των δεδομένων δεν βοηθάει πολύ στην εξαγωγή συμπερασμάτων.

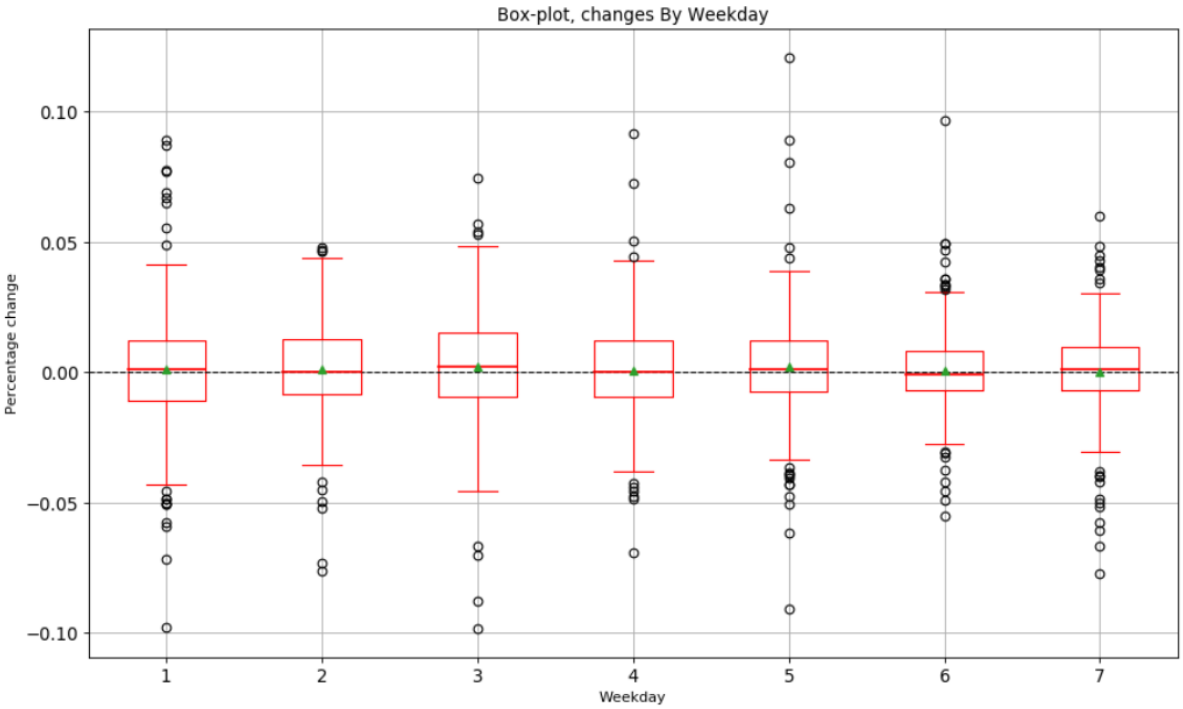
Για τον λόγο αυτό στο παρακάτω διάγραμμα αναπαριστάτε το ποσοστό διαφοροποίησης της τιμής από την μία μέρα στην άλλη, ανά τον μήνα (change volume).



Εικόνα 29: Box-Plot διάγραμμα ρυθμού αλλαγής τιμής ανά μήνα.

Με βάση το παραπάνω διάγραμμα παρατηρούμε την μεγάλη αστάθεια που παρουσιάζει το ποσοστό αλλαγής της τιμής στα μέσα και στο τέλους του έτους. Επιπλέον, παρατηρείτε η θετική αύξηση του ποσοστού της ημερήσιας διαφοράς της τιμής του μήνες 5 και 6 με θετικές τιμές των *Outliers*. Αρά, αυτή η παρατήρηση αυτή μπορεί να οδηγήσει σε μια επενδυτική απόφαση.

Ας παρατηρήσουμε όμως την κυκλικότητα της χρονοσειράς στο παρακάτω Box-Plot στο οποίο απεικονίζεται η διαφορά στην καθημερινή διακύμανση της τιμής ανά τις επτά μέρες της εβδομάδας. Με τον αριθμό 1 να αναφέρετε στην 'Δευτέρα'.

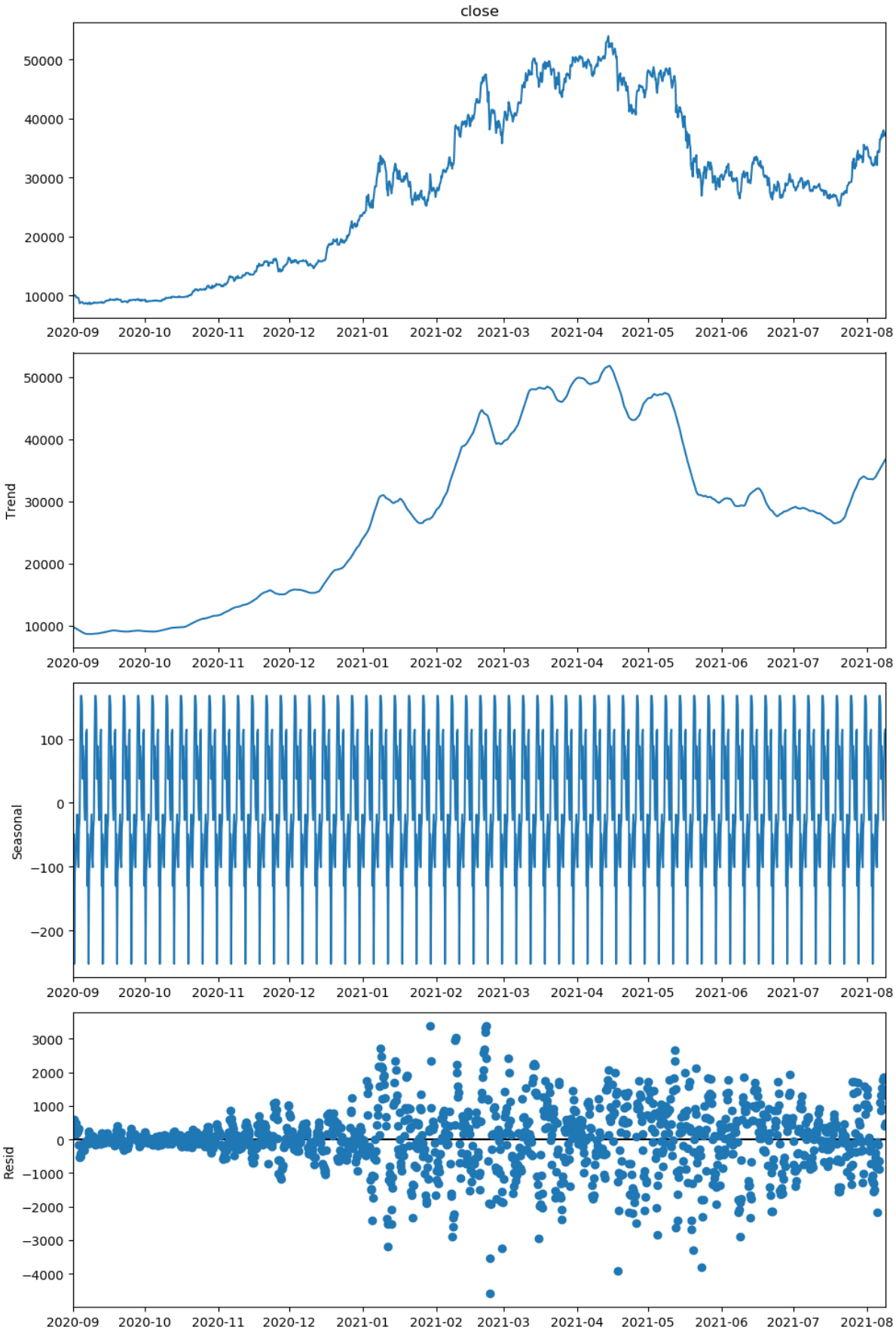


Εικόνα 30: Box-Plot ρυθμού μεταβολής τιμής ανά ημέρα.

Μπορούμε να παρατηρήσουμε με βάση το εύρος των τιμών την αστάθεια που προκύπτει την ημέρα Τρίτη, καθώς και τις αρνητικές ακραίες τιμές του Σαββάτου.

8.1 Αποσύνθεση της χρονοσειράς

Γνωρίζοντας ότι οι χρονοσειρές αποτελούνε μια σύνθεση χαρακτηριστικών (τάση, θόρυβος , κυκλικότητα) ο διαχωρισμός του βοηθήσει στην αναγνώριση των χαρακτηριστικών στα οποία πρέπει να επέμβουμε στο κομμάτι της πρόβλεψης.



Εικόνα 31: Αποσύνθεση χρονοσειράς BTC.

Παρατηρείται, η μεγάλη διασπορά των ακραίων τιμών κατά τον 2^{ο} και 3^{ο} μήνα η οποία δικαιολογείται από την απότομη αύξηση τις τιμής τους μήνες εκείνους. Επιπλέον, έχουμε περιοδικότητα από το διάγραμμα της εποχικότητας η οποία επαναλαμβάνεται ανά μήνα.

Από τα πρώτα στάδια της ανάλυσης μιας χρονοσειράς είναι η παρατήρηση της παρουσίας ή μη παρουσίας Τάσης (*trend*), ώστε να οδηγηθούμε στο συμπέρασμα αν η χρονοσειρά μας είναι στατική ή όχι. Για την επιβεβαίωση της υπόθεσης ότι η χρονοσειρά μας είναι στάσιμη θα χρησιμοποιήσουμε τα οικονομετρικά τεστ *Dickey – Fuller* και *Zivot-Andrews*. Τα αποτελέσματα των τεστ φαίνονται στο παρακάτω πίνακα:

Table 4: Τεστ Στασιμότητας Χρονοσειράς.

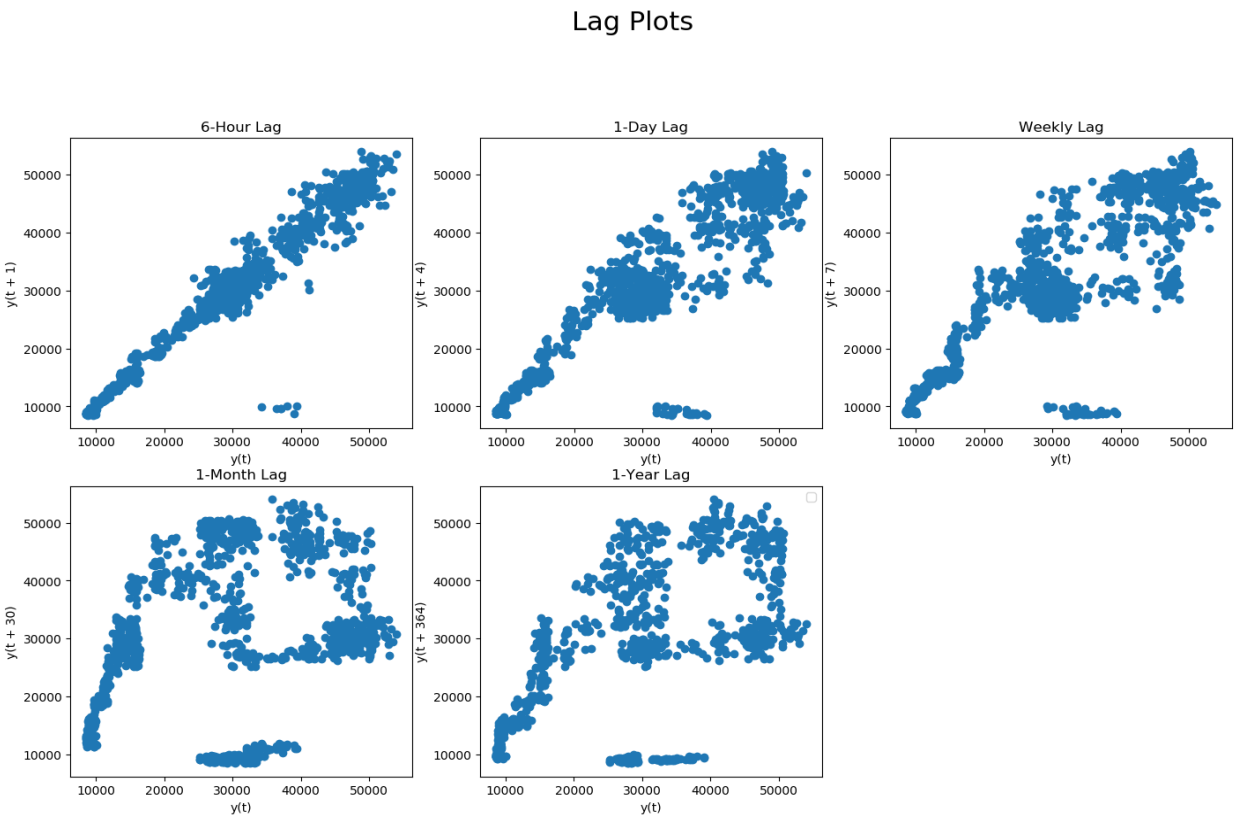
	ADF test	**Zivot-Andrew's test**
Test Results	-1.208795	-5.04
p-value	0.669801	0.02
Critical Values 1%	-3.435	-5.28
Critical Values 5%	-2.864	-4.81
Critical Values 10%	-2.568	-4.57

Όσο ποιο αρνητική η τιμή του τεστ τόσο ποιο εύκολα μπορούμε να απορρίψουμε τον ισχυρισμό για μη-στάσιμη χρονοσειρά. Επομένως, ο συνδυασμός του αποτελέσματος των τεστ και του διαστήματος εμπιστοσύνης μας οδηγεί στο συμπέρασμα ότι η χρονοσειρά μας δεν είναι στάσιμη.

8.1.1 Lag Plots

Με την βοήθεια των διαγραμμάτων υστέρησης μπορούμε να παρατηρήσουμε πρότυπα που δημιουργούνται τα οποία σχετίζονται με την «τυχαιότητα» των δεδομένων μας. Όσο ποιο αόριστο (δεν ακολουθεί κάποιο πρότυπο) είναι το διάγραμμα υστέρησης τόσο και μεγαλύτερη η τυχαιότητα στα δεδομένα μας.

Στα παρακάτω διαγράμματα υστέρησης διακρίνεται καθαρά ένα γραμμικό πρότυπο-ακολουθία (διάγραμμα 6 ωρών).



Εικόνα 32: Lag Plot Btc.

