ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Κρυπτονομίσματα: Πρόβλεψη τιμής βάσει χρονολογικών σειρών και συναισθηματικής ανάλυσης δεδομένων του Twitter με χρήση μηχανικής μάθησης

Διπλωματική Εργασία

Τσακλίδης Αργύριος

Επιβλέπουσα: Τουσίδου Ελένη

Βόλος 2022

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Κρυπτονομίσματα: Πρόβλεψη τιμής βάσει χρονολογικών σειρών και συναισθηματικής ανάλυσης δεδομένων του Twitter με χρήση μηχανικής μάθησης

Διπλωματική Εργασία

Τσακλίδης Αργύριος

Επιβλέπουσα: Τουσίδου Ελένη

Βόλος 2022

UNIVERSITY OF THESSALY

SCHOOL OF ENGINEERING

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

Cryptocurrencies: Price prediction based on time series and sentiment analysis of Twitter data using machine learning

Diploma Thesis

Tsaklidis Argyrios

Supervisor: Tousidou Eleni

Volos 2022

Εγκρίνεται από την Επιτροπή Εξέτασης:

Επιβλέπουσα **Τουσίδου Ελένη**
Ε.ΔΙ.Π., Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας

Μέλος **Βασιλακόπουλος Μιχαήλ**
Αναπληρωτής Καθηγητής, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας

Μέλος **Φεύγας Αθανάσιος**
Ε.ΔΙ.Π., Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας

Ημερομηνία έγκρισης: 20-2-2022

Ευχαριστίες

Ολοκληρώνοντας την διπλωματική μου εργασία, θα ήθελα αρχικά να απευθύνω θερμές και ειλικρινείς ευχαριστίες στην επιβλέπουσα καθηγήτρια κυρία Ελένη Τουσίδου για την άρτια συνεργασία μας, για την πολύτιμη και έγκαιρη επικοινωνία που είχαμε όλο αυτό το διάστημα και για τις καίριες παρατηρήσεις και διορθώσεις που μου υπέδειξε. Ευχαριστώ επίσης τα υπόλοιπα μέλη της επιτροπής, τον κύριο Βασιλακόπουλο και τον κύριο Φεύγα.

Τέλος, ευχαριστώ βαθύτατα τους γονείς μου που με στηρίζουν με κάθε πιθανό τρόπο, τον αδερφό μου, τους φίλους μου την Ανθή και την Χρύσα για την αμέριστη συμπαράστασή τους σε κάθε δυσκολία και τους είμαι ευγνώμων, γιατί οι στιγμές που μοιραστήκαμε όλα αυτά τα χρόνια έδωσαν την μεγαλύτερη αξία σε αυτό το κεφάλαιο της ζωής μου.

ΥΠΕΥΘΥΝΗ ΔΗΛΩΣΗ ΠΕΡΙ ΑΚΑΔΗΜΑΪΚΗΣ ΔΕΟΝΤΟΛΟΓΙΑΣ ΚΑΙ ΠΝΕΥΜΑΤΙΚΩΝ ΔΙΚΑΙΩΜΑΤΩΝ

«Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, δηλώνω ρητά ότι η παρούσα διπλωματική εργασία, καθώς και τα ηλεκτρονικά αρχεία και πηγαίοι κώδικες που αναπτύχθηκαν ή τροποποιήθηκαν στα πλαίσια αυτής της εργασίας, αποτελεί αποκλειστικά προϊόν προσωπικής μου εργασίας, δεν προσβάλλει κάθε μορφής δικαιώματα διανοητικής ιδιοκτησίας, προσωπικότητας και προσωπικών δεδομένων τρίτων, δεν περιέχει έργα/εισφορές τρίτων για τα οποία απαιτείται άδεια των δημιουργών/δικαιούχων και δεν είναι προϊόν μερικής ή ολικής αντιγραφής, οι πηγές δε που χρησιμοποιήθηκαν περιορίζονται στις βιβλιογραφικές αναφορές και μόνον και πληρούν τους κανόνες της επιστημονικής παράθεσης. Τα σημεία όπου έχω χρησιμοποιήσει ιδέες, κείμενο, αρχεία ή/και πηγές άλλων συγγραφέων, αναφέρονται ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή. Δηλώνω επίσης ότι τα αποτελέσματα της εργασίας δεν έχουν χρησιμοποιηθεί για την απόκτηση άλλου πτυχίου. Αναλαμβάνω πλήρως, ατομικά και προσωπικά, όλες τις νομικές και διοικητικές συνέπειες που δύναται να προκύψουν στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δεν μου ανήκει διότι είναι προϊόν λογοκλοπής».

Ο/Η Δηλών/ούσα

Τσακλίδης Αργύριος

Περίληψη

Η πρόβλεψη της τιμής ενός κρυπτονομίσματος μπορεί να αποτελέσει μια απαιτητική δοκιμασία μιας και οι παράγοντες που την καθορίζουν είναι πολλοί και ιδιαίτερα ασαφείς. Ο κύριος στόχος της παρούσας πτυχιακής είναι η πρόβλεψη της εξέλιξης των τιμών δημοφιλών κρυπτονομισμάτων, με έμφαση στο δημοφιλέστερο αυτή τη στιγμή κρυπτονόμισμα, το Bitcoin. Ως βασικοί παράγοντες καθορισμού της τιμής χρησιμοποιούνται οι οικονομικοί δείκτες του νομίσματος καθώς και δεδομένα που το αφορούν και που έχουν εξορυχθεί απο το Twitter.

Αρχικά αναλύονται οι παρελθοντικές τιμές ως χρονοσειρές και εφαρμόζονται μετασχηματισμοί με στόχο την σταθεροποίησή τους. Έπειτα, βάσει αυτών των χρονοσειρών γίνεται η πρώτη προσπάθεια πρόβλεψης με τις μεθόδους Facebook Prophet, ARIMA, SARIMAX και με Επαναλαμβανόμενα Νευρωνικά Δίκτυα. Με στόχο την βελτιστοποίηση της πρόβλεψης εισάγονται σε αυτά τα μοντέλα νέες παράμετροι που αφορούν δεδομένα από το Twitter σχετικά με το Bitcoin. Κατά την επεξεργασία αυτών των δεδομένων, φιλτράρεται το κείμενο του κάθε tweet και ορίζονται συντελεστές βαρύτητας βάσει της απήχησης του. Ακολουθεί η συναισθηματική ανάλυσή τους, με τη χρήση του εργαλείου VADER, απο την οποία προκύπτουν και οι νέες παράμετροι της πρόβλεψης. Το συνολικό σφάλμα της πρόβλεψης μετρήθηκε με RMSE = 312.09 USD για την πρόβλεψη με μια παράμετρο και με RMSE = 298.36 USD για την πρόβλεψη με την χρήση και της σύνθετης πολικότητας που αποδείχθηκε η βέλτιστη πρόσθετη παράμετρος.

Abstract

Predicting the price of a cryptocurrency can be a challenging task as the factors that determine it are many and very vague. The main goal of this thesis is to predict the evolution of the prices of popular cryptocurrencies, with emphasis on the currently most popular cryptocurrency, Bitcoin. Economic indicators and related data extracted from Twitter are used as key price determinants.

Initially, past values are analyzed as time series and transformations are applied in order to stabilize them. Then, based on the time series, the first prediction attempt is made with the methods Facebook Prophet, ARIMA SARIMAX and with Recurrent Neural Networks. In order to optimize the forecast, data relevant to Bitcoin are extracted from Twitter and are introduced as new parameters in the previous models. This data is processed by filtering the text of each tweet and setting weighting factors based on its impact. The next step is tweets' emotional analysis, using the VADER tool, which produces the new forecast paremeters. The total forecast error was measured with RMSE = 312.09 USD for the one-parameter forecast and with RMSE = 298.36 USD for the forecast using the compound polarity that proved to be the optimal additional parameter.

Πίνακας περιεχομένων

Κατάλογος σχημάτων

Κατάλογος πινάκων

Συντομογραφίες

RNN	Recurrent Neural Network
ARIMA	AutoRegressive Integrated Moving Average
SARIMAX	Seasonal AutoRegressive Integrated Moving Average with eXternalor exogenous regressors
DFT	Dickey-Fuller Test
ADFT	Augmented Dickey-Fuller Test
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
GRU	Gated Recurrent Units
BTC	Bitcoin
AIC	Akaike Information Criterion
RMSE	Root Mean Squared Error
MAPE	Mean Absolute Percentage Error
STL	Seasonal-Trend decomposition using LOESS
URL	Uniform Resource Locator
VADER	Valence Aware Dictionary for sEntiment Reasoning

Κεφάλαιο 1

Εισαγωγή

Οι λέξεις κοινωνικά δίκτυα, κρυπτονομίσματα και μηχανική μάθηση αποσπούν ολοένα και περισσότερο ενδιαφέρον. Οι εφαρμογές και οι επεκτάσεις τους έχουν ξεφύγει από τα στενά όρια της οικονομικής επιστήμης και των τεχνολογικών μελετών και πλέον διαδραματίζουν όλο και πιο καθοριστικό ρόλο ακόμα και σε καθημερινές συζητήσεις.

Όσον αφορά την εξέλιξη των κρυπτονομισμάτων, είναι ένα ζήτημα που απασχολεί σχεδόν όλο το κοινό των επενδυτών του συγκεκριμένου κλάδου μιας και μια εύστοχη πρόβλεψη μπορεί να αποφέρει μεγάλα κέρδη. Τις περισσότερες φορές όμως η πρόβλεψη αυτή καθίσταται ιδιαίτερα δύσκολη λόγω του πλήθους των παραγόντων που την καθορίζουν. Στην αντιμετώπιση αυτού του προβλήματος χρήσιμο εργαλείο θα μπορούσε να φανεί η ανάλυση δεδομένων με μηχανική μάθηση η οποία έχει αποδείξει την αξία της σε πολλούς τομείς πρόβλεψης.

Παράλληλα το Twitter αποτελεί μια ανεξάντλητη πηγή τέτοιων δεδομένων προς επεξεργασία. Λαμβάνοντας υπόψιν την δημοφιλία του ως κοινωνικό δίκτυο δεν πρέπει να προκαλεί έκπληξη το γεγονός οτι χιλιάδες απόψεις και ιδέες δημοσιεύονται και κοινοποιούνται σε αυτό κάθε δευτερόλεπτο. Όλος αυτός ο όγκος πληροφοριών μπορει να φανεί ιδιαίτερα χρήσιμος στην προσπάθεια εξόρυξης συμπερασμάτων τόσο για την κοινή γνώμη αλλά και ακόμα πιο εξειδικευμένα για συγκεκριμένες ομάδες ανθρώπων με καθορισμένα επιθυμητά χαρακτηριστικά.

1.1 Αντικείμενο της Διπλωματικής

Σκοπός της παρούσας διπλωματικής είναι η πρόβλεψη των τιμών του κρυπτονομίσματος Bitcoin με την χρήση τεχνικών μηχανικής μάθησης. Γίνεται ανάλυση των παρελθοντικών τιμών του ως χρονοσειρά και με την χρήση των μεθόδων Prophet, ARIMA, SARIMAX και RNN προβλέπονται οι μελλοντικές του τιμές. Στη συνέχεια χρησιμοποιούνται δεδομένα από το δίκτυο του Twitter τα οποία, αφού υποστούν κάποια προ-επεξεργασία, εισάγονται στο εργαλείο VADER με στόχο τη συναισθηματική τους ανάλυση. Από αυτή την ανάλυση προκύπτουν δείκτες οι οποίοι χρησιμοποιούνται ως πρόσθετη είσοδος σε ένα νευρωνικό δίκτυο σε συνδυασμό με τις παρελθοντικές τιμές του κρυπτονομίσματος.

1.2 Οργάνωση του τόμου

Η εργασία αποτελείται από το παρόν κεφάλαιο της εισαγωγής και επτά ακόμα κεφάλαια, το περιεχόμενο των οποίων παρουσιάζεται συνοπτικά παρακάτω.

Το κεφάλαιο 2 αφορά εργασίες που σχετίζονται θεματικά με την παρούσα και θα μπορούσαν να αποτελέσουν παραπομπές για τον αναγνώστη.

Το κεφάλαιο 3 αποτελεί το θεωρητικό υπόβαθρο της εργασίας και περιέχει βασικές πληροφορίες που χρειάζονται για την πλήρη κατανόηση των πειραμάτων που ακολουθούν. Αναλύονται οι όροι στους οποίους βασίζεται η εργασία, οι μεθοδολογίες που ακολουθήθηκαν και τα εργαλεία που χρησιμοποιήθηκαν κατα την διάρκεια των πειραμάτων.

Ακολουθεί το κεφάλαιο 4 όπου γίνεται λεπτομερής ανάλυση των μεθόδων πρόβλεψης που χρησιμοποιούνται.

Στο κεφάλαιο 5 γίνεται προσπάθεια πρόβλεψης της τιμής του κρυπτονομίσματος βάσει χρονοσειρών. Αρχικά περιγράφεται το σύνολο δεδομένων και γίνεται πρόβλεψη με τον αλγόριθμο Facebook Prophet. Έπειτα ακολουθούν οι προβλέψεις με τους αλγορίθμους ARIMA και SARIMAX όπου γίνεται τροποποίηση της χρονοσειράς με στόχο την σταθεροποίησή της. Το τελευταίο κομμάτι του κεφαλαίου αφορά την πρόβλεψη με νευρωνικά δίκτυα. Δοκιμάζοντας διάφορους συνδυασμούς νευρωνικών δικτύων, τροποποιώντας κάθε φορά τον αριθμό και το είδος των στρωμάτων που το αποτελούν, επιλέγεται το βέλτιστο δίκτυο. Στη συνέχεια με τη χρήση αυτού του δικτύου γίνεται πρόβλεψη και για άλλα κρυπτονομίσματα (Cardano, Cosmos, Crypto.com, Dogecoin, Polkadot, Ethereum, Solana, USD Coin, Tether και XRP).

Το κεφάλαιο 6 περιλαμβάνει την ανάλυση των δεδομένων του Twitter, δηλαδή την ανά-

λυση και οπτικοποίηση των αρχικών δεδομένων (tweets, hashtags, λέξεις-κλειδιά και χρήστες), την επεξεργασία του κειμένου του tweet, την ανάλυση της γλώσσας και τέλος την συναισθηματική ανάλυση με την χρήση του εργαλείου VADER. Αποτέλεσμα αυτής της ανάλυσης είναι η δημιουργία χρήσιμων δεικτών οι οποίοι θα αξιοποιηθούν στη συνέχεια, στο κεφάλαιο 7, με στόχο τη βελτίωση της πρόβλεψης των νευρωνικών δικτύων.

Στο τελευταίο κεφάλαιο (κεφάλαιο 8) γίνεται σύνοψη των πειραμάτων, εξάγονται συμπεράσματα και προτείνονται μελλοντικές επεκτάσεις της μελέτης αυτής.

Κεφάλαιο 2

Συναφείς Εργασίες

2.1 Ανάλυση και πρόβλεψη χρονοσειρών

Σύμφωνα με την επιστημονική βιβλιογραφία το μοντέλο ARIMA έχει αρκετά ενεργό ρόλο ως μέθοδος πρόβλεψης. Στο άρθρο των Yenidoğan et al. [2] συγκρίνονται οι μέθοδοι Facebook Prophet και ARIMA ως προς την απόδοσή τους στην πρόβλεψη της τιμής του Bitcoin, χρησιμοποιώντας δεδομένα μεταξύ Μαΐου 2016 και Μαρτίου 2018, ενώ το μοντέλο ARIMA παρουσιάζεται ως μέθοδος ανάλυσης χρονοσειρών και από τους Hirata et al [3].

Σημαντική είναι επίσης και η παρουσία των νευρωνικών δικτύων στην πρόβλεψη χρονοσειρών. Οι Ho, Xie και Goh [4] δημοσίευσαν μια συγκριτική μελέτη της γενικότερης αποτελεσματικότητας των μεθόδων ARIMA και των νευρωνικών δικτύων ως προς την δυνατότητά τους να χρησιμοποιηθούν στην πρόβλεψη χρονοσειρών. Οι Suhwan et al. [5] συνέκριναν ενα βαθύ νευρωνικό δίκτυο, ένα μοντέλο LSTM, ένα συνελικτικό νευρωνικό δίκτυο και διάφορους συνδυασμούς τους για την πρόβλεψη του κρυπτονοσμίσματος Bitcoin. Ταυτόχρονα, σε άρθρο τους οι Dutta, Kumar και Basu [6] προτείνουν μια προσέγγιση στην πρόβλεψη του Bitcoin με την χρήση ενός μοντέλου GRU ενώ η μελέτη των Connor, Martin και Atlas [7] προτείνει μια νέα μέθοδο στην πρόβλεψη χρονοσειρών με βάση τα επαναλαμβανόμενα νευρωνικά δίκτυα με εφαρμογή στην χρονοσειρά ζήτησης ηλεκτρικού ρεύματος γνωστής εταιρίας ενέργειας.

Τέλος, σε ακόμη ένα άρθρο, οι Derbentsev, Matviychuk και Soloviev [8] κάνουν πρόβλεψη της εξέλιξης τριών διαφορετικών κρυπτονομισμάτων (Bitcoin, Ethereum, Ripple) με την χρήση ενός μοντέλου δυαδικού αυτοπαλινδρομικού δέντρου (Binary Autoregressive Tree - BART) και νευρωνικών δικτύων (multilayer perceptron - MLP).

2.2 Συναισθηματική ανάλυση σε δεδομένα του Twitter

Όσον αφορά τις μεθόδους συναισθηματικής ανάλυσης, το VADER όπως και άλλες μέθοδοι μηχανικής μάθησης επιδεικνύουν ικανοποιητικά αποτελέσματα. Στην εργασία του ο Μουλοσιώτης [9] παρουσιάζει μια μεθοδολογία ανάλυσης συναισθήματος σε tweets σχετικά με τις εκλογές του 2020 στις ΗΠΑ, χρησιμοποιώντας το εργαλείο VADER και άλλες μεθόδους μηχανικής μάθησης, μελέτη που μπορεί να αξιοποιηθεί ως μέσο πρόβλεψης του εκλογικού αποτελέσματος. Αντίστοιχα οι Elbagir και Yang [10] χρησιμοποιούν το εργαλείο VADER με σκοπό να αναλύσουν το συναίσθημα σε Tweets σχετικά με τις αμερικανικές εκλογές το 2016 και να τα κατηγοριοποιήσουν με βάση το είδος του συναισθήματος που περιέχουν. Επίσης βασιζόμενος στη μέθοδο VADER, ο Τζαβέλλος [11] εκτελεί πρόβλεψη της τιμής του Bitcoin με βάση την ανάλυση συναισθήματος σε δεδομένα που έχουν συλλεχθεί από το Twitter. Παράλληλα ο Πλέσσας [12], στην διπλωματική εργασία του, μελετά την εξέλιξη των κρουσμάτων της Covid-19, εκτελεί προβλέψεις σχετικά με την πορεία της υγείας των ασθενών με την βοήθεια αλγορίθμων μηχανικής μάθησης και αναλύει άρθρα και tweets μέσω συναισθηματικής ανάλυσης με σκοπό την κατηγοριοποίησή τους.

Όσον αφορά προβλέψεις που αφορούν το χρηματηστήριο οι Mittal και Goel [13] πραγματοποιούν συναισθηματική ανάλυση σε δεδομένα του Twitter με την βοήθεια νευρωνικών δικτύων προκειμένου να προβλέψουν, βάσει αυτών, κινήσεις μετοχών. Οι Rao et al. [14] αναλύουν το συναίσθημα σε tweets σχετικά με τις DJIA, NASDAQ-100 και 13 άλλες μετοχές όπου τελικά συμπεραίνουν την ύπαρξη μεγάλης συσχέτισης μεταξύ της εξέλιξης του συναισθήματος και της τιμής των μετοχών.

Οι Bao et al. [15] περιγράφουν τον τρόπο με τον οποίο πρέπει να εκτελείται η προεπεξεργασία του κειμένου του tweet προκειμένου να γίνει με μεγαλύτερη ακρίβεια η συναισθηματική ανάλυσή του ενώ, παράλληλα, σε άρθρο των Koto και Adriani [16] γίνεται συγκριτική μελέτη μεταξύ των χαρακτηριστικών τα οποία αποδίδουν καλύτερα στην εκτέλεση της συναισθηματικής ανάλυσης σε tweets.

Τέλος σε μια αρκετά πρωτοποριακή μελέτη των Zhou και Tao [17] δημιουργείται ένα νέο μοντέλο συναισθηματικής ανάλυσης σε tweets που αφορούν διάφορα κοινωνικά γεγονότα και είναι σε θέση να εντοπίσει το ενδιαφέρον και τις απόψεις του κόσμου σχετικά με ένα δεδομένο κοινωνικό γεγονός.

Κεφάλαιο 3

Θεωρητικό υπόβαθρο

3.1 Εισαγωγή

Στο κεφάλαιο αυτό γίνεται αναλυτική περιγραφή των βασικών εννοιών της εργασίας. Περιλαμβάνει τις απαρίτητες πληροφορίες, τους ορισμούς, τα εργαλεία, τις μεθόδους, τους αλγορίθμους και τις πλατφόρμες που χρησιμοποιήθηκαν κατα την διάρκεια των πειραμάτων.

3.2 Κρυπτονομίσματα

Κρυπτονόμισμα ονομάζεται ένα ψηφιακό ή εικονικό νόμισμα, ασφαλισμένο μέσω κρυπτογραφίας, γεγονός που καθιστά σχεδόν αδύνατη την παραχάραξη του ή την περίπτωση το ίδιο κρυπτονόμισμα να χρησιμοποιηθεί ταυτόχρονα σε δυο συναλλαγές. Πολλά κρυπτονομίσματα είναι αποκεντρωμένα δίκτυα που βασίζονται στην τεχνολογία blockchain - ένα κατανεμημένο αρχείο συναλλαγών που συντηρείται από ένα δίκτυο διαφορετικών υπολογιστών. Καθοριστικό χαρακτηριστικό τους είναι ότι δεν εκδίδονται από καμία κεντρική αρχή, γεγονός που τα καθιστά θεωρητικά απρόσβλητα σε κρατικές παρεμβάσεις ή χειραγώγηση.

Το πρώτο συνθετικό της λέξης (κρυπτο-) αναφέρεται στους διάφορους αλγόριθμους κρυπτογράφησης και τις κρυπτογραφικές τεχνικές που προστατεύουν αυτές τις εγγραφές, όπως η κρυπτογράφηση ελλειπτικής καμπύλης (elliptical curve encryption), τα ζεύγη δημόσιου-ιδιωτικού κλειδιού (public-private key pairs) και οι συναρτήσεις κατακερματισμού (hashing functions).

Τα κρυπτονομίσματα μπορούν να εξορυχθούν ή να αγοραστούν από ανταλλακτήρια κρυπτονομισμάτων. Η εξόρυξη ενος κρυπτονομίσματος είναι η διαδικασία με την οποία ενα νέο

κρυπτονόμισμα τίθεται σε κυκλοφορία. Είναι επίσης ο τρόπος με τον οποίο το δίκτυο επιβεβαιώνει τις νέες συναλλαγές και αποτελεί κρίσιμο στοιχείο της συντήρησης και ανάπτυξης ολόκληρου του blockchain. Η εξόρυξη εκτελείται με χρήση εξελιγμένων υπολογιστικών μηχανημάτων που λύνουν ένα εξαιρετικά πολύπλοκο υπολογιστικό μαθηματικό πρόβλημα. Ο πρώτος υπολογιστής που θα βρει τη λύση στο πρόβλημα λαμβάνει το επόμενο μπλοκ bitcoin και αναλαμβάνεται η επίλυση του επόμενου προβλήματος.

Την παρούσα στιγμή οι πληρωμές με τη χρήση κρυπτονομισμάτων δεν είναι αποδεκτές στις περισσότερες συναλλαγές. Στην πραγματικότητα, τα κρυπτονομίσματα, ακόμη και τα πιο δημοφιλή όπως το Bitcoin, δεν χρησιμοποιούνται σχεδόν καθόλου για καθημερινές συναλλαγές λιανικής. Ωστόσο, η εκτίναξη της αξίας των κρυπτονομισμάτων τα καθιστά όλο και πιο δημοφιλή ως μέσα εμπορικών συναλλαγών.

3.2.1 Το Bitcoin και άλλα δημοφιλή κρυπτονομίσματα

Επι του παρόντος το Bitcoin είναι το πιο δημοφιλές και πολύτιμο κρυπτονόμισμα. Επινοήθηκε και παρουσιάστηκε στο κοινό από ένα ανώνυμο άτομο με το ψευδώνυμο Satoshi Nakamoto το 2008. Διατέθηκε στο κοινό το 2009 και από τότε παραμένει το κρυπτονόμισμα με τις περισσότερες συναλλαγές και την μεγαλύτερη κάλυψη. Τον Νοέμβριο του 2021, υπήρχαν περισσότερα από 18,8 εκατομμύρια Bitcoin σε κυκλοφορία με συνολική αξία αγοράς περίπου 1,2 τρισεκατομμύρια δολάρια και η τιμή της μονάδας έχει φτάσει να ξεπερνά τις 60 χιλιάδες αμερικανικά δολάρια πριν μερικούς μήνες.

Στον απόηχο της επιτυχίας του Bitcoin, κυκλοφόρησαν και πολλά άλλα κρυπτονομίσματα, γνωστά ως “altcoins”. Μερικά από αυτά είναι κλώνοι του Bitcoin, ενώ άλλα είναι εντελώς νέα νομίσματα που δημιουργήθηκαν από την αρχή. Κάθε κρυπτονόμισμα ισχυρίζεται ότι έχει διαφορετική λειτουργία και προδιαγραφές. Για παράδειγμα, το κρυπτονόμισμα Ethereum (ETH) διαφημίζεται ως το ανερχόμενο μέσο στην επικείμενη πλατφόρμα έξυπνων συμβολαίων και βρίσκεται πίσω από την έξαρση της τεχνολογίας των NFTs, ψηφιακές εκδόσεις τέχνης ή συλλεκτικά αντικείμενα που συνδέονται μέσω blockchain και είναι μοναδικά στο είδος τους. Το XRP της Ripple χρησιμοποιείται από τις τράπεζες για διευκόλυνση στις μεταφορές μεταξύ διαφορετικών γεωγραφικών περιοχών. Το Cardano (ADA) παρουσιάζεται ως μια πλατφόρμα blockchain τρίτης γενιάς. Βασίζεται στο proof-of-stake (PoS), που σημαίνει ότι οι περίπλοκοι υπολογισμοί proof-of-work (PoW) και η υψηλή χρήση ηλεκτρικής ενέργειας που απαιτούνται για την εξόρυξη νομισμάτων όπως το Bitcoin δεν είναι

απαραίτητες, καθιστώντας ενδεχομένως το δίκτυό του πιο αποτελεσματικό και βιώσιμο.