8.2 Εκπαίδευση και σύγκριση μοντέλων

Στη παράγραφο αυτήν θα εφαρμόσουμε και θα συγκρίνουμε τα αποτελέσματα διαφορετικών προβλεπτικών μοντέλων .Τα μοντέλα τα οποία θα πάρουν μέρος στην σύγκριση είναι τα:

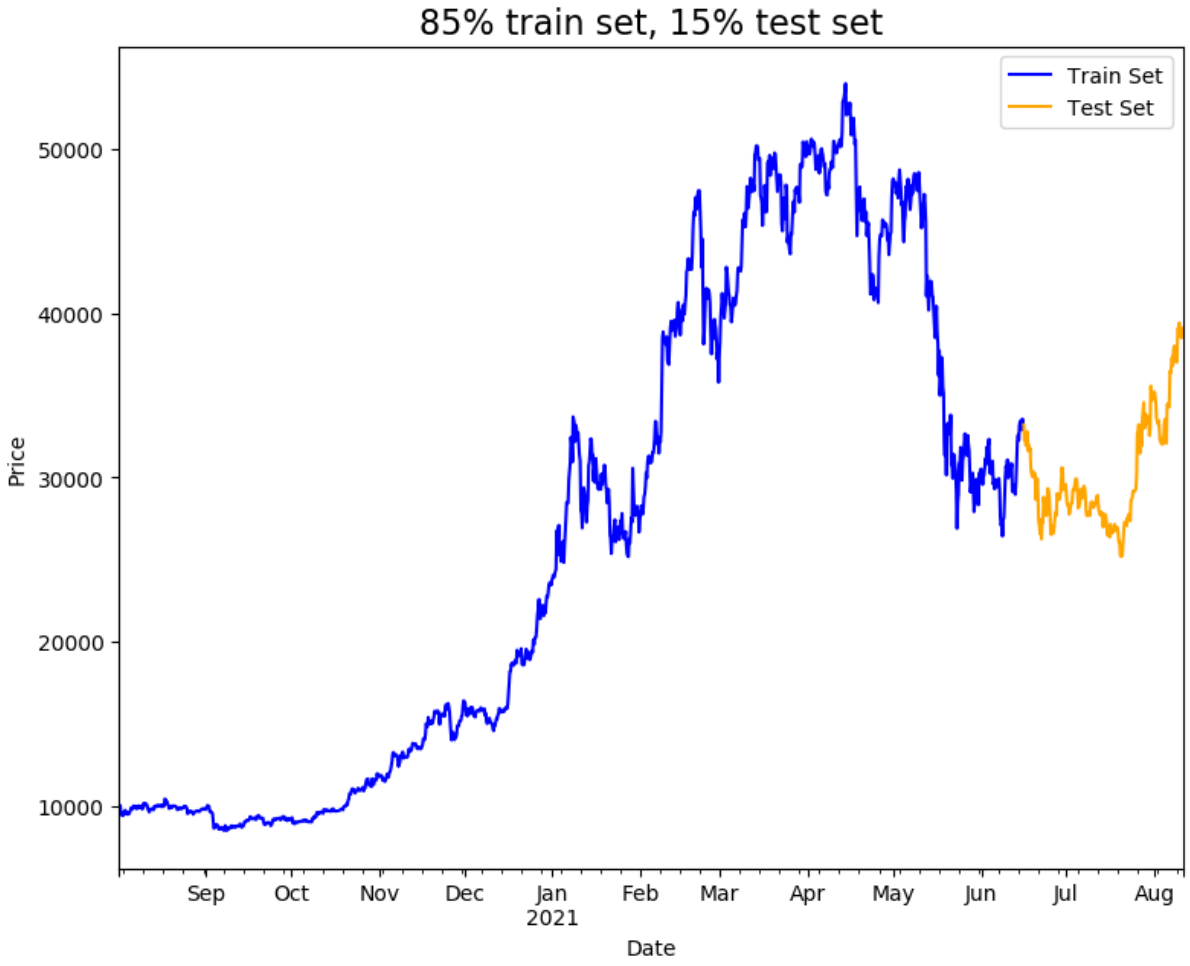
1) Naïve Method
2) Holt
3) Simple Exponential Smoothing
4) ARIMA
5) SARIMAX
6) Bayesian NN (BNN)
7) K-Nearest Neighbor regression
8) Long Short-Term Memory (LSTM)

Τα μοντέλα εξετάστηκαν τόσο για μακροχρόνιες προβλέψεις όσο και για την πρόβλεψη της επόμενης παρατήρησης.

8.2.1 Naïve Method

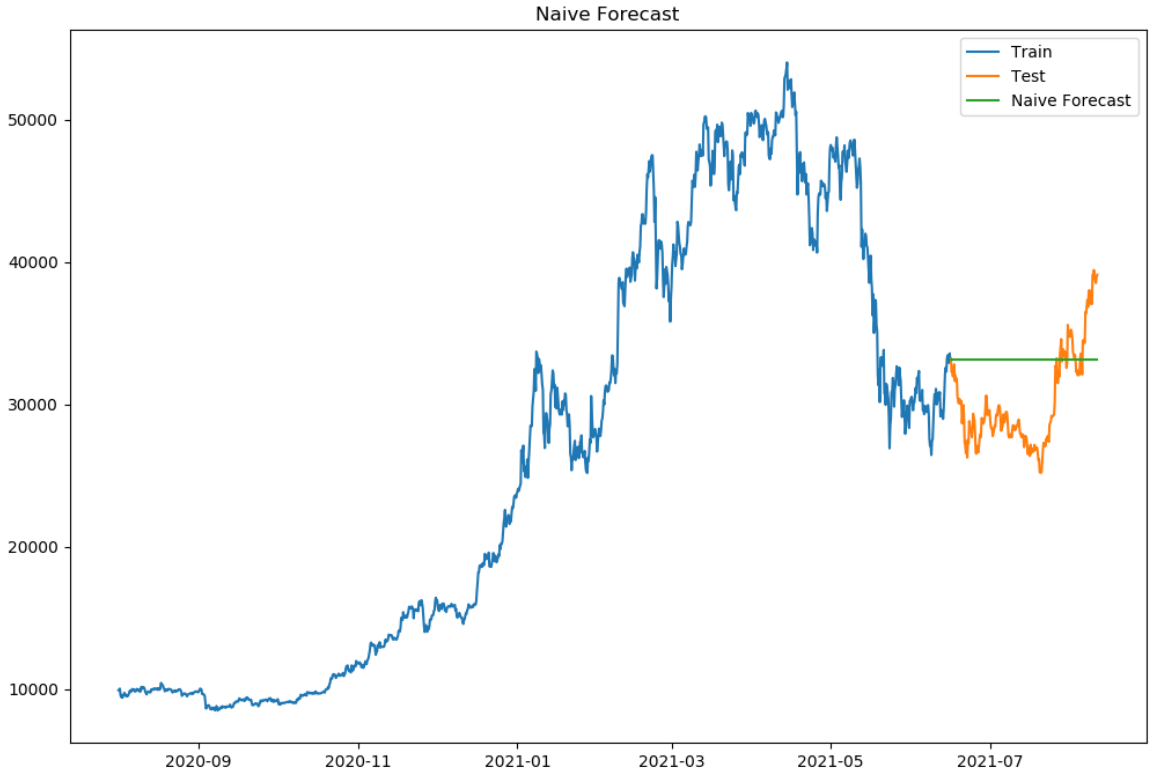
Σε περιπτώσεις όπου η χρονοσειρά είναι στάσιμη, μπορούμε να προβλέψουμε ότι η επόμενη παρατήρηση ισούται με την τελευταία παρατήρηση.

Ο διαχωρισμός των δεδομένων σε δεδομένα εκπαίδευσης, επικύρωσης και πρόβλεψης θα έχει ως εξής, *85%* δεδομένα εκπαίδευσης, *15%* δεδομένα πρόβλεψη.



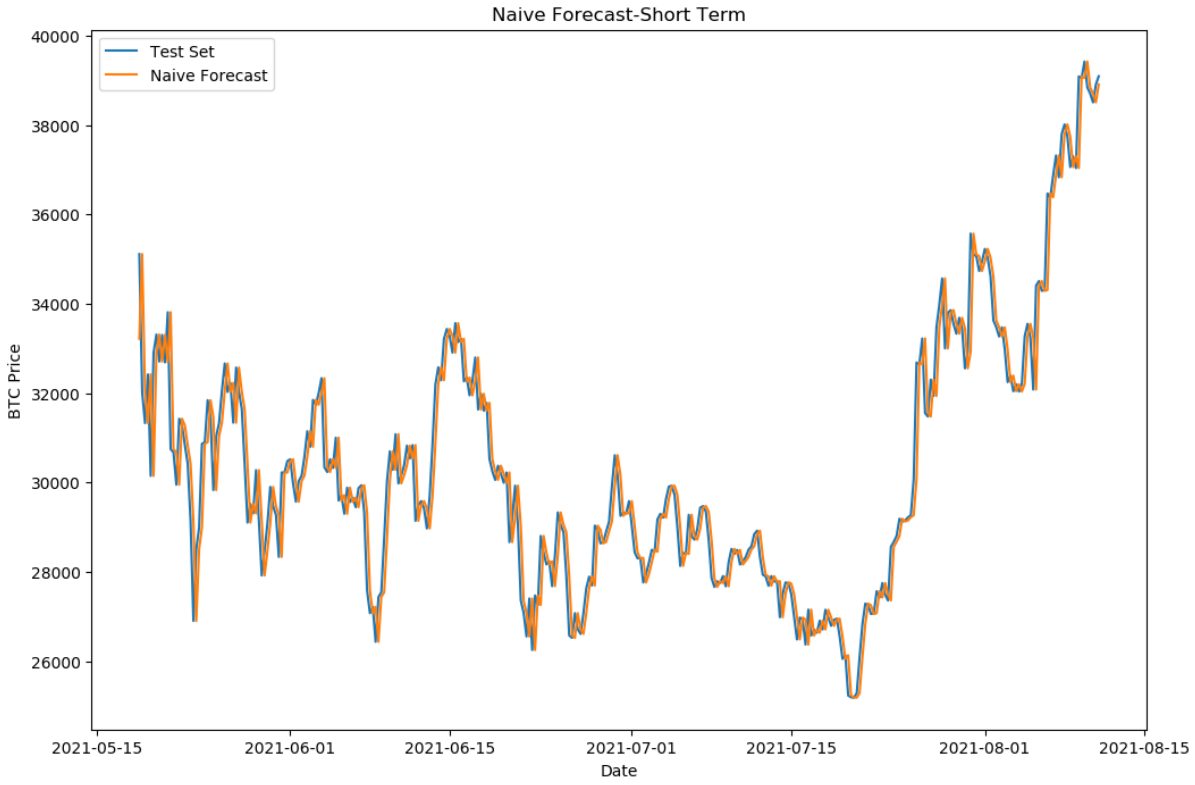
Εικόνα 33: Σύνολο Εκπαίδευσης και Πρόβλεψης.

Στο παρακάτω διάγραμμα παρατηρούμε την γραμμή πρόβλεψης. Όπως ήτανε αναμενόμενο η τιμή είναι μια ευθεία γραμμή, η οποία αποτελεί προέκταση της τελευταίας παρατήρησης.



Εικόνα 34: Naive Forecast

Αντίστοιχα η πρόβλεψη της επόμενης κάθε φορά παρατήρησης παρουσιάζεται στο παρακάτω σχήμα.

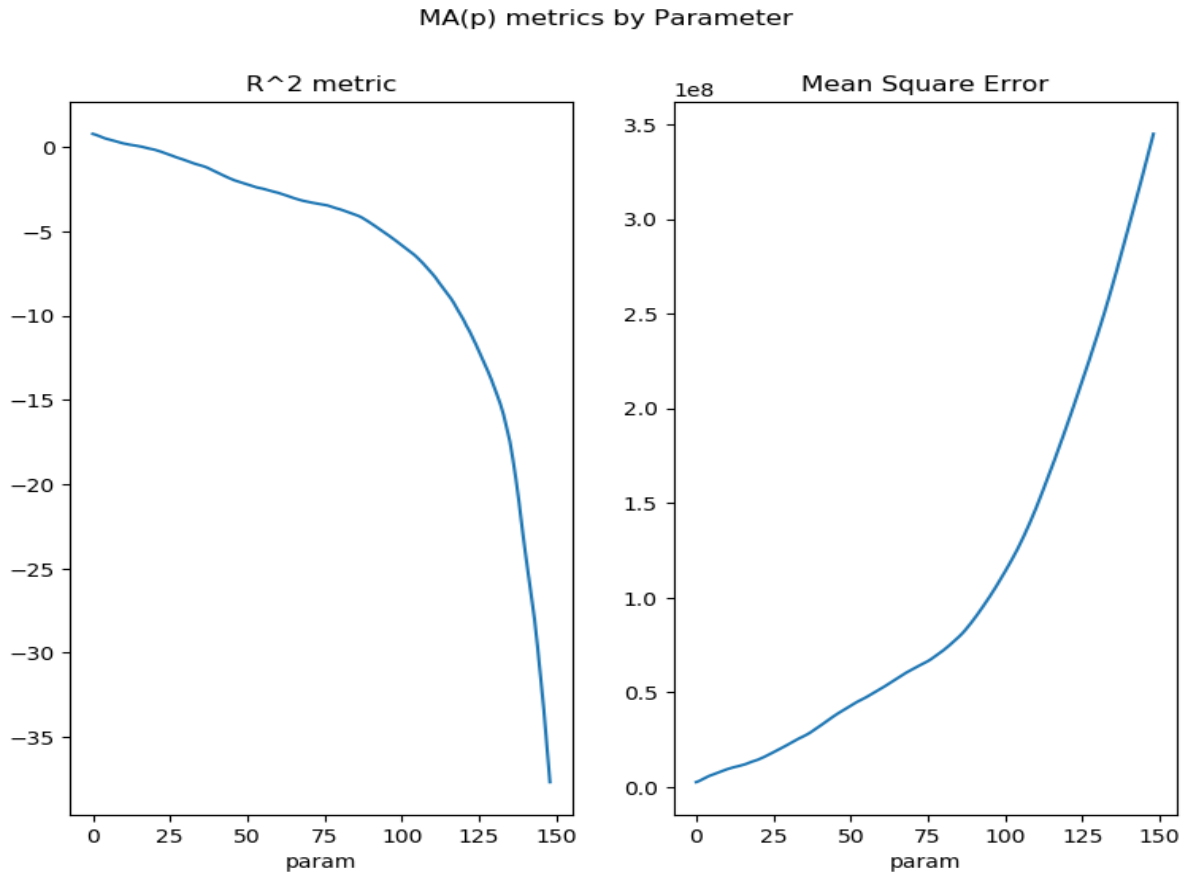


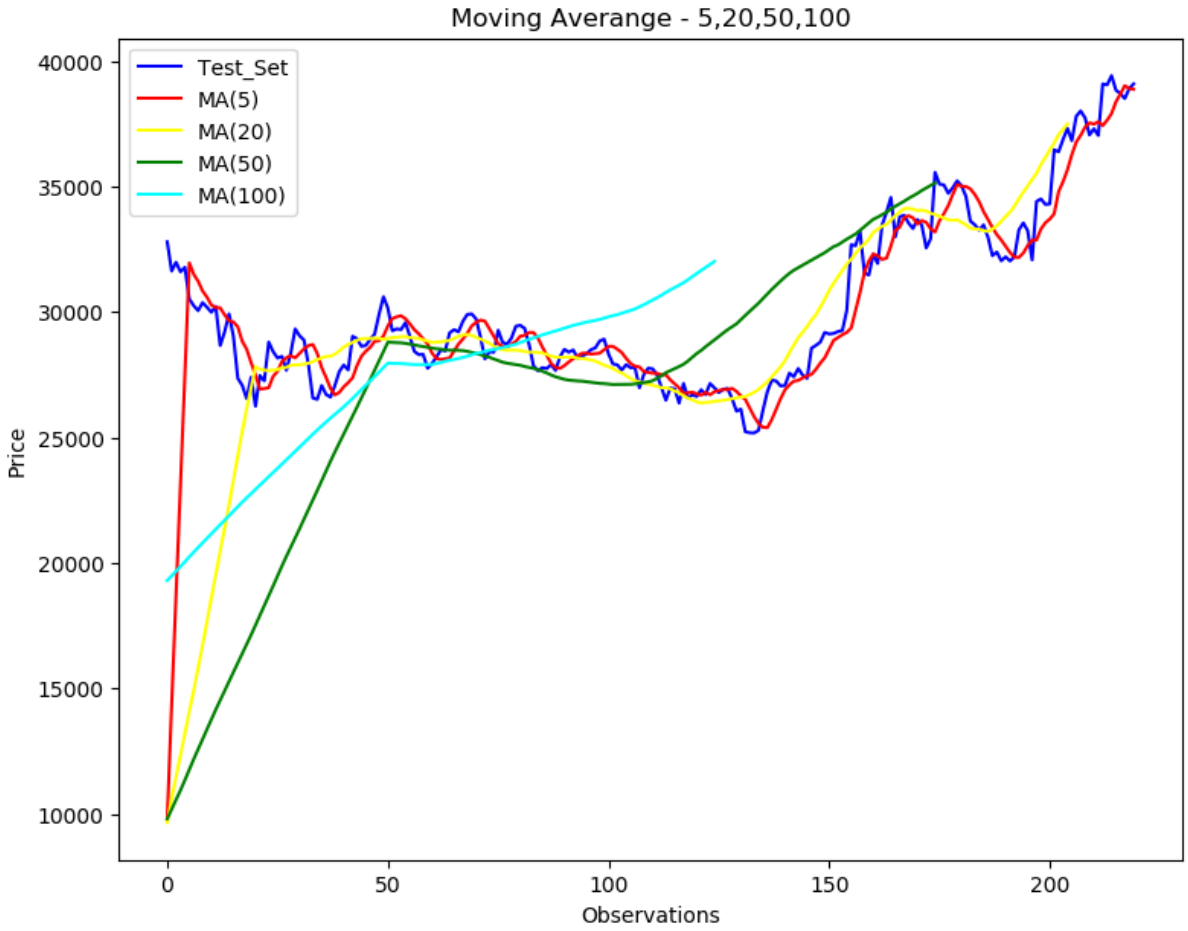
Εικόνα 35: Naive Forecast -Next step.