Μέχρι σήμερα έχουν κυκλοφορήσει χιλιάδες κρυπτονομίσματα στην αγορά, η συνολική αξία των οποίων έχει ξεπεράσει τα 2,1 τρισεκατομμύρια δολάρια, με το Bitcoin να αντιπροσωπεύει περίπου το 41% αυτής της συνολικής αξίας. [18]

3.3 Μηχανική Μάθηση

Σύμφωνα με τους Mohri, Rostamizadeh και Talwalkar [19] Μηχανική Μάθηση αποτελούν οι υπολογιστικές μέθοδοι που χρησιμοποιούν την εμπειρία τους για την βελτιστοποίηση ή την πραγματοποίηση προβλέψεων με ακρίβεια. Ως εμπειρία εννοείται η χρήση προηγούμενων πληροφοριών από το εκπαιδευόμενο σύστημα. Η ποιότητα και το μέγεθος αυτών των δεδομένων είναι καθοριστικοί παράγοντες για την επιτυχία και την ακρίβεια των προβλέψεων.

3.3.1 Εφαρμογές

Οι πρακτικές εφαρμογές της μηχανικής μάθησης είναι σε θέση να λύσουν ένα μεγάλο πλήθος προβλημάτων όπως:

- Ταξινόμηση κειμένων ή εγγράφων. Αυτό περιλαμβάνει προβλήματα όπως την αντιστοίχιση ενός θέματος με ένα κείμενο ή ένα έγγραφο ή την αυτόματη αναγνώριση για το εάν μια ιστοσελίδα περιέχει ακατάλληλο περιεχόμενο και τον εντοπισμό ανεπιθύμητων μηνυμάτων.

- Επεξεργασία φυσικής γλώσσας (Natural Language Processing). Οι περισσότερες εργασίες σε αυτό το πεδίο, συμπεριλαμβανομένης της αναγνώρισης μιας λέξης ως μέρος του λόγου (part-of-speech tagging), της ανάλυσης κειμένου χωρίς γενικότερο πλαίσιο (context-free parsing) ή της ανάλυσης εξάρτησης (dependency parsing), παρουσιάζονται ως προβλήματα μάθησης. Στο κεφάλαιο 6 της παρούσας εργασίας παρουσιάζεται μια εφαρμογή του συγκεκριμένου τύπου. Σε αυτά τα προβλήματα, οι προβλέψεις αναγνωρίζουν κάποια γενικότερη δομή. Για παράδειγμα, στην αναγνώριση μιας λέξης ως μέρος του λόγου, η πρόβλεψη για μια πρόταση είναι μια ακολουθία ετικετών μερών του λόγου που επισημαίνουν κάθε λέξη. Στην ανάλυση κειμένου χωρίς γενικότερο πλαίσιο, η πρόβλεψη έχει την μορφή δέντρου απόφασης (decision tree). Αυτές είναι

περιπτώσεις εμπλουτισμένων μαθησιακών προβλημάτων γνωστών ως δομημένα προβλήματα πρόβλεψης.

- Εφαρμογές επεξεργασίας λόγου. Αυτή η κατηγορία περιλαμβάνει την αναγνώριση ομιλίας, τη σύνθεση ομιλίας, την επαλήθευση και αναγνώριση ομιλητή, καθώς και υποπροβλήματα όπως η μοντελοποίηση γλώσσας και η ακουστική μοντελοποίηση.

- Εφαρμογές όρασης υπολογιστών (computer vision). Εδώ περιλαμβάνεται η αναγνώριση και ταυτοποίηση αντικειμένων, η ανίχνευση προσώπου, η οπτική αναγνώριση χαρακτήρων (optical character recognition - OCR), η ανάκτηση εικόνας βάσει περιεχομένου ή η εκτίμηση της στάσης σώματος.

- Εφαρμογές Υπολογιστικής Βιολογίας. Σε αυτή την κατηγορία περιλαμβάνεται η πρόβλεψη πρωτεϊνικής λειτουργίας, η αναγνώριση θέσεων-κλειδιών και η ανάλυση δικτύων γονιδίων και πρωτεϊνών.

- Πολλά άλλα προβλήματα όπως η ανίχνευση απάτης σε πιστωτικές κάρτες, τηλέφωνα ή ακόμα και σε ασφαλιστικές εταιρείες, οι δικτυακές εισβολές, η εκμάθηση παιχνιδιών όπως σκάκι ή τάβλι, ο μη υποβοηθούμενος έλεγχος οχημάτων όπως ρομπότ ή αυτοκίνητα, η ιατρική διάγνωση, οι μηχανές αναζήτησης ή τα συστήματα εξαγωγής πληροφοριών αντιμετωπίζονται με τη χρήση τεχνικών μηχανικής μάθησης.

3.3.2 Προβλήματα

Τα περισσότερα προβλήματα πρόβλεψης μπορούν να χαρακτηριστούν ως προβλήματα μάθησης και το πεδίο πρακτικής εφαρμογής της μηχανικής μάθησης επεκτείνεται διαρκώς. Μερικές από τις βασικές μεθόδους της μηχανικής μάθησης είναι οι παρακάτω:

- Κατηγοριοποίηση (Classification). Αυτό είναι το πρόβλημα της ανάθεσης κάθε στοιχείου σε μια κατηγορία. Ο αριθμός των κατηγοριών σε τέτοια προβλήματα είναι συχνά μικρότερος από μερικές εκατοντάδες, αλλά μπορεί να είναι πολύ μεγαλύτερος σε ορισμένες δύσκολες εργασίες και ακόμη και απεριόριστος όπως στην περίπτωση της οπτικής αναγνώρισης χαρακτήρων, την ταξινόμηση κειμένου ή την αναγνώριση ομιλίας.

- Παλινδρόμηση (Regression). Πρόκειται για το πρόβλημα της πρόβλεψης μιας πραγματικής τιμής για κάθε στοιχείο. Σε αυτή την κατηγορία προβλημάτων ανήκει και η

πρόβλεψη της τιμής του Bitcoin που πραγματεύεται η παρούσα εργασία. Άλλα παραδείγματα παλινδρόμησης αποτελούν η πρόβλεψη της αξίας των αποθεμάτων ή η πρόβλεψη της διακύμανσης άλλων οικονομικών μεταβλητών. Στην παλινδρόμηση, η ποινή για μια εσφαλμένη πρόβλεψη εξαρτάται από το μέγεθος της διαφοράς μεταξύ των αληθινών και των προβλεπόμενων τιμών, σε αντίθεση με το πρόβλημα της ταξινόμησης, όπου συνήθως δεν υπάρχει έννοια εγγύτητας μεταξύ των διαφόρων κατηγοριών.

- Ταξινόμηση (Ranking). Αυτό είναι το πρόβλημα της τοποθέτησης αντικειμένων σε σειρά σύμφωνα με κάποιο κριτήριο. Για παράδειγμα κατά την αναζήτηση στον Ιστό, η επιστροφή ιστοσελίδων σχετικών με ένα ερώτημα αναζήτησης, είναι ένα παράδειγμα ταξινόμησης. Πολλά άλλα παρόμοια προβλήματα ταξινόμησης προκύπτουν κατά τον σχεδιασμό συστημάτων εξαγωγής πληροφοριών ή της επεξεργασίας της φυσικής γλώσσας.

- Ομαδοποίηση (Clustering). Είναι το πρόβλημα της κατάτμησης ενός συνόλου στοιχείων σε ομοιογενή υποσύνολα. Η ομαδοποίηση χρησιμοποιείται συχνά για την ανάλυση συνόλων δεδομένων μεγάλου όγκου. Για παράδειγμα, στο πλαίσιο της ανάλυσης των κοινωνικών δικτύων, οι αλγόριθμοι ομαδοποίησης προσπαθούν να προσδιορίσουν φυσικές “κοινότητες” μέσα σε μεγάλες ομάδες ανθρώπων.

- Μείωση Διαστάσεων ή Πολλαπλή Εκμάθηση (Dimensionality Reduction or Manifold Learning). Αυτό το πρόβλημα αποτελείται από την μετατροπή μιας αρχικής αναπαράστασης στοιχείων σε μια αναπαράσταση χαμηλότερης διάστασης, διατηρώντας παράλληλα ορισμένες ιδιότητες της αρχικής αναπαράστασης. Σύνηθες παράδειγμα περιλαμβάνει την προεπεξεργασία ψηφιακών εικόνων σε εργασίες όρασης υπολογιστή.

3.3.3 Βασικοί όροι

Προκειμένου να επεξηγηθούν τα βασικά στάδια της μηχανικής μάθησης θα χρησιμοποιηθεί το παράδειγμα της ανίχνευσης ανεπιθύμητων μηνυμάτων (spam detection). Εντοπισμός ανεπιθύμητων μηνυμάτων είναι το πρόβλημα της εκμάθησης της αυτόματης ταξινόμησης μηνυμάτων email είτε ως επιθυμητό (non-spam) είτε ως ανεπιθύμητο (spam). Ακολουθεί μια λίστα ορισμών και ορολογίας που χρησιμοποιούνται συνήθως στη Μηχανική Μάθηση:

- Παραδείγματα (Examples): Στοιχεία ή περιπτώσεις δεδομένων που χρησιμοποιούνται για μάθηση ή αξιολόγηση. Στο πρόβλημα με τα ανεπιθύμητα μηνύματα, αυτά τα παραδείγματα αντιστοιχούν στη συλλογή μηνυμάτων email που θα χρησιμοποιήσουμε για εκμάθηση και δοκιμή.

- Χαρακτηριστικά (Features): Το σύνολο των χαρακτηριστικών (συχνά αντιπροσωπεύεται ως διάνυσμα) που σχετίζεται με ένα παράδειγμα. Στην περίπτωση των μηνυμάτων email, ορισμένα σχετικά χαρακτηριστικά μπορεί να περιλαμβάνουν το μήκος του μηνύματος, το όνομα του αποστολέα, διάφορα χαρακτηριστικά της κεφαλίδας, την παρουσία ορισμένων λέξεων-κλειδιών στο σώμα του μηνύματος κ.λπ.

- Ετικέτες (Labels): Τιμές ή κατηγορίες που αντιστοιχίζονται σε παραδείγματα. Στα προβλήματα ταξινόμησης, στα παραδείγματα εκχωρούνται συγκεκριμένες κατηγορίες, για παράδειγμα, οι κατηγορίες spam και non-spam στο πρόβλημα δυαδικής ταξινόμησης. Στην παλινδρόμηση, τα στοιχεία αποδίδονται σε ετικέτες πραγματικών τιμών.

- Υπερπαράμετροι (Hyperparameters): Ελεύθερες παράμετροι που δεν καθορίζονται από τον αλγόριθμο εκμάθησης, αλλά μάλλον καθορίζονται ως είσοδοι στον αλγόριθμο εκμάθησης.

- Δείγμα εκπαίδευσης (Training sample): Παραδείγματα που χρησιμοποιούνται για την εκπαίδευση ενός αλγόριθμου εκμάθησης. Στο πρόβλημα με τα ανεπιθύμητα μηνύματα, το δείγμα εκπαίδευσης αποτελείται από ένα σύνολο μηνυμάτων ηλεκτρονικού ταχυδρομείου μαζί με τις σχετικές ετικέτες τους. Το δείγμα εκπαίδευσης ποικίλλει για διαφορετικά σενάρια μάθησης.

- Δείγμα επικύρωσης (Validation sample): Παραδείγματα που χρησιμοποιούνται για τη ρύθμιση των παραμέτρων ενός αλγορίθμου εκμάθησης κατά την εργασία με δεδομένα με ετικέτα. Το δείγμα επικύρωσης χρησιμοποιείται για την επιλογή κατάλληλων τιμών για τις ελεύθερες παραμέτρους του αλγορίθμου εκμάθησης (υπερπαράμετροι).

- Δείγμα δοκιμής (Test sample): Παραδείγματα που χρησιμοποιούνται για την αξιολόγηση της απόδοσης ενός αλγορίθμου μάθησης. Το δείγμα δοκιμής είναι ξεχωριστό από τα δεδομένα εκπαίδευσης και επικύρωσης και δεν διατίθεται στο στάδιο εκμάθησης. Στο πρόβλημα ανεπιθύμητης αλληλογραφίας, το δείγμα δοκιμής αποτελείται από μια συλλογή μηνυμάτων email για τα οποία ο αλγόριθμος εκμάθησης πρέπει να προβλέπει

ετικέτες. με βάση τα χαρακτηριστικά. Στη συνέχεια, αυτές οι προβλέψεις συγκρίνονται με τις ετικέτες του δείγματος δοκιμής για να μετρηθεί η απόδοση του αλγορίθμου.

- Συνάρτηση απώλειας (Loss function): Μια συνάρτηση που μετρά τη διαφορά μεταξύ μιας προβλεπόμενης ετικέτας και μιας αληθινής ετικέτας.

- Υπερπροσαρμογή και υποπροσαρμογή (Overfitting and underfitting): Η ποιότητα του μοντέλου μηχανικής μάθησης μπορεί να προσδιοριστεί εξετάζοντας την προσαρμογή (fit) του μοντέλου στα δεδομένα. Η υπερπροσαρμογή και η υποπροσαρμογή είναι έννοιες που σχετίζονται με την προσαρμογή του μοντέλου στα δεδομένα εκπαίδευσης και οδηγούν σε ενέργειες βελτίωσης του μοντέλου. Για παράδειγμα, η υπερπροσαρμογή συμβαίνει όταν το μοντέλο ταιριάζει καλά στα δεδομένα προπόνησης αλλά δεν μπορεί να γενικευτεί σε καινούρια δεδομένα ή δεδομένα δοκιμής. Η υπερπροσαρμογή συνήθως καταπολεμάται αυξάνοντας την ποσότητα δεδομένων ή καθιστώντας το μοντέλο λιγότερο περίπλοκο. Από την άλλη πλευρά, η υποπροσαρμογή συμβαίνει όταν το μοντέλο δεν διαθέτει την απαραίτητη εκφραστική δύναμη να συλλάβει τον επιθυμητό στόχο. Η υποπροσαρμογή μπορεί να διορθωθεί με την αύξηση της πολυπλοκότητας του μοντέλου.

- Προετοιμασία και καθαρισμός δεδομένων (Data preparation and cleaning): Αφού οι παρατηρήσεις και τα χαρακτηριστικά αποκτηθούν από τα ακατέργαστα δεδομένα, πρέπει να διασφαλιστεί ότι τα δεδομένα είναι έτοιμα να τροφοδοτηθούν στον αλγόριθμο. Αυτό το βήμα συνήθως περιλαμβάνει τον εντοπισμό (και την εξάλειψη) λανθασμένων και ακραίων τιμών, τη συμπλήρωση τιμών που λείπουν και τον χωρισμό των δεδομένων σε δεδομένα εκπαίδευσης και επικύρωσης. Τέλος, σε ορισμένες περιπτώσεις μοντέλων τα δεδομένα μπορεί να χρειαστεί να ανακατευτούν τυχαία για να εξασφαλιστεί η σωστή σύγκλιση του αλγορίθμου. Αυτό συμβαίνει ειδικά σε περιπτώσεις αλγορίθμων που μαθαίνουν από μία παρατήρηση τη φορά αντίθετα με τους αλγορίθμους που μαθαίνουν από ένα μεγαλύτερο σύνολο παραδειγμάτων εκπαίδευσης σε κάθε βήμα.

- Οπτικοποίηση δεδομένων (Data visualization): Η κατανόηση των δεδομένων είναι απαραίτητη για την ανάπτυξη μια στρατηγικής λύσης του προβλήματος. Η οπτική επιθεώρηση των δεδομένων βοηθάει στην ανάπτυξη τέτοιων στρατηγικών. Η οπτικοποίηση συνήθως περιλαμβάνει την κατάτμηση (slicing and dicing) των δεδομένων σε

πολλαπλές διαστάσεις ώστε να σχεδιαστούν οι κατανομές των δεδομένων και τα ιστογράμματα των διάφορων χαρακτηριστικών των δεδομένων και να προσδιοριστούν οι συσχετίσεις μεταξύ αυτών και του στόχου. Η οπτικοποίηση μπορεί επίσης να παρέχει ενδείξεις για δεδομένα που λείπουν ή είναι λανθασμένα. [20]

3.3.4 Κατηγορίες

Οι κατηγορίες μηχανικής μάθησης διαφέρουν ως προς τους τύπους των δεδομένων εκπαίδευσης που είναι διαθέσιμα στο εκπαιδευόμενο μοντέλο, τη σειρά και τη μέθοδο με την οποία λαμβάνονται τα δεδομένα εκπαίδευσης και τα δεδομένα δοκιμής που χρησιμοποιούνται για την αξιολόγηση του αλγόριθμου εκμάθησης. Οι βασικές κατηγορίες μηχανικής μάθησης είναι οι εξής:

- Εποπτευόμενη μάθηση (Supervised learning): Το εκπαιδευόμενο μοντέλο λαμβάνει ένα σύνολο επισημασμένων παραδειγμάτων ως δεδομένα εκπαίδευσης και κάνει προβλέψεις για όλα τα αόρατα σημεία. Αυτό είναι το πιο κοινό σενάριο που σχετίζεται με προβλήματα ταξινόμησης και παλινδρόμησης. Το πρόβλημα ανίχνευσης ανεπιθύμητης αλληλογραφίας που συζητήθηκε και το πρόβλημα της πρόβλεψης της τιμής του Bitcoin είναι περιπτώσεις εποπτευόμενης εκμάθησης.

- Μάθηση χωρίς επίβλεψη (Unsupervised learning): Το εκπαιδευόμενο μοντέλο λαμβάνει αποκλειστικά δεδομένα εκπαίδευσης χωρίς ετικέτα και κάνει προβλέψεις για όλα τα μη ορατά σημεία. Δεδομένου ότι γενικά δεν υπάρχει διαθέσιμο παράδειγμα με ετικέτα σε αυτόν τον τύπο μηχανικής μάθησης, μπορεί να είναι δύσκολο να αξιολογηθεί ποσοτικά η απόδοσή του. Η ομαδοποίηση και η μείωση των διαστάσεων όπως και η συναισθηματική ανάλυση των tweets είναι παραδείγματα μαθησιακών προβλημάτων χωρίς επίβλεψη.

- Ημι-εποπτευόμενη μάθηση (Semi-supervised learning): Το εκπαιδευόμενο μοντέλο λαμβάνει ένα δείγμα εκπαίδευσης που αποτελείται από δεδομένα με ετικέτα και χωρίς ετικέτα και κάνει προβλέψεις για όλα τα μη ορατά σημεία. Η ημι-εποπτευόμενη μάθηση χρησιμοποιείται συνήθως σε περιβάλλοντα όπου τα δεδομένα χωρίς ετικέτα είναι εύκολα προσβάσιμα, αλλά οι ετικέτες είναι κοστοβόρο να προσδιοριστούν.

3.3.5 Υπολογισμός σφάλματος

Παράμετρος καθορισμού της ποιότητας ενός μοντέλου μηχανικής μάθησης είναι μεταξύ άλλων και η απόκλιση των προβλεπόμενων τιμών από τις πραγματικές, δηλαδή το σφάλμα της πρόβλεψης. Οι αλγόριθμοι υπολογισμού του σφάλματος των προβλέψεων που χρησιμοποιήθηκαν είναι οι εξής:

- Σφάλμα Ρίζας Μέσου Τετραγώνου (Root Mean Squared Error - RMSE): Η χρήση του RMSE είναι πολύ συνηθισμένη και θεωρείται μια εξαιρετική μέτρηση σφάλματος γενικής χρήσης για αριθμητικές προβλέψεις. Είναι η τετραγωνική ρίζα του μέσου όρου του τετραγώνου του συνόλου του σφάλματος:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(S_i - O_i)^2} \tag{3.1}$$

  όπου O_i είναι οι παρατηρήσεις, S_i προβλεπόμενες τιμές μιας μεταβλητής και n ο αριθμός των παρατηρήσεων που είναι διαθέσιμες για ανάλυση. Το RMSE είναι ένα καλό μέτρο ακρίβειας, αλλά μόνο για τη σύγκριση σφαλμάτων πρόβλεψης διαφορετικών μοντέλων για μια συγκεκριμένη μεταβλητή και όχι μεταξύ μεταβλητών, καθώς εξαρτάται από την κλίμακα [21].

- Κανονικοποιημένο ή Σχετικό Σφάλμα Ρίζας Μέσου Τετραγώνου (Normalized or Relative Root Mean Squared Error - nRMSE): Είναι η τετραγωνική ρίζα του μέσου όρου του τετραγώνου του συνόλου του σφάλματος προς την μέση τιμή των παρατηρήσεων πολλαπλασιασμένη με το εκατο:

$$nRMSE = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(S_i - O_i)^2}}{\overline{M}} * 100 \tag{3.2}$$

  όπου O_i είναι οι παρατηρήσεις, S_i οι προβλεπόμενες τιμές μιας μεταβλητής, n ο αριθμός των παρατηρήσεων που είναι διαθέσιμες για ανάλυση και $\overline{M}$ η μέση τιμή των παρατηρήσεων. Το κανονικοποιημένο μέσο τετραγωνικό σφάλμα ρίζας (nRMSE) επιτρέπει την αξιολόγηση της διασποράς των προβλεπόμενων τιμών σε σχέση με τις μετρούμενες τιμές [22].

- Μέσο Απόλυτο Ποσοστό Σφάλματος (Mean Absolute Percentage Error- MAPE): Είναι η τετραγωνική ρίζα του μέσου όρου του τετραγώνου του συνόλου του σφάλματος προς την μέση τιμή των παρατηρήσεων πολλαπλασιασμένη με το εκατο:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|O_i - S_i|}{O_i} * 100 \tag{3.3}$$

όπου O_i είναι οι παρατηρήσεις, S_i οι προβλεπόμενες τιμές μιας μεταβλητής και n ο αριθμός των παρατηρήσεων που είναι διαθέσιμες για ανάλυση [23].

3.4 Συναισθηματική ανάλυση

Η Συναισθηματική Ανάλυση ή αλλιώς Εξόρυξη Γνώμης (Opinion Mining) είναι το πεδίο μελέτης που αναλύει τις απόψεις, τα συναισθήματα, τις αξιολογήσεις, τις στάσεις και τα συναισθήματα των ανθρώπων από τη γραπτή γλώσσα. Είναι ένας από τους πιο ενεργούς ερευνητικούς τομείς στην επεξεργασία φυσικής γλώσσας και έχει επίσης μελετηθεί ευρέως στην εξόρυξη δεδομένων, την εξόρυξη δεδομένων Ιστού και την εξόρυξη κειμένου. Η αυξανόμενη σημασία της συναισθηματικής ανάλυσης συμπίπτει με την ανάπτυξη των μέσων κοινωνικής δικτύωσης, όπως οι κριτικές, οι συζητήσεις σε φόρουμ, τα ιστολόγια, τα μικροιστολόγια, το Twitter και τα κοινωνικά δίκτυα. Για πρώτη φορά στην ανθρώπινη ιστορία, υπάρχει τόσο μεγάλος όγκος δεδομένων, τα οποία εκφράζουν απόψεις και ιδέες, που καταγράφονται σε ψηφιακή μορφή για ανάλυση. Τα συστήματα συναισθηματικής ανάλυσης εφαρμόζονται σχεδόν σε κάθε επιχειρηματικό και κοινωνικό τομέα, επειδή οι απόψεις είναι το κεντρικό στοιχείο σχεδόν όλων των ανθρώπινων δραστηριοτήτων και είναι βασικοί παράγοντες επιρροής των ανθρώπινων συμπεριφορών. [24]

3.4.1 Συναισθηματική αναλυση σε δεδομένα του Twitter

Το Twitter είναι μια από τις πιο δημοφιλείς πλατφόρμες microblogging που χρησιμοποιείται ευρέως από τους ανθρώπους για να εκφράσουν απόψεις μέσω των οποίων ανακλώνται διάφορα συναισθήματα σε διαφορετικές περιπτώσεις. [25]

Σκοπός της συναισθηματικής ανάλυσης σε δεδομένα του Twitter είναι να εξαχθούν αυτά τα συναισθήματα από τα tweets των χρήστών. Τις τελευταίες δεκαετίες, η ανάγκη για έρευνα στον τομέα αυτό έχει αυξηθεί ιδιαίτερα λόγω της πολυπλοκότητας του λόγου και της μορφής των tweets (χρήση αργκό, συντομογραφίες κ.λπ.) που καθιστούν δύσκολη την επεξεργασία τους. Η έκταση ενός tweet είναι πολύ μικρή κάτι που δημιουργεί μια ακόμα δυσκολία στο ήδη πολύπλοκο πρόβλημα. [26]

3.5 Η γλώσσα Python

Η Python είναι μια από τις πιο δημοφιλείς γλώσσες προγραμματισμού για την Επιστήμη των Δεδομένων (Data Science) και ως εκ τούτου προσφέρει μεγάλο αριθμό χρήσιμων πρόσθετων βιβλιοθηκών που αναπτύχθηκαν από τη μεγάλη κοινότητά της.

Αν και η απόδοση γλωσσών ερμηνείας (interpreted languages), όπως η Python, για εργασίες εντατικών υπολογισμών είναι κατώτερη από τις γλώσσες προγραμματισμού χαμηλότερου επιπέδου, έχουν αναπτυχθεί βιβλιοθήκες επεκτάσεων όπως οι NumPy και SciPy που βασίζονται σε εφαρμογές Fortran και C κατώτερου επιπέδου για γρήγορες και διανυσματικές λειτουργίες σε πολυδιάστατους πίνακες.

Για τις εργασίες προγραμματισμού μηχανικής μάθησης, χρησιμοποιούνται κυρίως εργαλεία όπως το scikit-learn, μια από τις πιο δημοφιλείς και προσβάσιμες βιβλιοθήκες ανοιχτού κώδικα στον τομέα της μηχανικής μάθησης. [27]

3.6 Jupyter Notebook, Google Colab

Τα σημειωματάρια Jupyter και Google Colab είναι εργαλεία ανοιχτού κώδικα, που βασίζονται σε πρόγραμμα περιήγησης και λειτουργούν ως εικονικά τετράδια εργαστηρίου για υποστήριξη εργασιών, κώδικα, δεδομένων και απεικονίσεις βοηθώντας στην λεπτομερέστερη περιγραφή της ερευνητικής διαδικασίας. Είναι αναγνώσιμα και από την μηχανή και από τον άνθρωπο, κάτι που διευκολύνει την διαλειτουργικότητα και την διεπιστημονική επικοινωνία. Τα σημειωματάρια μπορούν να αποθηκευθούν σε διαδικτυακά αποθετήρια και να παρέχουν συνδέσεις με ερευνητικά αντικείμενα όπως σύνολα δεδομένων, κώδικα, μεθόδους, έγγραφα, εργασίες και δημοσιεύσεις που βρίσκονται κάπου αλλού.