8.3 Κινητός Μέσος

Για την μέθοδο του κινητού μέσου έγινε χρήση πολλών διαφορετικών «περιόδων». Στο παρακάτω διάγραμμα παραθέτουνε (ως καλύτερα αποτελέσματα) την εφαρμογή του κινητού μέσου στην χρονοσειρά μας με βήμα 20, 50 και 100. Όπως είναι κατανοητό η επιλογή μερικών περιόδων ≤ 20 συμβολίζουν την ταχύτερη εφαρμογή της χρονοσειράς που προκύπτει από την Κινητό μέσο στην χρονοσειρά μας και είναι κατάλληλη για πρόβλεψη μικρών χρονικών διαστημάτων. Αντίθετα η χρήση του κινητού μέσου με περίοδο 100 που χρησιμοποιήθηκε έχει ποιο αργή εφαρμογή στα δεδομένα μας και ενδείκνυται για πρόβλεψη μεγαλύτερων χρονικών διαστημάτων.

Με σκοπό την επιλογή της καλύτερης περιόδου για την εφαρμογή του κινητού μέσου μελετήθηκε η αποτελεσματικότητα του μοντέλου σε πολλές διαφορετικές περιόδους. Η επιλογή της καλύτερης περιόδου έγινε με βάση την μέση τιμή του νομίσματος σε παράθυρο 150 παρατηρήσεων καθώς και του συντελεστή προσδιορισμού R^2. Τα αποτέλεσμα της αναζήτησης παρουσιάζονται στα παρακάτω διαγράμματα.



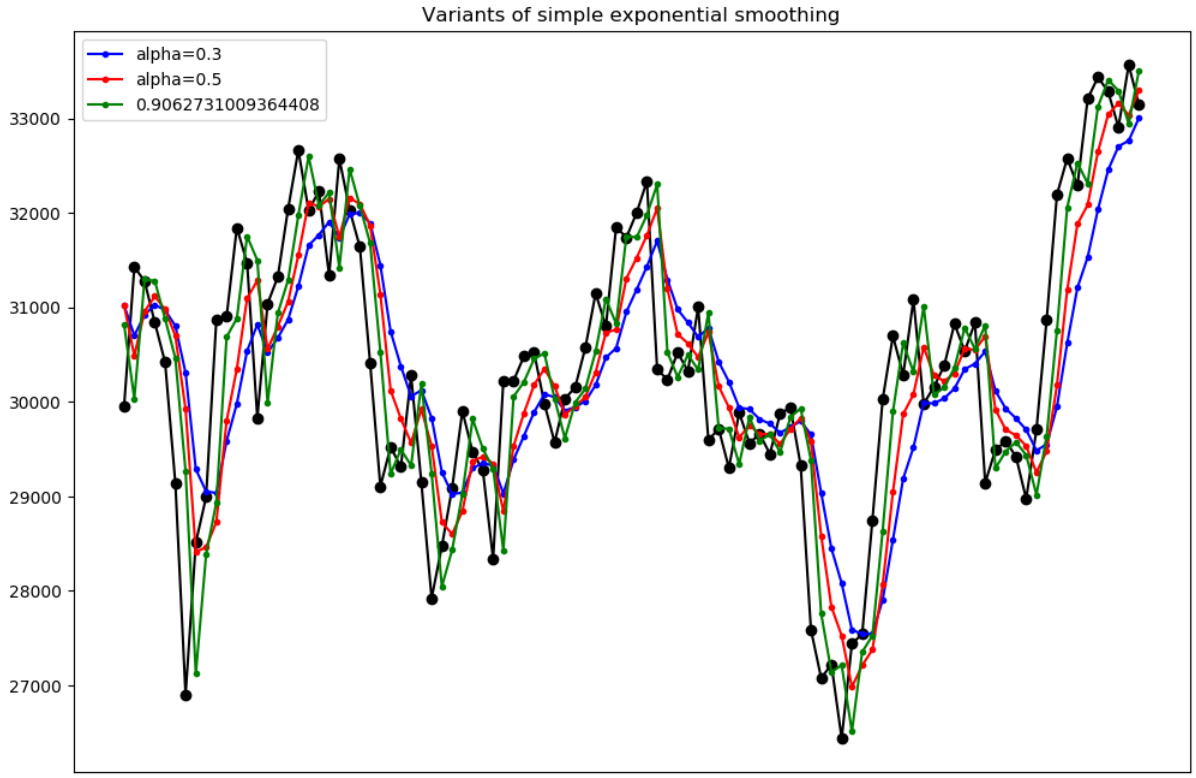


Εικόνα 36 : Moving Average- multiple steps.

8.4 Simple Exponential Smoothing

Όπως και στο προηγούμενο μοντέλο, χρησιμοποιήθηκε το ίδιο πλήθος παρατηρήσεων για εκπαίδευση και πρόβλεψη. Έγινε εκπαίδευση και πρόβλεψη με διαφορετικά μεγέθη της παραμέτρου *alpha*. Στο παρακάτω διάγραμμα παρατηρείτε η εφαρμογή του μοντέλου σε με διαφορετικές τιμές τις παραμέτρου (εξομάλυνσης) *alpha.*

Η παράμετρος *α* (alpha) αναπαριστά τον βαθμό επιρροής του μοντέλου από τις προηγούμενες παρατηρήσεις. Η παράμετρος παίρνει τιμές μεταξύ 0 και 1. Τιμές της παραμέτρου κοντά στο 1 υποδηλούμουν μεγαλύτερά εξάρτηση-επιρροή του μοντέλου από τις πρόσφατες παρατηρήσεις της χρονοσειράς. Αντίθετα, τιμές κοντά στο 0, κάνουν το μοντέλο να επηρεάζεται περισσότερο από τις ιστορικές τιμές [43].



Εικόνα 37: Exponential Smoothing fit with different alpha parameter.

Με βάση της μετρικές που χρησιμοποιούνται στην παρούσα εργασία, επιλέχθηκε ως παράμετρος εξομάλυνσης *alpha= 0.3*.

8.5 Holt's Method

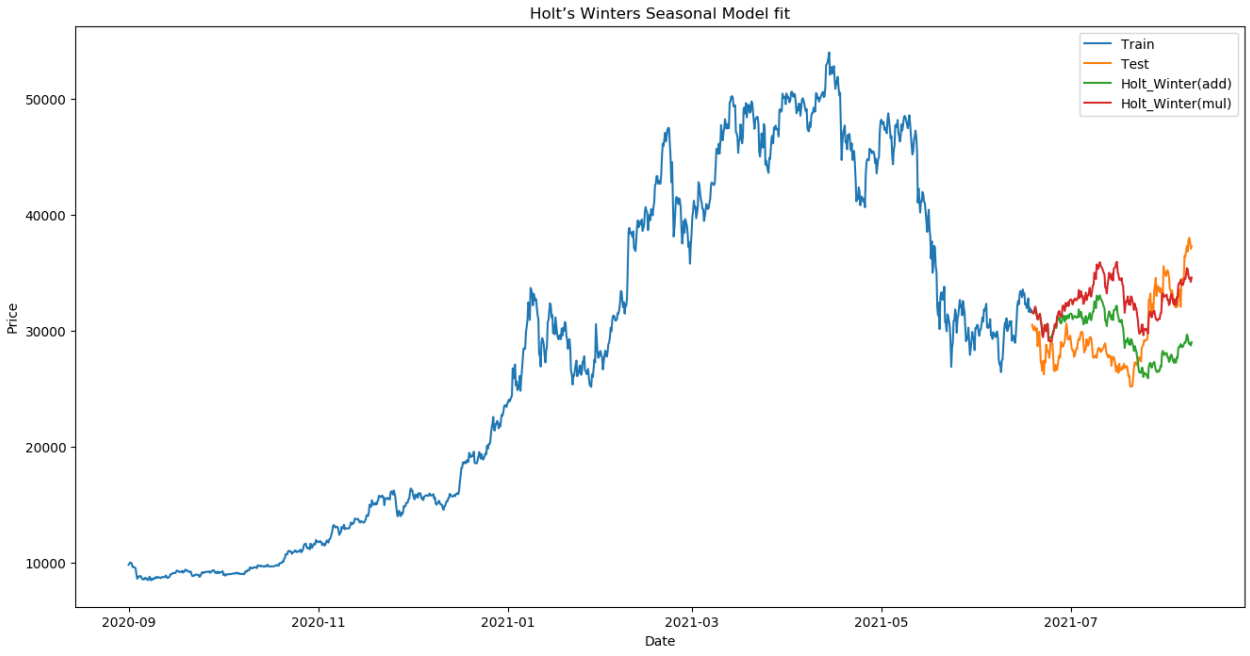
Έγινε χρήση του μοντέλου Holt σε διάφορες παραλλαγές. Χρησιμοποιήθηκε η εκθετική εξομάλυνση και μέθοδοι απόσβεσης. Οι παράμετροι που επιλέχθηκαν από το σύστημα ως ποιο κατάλληλοι με βάση το άθροισμα ελαχίστων τετραγώνων παρουσιάζονται στο παρακάτω πίνακα.

Table 5: Παράμετροι ανά σύστημα.

	SES	Holt's	Exponential	Additive	Multiplicative
α	1.000000	0.974308	9.776329e-01	0.978852	0.974891
β	NaN	0.000000	4.016578e-12	0.000000	0.000000
ϕ	NaN	NaN	NaN	0.980000	0.981637
l_0	263.917688	258.882600	2.603440e+02	257.357526	258.940454
b_0	NaN	5.010783	1.013780e+00	6.644741	1.038159
SSE	6761.350235	6004.138200	6.104195e+03	6036.555005	6081.995166

8.6 Holt Winter Seasonal Model

Η εφαρμογή του μοντέλου των Holt-Winter's έγινε με προσθήκη προσθετικής τάσης και πολλαπλασιαστικής τάσης. Αντίστοιχα, σε επίπεδο εποχικότητας. Η εποχικότητα ορίστηκε σε επίπεδο ημέρας. Στο παρακάτω γράφημα παρατηρείτε η εφαρμογή του μοντέλου στο σύνολο.



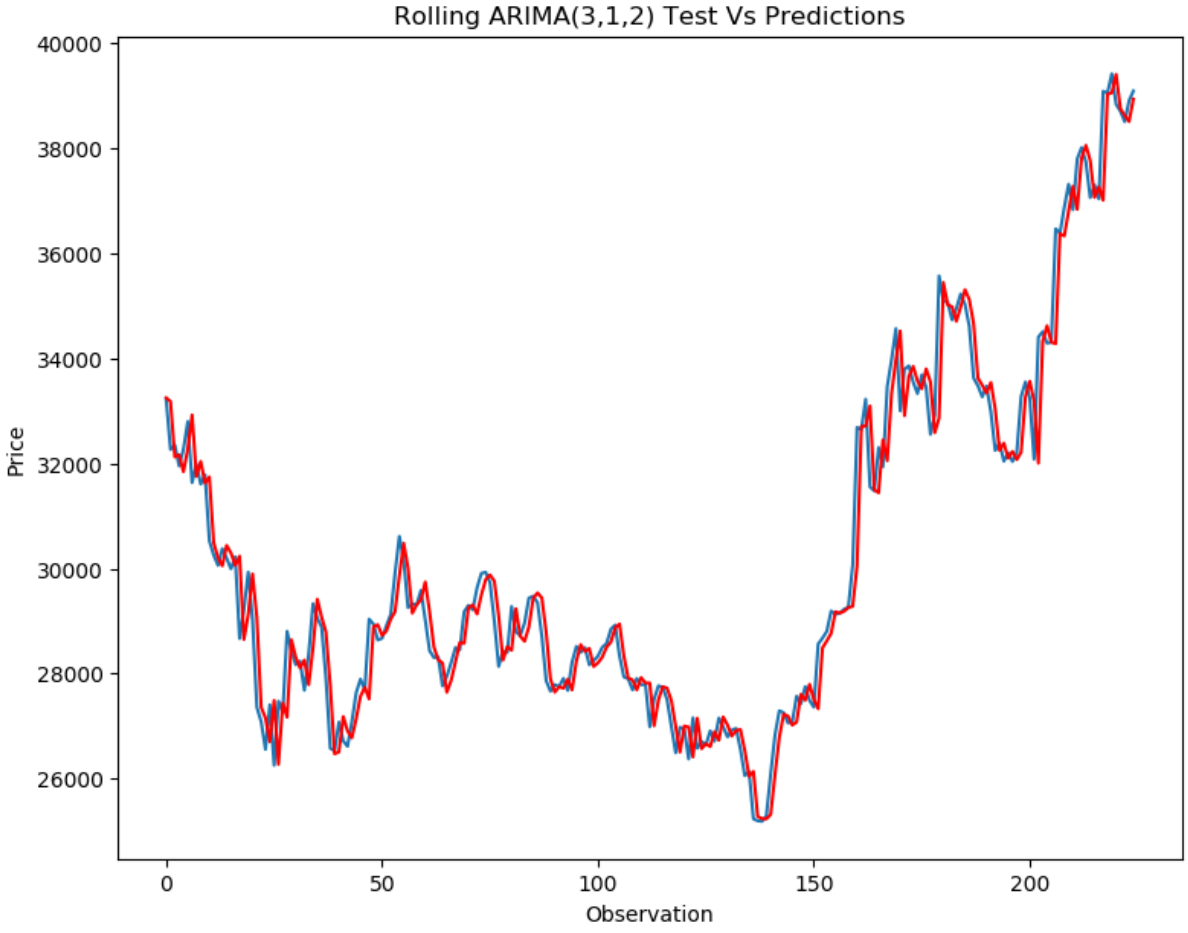
Εικόνα 38: Holts Winters Seasonal Model Fit results.

8.7 ARIMA

Για την εύρεση του καλυτέρου συνδυασμού των παραμέτρων (αυτοπαλινδρόμηση *p,* διαφόριση *q,* κινητός μέσος *d)* του μοντέλου έγινε η χρήση του μοντέλου με διαφορετικούς συνδυασμούς των μεταβλητών του και η αξιολόγηση του με βάση την μετρική *AIC.* Καλύτερα αποτελέσματα είχε το μοντέλο με σειρά (3,1,2).

Η πρόβλεψη με το μοντέλο έγινε με την τεχνική της κυλιόμενης πρόβλεψης. Δηλαδή το μοντέλο μας έκανε πρόβλεψη για την επόμενη παρατήρηση του συνόλου πρόβλεψης (test set) και στην συνέχεια προστίθεται στο τεστ εκπαίδευσης η τιμή αυτή (επόμενη παρατήρηση του συνόλου πρόβλεψης) και το μοντέλο ξαναεκπαιδεύεται και κάνει πρόβλεψη για την επόμενη «άγνωστη» παρατήρηση.

Παρακάτω είναι τα αποτελέσματα της εκπαίδευσης και της πρόβλεψης με την τεχνική αυτή.

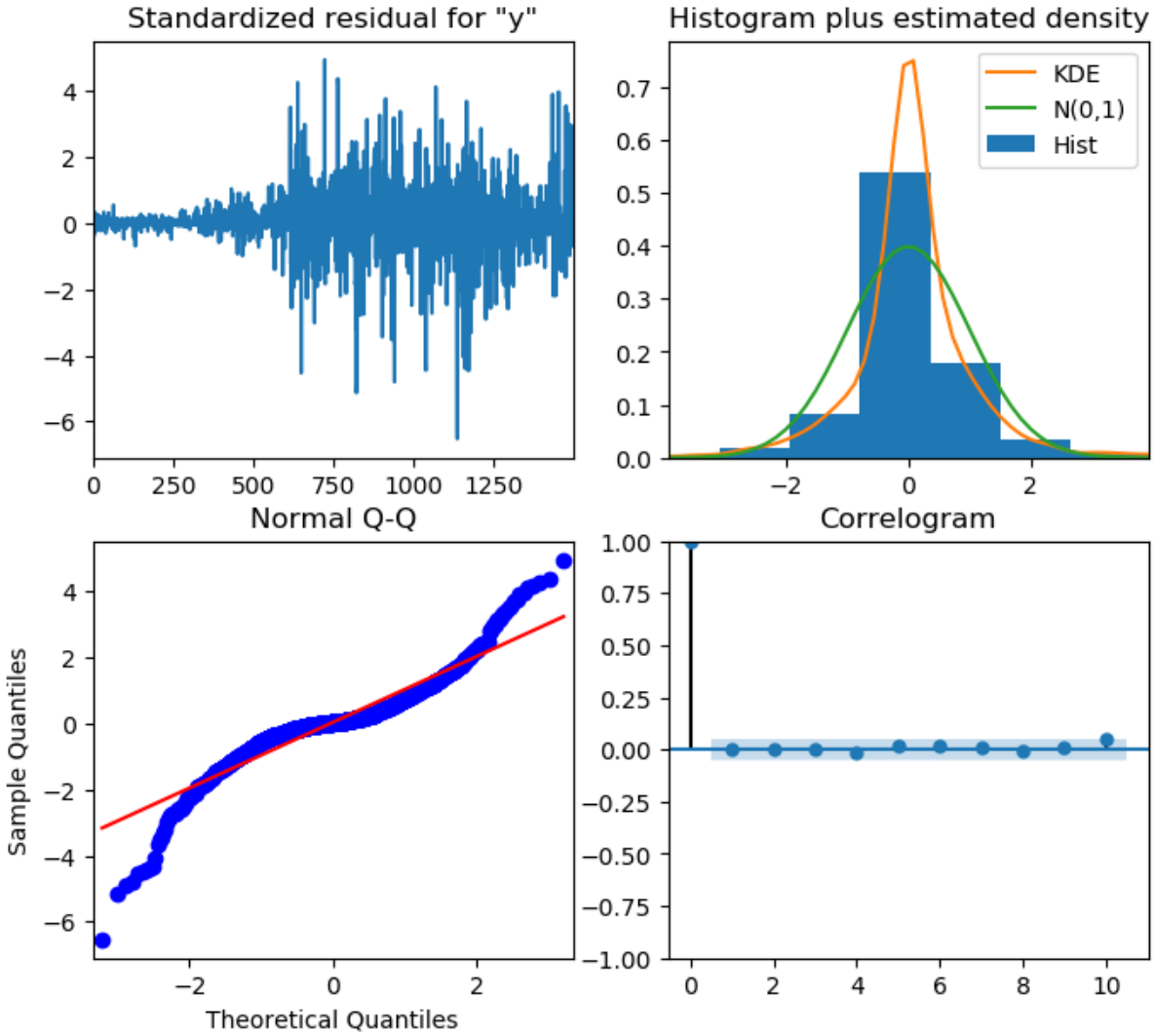


Εικόνα 39: ARIMA(3,1,2) Fit.

Στο παρακάτω διάγραμμα παρατηρούμε:

- Πάνω αριστερά: Η κανονικοποιημένη διασπορά των ακραίων τιμών. Μια ομοιόμορφη διασπορά γύρω από το 0 δείχνει ότι το μοντέλο μας δεν εχει προκατάληψη ως προς το αποτέλεσμα.
- Πάνω δεξιά: Το ιστόγραμμα παρουσιάζει μία κανονική κατανομή γύρω από το μηδέν. Η KDE καμπύλη δείχνει να ακολουθεί την κανονική κατανομή- ένδειξη καλής προσαρμογής του μοντέλου.

- Κάτω αριστερά: Ιδανικά οι παρατηρούμενες τιμές πρέπει να εφαρμόζουν στην κόκκινη γραμμή. Οι αποκλίσεις από την κόκκινη γραμμή είναι ένδειξη ασσυμετρίας στην κατανομή.
- Κάτω δεξιά: Το κορελόγραμμα ή αλλιώς διάγραμμα μερικής αυτόσυσχέτισης δείχνει ότι δεν υπάρχει συσχέτιση μεταξύ σφαλμάτων.



Εικόνα 40: Διαγνωστικά διαγράμματα αποτελεσμάτων ARIMA μοντέλου.

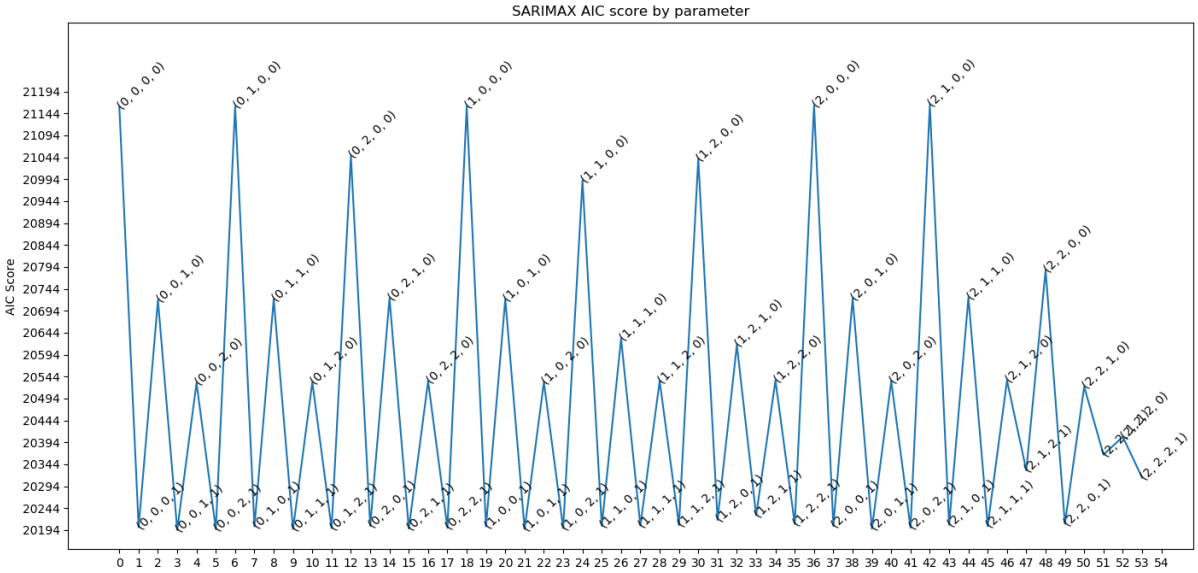
8.8 Εποχικά μοντέλα ARIMA

Τα μοντέλα ARIMA έχουν την δυνατότητα να χειριστούν και δεδομένα τα οποία παρουσιάζουν εποχικότητα. Ένα εποχικό μοντέλο SARIMA αποτελείται από τους όρους που δεν αφορούν την εποχικότητα (p, d, q), και τους εποχικούς όρους $(P, D, Q)m$, όπου m ο αριθμός των παρατηρήσεων ανά χρόνο.

Οι εποχικές συνιστώσες του μοντέλου αποτελούν παρόμοιους όρους με τους όρους από τις μη εποχικές συνιστώσες του μοντέλου, με μετατοπίσεις όμως της εποχικής περιόδου [46].

Οι τιμές των εποχικών όρων μπορούν να παρατηρηθούν μέσα από τις χρονικές υστερήσεις των διαγραμμάτων αυτόσυσχέτισης [1].

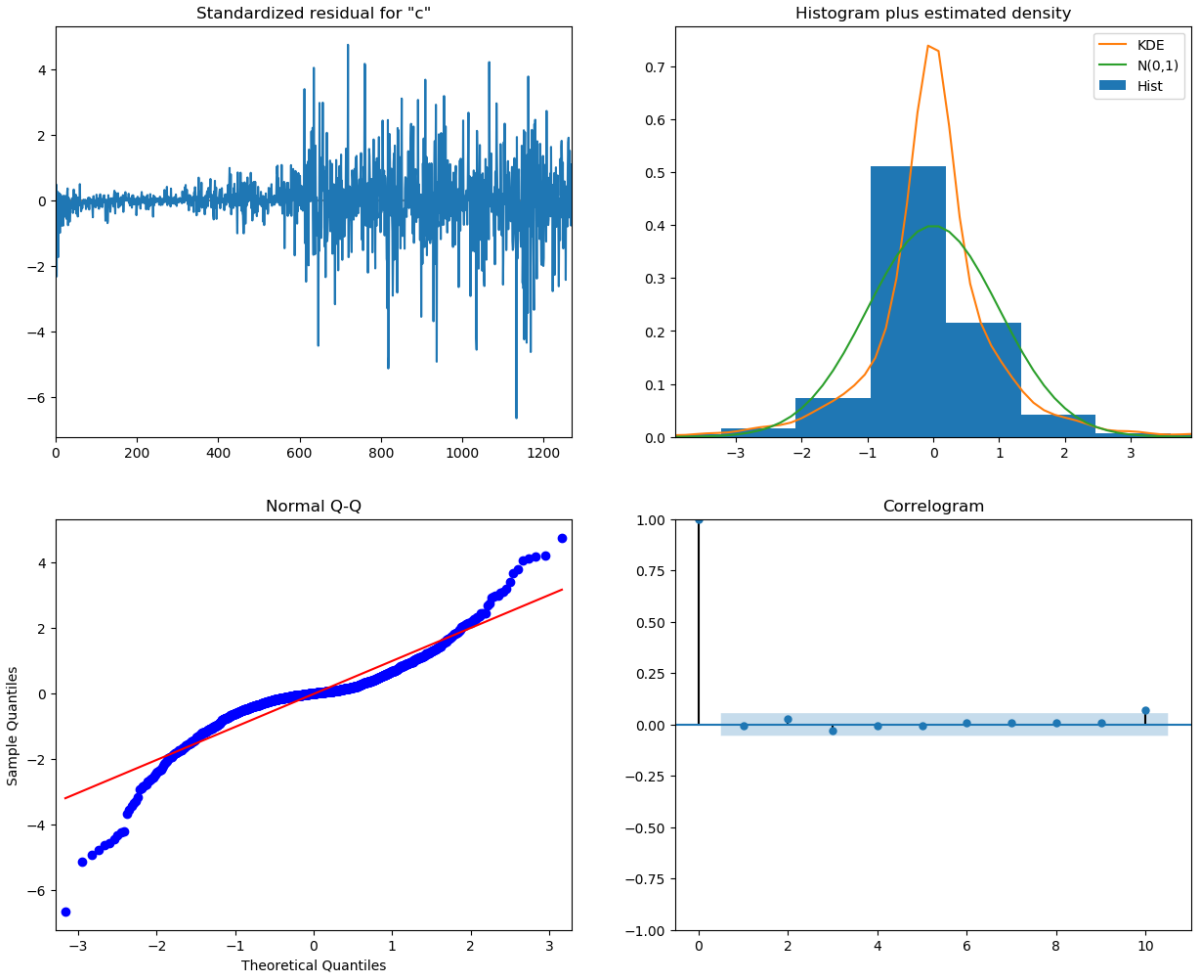
Στην παρακάτω εικόνα παρατηρούμε την τιμή του κριτηρίου *Akaike Information criterion* (AIC), σε σχέση με τους διάφορους εποχικούς όρους που «δοκιμάστηκαν» κατά την εφαρμογή του μοντέλου στα δεδομένα μας (χρονοσειρά κρυπτονομίσματος *BTC*).



Εικόνα 41: Μέτρο AIC ανά συνδυασμό όρων μοντέλου SARIMAX.

Στον διαγνωστικό έλεγχο υπολοίπων παρατηρείτε:

- Πάνω αριστερά: Η κανονικοποιημένη διασπορά των ακραίων τιμών. Μια ομοιόμορφη διασπορά γύρω από το 0 δείχνει ότι το μοντέλο μας δεν έχει προκατάληψη ως προς το αποτέλεσμα.
- Πάνω δεξιά: Το ιστόγραμμα παρουσιάζει μία κανονική κατανομή γύρω από το μηδέν. Η KDE καμπύλη δείχνει να ακολουθεί την κανονική κατανομή- ένδειξη καλής προσαρμογής του μοντέλου.
- Κάτω αριστερά: Ιδανικά οι παρατηρούμενες τιμές πρέπει να εφαρμόζουν στην κόκκινη γραμμή. Οι αποκλίσεις από την κόκκινη γραμμή είναι ένδειξη ασσυμετρίας στην κατανομή.
- Κάτω δεξιά: Το κορελόγραμμα ή αλλιώς διάγραμμα μερικής αυτόσυσχέτισης δείχνει ότι δεν υπάρχει συσχέτιση μεταξύ σφαλμάτων.



Εικόνα 42: Διαγνωστικά διαγράμματα αποτελεσμάτων SARIMA μοντέλου.

8.9 LSTM

Με την βοήθεια των λογισμικών Keras και Tensorflow στη Python δημιουργήθηκε ένα διαδοχικό μοντέλο LSTM νευρωνικού δικτύου.

Table 6: Αρχιτεκτονική LSTM.