Κεφάλαιο 4

Μέθοδοι Πρόβλεψης

4.1 Facebook Prophet

Το μοντέλο Facebook Prophet είναι μια διαδικασία πρόβλεψης δεδομένων χρονοσειρών που βασίζεται σε ένα προσθετικό μοντέλο όπου οι μη γραμμικές τάσεις προσαρμόζονται μέσω της ετήσιας, εβδομαδιαίας και της ημερήσιας εποχικότητας, συμπεριλαμβανομένων και των περιπτώσεων διακοπών. Λειτουργεί καλύτερα με χρονοσειρές που έχουν ισχυρή εποχικότητα και μεγάλο αριθμό ιστορικών δεδομένων. Το Prophet είναι ανθεκτικό σε δεδομένα που λείπουν και σε αλλαγές στην τάση και συνήθως χειρίζεται καλά τις ακραίες τιμές. [28]

Η πρόβλεψή του βασίζεται σε τρία βασικά στοιχεία: τις τάσεις $g(t)$, την εποχικότητα $s(t)$ και τις αργίες $h(t)$, τα οποία συντίθενται ως εξής:

$$y(t) = g(t) + s(t) + h(t) + \varepsilon(t) \tag{4.1}$$

όπου $\varepsilon(t)$ είναι όρος σφάλματος. Υπάρχουν διάφορα λειτουργικά οφέλη από αυτή την προσέγγιση. Λόγω εβδομαδιαίας και ετήσιας εποχικότητας, η εποχιακή συνιστώσα $s(t)$ παρέχει ένα ευέλικτο μοντέλο περιοδικών αλλαγών. Το τμήμα $h(t)$ αντικατοπτρίζει τις προβλέψιμες ετήσιες μη φυσιολογικές ημέρες συμπεριλαμβανομένων εκείνων που συμβαίνουν βάσει ακανόνιστων χρονοδιαγραμμάτων. Ο όρος σφάλματος, $\varepsilon(t)$ αντικατοπτρίζει τις πληροφορίες που δεν εκφράζονται από το μοντέλο. Τυπικά μοντελοποιείται ως κανονικά κατανεμημένος θόρυβος. [29]

4.2 ARIMA

Ένα από τα πιο σημαντικά και ευρέως χρησιμοποιούμενα μοντέλα πρόβλεψης χρονοσειρών είναι το μοντέλο αυτοπαλίνδρομου ολοκληρωμένου κινητού μέσου όρου (autoregressive integrated moving average - ARIMA). Η δημοτικότητα του μοντέλου ARIMA οφείλεται στις στατιστικές του ιδιότητες καθώς και στη μεθοδολογία Box–Jenkins στη διαδικασία κατασκευής μοντέλων. Επιπλέον, διάφορα μοντέλα εκθετικής εξομάλυνσης μπορούν να εφαρμοστούν από τα μοντέλα ARIMA. Αν και τα μοντέλα ARIMA είναι αρκετά ευέλικτα καθώς μπορούν να αντιπροσωπεύουν πολλούς διαφορετικούς τύπους χρονοσειρών, π.χ. καθαρή αυτοπαλινδρομική (autoregressive - AR), καθαρό κινητό μέσο όρο (moving average - MA) και συνδυασμένη σειρά AR και MA (ARMA), ο κύριος περιορισμός τους είναι η προϋποθέση της γραμμικής μορφής του μοντέλου. Δηλαδή υποτίθεται μια δομή γραμμικής συσχέτισης μεταξύ των τιμών χρονοσειρών και επομένως, δεν μπορούν να εντοπιστούν μη γραμμικά μοτίβα από το μοντέλο ARIMA. Η προσέγγιση των γραμμικών μοντέλων σε πολύπλοκα προβλήματα του πραγματικού κόσμου δεν είναι πάντα ικανοποιητική.

Σε ένα μοντέλο ARIMA, η μελλοντική τιμή μιας μεταβλητής θεωρείται ότι είναι μια γραμμική συνάρτηση πολλών προηγούμενων παρατηρήσεων και τυχαίων σφαλμάτων. Δηλαδή, η διαδικασία που δημιουργεί τις χρονοσειρές έχει τη μορφή:

$$y_t = \theta_0 + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + ... + \varphi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - ... - \theta_q \varepsilon_{t-q} \quad (4.2)$$

όπου y_t και ε_t είναι η πραγματική τιμή και το τυχαίο σφάλμα τη χρονική περίοδο t, αντίστοιχα και τα $\varphi_i (i = 1, 2, 1, ..., p)$ και $\theta_j (j = 0, 1, 2, ..., q)$ είναι παράμετροι του μοντέλου. Τα p και q είναι ακέραιοι αριθμοί και συχνά αναφέρονται ως τάξεις του μοντέλου. Τα τυχαία σφάλματα ε_t θεωρούνται ότι κατανέμονται ανεξάρτητα και πανομοιότυπα με μέσο όρο μηδέν και σταθερή διακύμανση σ^2. Η εξίσωση 4.2 περιλαμβάνει αρκετές σημαντικές ειδικές περιπτώσεις της οικογένειας μοντέλων ARIMA. Αν $q = 0$, τότε η 4.2 γίνεται μοντέλο AR τάξης p. Όταν $p = 0$, το μοντέλο μειώνεται σε μοντέλο MA τάξης q.

Βασικό βήμα για το χτίσιμο του μοντέλου ARIMA είναι να προσδιοριστεί η κατάλληλη σειρά μοντέλου (p, q). Στο βήμα αυτό, συχνά απαιτείται μετασχηματισμός δεδομένων για να καταστεί σταθερή (stationary) η χρονοσειρά. Η σταθερότητα είναι απαραίτητη προϋπόθεση για την κατασκευή ενός μοντέλου ARIMA. Μια σταθερή χρονοσειρά έχει την ιδιότητα ότι τα στατιστικά χαρακτηριστικά της, όπως ο μέσος όρος και η δομή αυτοσυσχέτισης, είναι σταθερά με την πάροδο του χρόνου. Όταν η παρατηρούμενη χρονοσειρά παρουσιάζει τάση

και ετεροσκεδαστικότητα (heteroscedasticity) [1], πριν προσαρμοστεί ένα μοντέλο ARIMA, εφαρμόζεται διαφοροποίηση (differencing) και άλλοι μετασχηματισμοί στα δεδομένα για την αφαίρεση της τάσης και τη σταθεροποίηση της διακύμανσης. [30].

Στην περίπτωση που η χρονοσειρά υποστεί διαφοροποίηση εισάγεται στο μοντέλο και η μεταβλητή d η οποία αφορά τον αριθμό των διαφοροποιήσεων που απαιτούνται ώστε η σειρά να γίνει σταθερή. Οπότε συνολικά το χτίσιμο του μοντέλου ARIMA απαιτεί τον προσδιορισμό των τριών μεταβλητών p, d και q.

4.2.1 Auto-ARIMA

Το εργαλείο auto-ARIMA βοηθάει στην αυτοματοποίηση του προσδιορισμού των παραμέτρων p, d και q προσαρμόζοντας στην χρονοσειρά πολλά μοντέλα ARIMA διαφορετικής αλληλουχίας p, d και q και επιλέγει αυτό που απαιτεί το χαμηλότερο AIC (Akaike Information Criterion). Το AIC αποτελεί ένα μέτρο της ποιότητας του μοντέλου που λαμβάνεται προσομοιώνοντας την κατάσταση στην οποία δοκιμάζεται το μοντέλο, σε διαφορετικό σύνολο δεδομένων. Αφού υπολογιστούν τα διαφορετικά μοντέλα, μπορούν να συγκριθούν χρησιμοποιώντας αυτό το κριτήριο. Σύμφωνα με τη θεωρία του Akaike [31], το μοντέλο με την μεγαλύτερη ακρίβεια έχει το μικρότερο AIC. [32]

4.2.2 SARIMAX

Το μοντέλο SARIMAX (Seasonal AutoRegressive Integrated Moving Average with eXternal or exogenous regressors) αποτελεί επέκταση του ARIMA και χρησιμοποιείται σε σύνολα δεδομένων που παρουσιάζουν εποχικούς κύκλους. Η διαφορά μεταξύ ARIMA και SARIMAX είναι η εποχικότητα και οι εξωγενείς παράγοντες. Το SARIMAX απαιτεί όχι μόνο τα ορίσματα p, d και q που απαιτεί το ARIMA, αλλά επίσης άλλο ένα σύνολο ορισμάτων p, d και q που αφορούν την εποχικότητα καθώς και ένα όρισμα που ονομάζεται s που είναι η περιοδικότητα του εποχιακού κύκλου των δεδομένων.

Η γενική μορφή ενός μοντέλου SARIMAX έχει την μορφή:

$$L_t = \beta_0 + \sum_{j=1}^{m} \beta_j X_{t,j} + [\varphi^*(\mathrm{B})]^{-1} \theta^*(\mathrm{B}) \alpha_t \tag{4.3}$$

[1] Ετεροσκεδαστικότητα ονομάζεται η ιδιότητα μιας χρονοσειράς της οποίας οι τιμές παρουσιάζουν μεταβαλλόμενη διακύμανση.

όπου B είναι ο συνήθης τελεστής μετατόπισης προς τα πίσω ($\mathrm{B}^j z_t = z_{t-j}$), ο ακέραιος s είναι η εποχιακή περίοδος και τα α_t είναι ανεξάρτητα και πανομοιότυπα κατανεμημένα τυχαία υπολείμματα με μηδενικό μέσο όρο, διακύμανση σ_α^2 και πεπερασμένη κύρτωση. Το σφάλμα ϵ_t εμφανίζεται ως $[\varphi^*(\mathrm{B})]^{-1}\,\theta^*(\mathrm{B})\alpha_t$. Στην πράξη, η υπόθεση της διεργασίας λευκού θορύβου αναπαρίσταται στο α_t αντί για το ϵ_t. [33]

4.2.3 Augmented Dickey-Fuller test (ADFT)

Το Augmented Dickey-Fuller test είναι μια μέθοδος ελέγχου ύπαρξης μονάδας ρίζας (unit root) στην χρονοσειρά, καθιστώντας την μη σταθερή. Όπως υποδηλώνει το όνομα, το ADFT είναι μια προέκταση του Dickey Fuller Test (DFT). Το DFT ελέγχει την αρχική υπόθεση ότι $\alpha = 1$ στην ακόλουθη εξίσωση μοντέλου (εξίσωση 4.4) όπου α είναι ο συντελεστής της πρώτης καθυστέρησης (lag) στο Y. Αρχική υπόθεση H_0: $\alpha = 1$.

$$y_t = c + \beta_t + \alpha y_{t-1} + \varphi_1 \Delta Y_{t-1} + e_t \tag{4.4}$$

όπου y_{t-1} είναι η πρώτη καθυστέρηση της χρονοσειράς και ΔY_{t-1} η πρώτη διαφορά της χρονοσειράς την χρονική στιγμή $(t-1)$.

Το ADFT επεκτείνει την εξίσωση του DFT για να συμπεριλάβει τη διαδικασία παλινδρόμησης υψηλής τάξης στο μοντέλο. Η εξίσωση του ADFT είναι η εξής:

$$y_t = c + \beta_t + \alpha y_{t-1} + \varphi_1 \Delta Y_{t-1} + \varphi_2 \Delta Y_{t-2} + ... + \varphi_p \Delta Y_{t-p} + e_t \tag{4.5}$$

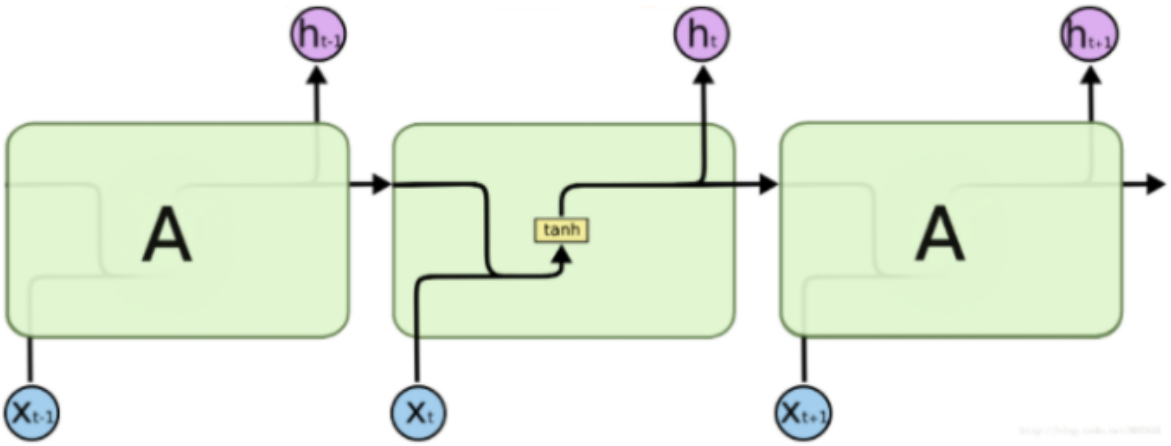
Εφόσον στην αρχική υπόθεση θεωρείται παρουσία μοναδιαίας ρίζας, δηλαδή $\alpha = 1$, η τιμή p που λαμβάνεται θα πρέπει να είναι μικρότερη από το επίπεδο σημαντικότητας (συνήθως 0.05) προκειμένου να απορριφθεί. Ως εκ τούτου, συμπεραίνεται ότι η σειρά είναι σταθερή. [34]

4.3 Recurrent Neural Networks

Ένα Επαναλαμβανόμενο Νευρωνικό Δίκτυο (Recurrent Neural Network - RNN) είναι μια κατηγορία Τεχνητών Νευρωνικών Δικτύων (Artificial Neural Network - ANN) όπου οι συνδέσεις μεταξύ των κόμβων σχηματίζουν ένα κατευθυνόμενο ή μη κατευθυνόμενο γράφημα

κατά μήκος μιας χρονικής ακολουθίας (σχήμα 4.1). Αυτό του επιτρέπει να επιδεικνύει χρονικά δυναμική συμπεριφορά. Προερχόμενα από τα Τροφοδοτικά Νευρωνικά Δίκτυα (Feed-forward Neural Networks), τα RNN μπορούν να χρησιμοποιήσουν την εσωτερική τους κατάσταση (μνήμη) για να επεξεργαστούν ακολουθίες μεταβλητού μήκους εισόδων. Ο όρος "αναδρομικό νευρωνικό δίκτυο" χρησιμοποιείται για να αναφερθεί στην κατηγορία δικτύων με άπειρη παλμική απόκριση. Ένα επαναλαμβανόμενο δίκτυο άπειρων παλμών είναι ένα κατευθυνόμενο κυκλικό γράφημα που δεν μπορεί να ξετυλιχτεί.

Τα επαναλαμβανόμενα δίκτυα άπειρων παλμών μπορούν να έχουν πρόσθετες αποθηκευμένες καταστάσεις και ο χώρος αποθήκευσης μπορεί να είναι υπό άμεσο έλεγχο από το νευρωνικό δίκτυο. Ο χώρος αποθήκευσης μπορεί επίσης να αντικατασταθεί από άλλο δίκτυο ή γράφημα εάν ενσωματώνει χρονικές καθυστερήσεις ή περιέχει βρόχους ανάδρασης. Τέτοιες ελεγχόμενες καταστάσεις αναφέρονται ως φραγμένη κατάσταση ή φραγμένη μνήμη και αποτελούν μέρος των Δικτύων Μακράς Βραχυπρόθεσμης Μνήμης (Long Short-Term Memory - LSTM) και των Φραγμένων Επαναλαμβανόμενων Μονάδων (Gated Recurrent Units - GRU).



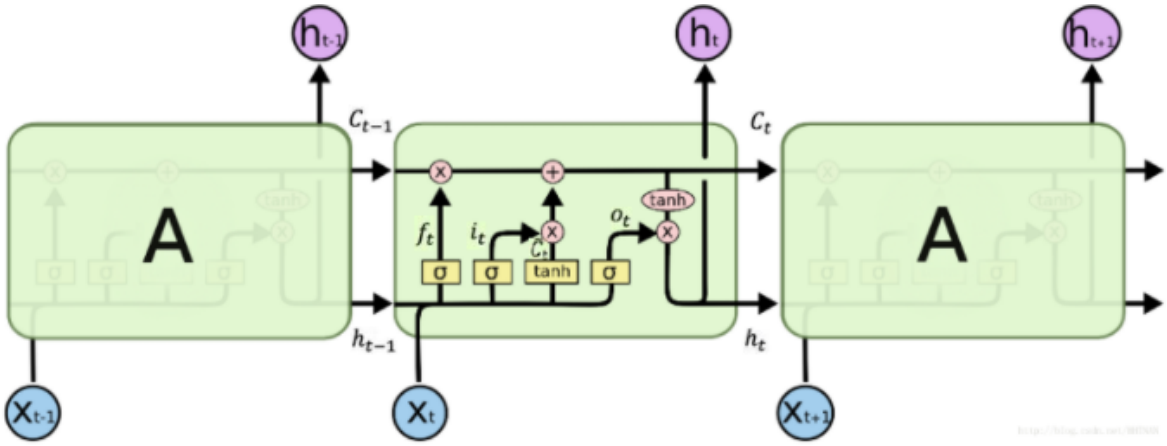
Σχήμα 4.1: Διαδοχικά τοποθετημένα κελιά RNN. [1]

4.3.1 LSTM

Μια μονάδα LSTM αποτελείται από ένα κελί (cell), μια πύλη εισόδου (input gate), μια πύλη εξόδου (output gate) και μια πύλη λήθης (forget gate) (σχήμα 4.2). Το κελί θυμάται τιμές σε αυθαίρετα χρονικά διαστήματα και οι τρεις πύλες ρυθμίζουν τη ροή των πληροφοριών από και προς το κελί.

Τα δίκτυα LSTM είναι κατάλληλα για ταξινόμηση, επεξεργασία και την πραγματοποίηση προβλέψεων με βάση δεδομένα χρονοσειρών, καθώς μπορεί να υπάρχουν καθυστερήσεις

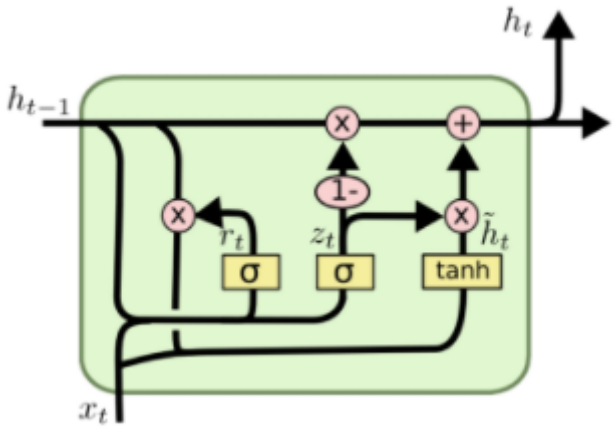
άγνωστης διάρκειας μεταξύ σημαντικών γεγονότων σε μια χρονοσειρά. Τα LSTM αναπτύχθηκαν για να αντιμετωπίσουν το πρόβλημα της εξαφάνισης της κλίσης (vanishing gradient) που μπορεί να παρουσιαστεί κατά την εκπαίδευση των παραδοσιακών RNN.



Σχήμα 4.2: Διαδοχικά τοποθετημένα κελιά LSTM. [1]

4.3.2 GRU

Το GRU είναι σαν ένα LSTM με μια πύλη λήθης αλλά έχει λιγότερες παραμέτρους, καθώς δεν διαθέτει πύλη εξόδου (σχήμα 4.3). Η απόδοση της GRU σε ορισμένες εργασίες μοντελοποίησης πολυφωνικής μουσικής, μοντελοποίησης σημάτων ομιλίας και επεξεργασίας φυσικής γλώσσας βρέθηκε να είναι παρόμοια με αυτή του LSTM. Οι GRU έχουν αποδειχθεί ότι παρουσιάζουν καλύτερη απόδοση σε ορισμένα μικρότερα και λιγότερο συχνά σύνολα δεδομένων.



Σχήμα 4.3: Δομή κελιού GRU. [1]

Κεφάλαιο 5

Ανάλυση και πρόβλεψη χρονοσειρών

5.1 Εισαγωγή

Το κεφάλαιο αυτό περιλαμβάνει πειράματα πρόβλεψης με τέσσερις διαφορετικές μεθόδους μηχανικής μάθησης, τα μοντέλα Facebook Prophet, ARIMA, SARIMAX και με νευρωνικά δίκτυα. Το κρυπτονόμισμα το οποίο μελετάται και με βάση το οποίο γίνεται η βελτιστοποίηση των μοντέλων, είναι το Bitcoin. Συγκεκριμένα οι προσπάθειες πρόβλεψης γίνονται μέσω μελέτης της τιμής κλεισίματος του κρυπτονομίσματος. Αφού επιλεχθεί το βέλτιστο μοντέλο γίνεται πρόβλεψη και για άλλα δημοφιλή κρυπτονομίσματα.

5.2 Περιγραφή δεδομένων

Τα δεδομένα [35] που αναλύθηκαν περιέχουν τις τιμές διάφορων δεικτών της εξέλιξης της τιμής 23 κρυπτονομισμάτων. Το διάστημα στο οποίο είναι διαθέσιμες οι τιμές για κάθε κρυπτονόμισμα και ο αριθμός των δεδομένων φαίνονται στον πίνακα 5.1. Η ημερομηνία από την οποία ξεκινά η διαθεσιμότητα των δεδομένων σχετίζεται με την ημερομηνία κυκλοφορίας του κάθε κρυπτονομίσματος στην αγορά, ενώ στις 7/7/2021 είναι η ημερομηνία συλλογής των δεδομένων. Συνεπώς κρυπτονομίσματα που κυκλφόρησαν πιο πρόσφατα από άλλα έχουν μικρότερο αριθμό διαθέσιμων τιμών. Τα περιεχόμενα του συνόλου δεδομένων περιλαμβάνουν τα εξής στοιχεία:

- Date: Η ημερομηνία κατά την οποία έγινε η παρατήρηση. Έχει την μορφή Έτος - Μήνας - Ημέρα (EEEE-MM-HH) και ακολουθείται από την χρονική στιγμή της καταγραφής στο 24ωρο σύστημα αναγραφής ώρας στη μορφή Ωρες : Λεπτά : Δευτερόλεπτα

Πίνακας 5.1: Διαθέσιμο διάστημα και αριθμός δεδομένων για τα 23 κρυπτονομίσματα

Κρυπτονόμισμα	**Δεδομένα από**	**Δεδομένα έως**	**Αριθμός δεδομένων**
Aave (AAVE)	6/11/2020	7/7/2021	275
Binance Coin (BNB)	27/7/2017	7/7/2021	1442
Bitcoin (BTC)	30/4/2013	7/7/2021	2991
Cardano (ADA)	3/10/2017	7/7/2021	1374
Chainlink (LINK)	22/9/2017	7/7/2021	1385
Cosmos (ATOM)	16/3/2019	7/7/2021	845
Crypto.com (CRO)	16/12/2018	7/7/2021	935
Dogecoin (DOGE)	13/12/2013	7/7/2021	2760
EOS (EOS)	3/7/2017	7/7/2021	1466
Ethereum (ETH)	9/8/2015	7/7/2021	2160
IOTA (MIOTA)	15/6/2017	7/7/2021	1484
Litecoin (LTC)	30/4/2013	7/7/2021	2991
Monero (XMR)	23/5/2014	7/7/2021	2602
NEM (XEM)	3/4/2015	7/7/2021	2288
Polkadot (DOT)	22/8/2020	7/7/2021	320
Solana (SOL)	12/4/2020	7/7/2021	452
Stellar (XLM)	7/8/2014	7/7/2021	2527
Tether (USDT)	27/2/2015	7/7/2021	2318
Tron (TRX)	15/9/2017	7/7/2021	1392
USD Coin (USDC)	10/10/2018	7/7/2021	1002
Uniswap (UNI)	19/9/2020	7/7/2021	292
Wrapped Bitcoin (WBTC)	1/2/2019	7/7/2021	888
XRP (XRP)	6/8/2013	7/7/2021	2893

(ΩΩ : ΛΛ : ΔΔ).

- High (Highest Price): Αφορά την μέγιστη τιμή του κρυπτονομίσματος που καταγράφηκε την δεδομένη ημέρα.

- Low (Lowest Price): Αφορά την ελάχιστη τιμή του κρυπτονομίσματος που κατάγράφηκε την δεδομένη ημέρα.

- Open (Opening Price): Η τιμή ανοίγματος είναι η τιμή στην οποία ανοίγει το κρυπτονόμισμα στην αγορά κατά την έναρξη των συναλλαγών την δεδομένη ημέρα. Υπάρχει περίπτωση να είναι διαφορετική από την τιμή κλεισίματος της προηγούμενης ημέρας. Για παράδειγμα το κρυπτονόμισμα μπορεί να ανοίξει σε υψηλότερη τιμή από την τιμή κλεισίματος λόγω υπερβολικής ζήτησης.

- Close (Closing Price): Η τιμή κλεισίματος αναφέρεται στην τελευταία τιμή στην οποία διαπραγματεύεται ενα κρυπτονόμισμα κατά τη διάρκεια της δεδομένης ημέρας. [36]

- Volume (Market Volume): Όγκος αγοράς είναι το άθροισμα των συναλλαγών που πραγματοποιούνται την δεδομένη ημέρα. Μεγαλύτερος όγκος συναλλαγών κρυπτονομισμάτων οδηγεί σε πιο δίκαιες τιμές κρυπτονομισμάτων και μειώνει την πιθανότητα στρεβλής τιμολόγησης. Αντίθετα χαμηλός όγκος ανταλλαγής κρυπτονομισμάτων σηματοδοτεί αναποτελεσματικές ή χαμηλές συναλλαγές, καθώς οι ζητούμενες τιμές των πωλητών αποτυγχάνουν να ανταποκριθούν στις προσφορές των πιθανών αγοραστών. Ο όγκος μπορεί να δείξει την κατεύθυνση και την κίνηση του κρυπτονομίσματος καθώς και μια πρόβλεψη της μελλοντικής τιμής και της ζήτησης του. [37]

- Market Cap (Market Capitalization): Η κεφαλαιοποίηση αγοράς είναι το άθροισμα της αξίας όλων των κρυπτονομισμάτων που έχουν παραχθεί την δεδομένη ημέρα και υπολογίζεται πολλαπλασιάζοντας τον αριθμό των κρυπτονομισμάτων που παράχθηκαν με την αξία του την δεδομένη στιγμή. Ως δείκτης μπορεί να αντιπροσωπεύσει την σταθερότητα ενός κρυπτονομίσματος. Κρυπτονομίσματα με μεγάλη κεφαλαιοποίηση αγοράς, συμπεριφέρονται πιο σταθερά στις μεταβολές, χωρίς μεγάλες πτώσεις ή αυξήσεις. Σύμφωνα με αυτόν το δείκτη τα κρυπτονομίσματα χωρίζονται σε:

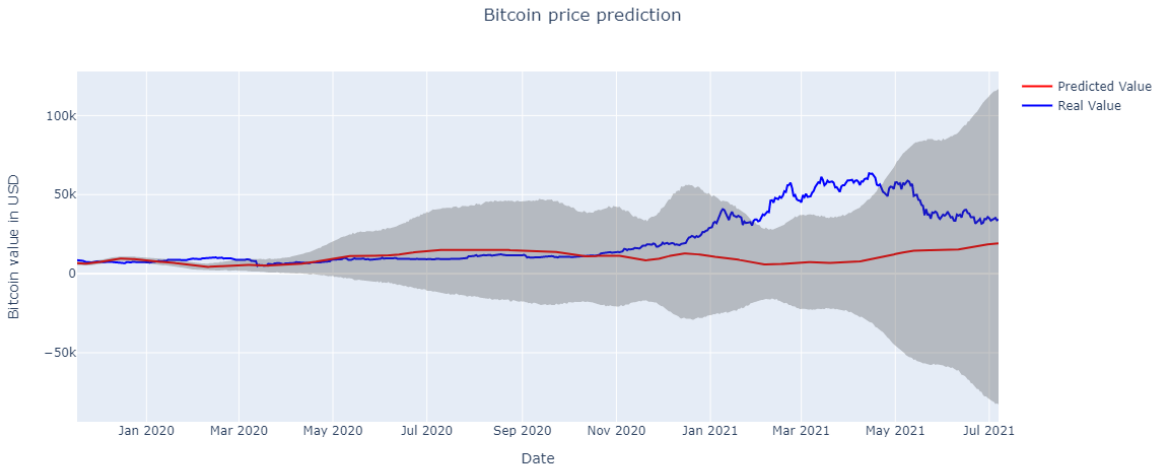
  – High-cap: Συμπεριλαμβανομένων των Bitcoin και Ethereum, έχουν κεφαλαιοποίηση άνω των 10 δισεκατομμυρίων δολαρίων.