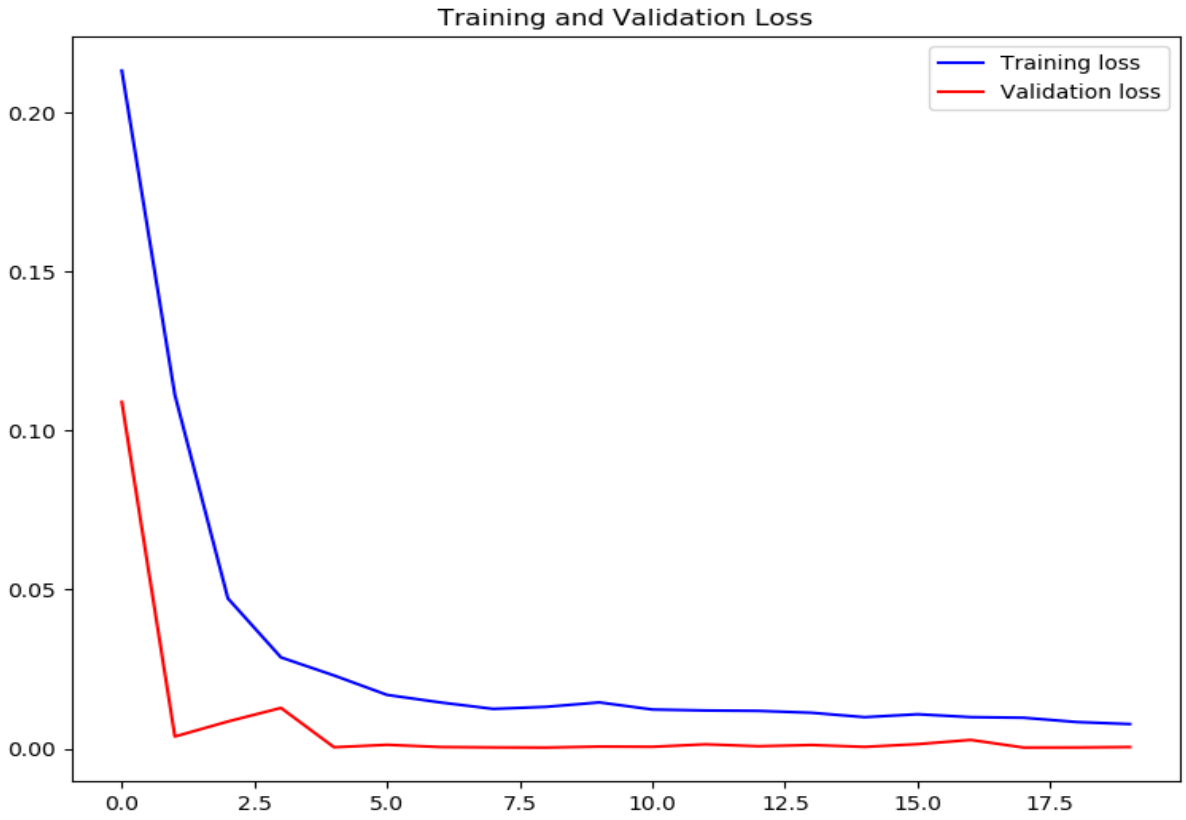
lstm_42 (LSTM)	(None, 35, 120)	62400
dropout_42 (Dropout)	(None, 35, 120)	0
lstm_43 (LSTM)	(None, 35, 80)	64320
dropout_43 (Dropout)	(None, 35, 80)	0
lstm_44 (LSTM)	(None, 35, 60)	33840
dropout_44 (Dropout)	(None, 35, 60)	0
lstm_45 (LSTM)	(None, 20)	6480
dropout_45 (Dropout)	(None, 20)	0
dense_9 (Dense)	(None, 1)	21

Total params: 167,061
Trainable params: 167,061
Non-trainable params: 0

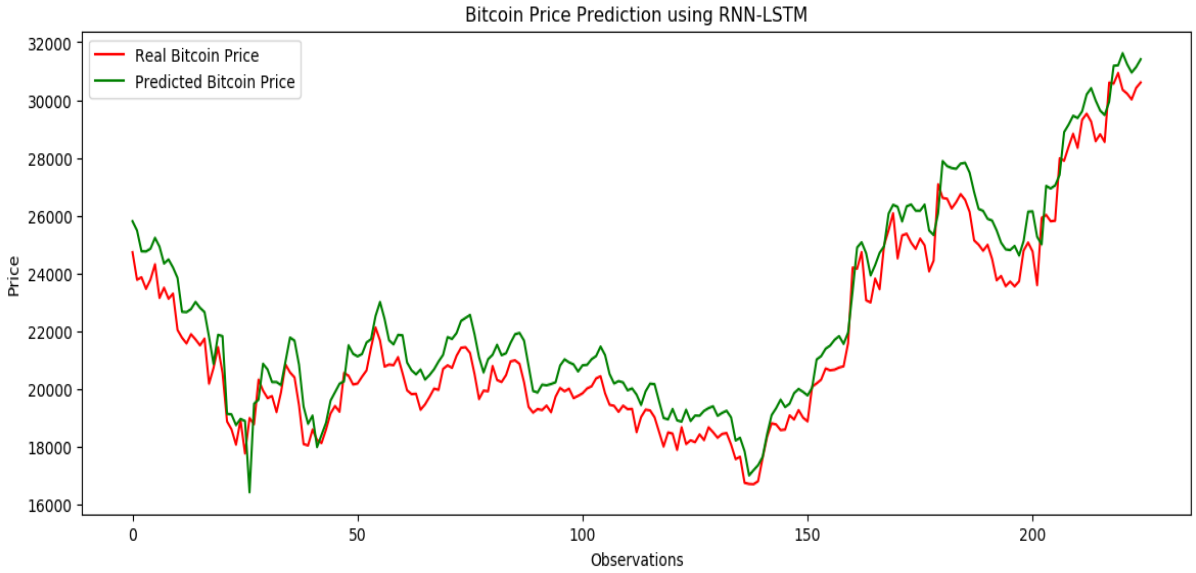
Το διαδοχικό μοντέλο που αναπτύχθηκε αποτελείται από πέντε επίπεδα νευρώνων. Ανάμεσα σε κάθε επίπεδο παρεμβάλετε ένα στρώμα (Dropout) για την αποφυγή του overfitting. Τα δεδομένα κανονικοποιήθηκαν πριν την είσοδο τους στο νευρωνικό δίκτυο.

Τα αποτελέσματα της εκπαίδευσης σε 20 εποχές είναι τα παρακάτω:

Το διάγραμμα σφάλματος μεταξύ συνόλου επικύρωσης και εκπαίδευσης υποδεικνύει ότι το μοντέλο μας δεν παρουσιάζει Overfitting ή Underfitting.



Εικόνα 43: Σφάλμα Εκπαίδευσής και Σφάλμα επικύρωσης μοντέλου.



Εικόνα 44: LSTM fit στο σύνολο πρόβλεψης.

Στον παρακάτω πίνακα παρουσιάζονται τα αποτελέσματα των προβλεπτικών μοντέλων για μεγάλο χρονικό διάστημα (250 περιόδων).

METHOD	R^2	RMSE	MSE	MAE	MAPE
Naïve – Without History	-0.0042	641.666	411736.2	423.455	1.0237
Naïve Short	0.5361	2305.22	5314081.8	590.32	0.019010
MA(20) – Without History	-0.1053	3558.7	12664	3154.6	0.1055
MA(50) – Without History	-0.0002	3385.32	11460	2812.6	0.0910
MA(150) – Without History	-0.0009	3386.5	11468	2828.2	0.0917
MA(5) -next step	0.50	2401.6	5768150	975.0	0.0316
MA(15) -next step	0.07	3350.05	11222843	1687.76	0.0559
MA(20) -next step	-0.1	3730.1	13913944	2046.9	0.068
MA(50) -next step	-2.174	6478.9	41977269	4391.5	0.144
MA(100) -next step	-5.68	10541.8	111131330	9852.8	0.329
MA(150) -next step	-37.66	18572.04	344921040	18483.9	0.549
SARIMAX -next step	0.961	653.1	426636	451.8	0.01
Simple Exponential Smoothing(Alpha = 0.3)	-0.227	4280.6	18324220.3	3669.4	0.1271
Holt-Winters (ADD)	-1.071	4872.1	23738226.1	4223.6	0.147
Holt-Winters (MUL)	-1.907	5772.1	33317850.1	4918.25	0.173
ARIMA(3,1,2)	-0.67267	4377.7	19164701.1	3892.6	0.1345
Rolling ARIMA(3,1,2)	0.9629	646.94	418538.5	447.0	0.0147
LSTM (20 epochs)	0.915	983.35	966992.7	902.95	0.041
LSTM (25 epochs)	0.949	758.1	574814.5	615.78	0.027
LSTM (60 epochs)	0.99	819	671041	636.2	2.7

Ως μία περίοδο ορίζουμε τις έξι ώρες (6 hours).

Μπορούμε να παρατηρήσουμε τις μεγάλες διαφορές όσο αναφορά την αποτελεσματικότητα των μοντέλων σε σχέση με τη παράθυρο του χρόνου πρόβλεψης.

9. Συνένωση πηγών και Συμπεράσματα

Σκοπός της εργασίας είναι η καλύτερη και αποτελεσματικότερη πρόβλεψη της τιμής ενός κρυπτονομίσματος καθώς και παρατήρηση της συμβολής των social media και διάφορων ενημερωτικών blog στην διακύμανση της τιμής αυτής.

- Καθορισμός χαρακτηριστικών,
- Normalization (Κανονικοποίηση)
- Heatmap
- Επιλογή παραθύρου τροφοδότησής (Prediction per window to show effective window)
- Δημιουργία-Επιλογή μοντέλου (Build NN)

9.1 Feature Extraction

Για την καλύτερη αξιοποίηση όλων των χαρακτηριστικών κατά την χρήση όλων των δεδομένων μας συνδυαστικά, μελετήθηκε η σημαντικότητα κάθε χαρακτηριστικού στην τελική πρόβλεψη. Τα διαθέσιμα χαρακτηριστικά είναι:

- Low (χαμηλότερη τιμή περιόδου)
- High (μέγιστη τιμή περιόδου)
- Open (τιμή ανοίγματος περιόδου)
- Close (τιμή κλεισίματος περιόδου)
- Volume_BTC (μέγεθος συναλλαγών περιόδου)
- Rate (ρυθμός μεταβολής τιμής)
- Vader_Tw (δείκτης Vader για το Tweeter)
- Vader_Blog (δείκτης αποτελέσματος Vader για δεδομένα BLOG)
- SVM_Tw (κατηγοριοποίηση αλγορίθμου SVM για το Twitter)
- SVM_Blog (κατηγοριοποίηση αλγορίθμου SVM για δεδομένα BLOG)
- WordNet_Tw (δείκτης WordNet για το Twitter)
- WordNet_Blog (δείκτης WordNet για δεδομένα BLOG)
- Volume_Blog (πλήθος δεδομένων περιόδου-συλλογή Blogs)
- Volume_Tw (πλήθος δεδομένων περιόδου-συλλογή Tw)

Τα αποτελέσματα των χαρακτηρισμένων κειμένων αθροιστήκαν στο εύρος κάθε περιόδου και κανονικοποιήθηκαν. Ως περίοδος ορίζονται οι έξι ώρες.

Τα δεδομένα τα οποία έλειπαν από το σύνολο των δεδομένων μας “missing values” , περίοδοι για τους οποίους δεν είχαμε κειμενικά δεδομένα (tweets/ Headlines), αντικαταστάθηκαν με την διάμεση τιμή του αντίστοιχου χαρακτηριστικού.

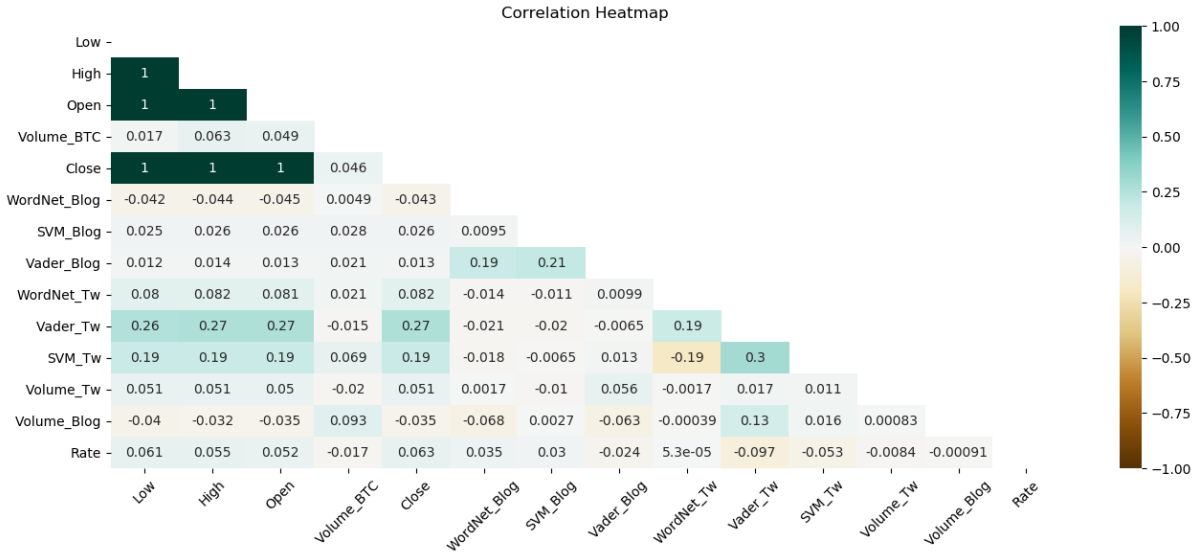
9.1.1 Επιλογή Χαρακτηριστικών

Η επιλογή χαρακτηριστικών επικεντρώνεται κυρίως στην απομάκρυνση των χαρακτηριστικών τα οποία δεν περιέχουν ή περιέχουν πλεονάζουσα πληροφορία για το μοντέλο μας [44] .

Η επιλογή χαρακτηριστικών μπορεί να γίνει με πολλούς τρόπους. Κάποιοι από αυτούς χρησιμοποιούν στατιστικές τεχνικές για να αξιολογήσουν την σχέση κάθε χαρακτηριστικού με το χαρακτηριστικό στόχο και οι στατιστικές αυτές τιμές συμβάλουν στην απόφαση της διατήρησης-διαγραφής. Ο έλεγχος της γραμμικής συσχέτισης αποτελεί έναν τέτοιο τρόπο.

9.1.2 Έλεγχος Γραμμικής συσχέτισης Χαρακτηριστικών

Υπολογίστηκε ο βαθμός της συσχέτισης Pearson για τον έλεγχο της γραμμικής συσχέτισης των χαρακτηριστικών.



Εικόνα 45: Πίνακας συσχέτισης Pearson.

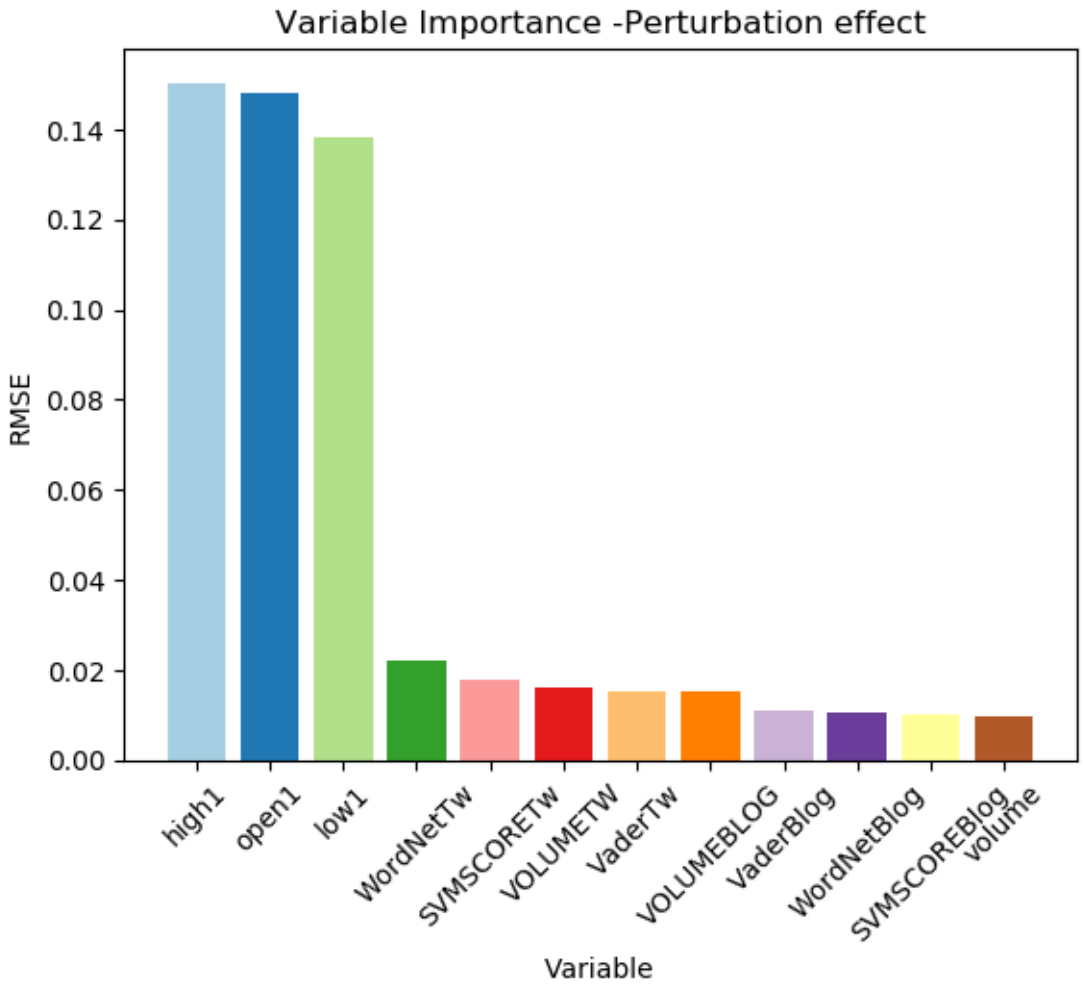
Στον παραπάνω πίνακα παρατηρούμε την γραμμική συσχέτιση των χαρακτηριστικών. Οι τιμές «ανοίγματος» (Open), «υψηλότερη» (high) και «χαμηλότερη» (low) τιμή φαίνεται ότι έχουν μεγαλύτερη γραμμική συσχέτιση με την τιμή στόχο (close). Παρατηρείτε επίσης συσχέτιση μεταξύ την κατηγοριοποίηση του συναισθήματος των δεδομένων που προέρχονται από το Twitter.

9.1.3 Σημαντικότητα Χαρακτηριστικών

Για να μετρήσουμε την σημαντικότητα των χαρακτηριστικών έγινε πρόβλεψη με χρήση του LSTM μοντέλου με ένα μέρος του συνόλου δεδομένου (500 παρατηρήσεις). Στην συνέχεια, σε κάθε ένα χαρακτηριστικό (κάθε φορά σε διαφορετικό) προστέθηκε ένας τυχαίος αριθμός από μια κανονική κατανομή. Τέλος, μετρήθηκε ο βαθμός στον οποίο επηρεάζει κάθε χαρακτηριστικό την πρόβλεψη. Για την μέτρηση αυτή χρησιμοποιήθηκε η ρίζα μέσου τετραγωνικού σφάλματος.

Όσο μεγαλύτερη η επίδραση τόσο ποιο «σημαντικό» θεωρείτε το χαρακτηριστικό για το μοντέλο μας.

Στον παρακάτω πίνακα παρατηρούμε την σημαντικότητα που προκύπτει στα χαρακτηριστικά μας:



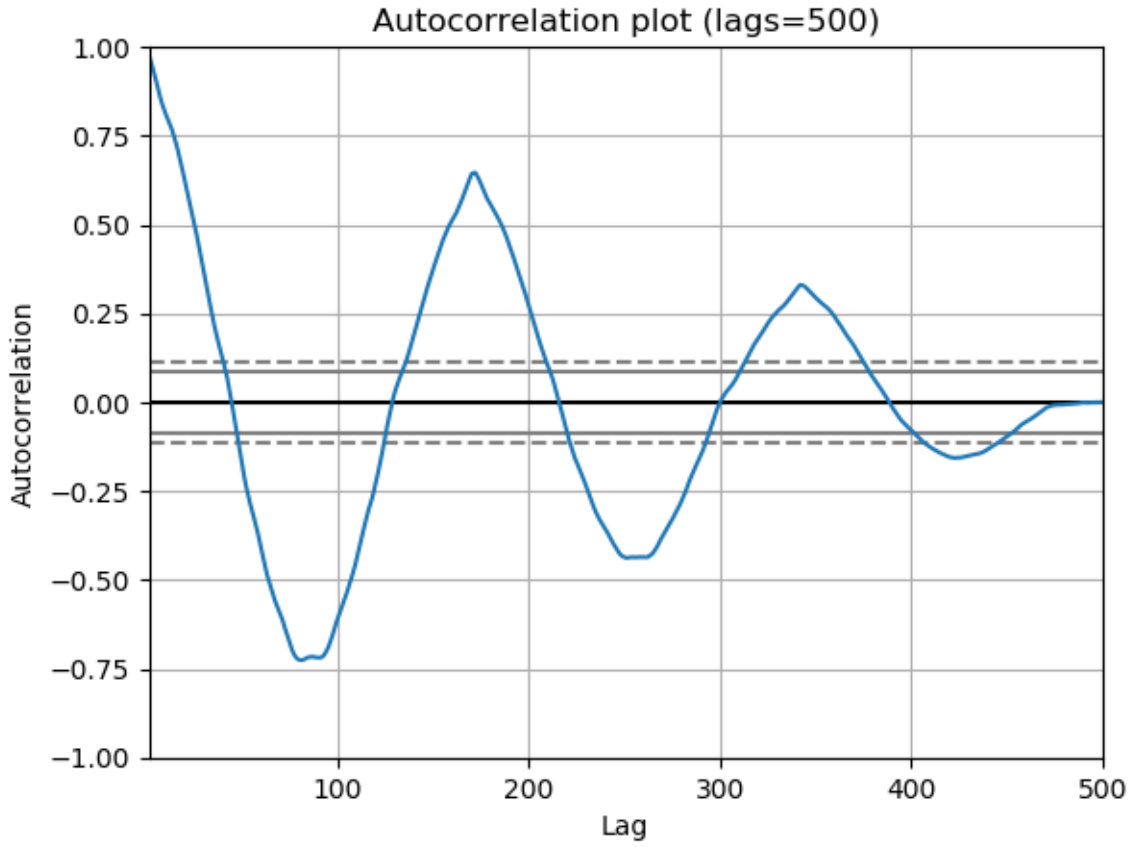
Εικόνα 46: Σημαντικότητα χαρακτηριστικών μέθοδος παραμετροποίησης.

Η παραπάνω εικόνα δείχνει πως η μέγιστη τιμή του κρυπτονομίσματος (χρονική περίοδος 6 hours), καθώς και οι τιμές «ανοίγματος» και η μικρότερη τιμή του νομίσματος (χρονική περίοδος 6 hours) επηρεάζουν σε μεγάλο βαθμό το μοντέλο μας. Στην συνέχεια ακολουθούν τα χαρακτηριστικά τα οποία σχετίζονται με την κατηγοριοποίηση των δεδομένων του Twitter κάτι που επαληθεύεται από τον πίνακα γραμμικής συσχέτισης. Τέλος, ακολουθούν τα χαρακτηριστικά που σχετίζονται με την κατηγοριοποίηση του συναισθήματος των δεδομένων που προέρχονται από τα blogs.

Συνδυάζοντας την γραμμική συσχέτιση των χαρακτηριστικών, τον βαθμό επιρροής των χαρακτηριστικών στο μοντέλο, καθώς και μετά από σειρά πειραμάτων αφαιρέθηκαν από το σύνολο δεδομένων μας τα χαρακτηριστικά WordNetBlog, Volume και SVMSCOREBlog.

9.2 Καθορισμός Χρονικού Πλαισίου

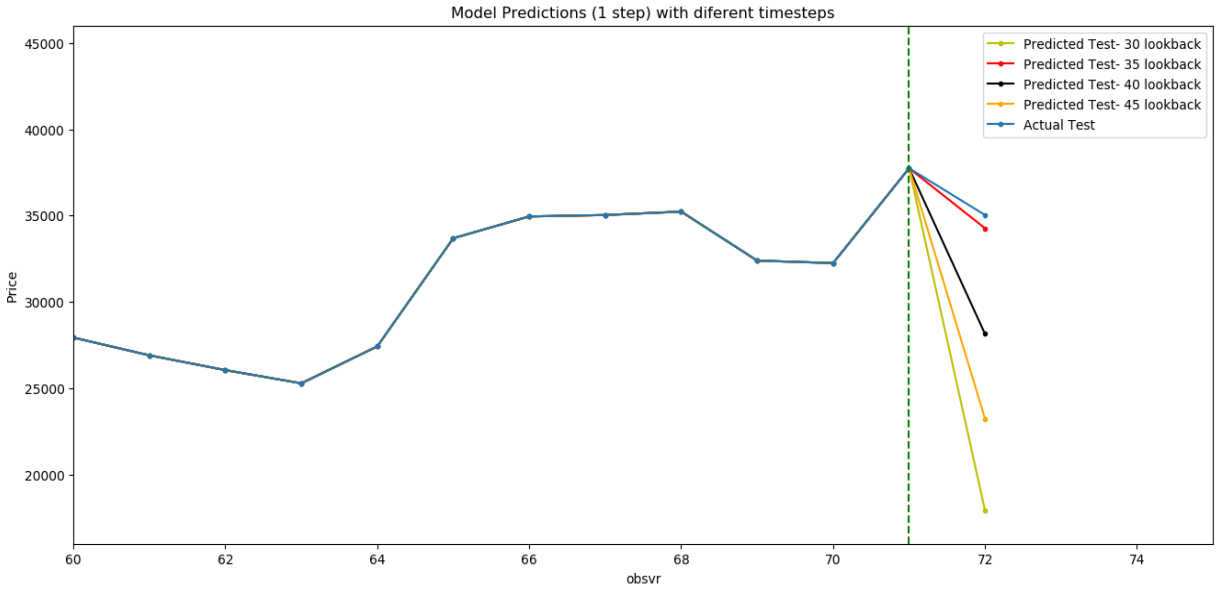
Μια από τις σημαντικότερες παραμέτρους των επαναληπτικών Νευρωνικών δίκτυών LSTM είναι ο καθορισμός του χρονικού παραθύρου (timesteps) που θα χρησιμοποιηθεί. Τα timesteps συμβάλλουν στην ανακάλυψη τόσο μακροχρόνιων όσο και βραχυχρόνιων εξαρτήσεων στα δεδομένα και αποτελεσματικά στην καλύτερη πρόβλεψη.



Εικόνα 47: Autocorrelation Btc.

Για την εύρεση του κατάλληλου παραθύρου που θα χρησιμοποιηθούν μελετήθηκε η αυτοσυσχέτιση των τιμών του κρυπτονομίσματος.

Ως στατιστικά σημαντικές παρελθοντικές παρατηρήσεις θεωρούμε εκείνες τις παρατηρήσεις των οποίων ο συντελεστής αυτόσυσχέτισης βρίσκεται πάνω από την αχνή γραμμή του διαγράμματος. Αυτό σημαίνει ότι υπάρχει συσχέτιση των παρατηρήσεων με τις προηγούμενες 50 παρατηρήσεις. Για τον λόγο αυτόν μελετήθηκαν οι τιμές 30, 35, 40 και 45, με την τιμή των 35 να μας δίνει τα καλύτερα αποτελέσματα στη πρόβλεψη της επόμενης τιμής.



Εικόνα 48: Πρόβλεψη επόμενης τιμής με χρήση διαφορετικών timesteps.

Μετα από πειράματα παρατηρούμε ότι για την πρόβλεψη της επόμενης τιμής το μοντέλο μας δίνει καλύτερα αποτελέσματα όταν τροφοδοτείτε με πληροφορία από τις προηγούμενες 35 παρατηρήσεις για κάθε ένα από τα 9 χαρακτηριστικά.

9.3 Καθορισμός Υπερπαραμέτρων

Η απόδοση των Νευρωνικών δικτύων και συνεπώς και του LSTM εξαρτάται από τις τιμές των Υπερπαραμέτρων τους. Για την κατασκευή του βέλτιστου μοντέλου εξετάστηκαν και οι παρακάτω παράμετροι:

- LSTM units
- LSTM layers
- Dropout
- Learning Rate
- Batch size
- Epochs
- optimizer

Σαν μετρική (loss function) χρησιμοποιήθηκε το mean squared error (mse) και ως συνάρτηση ενεργοποίησης η «relu».

Το βέλτιστο σύνολο υπερπαραμέτρων μετά από σειρά πειραμάτων, και είναι το ακόλουθο:

Table 7: Αρχιτεκτονική τελικού LSTM μοντέλου.

```
Model: "sequential_13"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm_62 (LSTM)               (None, 35, 20)            2400
_________________________________________________________________
dropout_62 (Dropout)         (None, 35, 20)            0
_________________________________________________________________
lstm_63 (LSTM)               (None, 35, 80)            32320
_________________________________________________________________
dropout_63 (Dropout)         (None, 35, 80)            0
_________________________________________________________________
lstm_64 (LSTM)               (None, 35, 100)           72400
_________________________________________________________________
dropout_64 (Dropout)         (None, 35, 100)           0
_________________________________________________________________
lstm_65 (LSTM)               (None, 35, 120)           106080
_________________________________________________________________
dropout_65 (Dropout)         (None, 35, 120)           0
_________________________________________________________________
lstm_66 (LSTM)               (None, 35, 240)           346560
_________________________________________________________________
dropout_66 (Dropout)         (None, 35, 240)           0
_________________________________________________________________
dense_13 (Dense)             (None, 35, 1)             241
=================================================================
Total params: 560,001
Trainable params: 560,001
Non-trainable params: 0
```

Το νευρωνικό αναπτύχθηκε και εκπαιδεύτηκε με την βοήθεια της βιβλιοθήκης tensorflow.keras. Κατά την εκπαίδευση του δικτύου χρησιμοποιήθηκε [48]:

- «Early stopping» μέθοδος του API Keras. Η Early Stopping είναι μέθοδος η οποία σταματάει την εκπαίδευση του νευρωνικού όταν η αποτελεσματικότητα του μοντέλου δεν βελτιώνεται στο σύνολο επικύρωσης.
- «ReduceLROnPlateau» μέθοδος του API Keras. Η μέθοδος αυτή παρακολουθεί μία ποσότητα (σφάλμα εκπαίδευσης - training loss) και μειώνει τον βαθμό μάθησης (learning rate) όταν η ποσότητα αυτή σταματήσει να μειώνετε (ή αυξάνεται ανάλογα με την ποσότητα που παρακολουθείτε).
- «ModelCheckpoint» μέθοδος του API Keras. Η μέθοδος αυτή επιτρέπει την αποθήκευση των βαρών “weights” του μοντέλου βασισμένο σε μία ποσότητα. Στα πειράματα για τον σκοπό της εργασίας η ποσότητα αυτή ήταν το σφάλμα εκπαίδευσης (training loss).