– Mid-cap: Έχουν κεφαλαιοποίηση μεταξύ 1 και 10 δισεκατομμυρίων δολαρίων.

– Low-cap: Έχουν κεφαλαιοποίηση κάτω του 1 δισεκατομμυρίου δολαρίων. [38]

5.3 Το μοντέλο Facebook Prophet

Η πρώτη προσπάθεια πρόβλεψης έγινε με το αυτοματοποιημένο εργαλείο Facebook Prophet [39], χρησιμοποιώντας μία μεταβλητή πρόβλεψης, την τιμή "Close" του συνόλου δεδομένων του κρυπτονομίσματος. Τα δεδομένα (2991 τιμές) χωρίστηκαν σε 80% δεδομένα εκπαίδευσης του μοντέλου και 20% δεδομένα ελέγχου. Αυτο σημαίνει πως το μοντέλο χρησιμοποίησε την τιμή κλεισίματος (Close) από τις 29-04-2013 έως τις 15-11-2019 (2392 τιμές) για να εκπαιδευτεί και προέβλεψε την εξέλιξη της τιμής από τις 16-11-2013 έως τις 06-07-2021 (599 τιμές). Τα αποτελέσματα της πρόβλεψης παράλληλα με τα πραγματικα δεδομένα για το διάστημα 29-04-2013 έως 15-11-2019 παρουσιάζονται στο σχήμα 5.1 όπου με μπλε χρώμα επισημαίνονται οι πραγματικές τιμές του Bitcoin, με κόκκινο οι τιμές που προέκυψαν από την πρόβλεψη και με γκρι χρώμα το εύρος του πιθανού σφάλματος της πρόβλεψης.



Σχήμα 5.1: Αποτελέσματα πρόβλεψης του αλγορίθμου Facebook Prophet

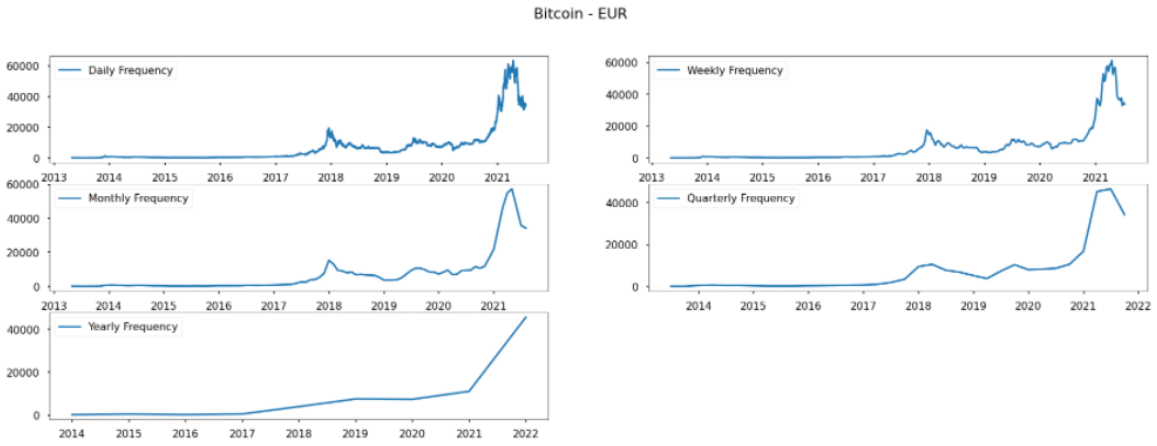
Ο αλγόριθμος πρόβλεψης Facebook Prophet δεν έδωσε ικανοποιητικά αποτελέσματα αφού δεν κατάφερε να προβλέψει σωστά τις αυξομειώσεις της τιμής του κρυπτονομίσματος (σφάλμα RMSE=20970.32). Αυτό πιθανότατα οφείλεται σε χαμηλή εποχικότητα των τιμών της χρονοσειράς.

5.4 Τα μοντέλα ARIMA και SARIMAX

Η επόμενη μέθοδος που χρησιμοποιήθηκε είναι το στατιστικό μοντέλο πρόβλεψης χρονοσειρών ARIMA και συγκεκριμένα το auto-ARIMA. Βασική προϋπόθεση για το ARIMA είναι η χρονοσειρά που προβλέπεται να είναι σταθερή, δηλαδή να μην παρουσιάζει τάσεις ή εποχικότητα. Το εργαλείο auto-ARIMA μετατρέπει αυτόματα την χρονοσειρά σε σταθερή και επιλέγει τις κατάλληλες παραμέτρους προκειμένου να χρησιμοποιηθεί βέλτιστα το ARIMA. Όπως και στο προηγούμενο μοντέλο έτσι και σε αυτή την περίπτωση μελετήθηκε η τιμή κλεισίματος του Bitcoin.

5.4.1 Δειγματοληψία

Προκειμένου να μελετηθεί η χρονοσειρά ως προς την τάση και την εποχικότητα, γίνεται εβδομαδιαία, μηνιαία, τριμηνιαία και ετήσια δειγματοληψία των τιμών. Οι χρονοσειρές που προκύπτουν από την δειγματοληψία παρουσιάζονται στο σχήμα 5.2.

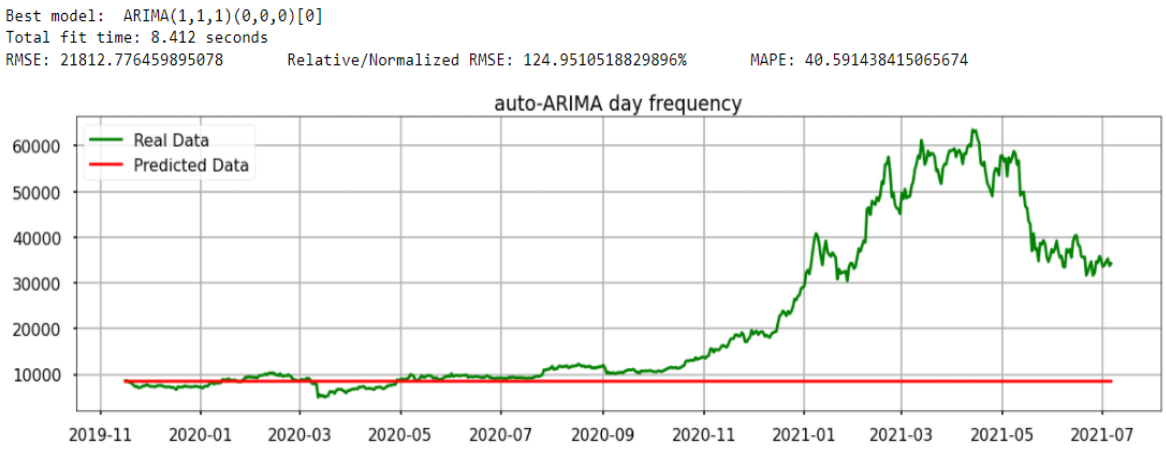


Σχήμα 5.2: Δειγματοληψία των τιμών κλεισίματος του Bitcoin

5.4.2 Αποτελέσματα πρόβλεψης με χρήση του μοντέλου auto-ARIMA

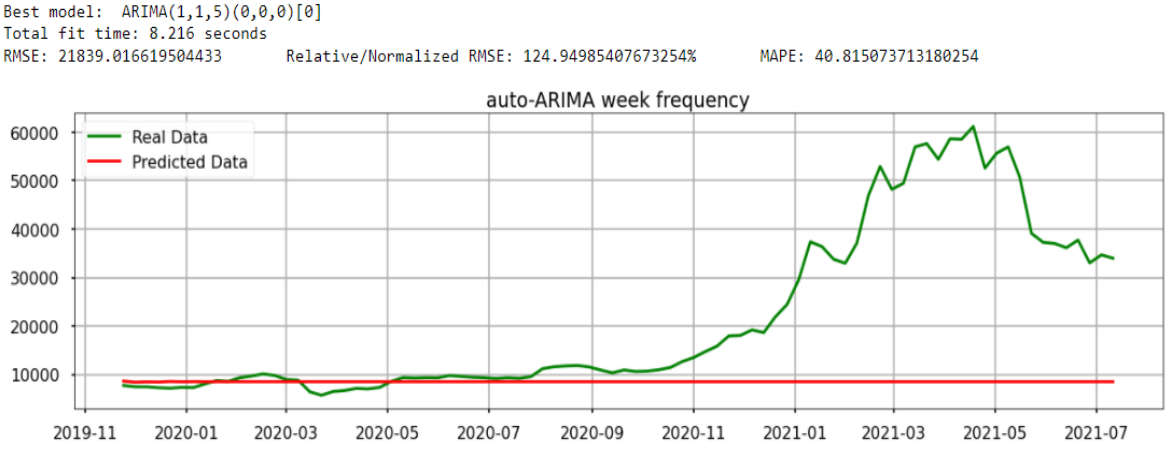
Ακολουθούν τα αποτελέσματα της πρόβλεψης για κάθε μια από τις χρονοσειρές που προέκυψαν από την δειγματοληψία. Έπειτα από επαναληπτικές δοκιμές, τα μοντέλα που πέτυχαν το χαμηλότερο AIC για κάθε χρονοσειρά είναι τα εξής:

- Χρονοσειρά ημερήσιας συχνότητας (σχήμα 5.3): Μοντέλο ARIMA(1,1,1)(0,0,0) με: AIC = 33171.355, RMSE = 21812.77, Relative/Normalized RMSE = 124.95% και MAPE = 40.59%.



Σχήμα 5.3: Πρόβλεψη ARIMA βάσει της χρονοσειράς ημερήσιας συχνότητας

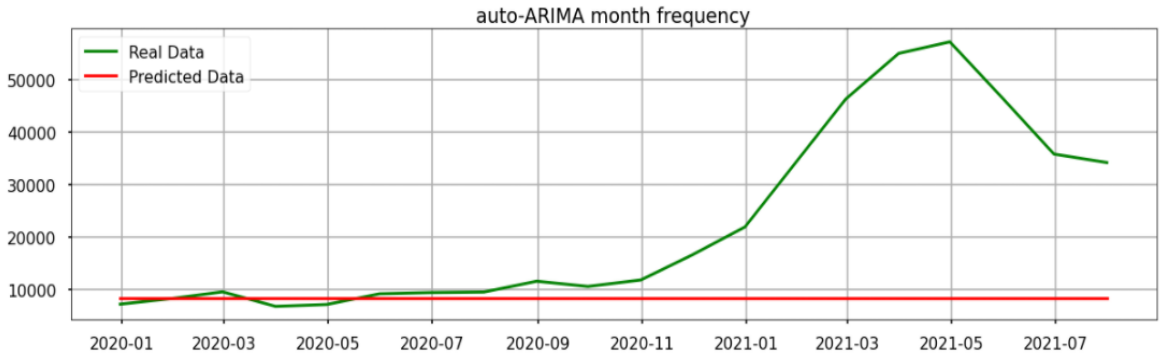
- Χρονοσειρά εβδομαδιαίας συχνότητας (σχήμα 5.4): Μοντέλο ARIMA(1,1,5)(0,0,0) με: AIC = 5186.360, RMSE = 21839.01, Relative/Normalized RMSE = 124.94% και MAPE = 40.81%.



Σχήμα 5.4: Πρόβλεψη ARIMA βάσει της χρονοσειράς εβδομαδιαίας συχνότητας

- Χρονοσειρά μηνιαίας συχνότητας (σχήμα 5.5): Μοντέλο ARIMA(0,1,1)(0,0,0) με: AIC = 1339.874, RMSE = 22143.64, Relative/Normalized RMSE = 126.12% και MAPE = 42.43%.

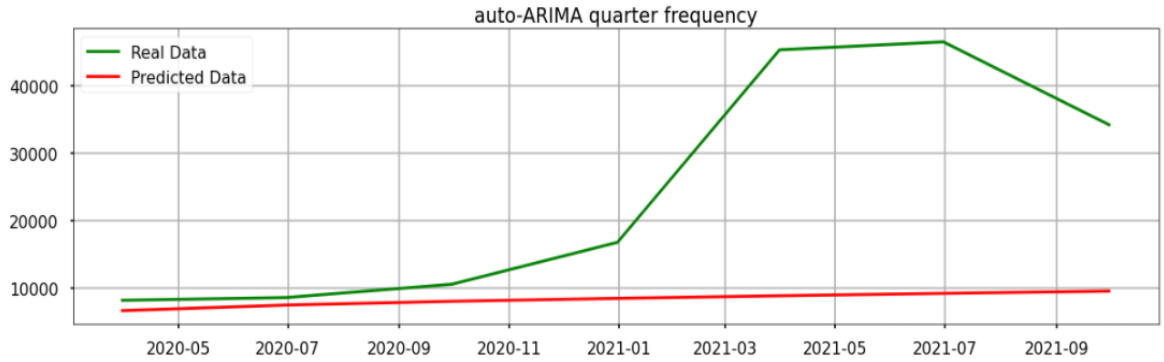
```
Best model:  ARIMA(0,1,1)(0,0,0)[0]
Total fit time: 0.643 seconds
RMSE: 22143.645033110173       Relative/Normalized RMSE: 126.12225492156162%       MAPE: 42.435747227700574
```



Σχήμα 5.5: Πρόβλεψη ARIMA βάσει της χρονοσειράς μηνιαίας συχνότητας

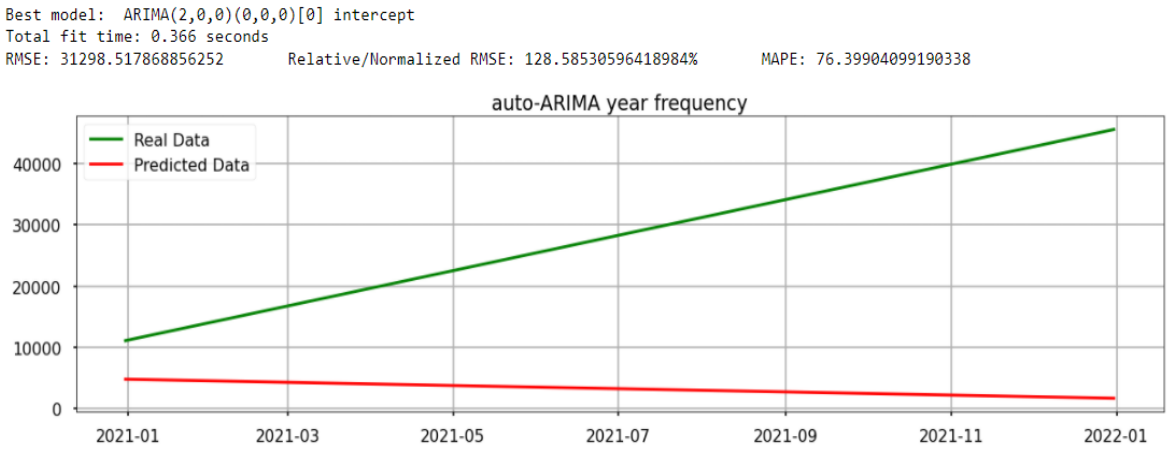
- Χρονοσειρά τριμηνιαίας συχνότητας (σχήμα 5.6): Μοντέλο ARIMA(1,1,2)(0,0,0) με: AIC = 459.538, RMSE = 22018.17, Relative/Normalized RMSE = 127.93% και MAPE = 48.07%.

```
Best model:  ARIMA(1,1,2)(0,0,0)[0] intercept
Total fit time: 1.270 seconds
RMSE: 22018.17701747769       Relative/Normalized RMSE: 127.93168721452612%       MAPE: 48.074825170366786
```



Σχήμα 5.6: Πρόβλεψη ARIMA βάσει της χρονοσειράς τριμηνιαίας συχνότητας

- Χρονοσειρά ετήσιας συχνότητας (σχήμα 5.7): Μοντέλο ARIMA(2,0,0)(0,0,0) με: AIC = 130.567, RMSE = 31298.51, Relative/Normalized RMSE = 128.58% και MAPE = 76.39%.

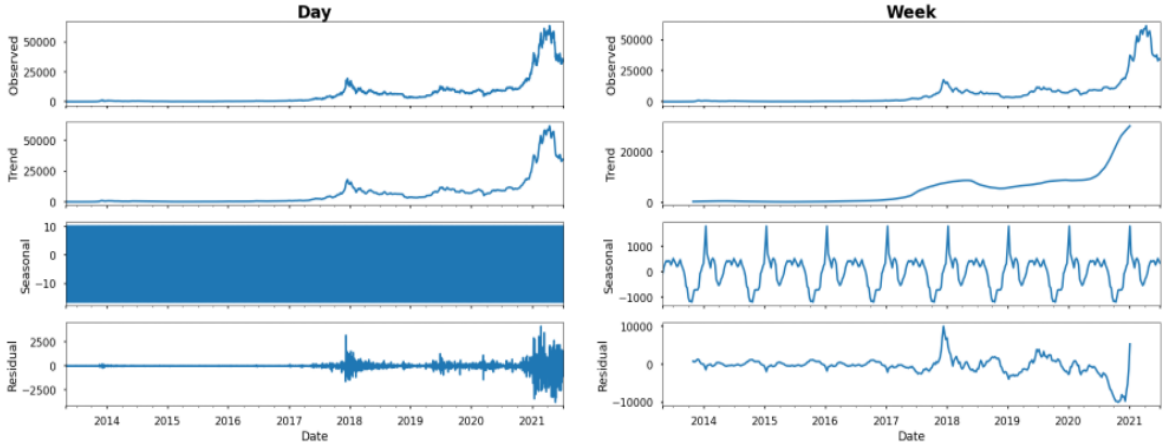


Σχήμα 5.7: Πρόβλεψη ARIMA βάσει της χρονοσειράς ετήσιας συχνότητας

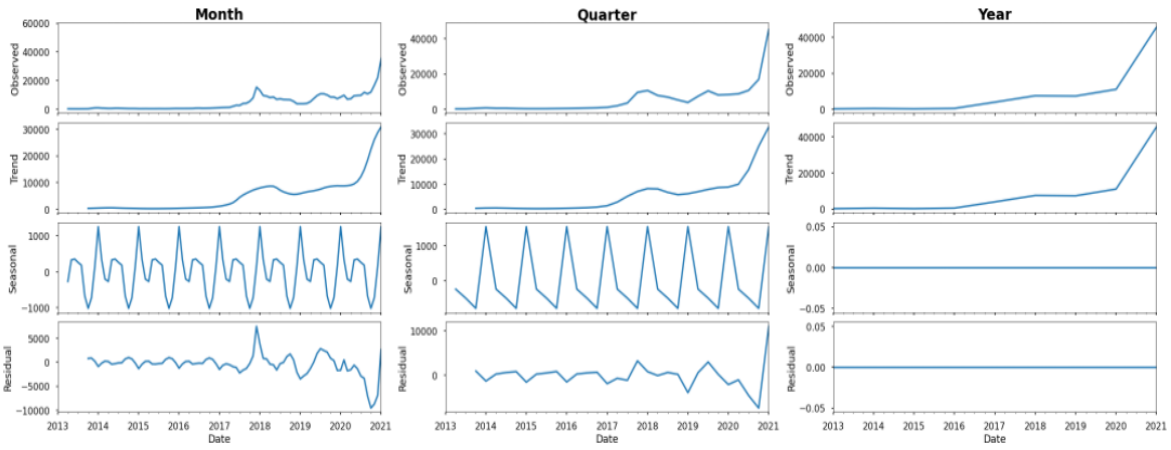
Τα αποτελέσματα των προβλέψεων δεν είναι ικανοποιητικά. Αυτό μπορεί να οφείλεται είτε στην ύπαρξη εποχικότητας στα δεδομένα την οποία ο αλγόριθμος ARIMA δεν κατάφέρνει να διαχειριστεί, είτε στην αποτυχία μετατροπής των χρονοσειρών σε σταθερές. Στις επόμενες ενότητες εξετάζεται η σταθερότητα της χρονοσειράς και γίνεται προσπάθεια σταθεροποίησής της.

5.4.3 Έλεγχος σταθερότητας χρονοσειρών

Στόχος σε αυτό το βήμα είναι να εξεταστεί αν υπάρχει εποχικότητα και τάση στις 5 χρονοσειρές, αρχικά από την ανάλυση STL(Seasonal-Trend decomposition using LOESS) και έπειτα από το Augmented Dickey-Fuller Test. Τα αποτελέσματα της αποσύνθεσης STL φαίνονται στα σχήματα 5.8 και 5.9 και τα αποτελέσματα του Augmented Dickey-Fuller Test στον πίνακα 5.2. Οι χρονοσειρές δεν είναι σταθερές και θα χρειαστούν περαιτέρω μετατροπές ώστε να μπορούν να χρησιμοποιηθούν στο μοντέλο ARIMA.



Σχήμα 5.8: Αποσύνθεση STL της αρχικής χρονοσειράς ημερήσιας και εβδομαδιαίας συχνότητας



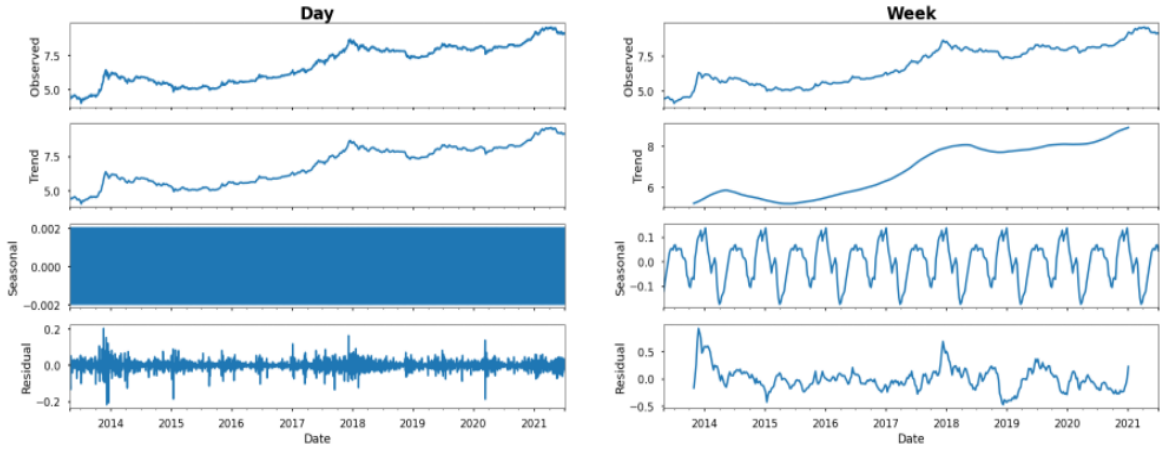
Σχήμα 5.9: Αποσύνθεση STL της αρχικής χρονοσειράς μηνιαίας,τριμηνιαίας και ετήσιας συχνότητας

Πίνακας 5.2: Αποτελέσματα κριτηρίου Dickey-Fuller για τις αρχικές χρονοσειρές

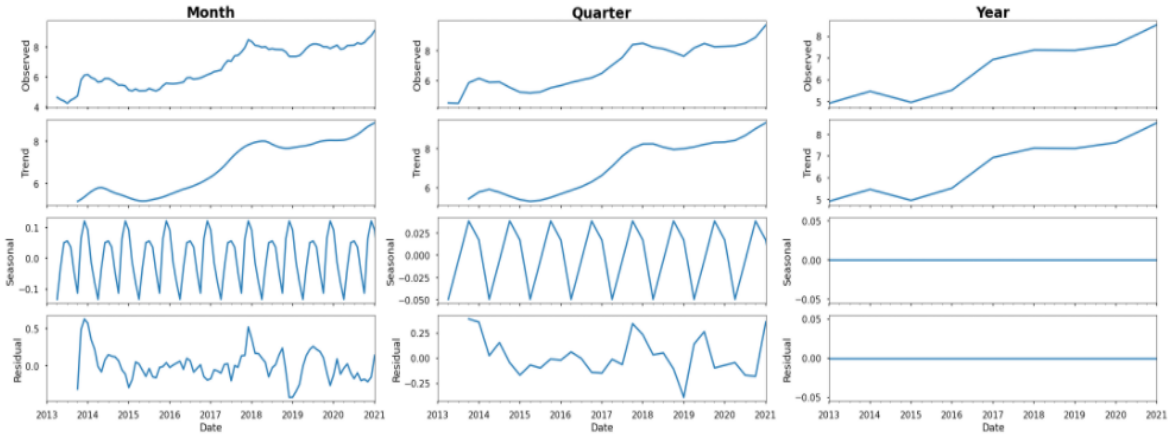
Συχνότητα	**Αποτελέσματα Augmented Dickey-Fuller test**
Ημερήσια	*p=0.819911*
Εβδομαδιαία	*p=0.984068*
Μηνιαία	*p=0.999028*
Τριμηνιαία	*p=0.998702*
Ετήσια	*p=1*

5.4.4 Μετασχηματισμός Box-Cox

Το πρώτο βήμα είναι ο μετασχηματισμός Box-Cox ο οποίος δεν κατάφέρνει να σταθεροποιήσει τις χρονοσειρές όπως προκύπτει από την ανάλυση STL (σχήματα 5.10 και 5.11) και το Augmented Dickey-Fuller Test (πίνακας 5.3).