Στον παρακάτω πίνακα παρουσιάζονται τα βέλτιστα σύνολα υπερπαραμέτρων.

Table 8: Καθορισμός Υπερπαραμέτρων LSTM.

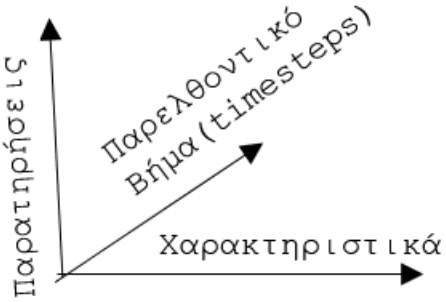
Υπερπαράμετρος	Σύνολο Πειραμάτων	Βέλτιστο
LSTM layers	1,2,3,4,5,6	5
LSTM units	20,60,80,100,120,240	20,80,100,120,240
Dropout	0.1 ,0.2 ,0.3	0.2
Batch size	8,16,32,64	32
optimizer	“adam”, “rmsprop”	adam
Learning Rate	0.0001, 0.01, 0.05, 0.1	0.01

9.4 Μετασχηματισμός Δεδομένων

Ο αλγόριθμος LSTM απαιτεί ως είσοδο δεδομένα τρισδιάστατης μορφής. Έτσι το σύνολο δεδομένων μετασχηματιστικέ ως εξής:

X train	(1333, 35, 9)	Y train	(1333, 1)
X validation	(260, 35, 9)	Y validation	(260, 1)
X test	(70, 35, 9)	Y test	(70,1)

Το σύνολο αποτελείτε από 1663 παρατηρήσεις οι οποίες αποτελούνται από 35 λίστες, 9 χαρακτηριστικών.



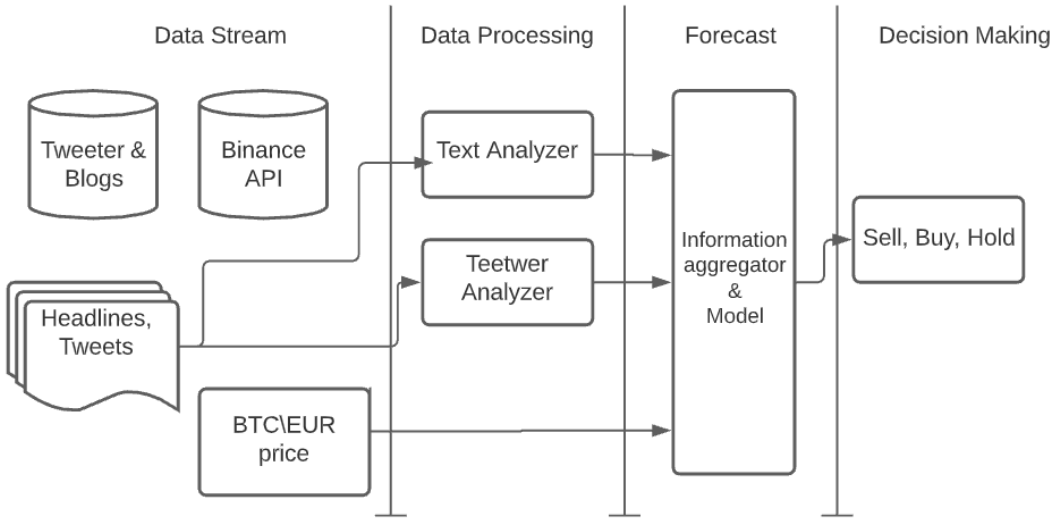
Εικόνα 49: 3-D Δεδομένα Εισόδου.

9.5 Αυτοματοποίηση της Διαδικασίας

Η παραπάνω ανάλυση της χρονοσειράς καθώς και των κειμένων εκτός από την εξαγωγή συμπερασμάτων και την καλύτερη κατανόηση του πεδίου μπορεί να αξιοποιηθεί και να εφαρμοστεί σε προβλέψεις του πραγματικού κόσμου. Για τον σκοπό αυτό με την χρήση του API του tweeter καθώς και του API της πλατφόρμας BINANCE επιτεύχθηκε η δημιουργία αυτοματοποιημένης διαδικασίας για την εξαγωγή και η αποθήκευση των δεδομένων. Επιπλέον, με χρήση των TASK SERVICES των WINDOWS εκτελείτε κώδικας python ο οποίος λαμβάνει δεδομένα από ιστοσελίδες.

Η προσωρινή αποθήκευση των δεδομένων γίνετε σε SQL Server Management Studio.

Στο παρακάτω διάγραμμα απεικονίζεται το *flow* της διαδικασίας.



Εικόνα 50: Διάγραμμα (flow chart) διαδικασίας.

Ορίστηκαν τρείς καταστάσεις – κατηγορίες για μετάφραση της πρόβλεψης που πραγματοποιήθηκε:

- *Sell*, αφορά την πώληση του κρυπτονομίσματος,
- *Buy*, αφορά την αγορά του νομίσματος,
- *Hold*, κατάσταση αδράνειας.

Έχοντας, εκ των προτέρων προβλέψει τις τιμές κλεισίματων του κρυπτονομίσματος ανά έξι ώρες, μπορούμε να χρησιμοποιήσουμε αυτές τις προβλέψεις για να πάρουμε «έξυπνες» αποφάσεις σχετικά με την αφορά ή την πώληση του νομίσματος.

9.6 Αποθήκευση Δεδομένων και Σχήμα Βάσης

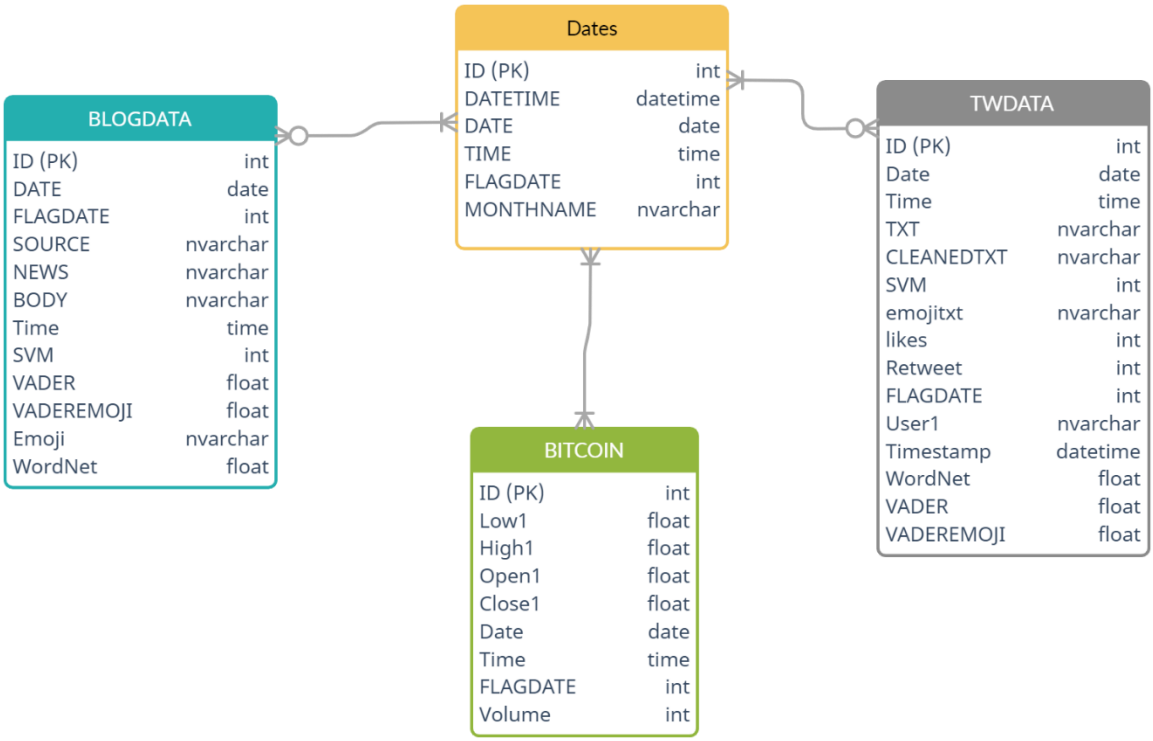
Για την αποθήκευση των αναγκαίων για την εκπαίδευση και πρόβλεψη δεδομένων χρησιμοποιήθηκε βάση δεδομένων SQL Server Management Studio 2018. Για σκοπούς της εργασίας αποθηκευτήκαν δεδομένα που αφορούν την τιμή του κρυπτονομίσματος Bitcoin σε περίοδο 6 ωρών, τα δεδομένα της πλατφόρμας Tweeter καθώς και τα κειμενικά δεδομένα των Blog (μέθοδος web-scrapping).

Το λογικό σχήμα της βάσης δημιουργήθηκε με το εργαλείο https://app.creately.com/ [47].

Στην βάση δημιουργήθηκαν αντικείμενα SQL Server Objects τα οποία φροντίζουν για

- την διαγραφή της διπλότυπης πληροφορίας
- την ενημέρωση των νέων τιμών στο πεδίο FLAGDATE το οποίο βασισμένο στην ώρα κατηγοριοποιεί κάθε ημέρα σε 4 υποσύνολα [0,1,2,3].

Η ενημέρωση της πληροφορίας από και προς την βάση έγινε με χρήση της βιβλιοθήκης pyodbc.



Εικόνα 51: Λογικό σχήμα SQL Server βάσης Δεδομένων.

Τα δεδομένα διαγράφηκαν μετά το πέρας της εργασίας.

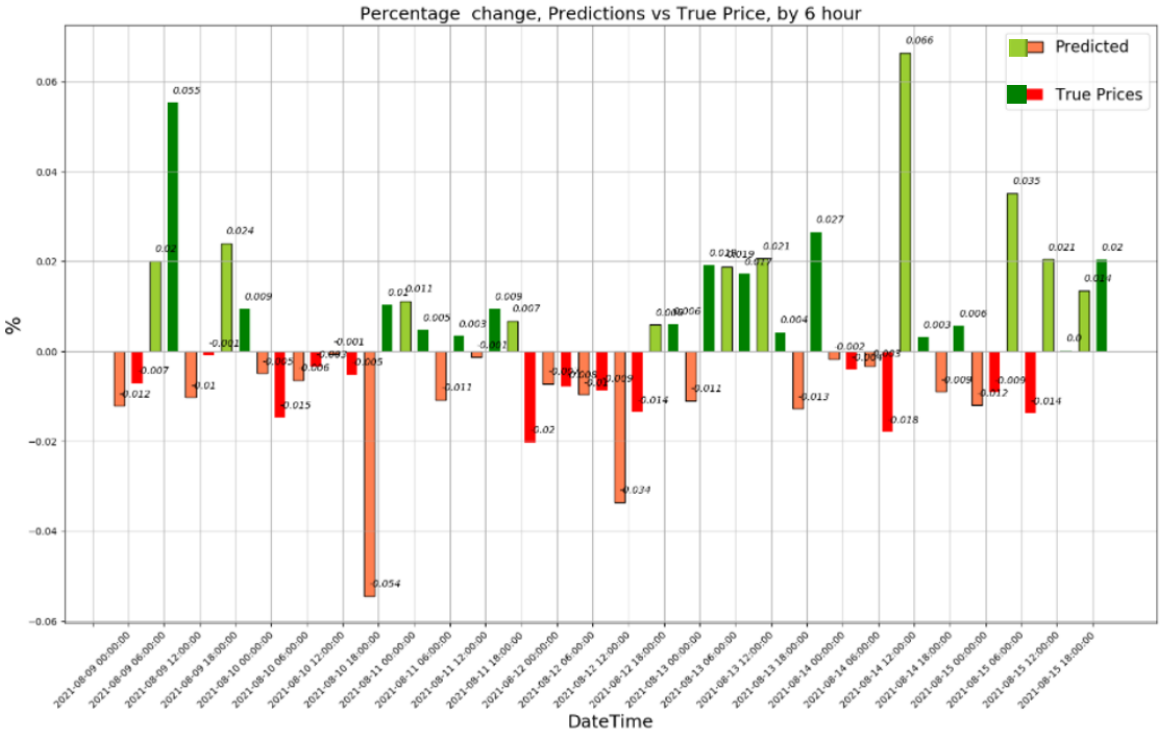
9.7 Εφαρμογή και Αποτελέσματα

Η πρόβλεψη της τιμής για χρονικό διάστημα των επόμενων ωρών αποτελεί ένα πρόβλημα ο οποίο έχει μελετηθεί εκτενώς στην βιβλιογραφία καθώς η γνώση αυτή μπορεί να αποφέρει σημαντικά οικονομικά, κοινωνικά και ενεργειακά οφέλη. Ωστόσο η ακριβής πρόβλεψη μίας τιμής είναι ένα αρκετά δύσκολο πρόβλημα και για τον λόγο αυτό το πρόβλημα μετατοπίζεται στο αν η τιμή αυτή αυξάνεται ή μειώνετε.

Η περίοδος πρόβλεψης αποτελείτε από 28 παρατηρήσεις και διήρκησε από 2021-08-09 00:00:00 έως 2021-08-15 18:00:00.

Κατά την περίοδο πρόβλεψης συλλέχθηκαν 1257 δεδομένα από την πλατφόρμα Tweeter και 51 κειμενικά δεδομένα από τα Blog.

Στην παρακάτω εικόνα παρατηρούμε το ποσοστό μεταβολής τη τιμής του κρυπτονομίσματος και τις τιμής πρόβλεψης.



Εικόνα 52: Ποσοστό μεταβολής τιμής (Πραγματική - LSTM).

Η αποτελεσματικότητα του αλγορίθμου εάν θεωρήσουμε ότι το πρόβλημα αποτελεί πρόβλημα κατηγοριοποίησης είναι *Acc* = 71.4%

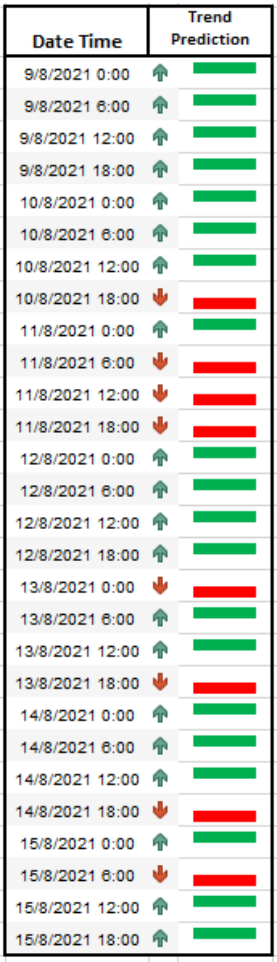
Στην διπλανή εικόνα με πράσινο και βελάκι ανόδου αναπαρίστανται οι παρατηρήσεις (ημέρα-ώρα) κατά τις οποίες το δίκτυο πρόβλεψε σωστά την κίνηση της τάσης της χρονοσειράς. Με κόκκινο και καθοδικό βελάκι έχουν σημαδευτεί οι παρατηρήσεις στις οποίες η πρόβλεψη του δικτύου είχε διαφορετική τάση από την πραγματική τιμή.

Παρατηρήθηκε ότι η αποτελεσματικότητα αυξάνεται όταν η περίοδος πρόβλεψης είναι μεγαλύτερη της μίας περιόδου.