Σχήμα 5.10: Αποσύνθεση STL της χρονοσειράς ημερήσιας και εβδομαδιαίας συχνότητας μετασχηματισμένης κατά Box-Cox



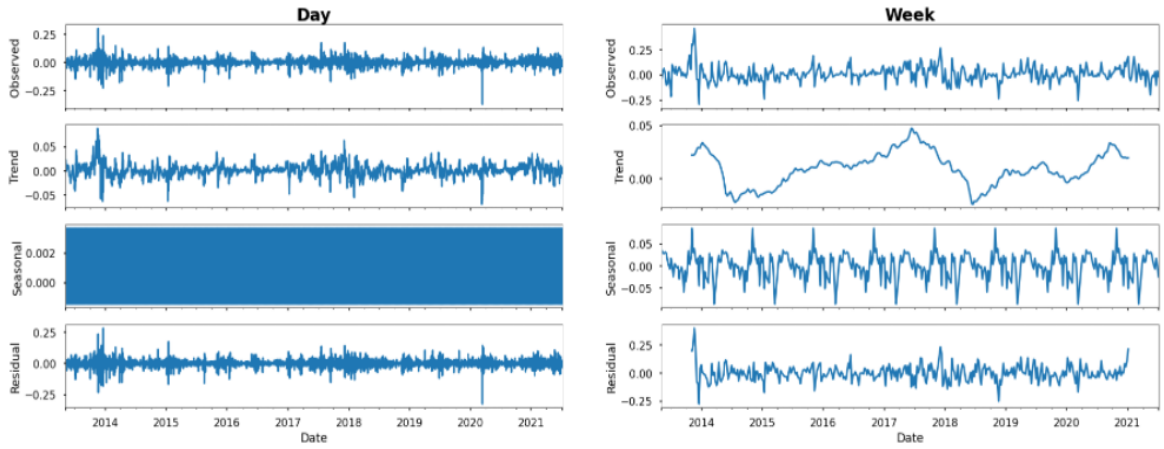
Σχήμα 5.11: Αποσύνθεση STL της χρονοσειράς μηνιαίας, τριμηνιαίας και ετήσιας συχνότητας μετασχηματισμένης κατά Box-Cox

Πίνακας 5.3: Αποτελέσματα κριτηρίου Dickey-Fuller για τις μετασχηματισμένες κατά Box-Cox χρονοσειρές

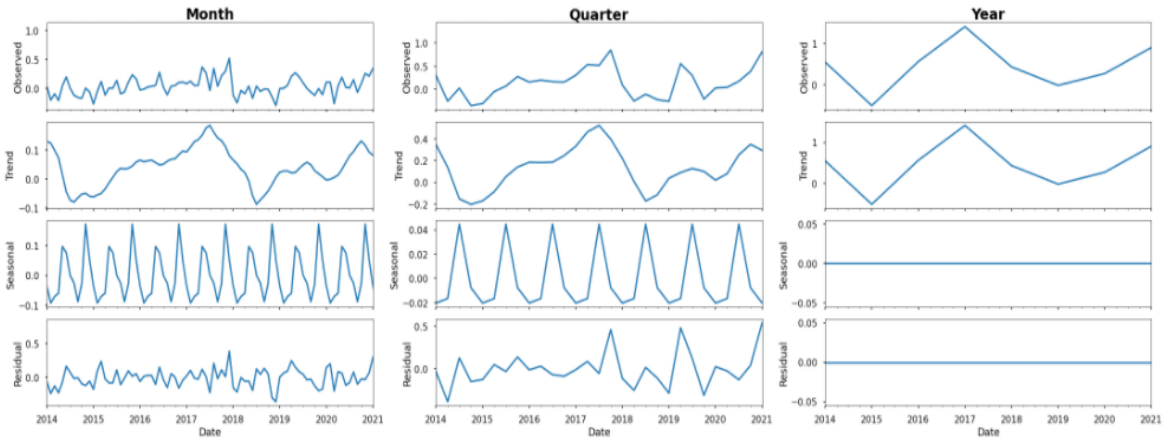
Συχνότητα	**Αποτελέσματα Augmented Dickey-Fuller test**
Ημερήσια	*p=0.801816*
Εβδομαδιαία	*p=0.771178*
Μηνιαία	*p=0.699045*
Τριμηνιαία	*p=0.576303*
Ετήσια	*p=0.772739*

5.4.5 Διαφοροποίηση των χρονοσειρών μετασχηματισμένων κατά Box-Cox

Ως δεύτερη προσπάθεια σταθεροποίησης εφαρμόζεται διαφοροποίηση στις χρονοσειρές που προέκυψαν από τον μετασχηματισμό Box-Cox. Στα σχήματα 5.12 και 5.13 παρουσιάζονται οι διαφοροποιημένες χρονοσειρές και τα αποτελέσματα του Augmented Dickey-Fuller Test στον πίνακα 5.4. Οι χρονοσειρές φαίνεται να παρουσιάζουν τάση και εποχικότητα όμως η τιμή p όλων των χρονοσειρών, εκτός αυτής με τριμηνιαία συχνότητα, είναι μικρότερη του 0.05 οπότε υπάρχει πιθανότητα να είναι σταθερές και να μπορούν να χρησιμοποιηθούν στο μοντέλο ARIMA.



Σχήμα 5.12: Αποσύνθεση STL της χρονοσειράς ημερήσιας και εβδομαδιαίας συχνότητας μετασχηματισμένης κατά Box-Cox και διαφοροποιημένης



Σχήμα 5.13: Αποσύνθεση STL της χρονοσειράς μηνιαίας, τριμηνιαίας και ετήσιας συχνότητας μετασχηματισμένης κατά Box-Cox και διαφοροποιημένης

Πίνακας 5.4: Αποτελέσματα κριτηρίου Dickey-Fuller για τις διαφοροποιημένες χρονοσειρές

Συχνότητα	**Αποτελέσματα Augmented Dickey-Fuller test**
Ημερήσια	*p=0*
Εβδομαδιαία	*p=0*
Μηνιαία	*p=0*
Τριμηνιαία	*p=0.002951*
Ετήσια	*p=0*

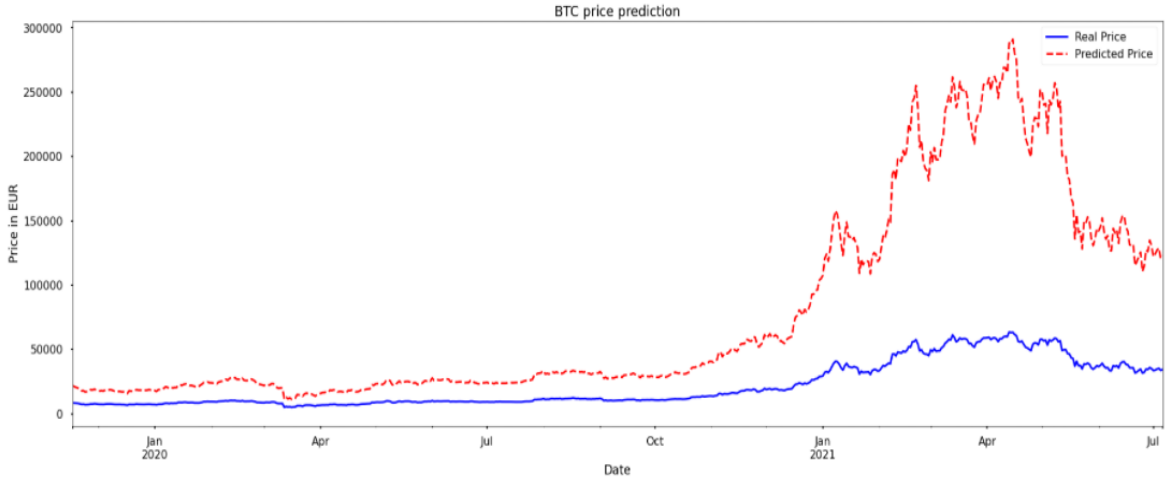
5.4.6 Αποτελέσματα πρόβλεψης με τη χρήση του μοντέλου SARIMAX

Το μοντέλο SARIMAX έχει την ίδια απαίτηση με το μοντέλο ARIMA ως προς την σταθερότητα αλλά χρησιμοποιείται για δεδομένα που παρουσιάζουν εποχικούς κύκλους. Θα δοκιμαστούν ξανά τα παραπάνω σύνολα δεδομένων χρησιμοποιώντας αυτή τη φορά τον αλγόριθμο SARIMAX. Οι χρονοσειρές τριμηνιαίας και ετήσιας συχνότητας παρουσίασαν πρόβλημα κατά την εκτέλεση του κώδικα το οποίο πιθανώς οφείλεται σε σφάλμα της υλοποίησης της βιβλιοθήκης scipy.linalg.schur. Ακολουθούν τα μοντέλα που επιλέχθηκαν και τα αποτελέσματα της πρόβλεψης για κάθε χρονοσειρά:

- Χρονοσειρά ημερήσιας συχνότητας διαφοροποιημένης και μετασχηματισμένης κατά Box-Cox (Augmented Dickey-Fuller Test δείκτης p=0<0.05). Τα στοιχεία του μοντέλου που επιλέχθηκε βρίσκονται στο σχήμα 5.14 και το αποτέλεσμα της πρόβλεψης του SARIMAX βρίσκεται στο σχήμα 5.15.

```
      parameters          aic
7   (0, 1, 0, 1) -8724.837396
9   (0, 1, 1, 1) -8723.827521
25  (1, 1, 0, 1) -8722.971386
13  (0, 2, 0, 1) -8722.825216
31  (1, 2, 0, 1) -8722.527312
                                SARIMAX Results
==========================================================================================
Dep. Variable:                    Close_box_diff   No. Observations:                 2392
Model:             SARIMAX(0, 1, 1)x(0, 1, 1, 52)   Log Likelihood                4365.419
Date:                            Sat, 08 Jan 2022   AIC                           -8724.837
Time:                                    20:49:31   BIC                           -8707.565
Sample:                                04-30-2013   HQIC                          -8718.545
                                     - 11-16-2019
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.9904      0.003   -332.665      0.000      -0.996      -0.985
ma.S.L52      -0.9921      0.059    -16.765      0.000      -1.108      -0.876
sigma2         0.0013   7.45e-05     17.349      0.000       0.001       0.001
===================================================================================
Ljung-Box (L1) (Q):                   0.20   Jarque-Bera (JB):              3626.93
Prob(Q):                              0.66   Prob(JB):                         0.00
Heteroskedasticity (H):               0.65   Skew:                            -0.43
Prob(H) (two-sided):                  0.00   Kurtosis:                         9.04
===================================================================================
```

Σχήμα 5.14: Επιλογή ιδανικού μοντέλου SARIMAX για την χρονοσειρά ημερήσιας συχνότητας διαφοροποιημένης και μετασχηματισμένης κατά Box-Cox



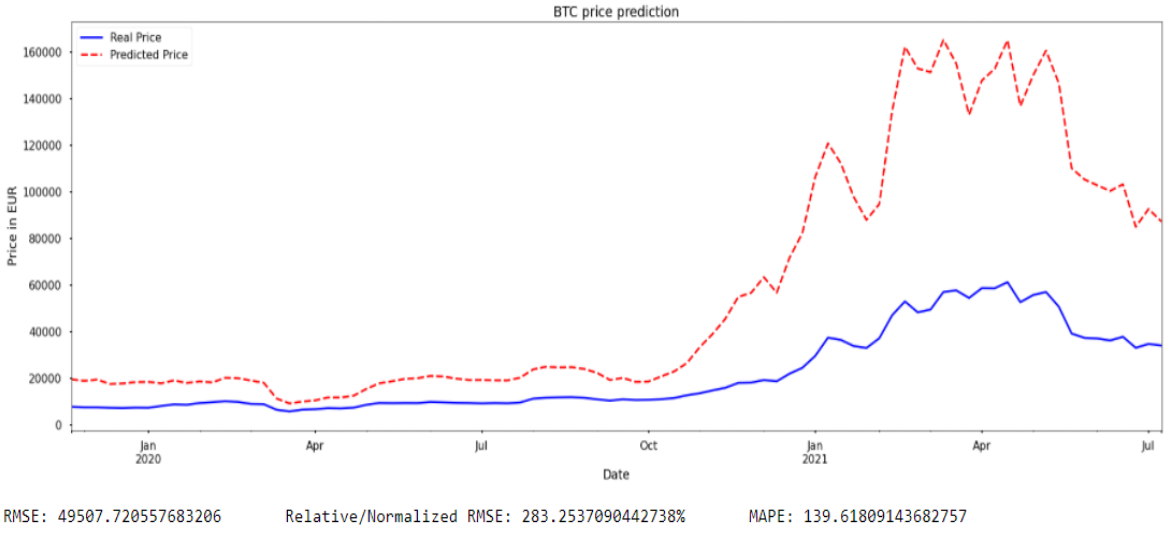
RMSE: 83663.36648049628 Relative/Normalized RMSE: 479.07506377132796% MAPE: 209.71480303282902

Σχήμα 5.15: Αποτέλεσμα πρόβλεψης SARIMAX βάσει της χρονοσειράς ημερήσιας συχνότητας διαφοροποιημένης και μετασχηματισμένης κατά Box-Cox

- Χρονοσειρά εβδομαδιαίας συχνότητας διαφοροποιημένης και μετασχηματισμένης κατά Box-Cox (Augmented Dickey-Fuller Test δείκτης p=0<0.05). Τα στοιχεία του μοντέλου που επιλέχθηκε βρίσκονται στο σχήμα 5.16 και το αποτέλεσμα της πρόβλεψης του SARIMAX βρίσκεται στο σχήμα 5.17.

```
      parameters         aic
25  (1, 1, 0, 1) -561.935507
51  (2, 2, 1, 1) -561.628806
27  (1, 1, 1, 1) -561.110947
33  (1, 2, 1, 1) -560.985735
29  (1, 1, 2, 1) -560.975706
                                SARIMAX Results
==========================================================================================
Dep. Variable:                    Close_box_diff   No. Observations:                  341
Model:             SARIMAX(1, 1, 1)x(0, 1, 1, 52)   Log Likelihood                284.968
Date:                            Sat, 08 Jan 2022   AIC                            -561.936
Time:                                    18:15:19   BIC                            -547.284
Sample:                                05-12-2013   HQIC                           -556.064
                                     - 11-17-2019
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.3660      0.057      6.443      0.000       0.255       0.477
ma.L1         -0.9820      0.022    -44.511      0.000      -1.025      -0.939
ma.S.L52      -0.9471      0.463     -2.044      0.041      -1.855      -0.039
sigma2         0.0060      0.003      2.287      0.022       0.001       0.011
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                21.88
Prob(Q):                              0.99   Prob(JB):                         0.00
Heteroskedasticity (H):               0.83   Skew:                            -0.50
Prob(H) (two-sided):                  0.35   Kurtosis:                         3.91
===================================================================================
```

Σχήμα 5.16: Επιλογή ιδανικού μοντέλου SARIMAX για την χρονοσειρά εβδομαδιαίας συχνότητας διαφοροποιημένης και μετασχηματισμένης κατά Box-Cox



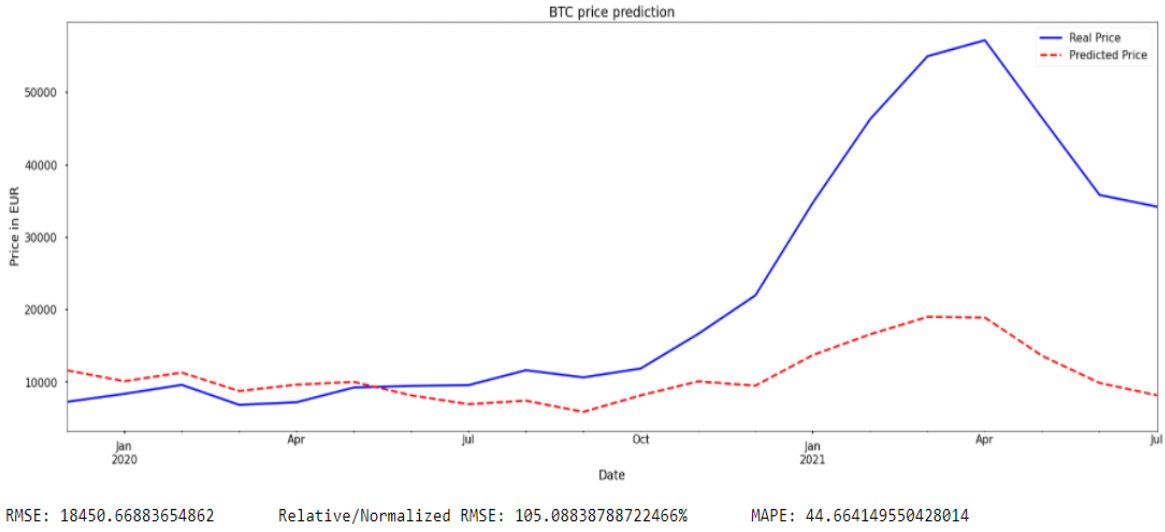
RMSE: 49507.720557683206 Relative/Normalized RMSE: 283.2537090442738% MAPE: 139.61809143682757

Σχήμα 5.17: Αποτέλεσμα πρόβλεψης SARIMAX βάσει της χρονοσειράς εβδομαδιαίας συχνότητας διαφοροποιημένης και μετασχηματισμένης κατά Box-Cox

- Χρονοσειρά μηνιαίας συχνότητας διαφοροποιημένης και μετασχηματισμένης κατά Box-Cox (Augmented Dickey-Fuller Test δείκτης p=0<0.05). Τα στοιχεία του μοντέλου που επιλέχθηκε βρίσκονται στο σχήμα 5.18 και το αποτέλεσμα της πρόβλεψης του SARIMAX βρίσκεται στο σχήμα 5.19.

```
      parameters         aic
0   (0, 0, 0, 0)    9.816804
42  (2, 1, 0, 0)   10.393353
24  (1, 1, 0, 0)   10.490418
30  (1, 2, 0, 0)   11.087765
48  (2, 2, 0, 0)   11.155050
                                SARIMAX Results
==========================================================================================
Dep. Variable:                    Close_box_diff   No. Observations:                   79
Model:             SARIMAX(0, 1, 0)x(0, 1, 0, 52)   Log Likelihood                -3.908
Date:                            Sat, 08 Jan 2022   AIC                            9.817
Time:                                    21:23:32   BIC                           11.075
Sample:                                05-31-2013   HQIC                          10.179
                                     - 11-30-2019
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
sigma2         0.0791      0.019      4.233      0.000       0.042       0.116
===================================================================================
Ljung-Box (L1) (Q):                   0.09   Jarque-Bera (JB):                 0.62
Prob(Q):                              0.76   Prob(JB):                         0.73
Heteroskedasticity (H):               0.23   Skew:                            -0.01
Prob(H) (two-sided):                  0.04   Kurtosis:                         3.76
===================================================================================
```

Σχήμα 5.18: Επιλογή ιδανικού μοντέλου SARIMAX για την χρονοσειρά μηνιαίας συχνότητας διαφοροποιημένης και μετασχηματισμένης κατά Box-Cox



Σχήμα 5.19: Αποτέλεσμα πρόβλεψης SARIMAX βάσει της χρονοσειράς μηνιαίας συχνότητας διαφοροποιημένης και μετασχηματισμένης κατά Box-Cox

5.4.7 Συμπεράσματα

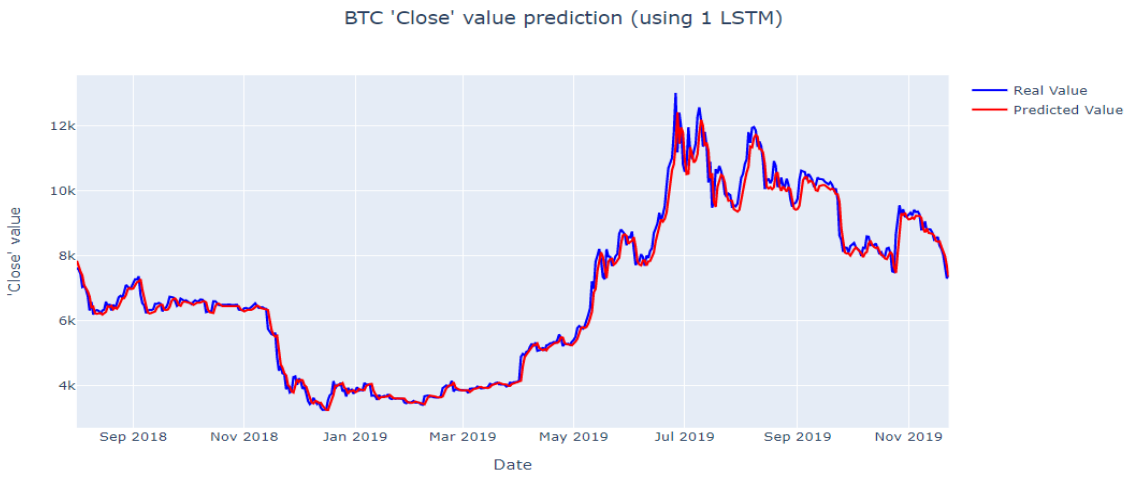
Τα αποτελέσματα του αλγορίθμου SARIMAX είναι σαφώς καλύτερα από αυτά του μοντέλου auto-ARIMA αλλά και από αυτά του Facebook Prophet. Ο αλγόριθμος αντιλαμβάνεται σε καλύτερο βαθμό τις αυξομειώσεις της τιμής του κρυπτονομίσματος και γι' αυτό είναι και σε θέση να προβλέψει σωστά τις περισσότερες φορές την τάση. Παρόλα αυτά το σφάλμα της πρόβλεψης εξακολουθεί να είναι αυξημένο αφού οι τιμές που έδωσε ο αλγόριθμος απέχουν αρκετά από τις πραγματικές. Η επόμενη προσπάθεια πρόβλεψης γίνεται με την βοήθεια νευρωνικών δικτύων τα οποία αποδεικνύονται πιο αποδοτικά στο να αντιλαμβάνονται τις μεταβολές των δεδομένων.

5.5 Νευρωνικά Δίκτυα

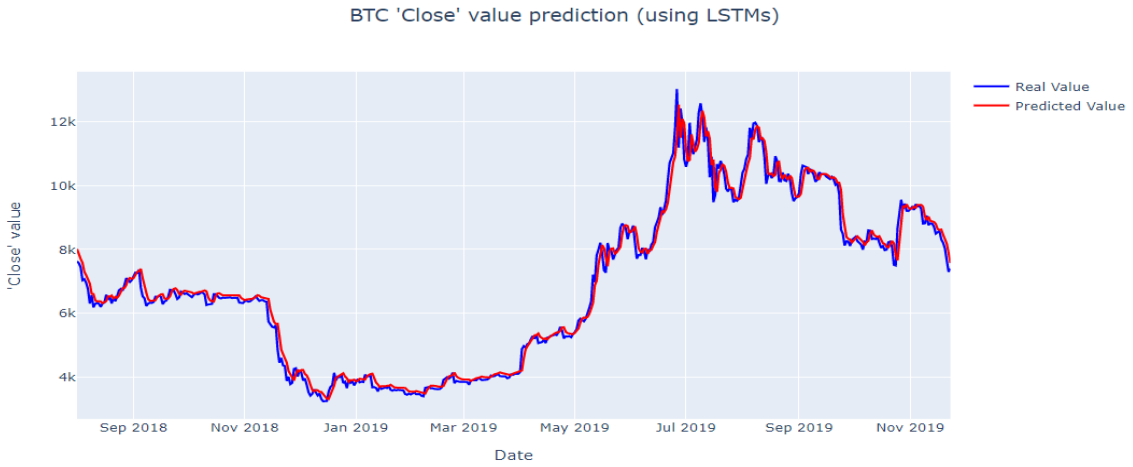
Σε μια προσπάθεια βελτιστοποίησης της πρόβλεψης του SARIMAX εισάγονται στη διαδικασία και τα νευρωνικά δίκτυα. Συγκεκριμένα επιλέγονται τα αναδρομικά Long Short Term Memory (LSTM) και Gated Recurrent Unit (GRU) δίκτυα. Βασικό στοιχείο τους είναι η δυνατότητα τους να συνδυάζουν παρελθοντικές πληροφορίες χρησιμοποιώντας μονάδες μνήμης, με συνέπεια να δίνουν καλύτερα αποτελέσματα από τον αλγόριθμο ARIMA, όπως δείχνουν και αποτελέσματα άλλων ερευνών [40].

5.5.1 Επιλογή στρωμάτων νευρωνικού δικτύου

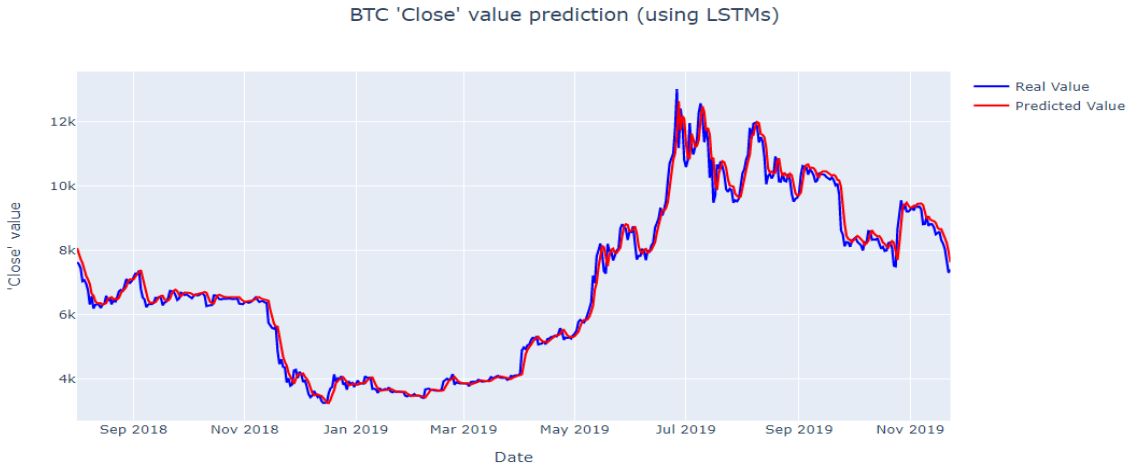
Ο αριθμός και το είδος του κάθε στρώματος του νευρωνικού δικτύου επιλέχθηκαν δοκιμάζοντας αρκετούς πιθανούς συνδυασμούς των στρωμάτων LSTM και GRU και επιλέγοντας αυτόν με το μικρότερο RMSE. Τα νευρωνικά που δοκιμάστηκαν αποτελούνται από στρώματα των 50 νευρώνων και είναι τα εξής: νευρωνικό δίκτυο 1 (1 στρώμα LSTM) (σχήμα 5.20), νευρωνικό δίκτυο 2 (2 στρώματα LSTM) (σχήμα 5.21), νευρωνικό δίκτυο 3 (3 στρώματα LSTM) (σχήμα 5.22), νευρωνικό δίκτυο 4 (1 στρώμα GRU) (σχήμα 5.25), νευρωνικό δίκτυο 5 (2 στρώματα GRU) (σχήμα 5.24), νευρωνικό δίκτυο 6 (3 στρώματα GRU) (σχήμα 5.25), νευρωνικό δίκτυο 7 (4 στρώματα GRU) (σχήμα 5.26), νευρωνικό δίκτυο 8 (1 στρώμα LSTM και 1 στρώμα GRU) (σχήμα 5.27), νευρωνικό δίκτυο 9 (1 στρώμα GRU και 1 στρώματα LSTM) (σχήμα 5.28) και νευρωνικό δίκτυο 10 (1 στρώμα GRU και 2 στρώματα LSTM) (σχήμα 5.29)



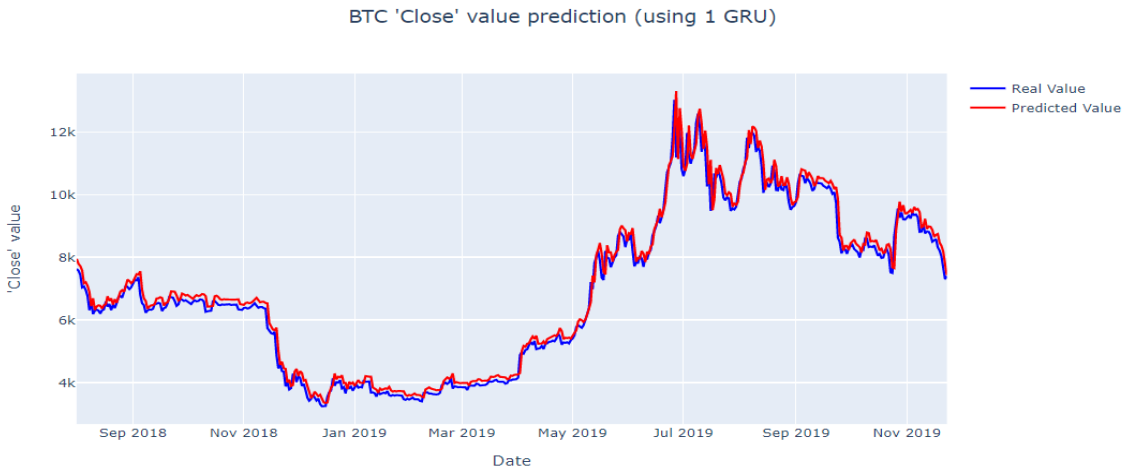
Σχήμα 5.20: Νευρωνικό δίκτυο 1: 1 στρώμα LSTM 50 νευρώνων



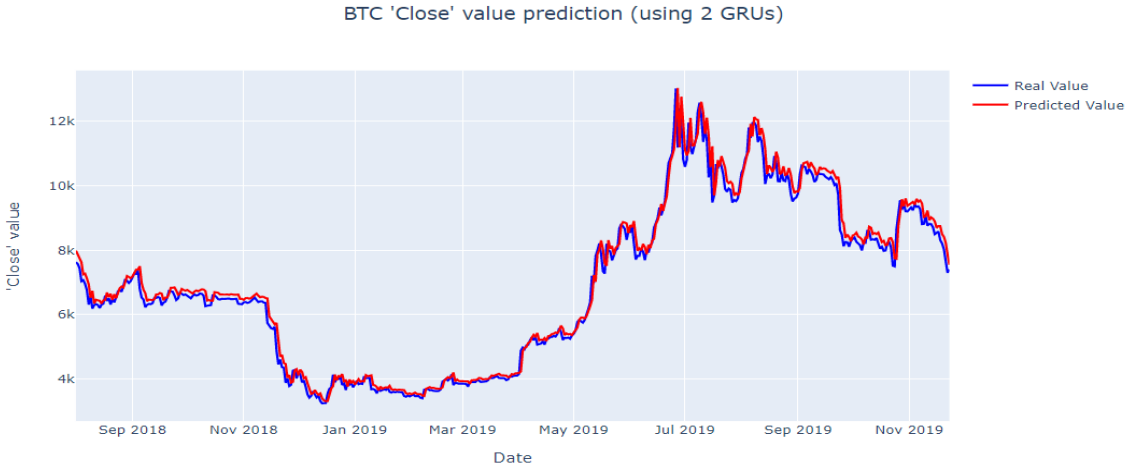
Σχήμα 5.21: Νευρωνικό δίκτυο 2: 2 στρώματα LSTM



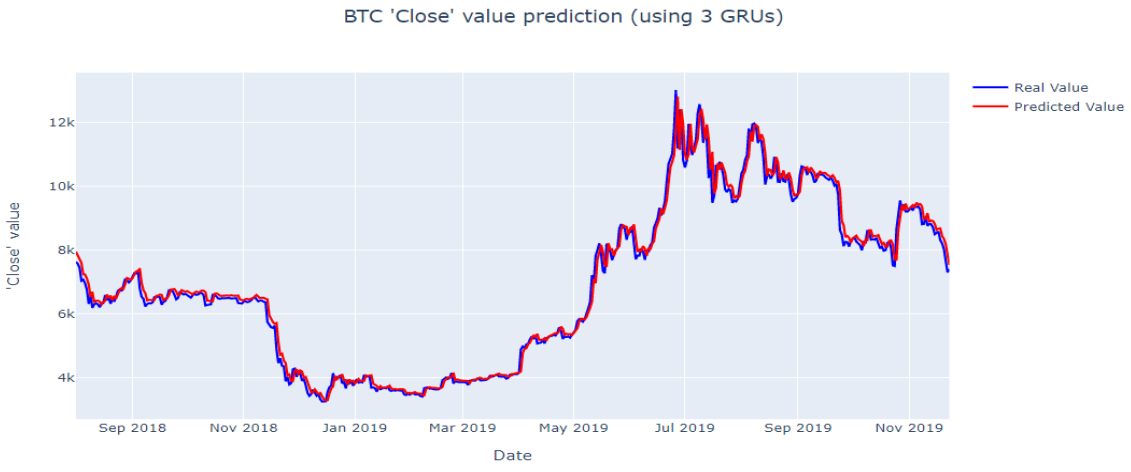
Σχήμα 5.22: Νευρωνικό δίκτυο 3: 3 στρώματα LSTM



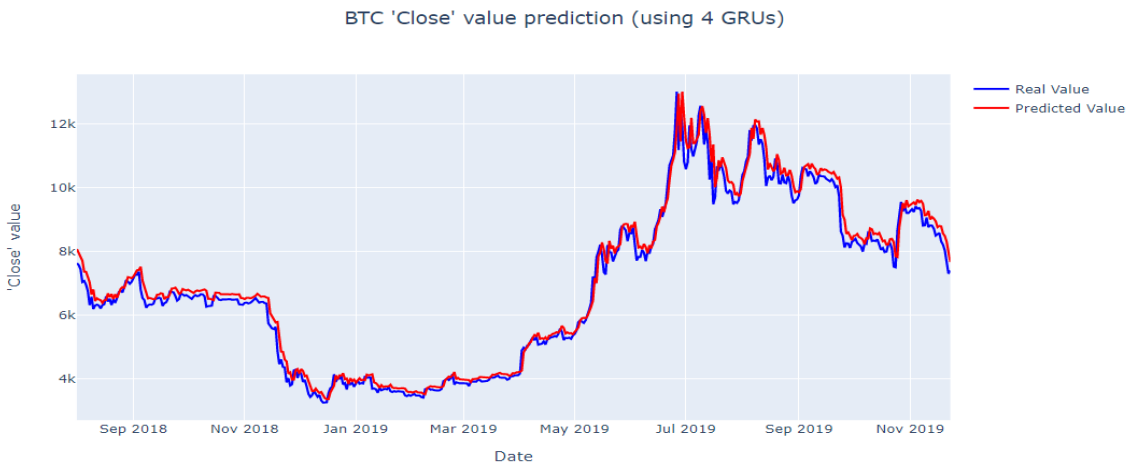
Σχήμα 5.23: Νευρωνικό δίκτυο 4: 1 στρώμα GRU



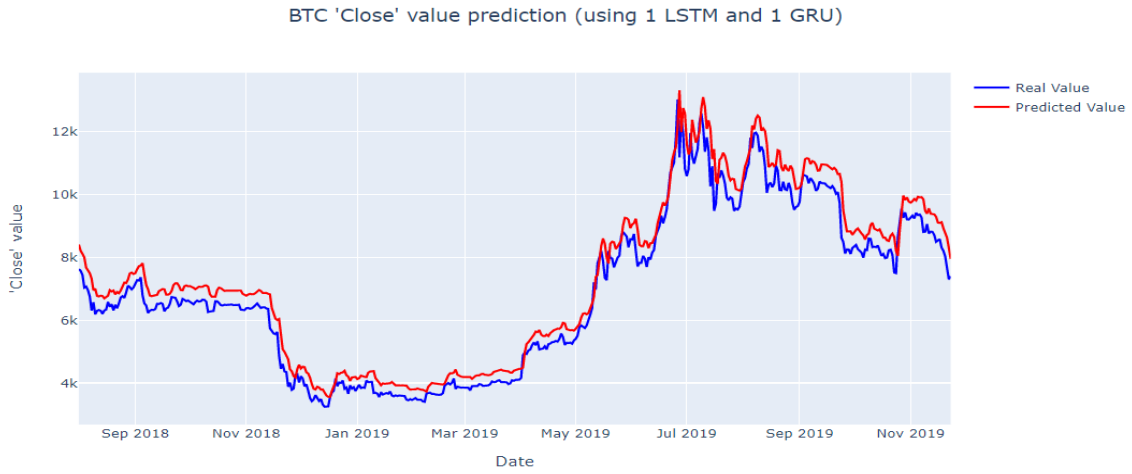
Σχήμα 5.24: Νευρωνικό δίκτυο 5: 2 στρώματα GRU



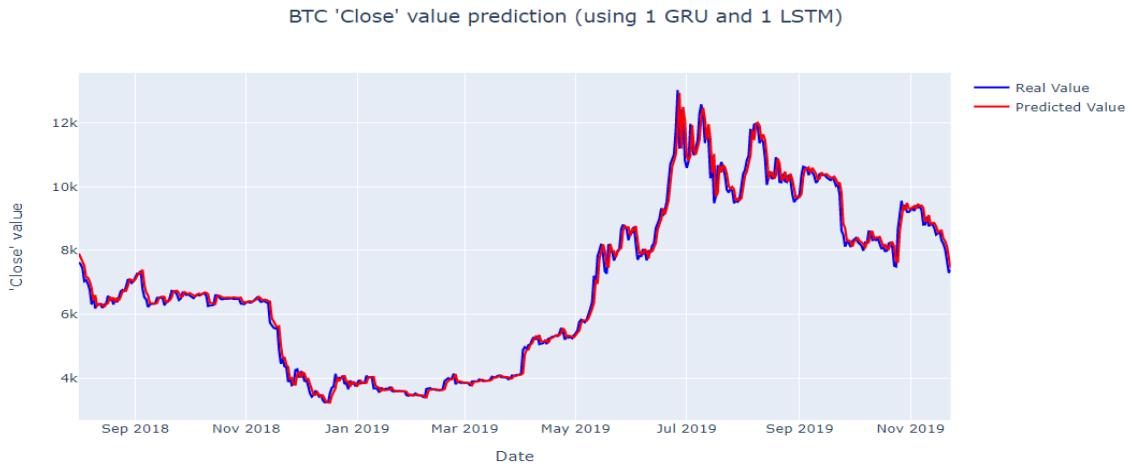
Σχήμα 5.25: Νευρωνικό δίκτυο 6: 3 στρώματα GRU



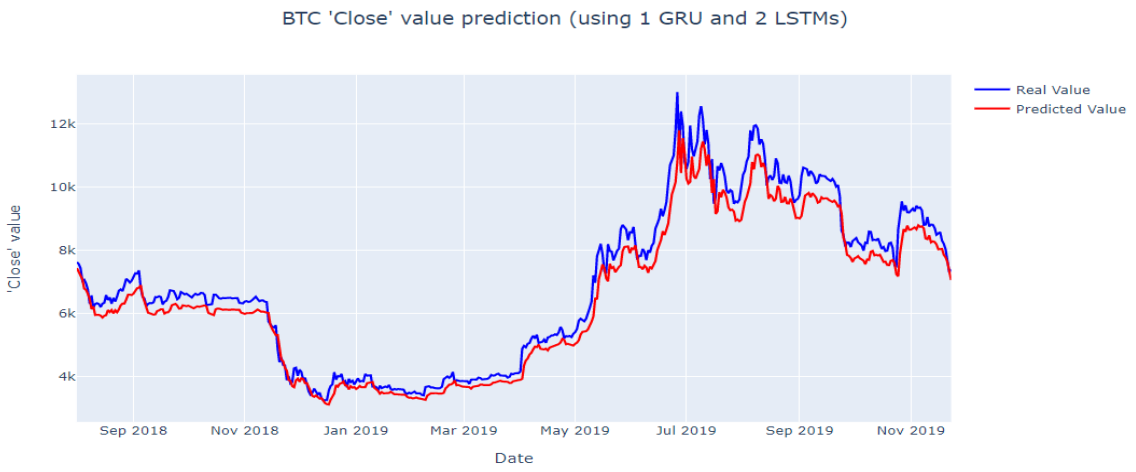
Σχήμα 5.26: Νευρωνικό δίκτυο 7: 4 στρώματα GRU



Σχήμα 5.27: Νευρωνικό δίκτυο 8: 1 στρώμα LSTM και 1 στρώμα GRU



Σχήμα 5.28: Νευρωνικό δίκτυο 9: 1 στρώμα GRU και 1 στρώμα LSTM



Σχήμα 5.29: Νευρωνικό δίκτυο 10: 1 στρώμα GRU και 2 στρώματα LSTM

Συνοπτικά οι επιδόσεις των δικτύων φαίνονται στον πίνακα 5.5. Το πιο ακριβές δίκτυο αναφορικά με την πρόβλεψη είναι το νευρωνικό δίκτυο 9 το οποίο αποτελείται από 1 στρώμα GRU και ένα στρώμα LSTM και προβλέπει τις μελλοντικές τιμές του κρυπτονομίσματος με σφάλμα RMSE=312.09.

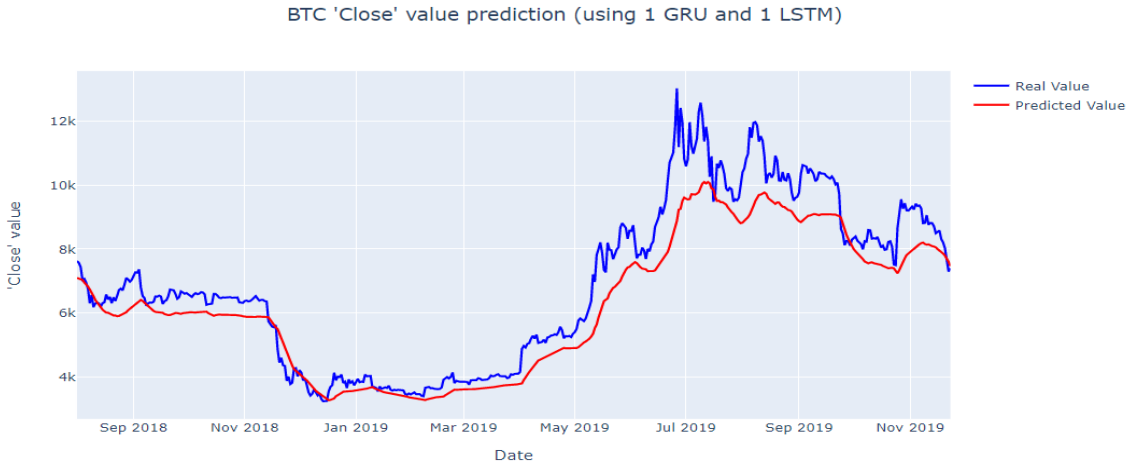
Πίνακας 5.5: Επιδόσεις των νευρωνικών δικτύων που εξετάστηκαν

Στρώματα	**Νευρώνες ανα στρώμα**	**RMSE**	**Relative RMSE**	**MAPE**
1 LSTM	50	318.84	12.52%	48.3%
2 LSTMs	50	319.84	12.76%	48.05%
3 LSTMs	50	322.8	12.8%	48.6%
1 GRU	50	331.99	13.16%	48.87%
2 GRUs	50	325.02	12.88%	48.98%
3 GRUs	50	313.92	12.35%	48.84%
4 GRUs	50	353.36	14.01%	49.24%
1 LSTM - 1 GRU	50	560.87	22.23%	51.19%
1 GRU - 1 LSTM	50	312.09	12.37%	47.27%
1 GRU - 2 LSTMs	50	543.38	21.54%	45.16%

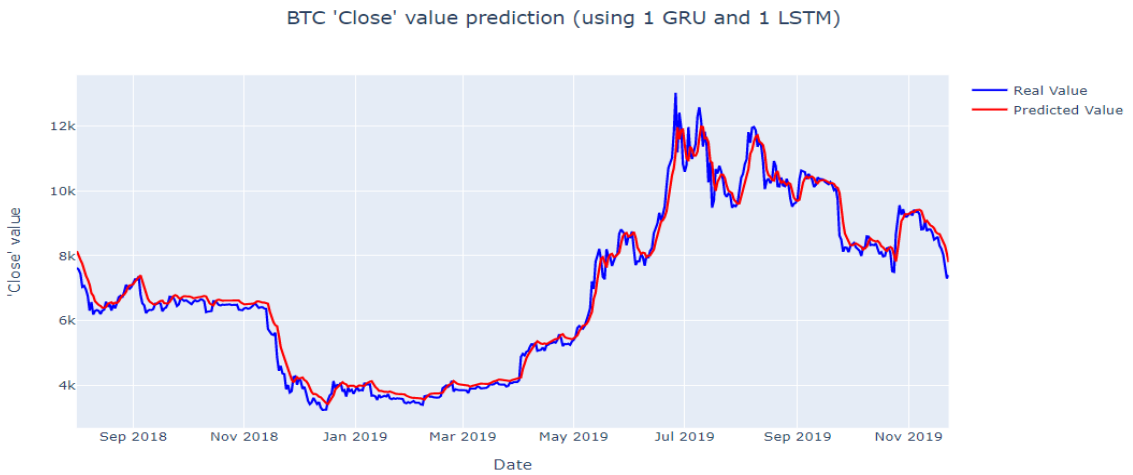
5.5.2 Επιλογή αριθμού νευρώνων στρωμάτων νευρωνικού δικτύου

Με αφετηρία το νευρωνικό δίκτυο 9, επιλέχθηκε επίσης βάσει ελαχίστου RMSE ο βέλτιστος αριθμός νευρώνων ανα στρώμα του νευρωνικού δικτύου. Ξεκινώντας με την δοκιμή

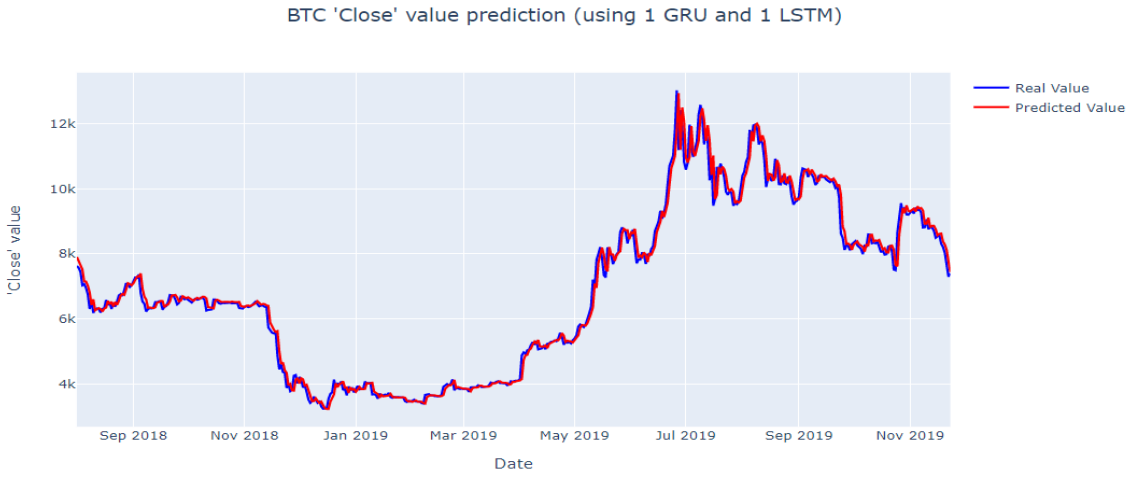
1 (1 νευρώνας ανα στρώμα) (σχήμα 5.30) παρατηρείται RMSE=927.1 οπότε δοκιμάζονται περισσότεροι νευρώνες με σκοπό την βελτίωση του σφάλματος. Ακολουθεί η δοκιμή 2 (10 νευρώνες ανα στρώμα) (σχήμα 5.31) και η δοκιμή 3 (50 νευρώνες ανα στρώμα) (σχήμα 5.32) όπου το σφάλμα μειώνεται σε RMSE=312.09. Στις δοκιμές 4 (100 νευρώνες ανα στρώμα) (σχήμα 5.33) και 5 (200 νευρώνες ανα στρώμα) (σχήμα 5.34) το σφάλμα αυξάνεται σημαντικά όμως στις επόμενες δοκιμές 6 (300 νευρώνες ανα στρώμα) (σχήμα 5.35) και 7 (400 νευρώνες ανα στρώμα) (σχήμα 5.36) το σφάλμα πλησιάζει αυτό της δοκιμής 3 η οποία μέχρι στιγμής παραμένει η καλύτερη. Τέλος η δοκιμή 8 (1000 νευρώνες ανα στρώμα) (σχήμα 5.37) δείχνει οτι το σφάλμα για 1000 νευρώνες ανα στρώμα αυξάνεται πάρα πολύ όπως και ο χρόνος εκτέλεσης της αναδρομικής διαδικασίας.



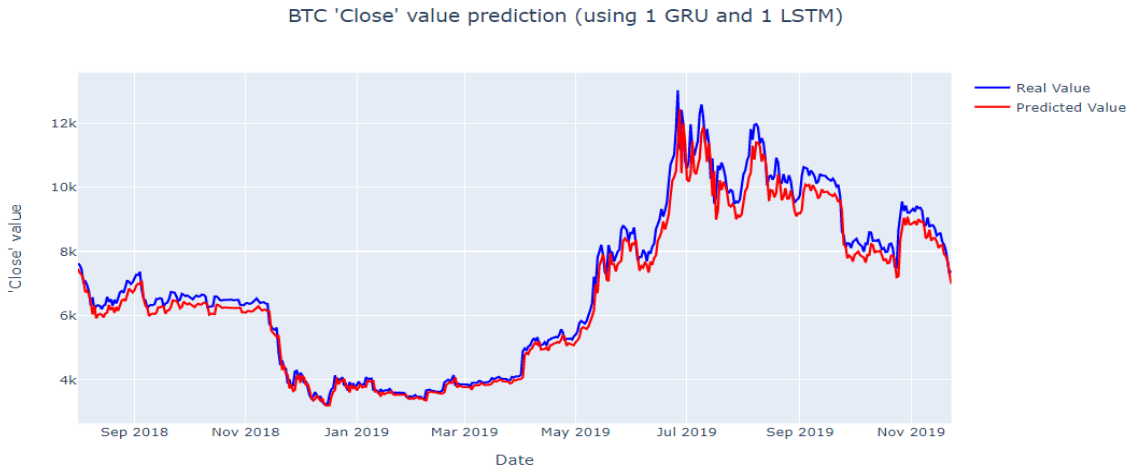
Σχήμα 5.30: Δοκιμή 1: 1 νευρώνας ανα στρώμα



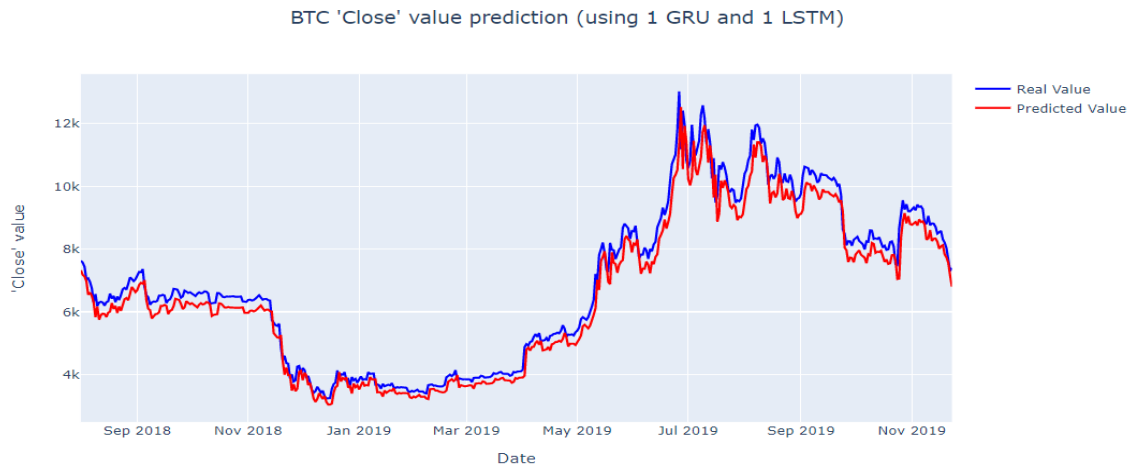
Σχήμα 5.31: Δοκιμή 2: 10 νευρώνες ανα στρώμα



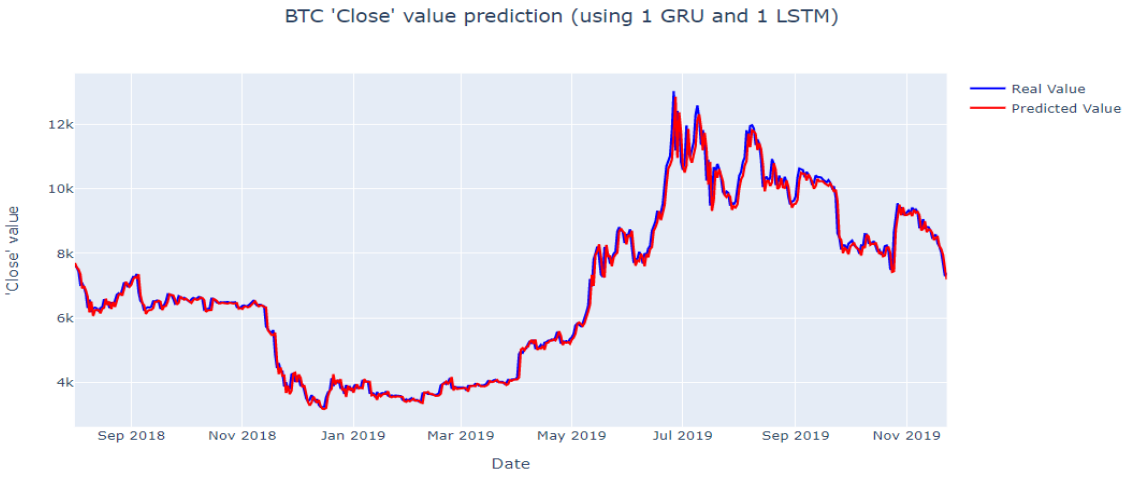
Σχήμα 5.32: Δοκιμή 3: 50 νευρώνες ανα στρώμα