Η αύξηση της πολυπλοκότητας του δικτύου δείχνει να αυξάνει την αποτελεσματικότητα (καλύτερη πρόβλεψη ανοδικής ή καθοδικής πορείας της τιμής). Ενδεικτικά, χρησιμοποιώντας όλες τις προαναφερθείσες υπερπαραμέτρους του δικτύου προκύπτει,

Αρχιτεκτονική (Layers)	Accuracy (Test Set)
[120,80,60,60,20]	46%
[120,100,80,60,20]	50%
[240,120,100,60,20]	64.2%

Όπως φαίνεται στον παραπάνω πίνακα, η αποτελεσματικότητα τόσο στην εκπαίδευση όσο και στην πρόβλεψη δείχνει να βελτιώνετε με την πρόσθεση επιπέδων (Layers). Η αύξηση της αποτελεσματικότητας δεν σημαίνει αναγκαστικά την βελτίωση του νευρωνικού μιας και η αύξηση της πολυπλοκότητας του δικτύου μειώνει την μεροληψία του (bias).



Εικόνα 53: Αποτέλεσμα Πρόβλεψής της τάσης ανά ημέρα.

9.8 Συμπεράσματα

Το πεδίο της ανάλυσης και πρόβλεψης της τιμής κρυπτονομισμάτων έχει ασχοληθεί εκτεταμένα με την εξερεύνηση μοντέλων που πετυχαίνουν μεγαλύτερη αποτελεσματικότητα και παραγόντων που μπορούν να επηρεάζουν την τιμή τους. Η παρούσα εργασία αξιοποίησε σε αρκετές περιπτώσεις την πληθώρα της πληροφορίας που υπάρχει στο συγκεκριμένο πεδίο.

Στην παρούσα εργασία επιχειρήθηκε:

- Η δημιουργία και αξιοποίηση κώδικα για την συλλογή δεδομένων από διαφορετικές πηγές του διαδικτύου και την δημιουργία συνόλων δεδομένων εκπαίδευσης και πρόβλεψης.

- Η αξιοποίηση διαφορετικών εργαλείων που αποσκοπούν στην ανάλυση του συναισθήματος. Συγκεκριμένα, χρησιμοποιήθηκαν εργαλεία μηχανικής μάθησης (Support Vector Machines) καθώς και τεχνικές βασισμένες σε Λεξικό (VADER) καθώς και Corpus-Based (WordNet)
- Η δημιουργία λίστας λέξεων-φράσεων βασισμένη σε λέξεις οι οποίες εμφανίζονται συχνά σε κειμενικά δεδομένα που αφορούν την αγορά των κρυπτονομισμάτων με σκοπό την δημιουργία κανόνων για την ενίσχυση των τεχνικών ανάλυσης συναισθήματος (μέθοδος ανταμοιβής-τιμωρίας)
- Πειράματα τα οποία είχαν ως σκοπό την ανάδειξη του βέλτιστου συνόλου υπερπαραμέτρων για την αποτελεσματικότερη πρόβλεψη της τιμής του κρυπτονομίσματος
- Ανάλυση και επιλογή των σημαντικότερων χαρακτηριστικών για την εκπαίδευση του μοντέλου βασισμένη τόσο σε στατιστικές μεθόδους όσο και σε πειράματα (μέθοδος παραμετροποίησης)
- Η δημιουργία, διαχείριση και συντήρηση βάσης δεδομένων (SQL Server Management Studio) για την αποθήκευση και καλύτερη αξιοποίηση των δεδομένων για σκοπούς της εργασίας
- Η πρόβλεψη της τιμής για την χρονική περίοδο μίας εβδομάδας σε φάσμα έξι ωρών. Για την απλοποίηση του προβλήματος έγινε η πρόβλεψη της τάσης της χρονοσειράς.
- Στο πλαίσιο της εργασίας δημιουργήθηκε μία αυτοματοποιημένη διαδικασία για την εκτέλεση ενός ολόκληρου κύκλου, από την συλλογή, την ανάλυση και την αποθήκευση έως και την πρόβλεψη της επόμενης τιμής του κρυπτονομίσματος.

9.8.1 Μελλοντικές επεκτάσεις

Σαν μελλοντική κατεύθυνση έρευνας το σύστημα μπορεί να τροποποιηθεί ώστε να λαμβάνει υπόψιν και στατιστικούς δείκτες. Επιπλέον, η συλλογή και ανάλυση μεγαλύτερων συνόλων δεδομένων για την κατηγοριοποίηση του συναισθήματος θα οδηγούσε αδιαμφησβήτητα σε συνολικότερη εικόνα επίδρασης των social media στην αγορά των κρυπτονομισμάτων καθώς και σε μεγαλύτερο σύνολο εκπαίδευσης, που ίσως οδηγήσει σε βελτιστοποίηση του νευρωνικού δικτύου. Τέλος, μια ακόμη κατεύθυνση είναι η τροποποίηση του αλγορίθμου ώστε να πραγματοποιεί real-time προβλέψεις μετά χρήση δεδομένων τα οποία προέρχονται από ζωντανή μετάδοση (live streaming). Κάτι τέτοιο βέβαιά απαιτεί συνεχή επανεκπαίδευση του μοντέλου ώστε να προσαρμόζεται στα νέα δεδομένα.

Βιβλιογραφία

[1] Hyndman, R.J., & Athanasopoulos, G. (2018). *Forecasting: principles and practice*, 2nd edition, OTexts: Melbourne, Australia. OTexts.com/fpp2

[2] NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/, Chapter 6, [viewed 8 Oct 2021]

[3] Halkos, George & Kevork, Ilias. (2005). Το υπόδειγμα τυχαίου περιπάτου με αυτοπαλίνδρομα σφάλματα. pp. 2, 17

[4] Assimakopoulos, V. & Nikolopoulos, K. (2000). *The Theta Model: A Decomposition Approach to Forecasting*. International Journal of Forecasting, 16(4), 521-530

[5] Peter J. Brockwell & Richard A. Davis (2002). *Introduction to Time Series and Forecasting*, Second Edition, Springer, pp. 7-35

[6] Αγγελική Παπάνα (2019),Κεφάλαιο 8. Γραμμικά στοχαστικά μοντέλα (Αυτοπαλίνδρομα μοντέλα), Σημειώσεις για το μάθημα Γραμμικά στοχαστικά μοντέλα, Πολυτεχνική σχολή, Α.Π.Θ. & Οικονομικό Τμήμα, Πανεπιστήμιο Μακεδονίας Ιανουάριος 2019

[7] Μιχαήλ Κυριακίδης (2018*), Τεχνικές Ανάλυσης και Πρόβλεψης Τηλεπικοινωνιακών Αγορών*, Μεταπτυχιακή διατριβή, pp.102-108,204-213

[8] Daniel T. Larose, Chantal D. Larose. (2015*), Data Mining and Predictive Analytics*, Wiley, 2015 (2nd Edition)

[9] Peña, D., Tiao, G. and Tsay, R., (2001). *A Course in Time Series Analysis*. 1st ed. Canada: A Wiley-Interscience Publication JOHN WILEY & SONS, INC.

[10]Stephanie Glen. "U Statistic: Definition, Different Types; Theil's U" From StatisticsHowTo.com: Elementary Statistics for the rest of us! https://www.statisticshowto.com/u-statistic-theils/

[12] Μιχαήλ Ε. Φιλιππάκης (2019), *Προβλεπτική αναλυτική*, Σημειώσεις για το μάθημα Προβλεπτική αναλυτική, Πανεπιστήμιο Πειραιώς, Ιανουάριος 2020

[13] Κουγιουμτζής Δημήτρης (2019), Ανάλυση Χρωνοσειρών, Σημειώσεις για το μάθημα Χρονοσειρές (ΤΗΜΜΥ), Πολυτεχνική σχολή ΑΠΘ, pp. 60-62

[15] Kostiantyn Kucher, Carita Paradis, Andreas Kerren (2017), *The State of the Art in Sentiment Visualization,* pp. 71-99

[16] Pang, Bo & Lee, Lillian. (2008). *Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval*. 2. 1-135. 10.1561/1500000011.

[17] Bo Pang & Lillian Lee (2008), *Opinion mining and sentiment analysis,* U.S.A, Chapter 4

[18] Keith Norambuena, Brian & Lettura, Exequiel & Villegas, Claudio. (2019). *Sentiment analysis and opinion mining applied to scientific paper reviews.* Intelligent Data Analysis. 23. 191-214. 10.3233/IDA-173807.

[19] Tie-Yan Liu, (2009), *Learning to Rank for Information Retrieval*, Vol. 3: No. 3, pp 225-331, Microsoft Research Asia, Sigma Center, P. R. China

[20] Socher et al., (2013) Socher et al., (2011) He, (2012) Amigó et al., (2013), Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank, pp. 2-4

[21] Hogenboom, A. C., Heerschop, B. M. W. T., Frasincar, F., Kaymak, U., & Jong, de, F. M. G. (2014). Multi-lingual support for lexicon-based sentiment analysis guided by semantics. Decision Support Systems, 62, 43-53. https://doi.org/10.1016/j.dss.2014.03.004

[22] Ekman, P. (1992). *An argument for basic emotions. Cognition and Emotion*, 6(3-4), pp. 169–200

[23] Russell, James & Mehrabian, Albert. (1977). *Evidence for a Three-Factor Theory of Emotions. Journal of Research in Personality.* 11. 273-294. 10.1016/0092-6566(77)90037-X

[24] Mohsen Farhadloo, Erik Rolland (2016), Fundamentals of Sentiment Analysis and Its Applications, Springer International Publishing

[25] Alnawas A. and Arıcı N., "The corpus based approach to sentiment analysis in modern standard Arabic and Arabic dialects: a literature review", Politeknik Dergisi, 21(2): 461-470, (2018).

[26] kumar, vinay. (2019). *Sentiment Analysis Techniques for Social Media Data: A Review*. Chapter 3-4

[27] Zainuddin, N., & Selamat, A. (2014), *Sentiment analysis using Support Vector Machine.* In International Conference on Computer, Communications, and Control Technology (I4CT), pp. 333-337, 2014, IEEE)

[28] Zhang, Lei & Wang, Shuai & Liu, Bing. (2018). *Deep Learning for Sentiment Analysis : A Survey*. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 8. 10.1002/widm.1253.

[29] Γεωργούλη, Α. (2015). Τεχνητή νοημοσύνη. [Προπτυχιακό εγχειρίδιο]. Αθήνα: Κάλλιπος, Ανοικτές Ακαδημαϊκές Εκδόσεις. Διαθέσιμο στο: http://hdl.handle.net/11419/3381

[30] ΠΑΝΟΣ ΑΡΓΥΡΑΚΗΣ, (2001), *ΤΕΧΝΗΤΗ ΝΟΗΜΟΣΥΝΗ – ΕΦΑΡΜΟΓΕΣ: Νευρωνικά Δίκτυα και Εφαρμογές,* ΕΛΛΗΝΙΚΟ ΑΝΟΙΚΤΟ ΠΑΝΕΠΙΣΤΗΜΙΟ, pp. 25,34,187-188

[31] Αλέξανδρος, Ν. Ζάχος, (2021), *Το νευρωνικό δίκτυο LSTM ως μοντέλο βροχής απορροής,* [ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ, ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ], pp. 13-18

[32] FELIX GER*,* (2001)*, Long Short-Term Memory in Recurrent Neural Networks,* [Master Thesis, Leibniz University], Deutschland, pp. 8-10

[33] Dishashree, G., (2017). Fundamentals of Deep Learning – Introduction to Recurrent Neural Networks. [Blog] analyticsvidhya, Available at: https://www.analyticsvidhya.com/blog/2017/12/introduction-to-recurrent-neural-networks/> [Accessed 4 June 2021].

[34] Olah, C., (2015). Understanding LSTM Networks. [Blog] colah's blog, Available at: <http://colah.github.io/posts/2015-08-Understanding-LSTMs/> [Accessed 10 April 2021]

[35] Hochreiter, Sepp & Schmidhuber, Jürgen. (1997*). Long Short-term Memory. Neural computation*. 9. 1735-80. 10.1162/neco.1997.9.8.1735.

[36] Sherratt, F.; Plummer, A.; Iravani, P. *Understanding LSTM Network Behaviour of IMU-Based Locomotion Mode Recognition for Applications in Prostheses and Wearables*. Sensors 2021, 21, 1264.https://doi.org/10.3390/s21041264

[37] Hutto, C. and Gilbert, E. (2014) VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text", Proceedings of the International AAAI Conference on Web and Social Media, 8(1), pp. 216-225. Available at: https://ojs.aaai.org/index.php/ICWSM/article/view/14550 (Accessed: 9 February 2022).

[38] Yasar Kaya, (2018), *Analysis of Cryptocurrency Market and Drivers of the Bitcoin Price: Understanding the price drivers of Bitcoin under speculative environment*, Master Thesis, Stockholm

[39] Hofmann, Markus. (2014). *Sentiment Versus Polarity within Tokens and Sentences*. Chapter 2

[40] Schütze, H., D. Manning, C. and Raghavan, P., 2008. Introduction to Information Retrieval. [ebook] Cambridge: Cambridge University Press, p.1. Available at: <https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html> [Accessed 17 June 2021].

[41] Hreinsson, E.M. and Blöndal, S.P. (2018). *The future of blockchain technology and cryptocurrencies*, [Doctoral dissertation, Reykjavik University]

[42] Gandal, N. and Halaburda, H. (2014). Competition in the cryptocurrency market. Bank of Canada Working Paper, No. 2014-33.

[43] Shmueli, G. and Lichtendahl, K., 2016. Practical time series forecasting with R. 2nd ed. Axelrod Schnall Publishers; 2nd edition, pp.97-100.

[44] Kuhn, M. and Johnson, K., 2019. *Applied predictive modeling*. New York: Springer, pp.487-519.

[45] ΠΟΛΙΤΗΣ Σ. ΑΓΗΣΙΛΑΟΣ (2020) Πρόβλεψη τιμής Ether με χρήση τεχνικών μηχανικής μάθησης, [ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ, ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ], pp. 45-49

[46] Χασιρτζόγλου Μάρκος, (2020), *ΑΝΑΛΥΣΗ ΚΑΙ ΠΡΟΒΛΕΨΗ ΧΡΟΝΟΣΕΙΡΩΝ ΜΕ ΜΟΝΤΕΛΑ ARIMA ΚΑΙ ΕΦΑΡΜΟΓΕΣ,* [ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ, Εθνικό Μετσόβιο Πολυτεχνείο], pp. 24-26

[47] Συμεωνίδης, Π., Γούναρης, Α. (2015). *Βάσεις, αποθήκες και εξόρυξη δεδομένων με τον SQL Server*. [Εργαστηριακός Οδηγός]. Αθήνα: Κάλλιπος, Ανοικτές Ακαδημαϊκές Εκδόσεις. Διαθέσιμο στο: http://hdl.handle.net/11419/276

[48] TensorFlow. 2022. Module: tf.keras | TensorFlow Core v2.8.0. [online] Available at: <https://www.tensorflow.org/api_docs/python/tf/keras/> [Accessed 4 July 2021].

[49] Ευάγγελος Σπηλιώτης, (2018), Ολοκληρωμένα Αυτοπαλινδρομικά Μοντέλα Κινητού Μέσου Όρου (ARIMA), Σημειώσεις για το μάθημα Τεχνικές Προβλέψεων ΣΗΜΜΥ ΕΜΠ, Εθνικό Μετσόβιο Πολυτεχνείο Σχολή Ηλεκτρολόγων Μηχ. κ Μηχ.