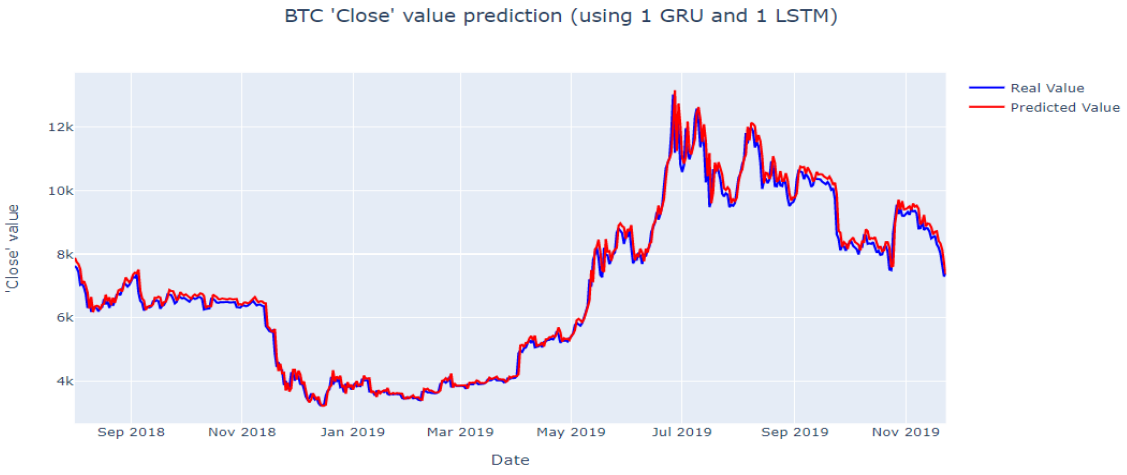
Σχήμα 5.33: Δοκιμή 4: 100 νευρώνες ανα στρώμα



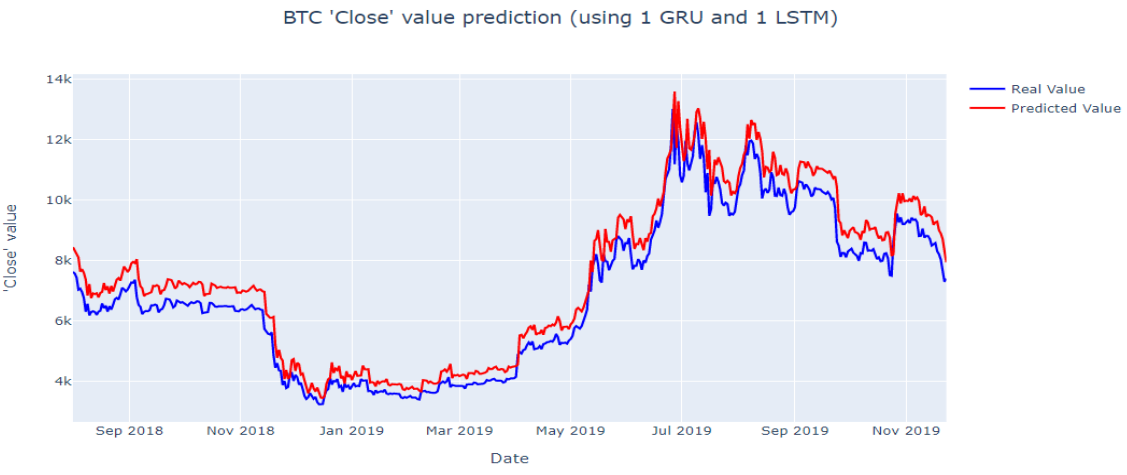
Σχήμα 5.34: Δοκιμή 5: 200 νευρώνες ανα στρώμα



Σχήμα 5.35: Δοκιμή 6: 300 νευρώνες ανα στρώμα



Σχήμα 5.36: Δοκιμή 7: 400 νευρώνες ανα στρώμα



Σχήμα 5.37: Δοκιμή 8: 1000 νευρώνες ανα στρώμα

Συνοπτικά οι επιδόσεις των δικτύων φαίνονται στον πίνακα 5.6. Ο βέλτιστος αριθμός νευρώνων για το νευρωνικό δίκτυο 9, το οποίο αποτελείται από 1 στρώμα GRU και ένα στρώμα LSTM, είναι 50 και το σφάλμα της πρόβλεψης του είναι RMSE=312.09.

Πίνακας 5.6: Επιδόσεις του νευρωνικού δικτύου 9 για διαφορετικό αριθμό νευρώνων ανα στρώμα

Στρώματα	**Νευρώνες ανα στρώμα**	**RMSE**	**Relative RMSE**	**MAPE**
1 GRU - 1 LSTM	1	927.1	36.75%	42.9%
1 GRU - 1 LSTM	10	361.61	14.34%	47.9%
1 GRU - 1 LSTM	50	312.09	12.37%	47.27%
1 GRU - 1 LSTM	100	419.86	16.64%	45.83%
1 GRU - 1 LSTM	200	471.58	18.69%	45.97%
1 GRU - 1 LSTM	300	313.82	12.73%	48.74%
1 GRU - 1 LSTM	400	312.85	12.4%	49%
1 GRU - 1 LSTM	1000	655.85	26%	52.27%

5.6 Σύγκριση των μεθόδων

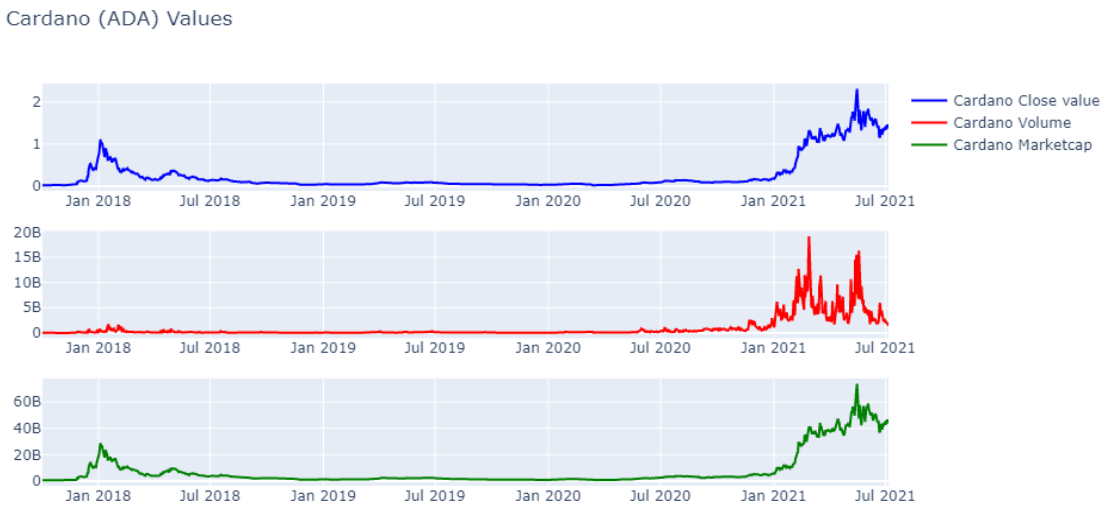
Σύμφωνα με το σφάλμα που υπολογίστηκε από κάθε μέθοδο συμπεραίνεται οτι τα νευρωνικά δίκτυα ανταποκρίνονται καλύτερα στις απαιτήσεις του προβλήματος πρόβλεψης της τιμής του Bitcoin. Αρχικά η μέθοδος Facebook Prophet έδωσε σφάλμα RMSE=20970.32, ακολούθησε η μέθοδος auto-ARIMA με σφάλμα RMSE=21812.77 και το SARIMAX που έδωσε σφάλμα RMSE=83663.37 προβλέποντας όμως καλύτερα τις αυξομειώσεις της τιμής. Η μέθοδος των νευρωνικών δικτύων πέτυχε τα καλύτερα αποτελέσματα με σφάλμα RMSE=312.09 για το νευρωνικό δίκτυο των 2 στρωμάτων (GRU-LSTM) και 50 νευρώνων ανα στρώμα. Τα αποτελέσματα παρουσιάζονται συνοπτικά στον πίνακα 5.7.

Πίνακας 5.7: Αποτελέσματα των μεθόδων πρόβλεψης

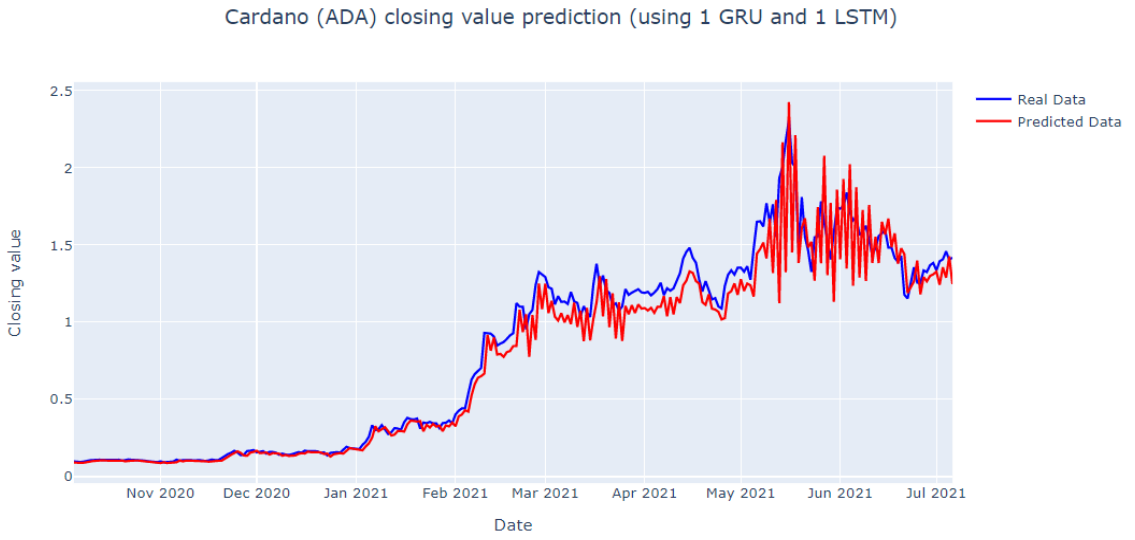
Μέθοδος	**RMSE**
Facebook Prophet	20970.32
auto-ARIMA	21812.77
SARIMAX	83663.37
Νευρωνικό δίκτυο 9	312.09

5.7 Εκτέλεση πρόβλεψης για άλλα κρυπτονομίσματα

Χρησιμοποιώντας το νευρωνικό δίκτυο της προηγούμενης ενότητας γίνεται προσπάθεια πρόβλεψης και για άλλα δημοφιλή κρυπτονομίσματα. Όπως δείχνουν οι μετρήσεις, υπάρχει μεγάλη απόκλιση μεταξύ των σφαλμάτων των κρυπτονομισμάτων. Κρυπτονομίσματα όπως το Cardano (σχήματα 5.38 και 5.39) και το Dogecoin (σχήματα 5.44 και 5.45) σημείωσαν μεγάλη ξαφνική αύξηση η οποία δεν προκύπτει από τις παρελθοντικές τιμές του νομίσματος αλλά από άλλους εξωγενείς παράγοντες οπότε το σφάλμα της πρόβλεψης είναι μεγάλο. Αντίθετα τα Solana (σχήματα 5.50 και 5.51), Polkadot (σχήματα 5.48 και 5.49), Cosmos (σχήματα 5.40 και 5.41) και Crypto.com Coin (σχήματα 5.42 και 5.43) σημείωσαν σταδιακές αυξήσεις και ακολούθησαν συγκεκριμένα μοτίβα εξέλιξης τα οποία το νευρωνικό δίκτυο κατάφερε να αναγνωρίσει και τελικά να προβλέψει με μικρότερο σφάλμα. Αναλύθηκαν επίσης κρυπτονομίσματα όπως το Tether (σχήματα 5.52 και 5.53) και το USD Coin (σχήματα 5.54 και 5.39) των οποίων η τιμή βρίσκεται σταθερα στο 1 δολάριο ΗΠΑ με πολύ μικρές αυξομειώσεις και το XRP (σχήματα 5.56 και 5.57) με μέγιστη τιμή τα 3 δολάρια ΗΠΑ, τα οποία παρουσιάζουν σχεδόν μηδενικό απόλυτο σφάλμα και σχετικά ελάχιστο MAPE. Συνοπτικά τα σφάλματα των προβλέψεων για κάθε κρυπτονόμισμα παρουσιάζονται στον πίνακα 5.8.

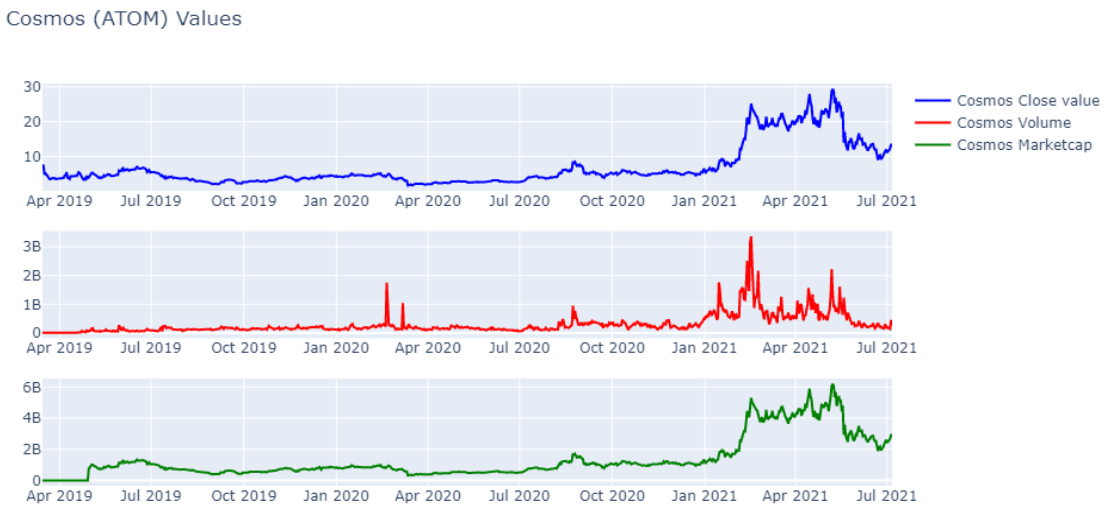


Σχήμα 5.38: Εξέλιξη οικονομικών δεικτών του Cardano (ADA)

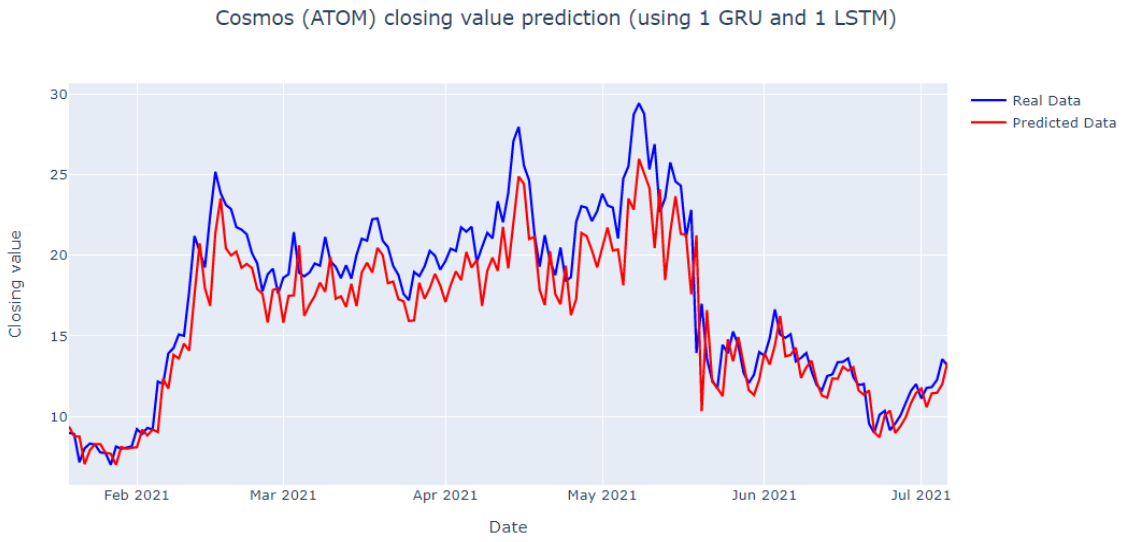



```
Prediction Variables
-----------------------------------
Amount of training data: 1099
Amount of layers: 2
Timesteps: 21
Neurons per layer: 50
Training/Test: 80% - 20%
RMSE: 0.15
Relative/Normalized RMSE: 24.95%
MAPE: 221.62%
Optimizer: adam
Activation Function: tanh
Epochs: 100
Based on: ['Close']
-----------------------------------
```

Σχήμα 5.39: Αποτέλεσμα πρόβλεψης για το κρυπτονόμισμα Cardano (ADA)



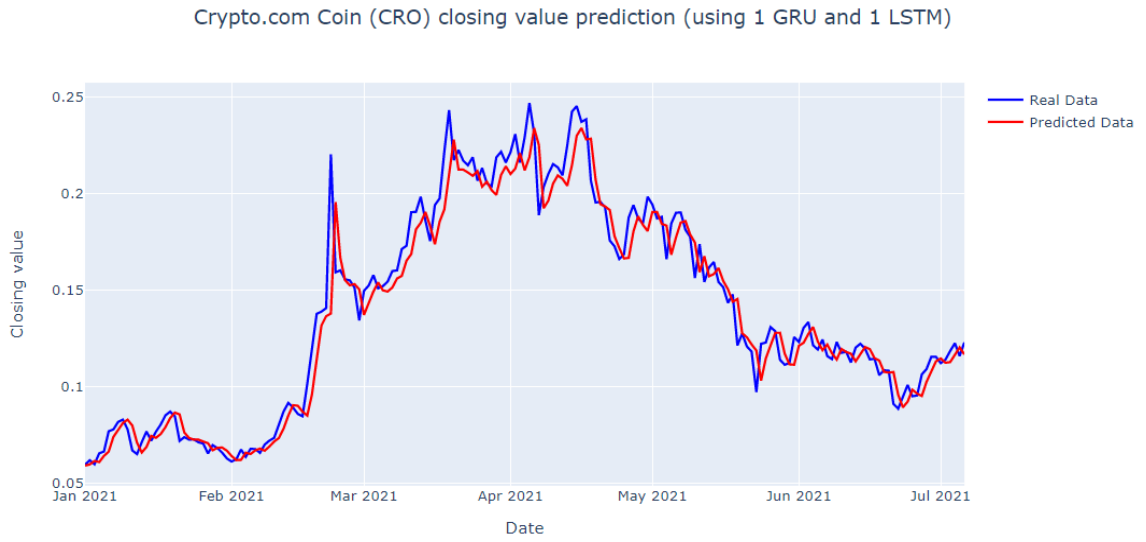
Σχήμα 5.40: Εξέλιξη οικονομικών δεικτών του Cosmos (ATOM)




```
Prediction Variables
-----------------------------------
Amount of training data: 676
Amount of layers: 2
Timesteps: 21
Neurons per layer: 50
Training/Test: 80% - 20%
RMSE: 2.33
Relative/Normalized RMSE: 42.57%
MAPE: 38.52%
Optimizer: adam
Activation Function: tanh
Epochs: 100
Based on: ['Close']
-----------------------------------
```

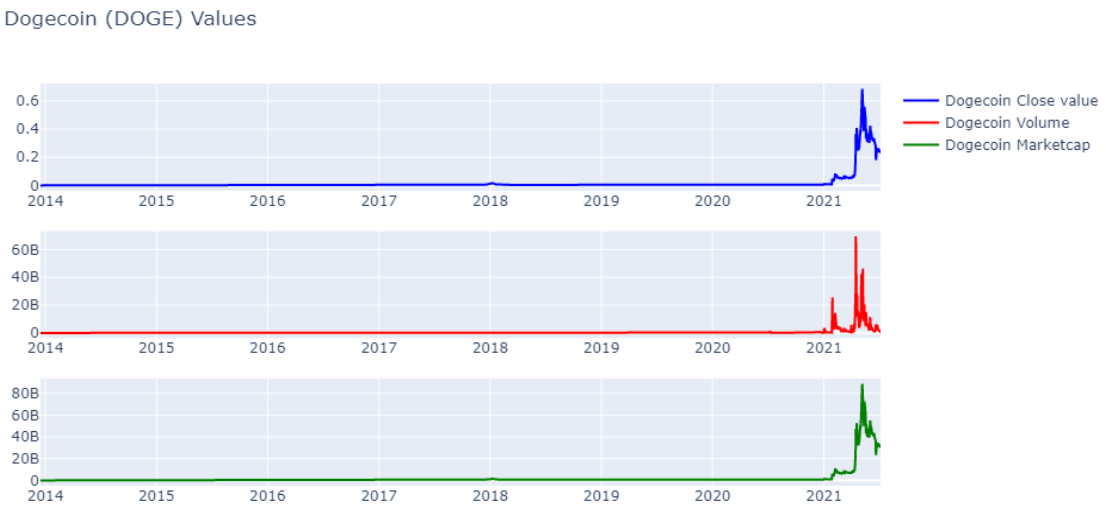
Σχήμα 5.41: Αποτέλεσμα πρόβλεψης για το κρυπτονόμισμα Cosmos (ATOM)

Σχήμα 5.42: Εξέλιξη οικονομικών δεικτών του Crupto.com Coin (CRO)

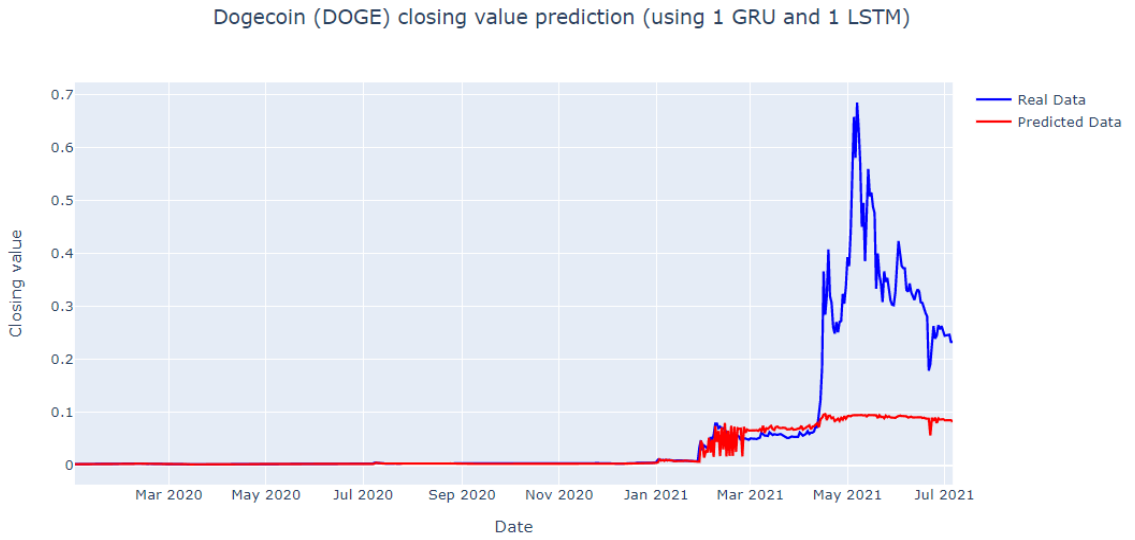



```
Prediction Variables
-----------------------------------
Amount of training data: 748
Amount of layers: 2
Timesteps: 21
Neurons per layer: 50
Training/Test: 80% - 20%
RMSE: 0.01
Relative/Normalized RMSE: 22.69%
MAPE: 50.86%
Optimizer: adam
Activation Function: tanh
Epochs: 100
Based on: ['Close']
-----------------------------------
```

Σχήμα 5.43: Αποτέλεσμα πρόβλεψης για το κρυπτονόμισμα Crupto.com Coin (CRO)

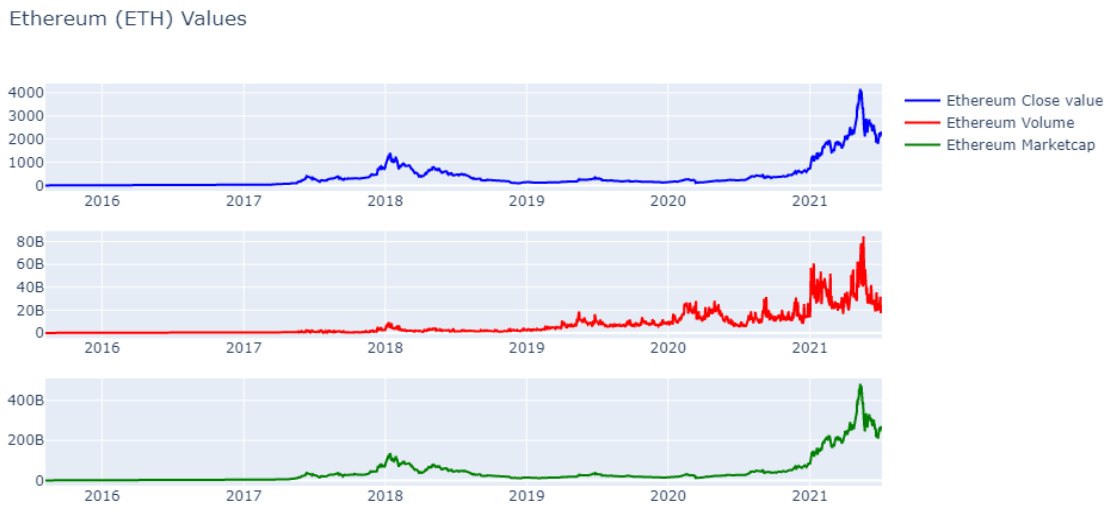


Σχήμα 5.44: Εξέλιξη οικονομικών δεικτών του Dogecoin (DOGE)

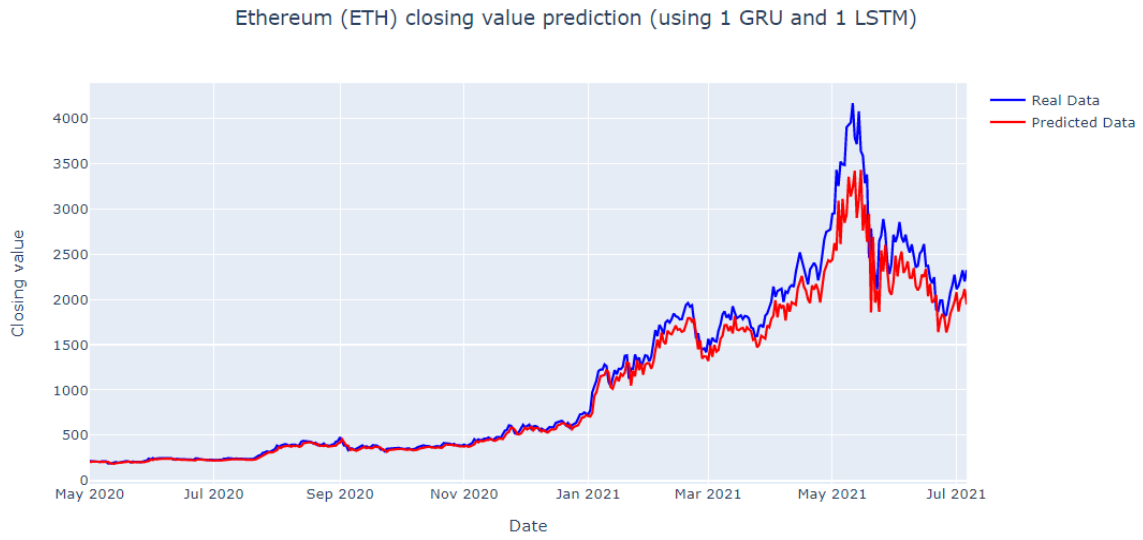



```
Prediction Variables
-----------------------------------
Amount of training data: 2208
Amount of layers: 2
Timesteps: 21
Neurons per layer: 50
Training/Test: 80% - 20%
RMSE: 0.11
Relative/Normalized RMSE: 84.11%
MAPE: 588.69%
Optimizer: adam
Activation Function: tanh
Epochs: 100
Based on: ['Close']
-----------------------------------
```

Σχήμα 5.45: Αποτέλεσμα πρόβλεψης για το κρυπτονόμισμα Dogecoin (DOGE)

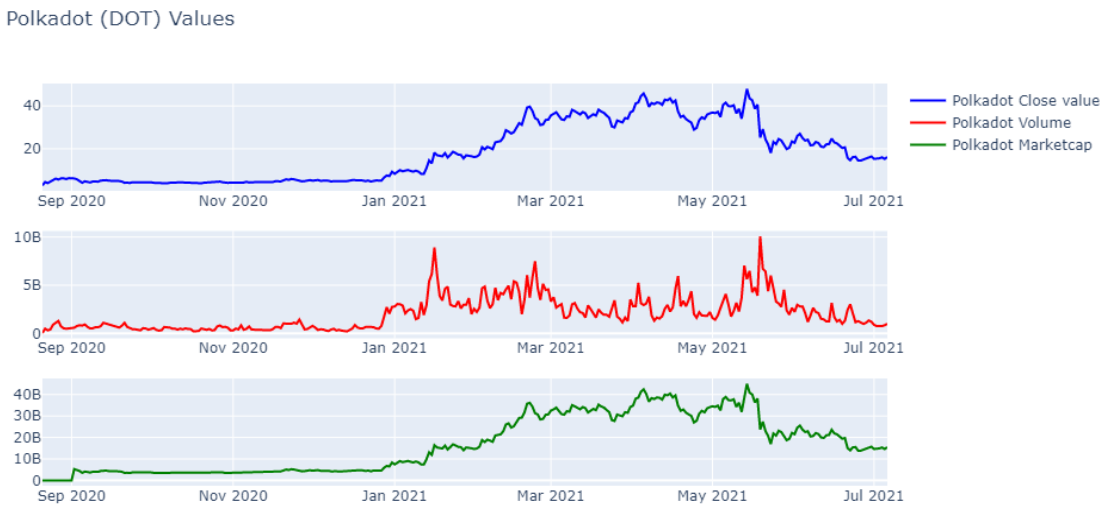


Σχήμα 5.46: Εξέλιξη οικονομικών δεικτών του Ethereum (ETH)

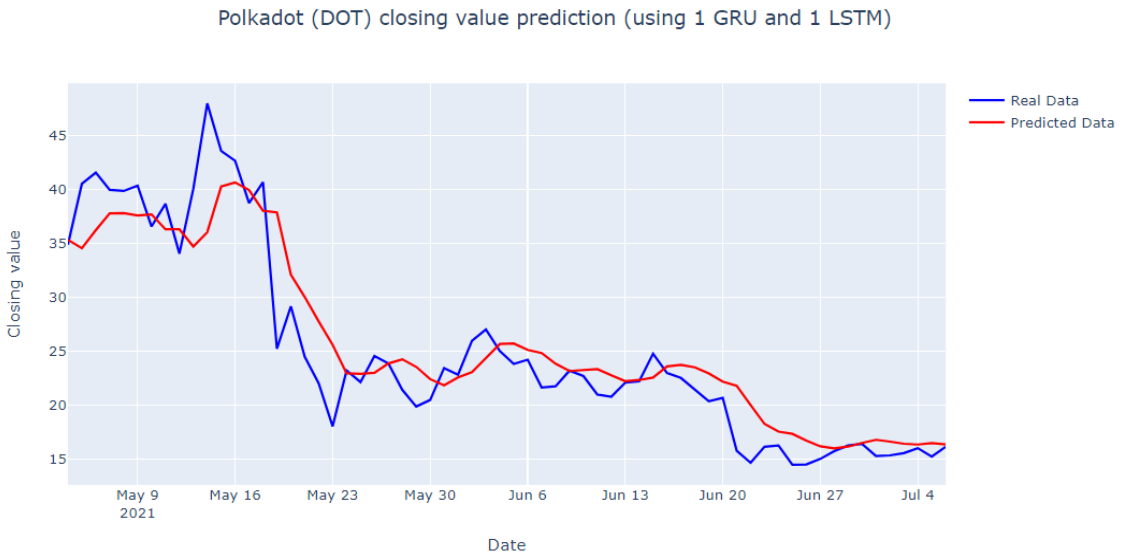



```
Prediction Variables
-----------------------------------
Amount of training data: 1728
Amount of layers: 2
Timesteps: 21
Neurons per layer: 50
Training/Test: 80% - 20%
RMSE: 202.27
Relative/Normalized RMSE: 20.97%
MAPE: 155.84%
Optimizer: adam
Activation Function: tanh
Epochs: 100
Based on: ['Close']
-----------------------------------
```

Σχήμα 5.47: Αποτέλεσμα πρόβλεψης για το κρυπτονόμισμα Ethereum (ETH)



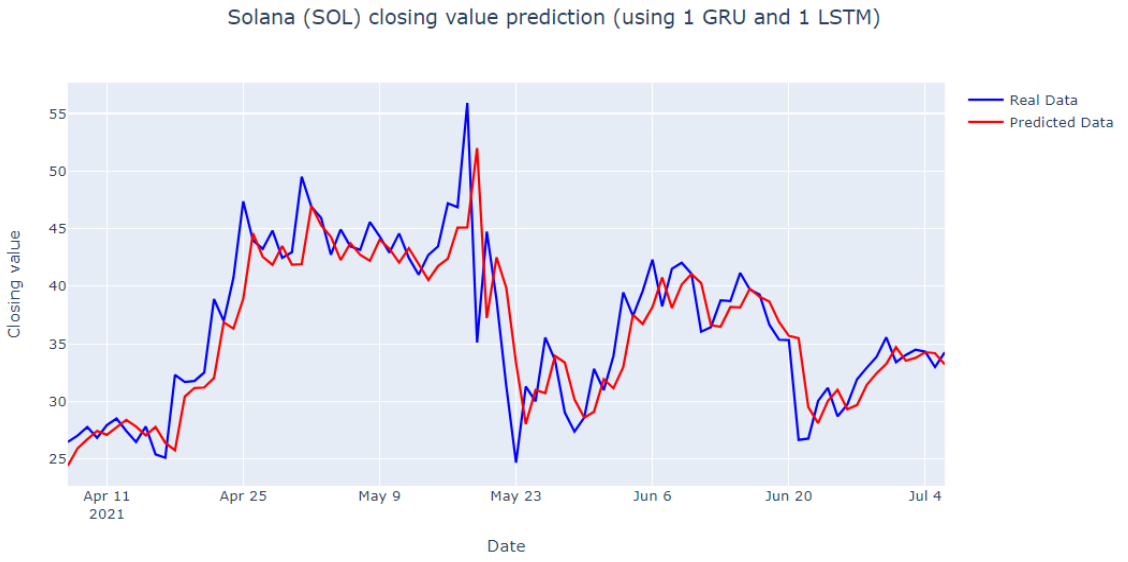
Σχήμα 5.48: Εξέλιξη οικονομικών δεικτών του Polkadot (DOT)




```
Prediction Variables
-----------------------------------
Amount of training data: 256
Amount of layers: 2
Timesteps: 21
Neurons per layer: 50
Training/Test: 80% - 20%
RMSE: 3.45
Relative/Normalized RMSE: 37.74%
MAPE: 39.11%
Optimizer: adam
Activation Function: tanh
Epochs: 100
Based on: ['Close']
-----------------------------------
```

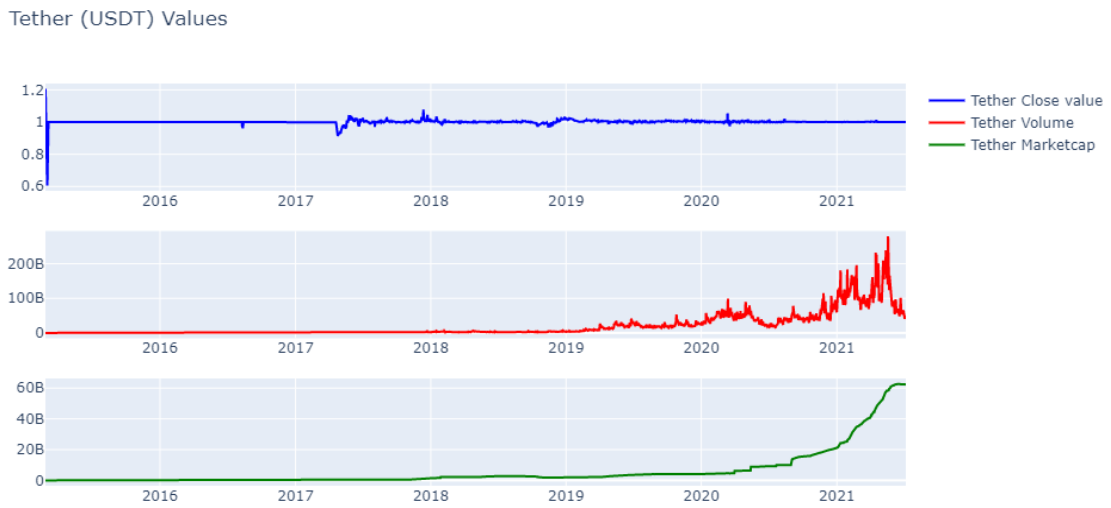
Σχήμα 5.49: Αποτέλεσμα πρόβλεψης για το κρυπτονόμισμα Polkadot (DOT)

Σχήμα 5.50: Εξέλιξη οικονομικών δεικτών του Solana (SOL)

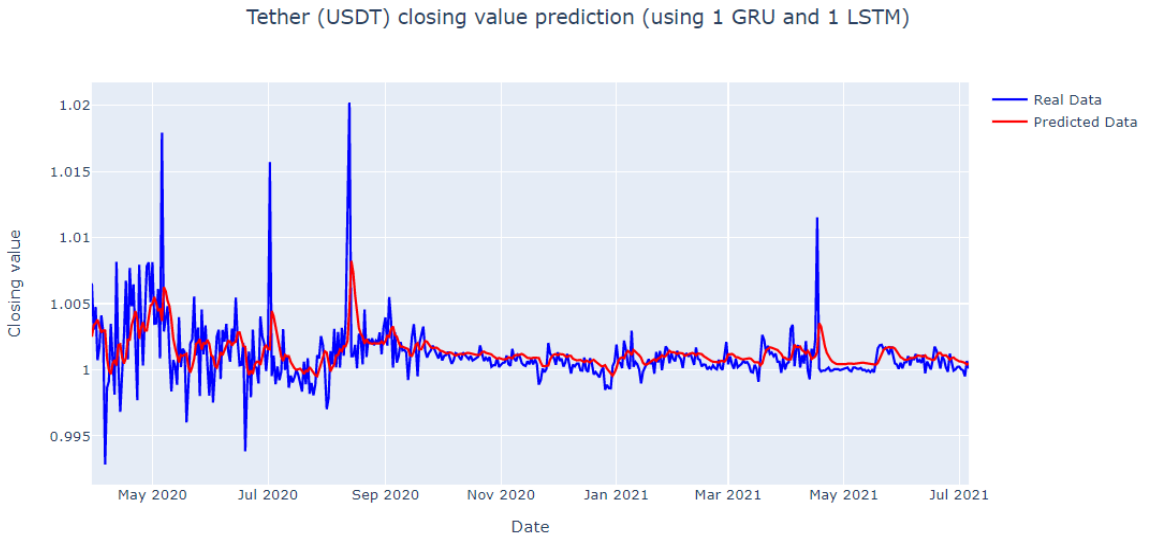



```
Prediction Variables
-----------------------------------
Amount of training data: 361
Amount of layers: 2
Timesteps: 21
Neurons per layer: 50
Training/Test: 80% - 20%
RMSE: 3.76
Relative/Normalized RMSE: 54.69%
MAPE: 21.08%
Optimizer: adam
Activation Function: tanh
Epochs: 100
Based on: ['Close']
-----------------------------------
```

Σχήμα 5.51: Αποτέλεσμα πρόβλεψης για το κρυπτονόμισμα Solana (SOL)

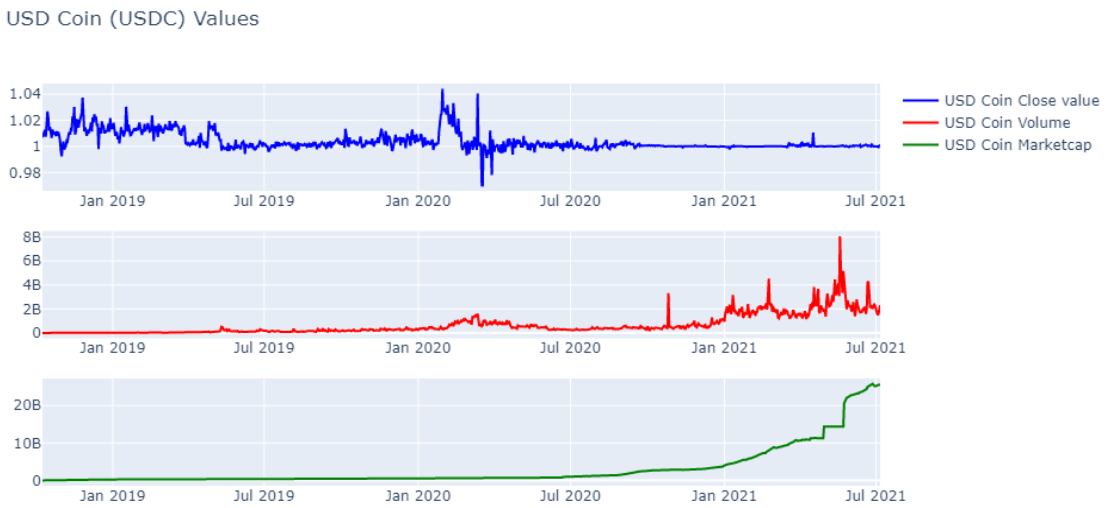


Σχήμα 5.52: Εξέλιξη οικονομικών δεικτών του Tether (USDT)

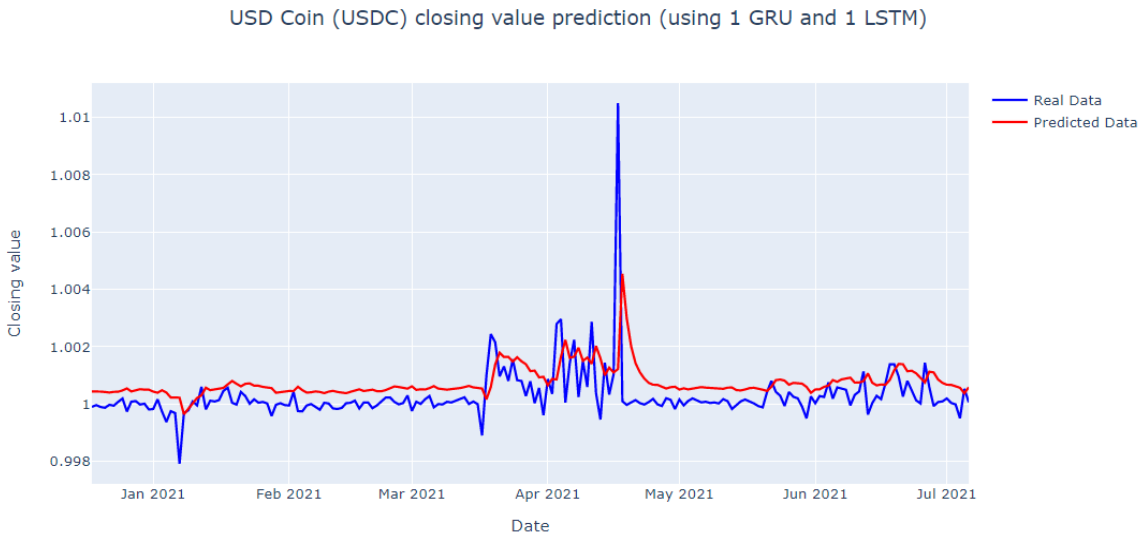



```
Prediction Variables
-----------------------------------
Amount of training data: 1854
Amount of layers: 2
Timesteps: 21
Neurons per layer: 50
Training/Test: 80% - 20%
RMSE: 0.0
Relative/Normalized RMSE: 95.49%
MAPE: 0.16%
Optimizer: adam
Activation Function: tanh
Epochs: 100
Based on: ['Close']
-----------------------------------
```

Σχήμα 5.53: Αποτέλεσμα πρόβλεψης για το κρυπτονόμισμα Tether (USDT)

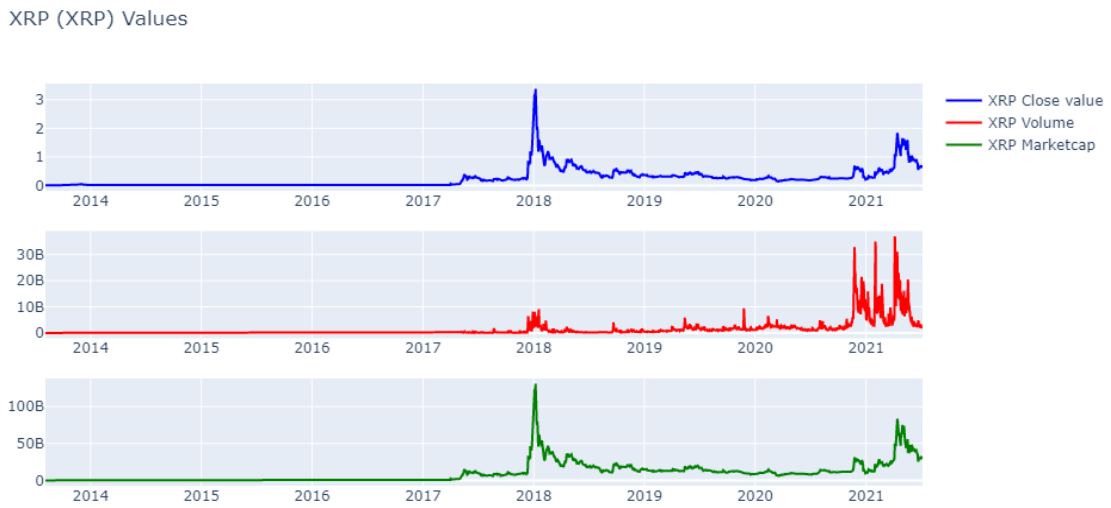


Σχήμα 5.54: Εξέλιξη οικονομικών δεικτών του USD Coin (USDC)

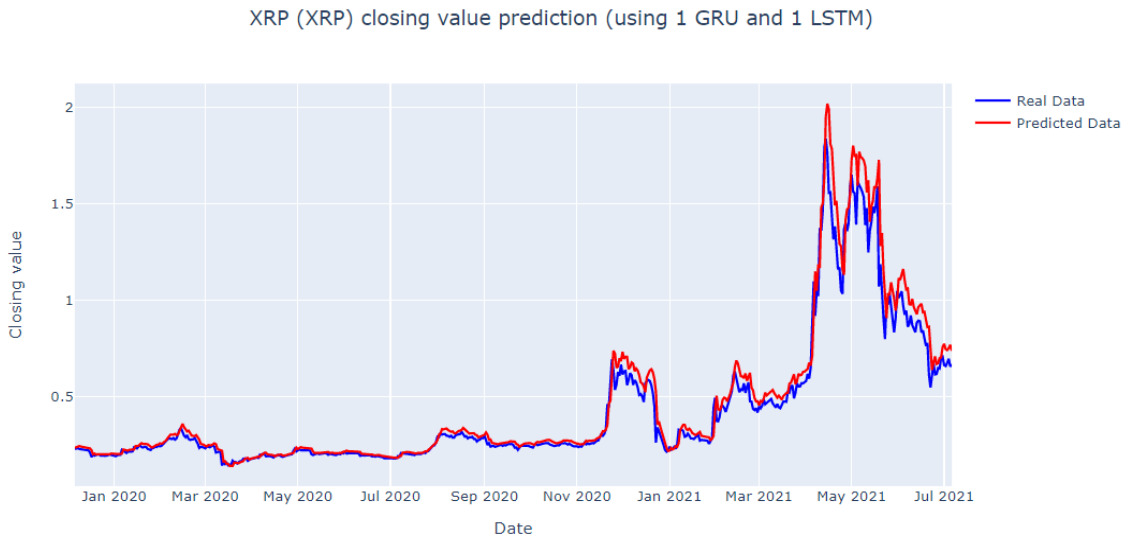



```
Prediction Variables
-----------------------------------
Amount of training data: 801
Amount of layers: 2
Timesteps: 21
Neurons per layer: 50
Training/Test: 80% - 20%
RMSE: 0.0
Relative/Normalized RMSE: 110.72%
MAPE: 0.08%
Optimizer: adam
Activation Function: tanh
Epochs: 100
Based on: ['Close']
-----------------------------------
```

Σχήμα 5.55: Αποτέλεσμα πρόβλεψης για το κρυπτονόμισμα USD Coin (USDC)



Σχήμα 5.56: Εξέλιξη οικονομικών δεικτών του XRP (XRP)




```
Prediction Variables
-----------------------------------
Amount of training data: 2314
Amount of layers: 2
Timesteps: 21
Neurons per layer: 50
Training/Test: 80% - 20%
RMSE: 0.08
Relative/Normalized RMSE: 22.72%
MAPE: 96.58%
Optimizer: adam
Activation Function: tanh
Epochs: 100
Based on: ['Close']
-----------------------------------
```

Σχήμα 5.57: Αποτέλεσμα πρόβλεψης για το κρυπτονόμισμα XRP (XRP)

Πίνακας 5.8: Σύνοψη των σφαλμάτων των προβλέψεων για τα κρυπτονομίσματα που αναλύθηκαν

Cryptocurrency	**Αριθμός δεδομένων (εκπαίδευση - επαλήθευση)**	**RMSE**	**Relative RMSE**	**MAPE**
Cardano (ADA)	1374 (1099 - 275)	0.15	24.95%	221.62%
Cosmos (ATOM)	845 (676 - 169)	2.33	42.57%	38.52%
Bitcoin (BTC)	2991 (2392 - 598)	312.09	12.37%	47.27%
Crypto.com (CRO)	935 (748 - 187)	0.01	22.69%	50.86%
Dogecoin (DOGE)	2760 (2208 - 552)	0.11	84.11%	588.69%
Polkadot (DOT)	320 (256 - 64)	3.45	37.74%	39.11%
Ethereum (ETH)	2160 (1728 - 432)	202.27	20.97%	155.84%
Solana (SOL)	452 (361 - 91)	3.76	54.69%	21.08%
USD Coin (USDC)	1002 (801 - 201)	≈ 0	110.72%	0.08%
Tether (USDT)	2318 (1854 - 464)	≈ 0	95.49%	0.16%
XRP (XRP)	2893 (2314 - 579)	0.08	22.72%	96.58%

Κεφάλαιο 6

Ανάλυση δεδομένων Twitter

6.1 Εισαγωγή

Το κεφάλαιο αυτό αφορά την ανάλυση δεδομένων από το Twitter σχετικά με το κρυπτονόμισμα Bitcoin. Γίνεται εξόρυξη πληροφορίων που αφορούν το κείμενο, τα likes, comments, retweets και τους χρήστες που τα δημοσίευσαν. Στη συνέχεια γίνεται συναισθηματική ανάλυση των tweets και με βάση τα αποτελέσματα δημιουργούνται κάποιοι δείκτες οι οποίοι στο επόμενο κεφάλαιο χρησιμοποιούνται ως πρόσθετοι παράγοντες στην πρόβλεψη της τιμής του κρυπτονομίσματος.

6.2 Περιγραφή αρχικού συνόλου δεδομένων

Τα δεδομένα [41] που αναλύθηκαν αφορούν συνολικά 20165013 tweets που έχουν συλλεχθεί από τις 19-04-2007 έως και τις 23-11-2019 με τη χρήση των Twitter API και Tweepy. Το κριτήριο επιλογής τους ήταν, να περιέχεται στο κείμενο του tweet είτε η λέξη “bitcoin” είτε η συντομογραφία “BTC” χωρίς ευαισθησία πεζών-κεφαλαίων χαρακτήρων. Πέρα από το κείμενο του tweet διατίθενται και πληροφορίες για το username, το id και το πραγματικό όνομα του χρήστη που έκανε το κάθε tweet, τον αριθμό των likes και comments που συγκέντρωσε το κάθε tweet ενώ για κάποια tweets υπάρχει διαθέσιμο και το URL που οδηγεί σε αυτά στην πλατφόρμα του Twitter. Στην εικόνα 6.1 απεικονίζεται ένα δείγμα του αρχικού συνόλου δεδομένων.

	id	user	fullname	url	timestamp	replies	likes	retweets	text
0	1132977055300300800	KamdemAbdiel	Abdiel kamdem	NaN	2019-05-27 11:49:14+00	0.0	0.0	0.0	È appena uscito un nuovo video! LES CRYPTOMONNAIES QUI PULVÉRISENT...
1	1132977073402736640	bitcointe	Bitcointe	NaN	2019-05-27 11:49:18+00	0.0	0.0	0.0	Cardano: Digitize Currencies; EOS https://t.co/1kTKqKEBIS 6500% RO...
2	1132977023893139456	3eyedbran	Bran - 3 Eyed Raven	NaN	2019-05-27 11:49:06+00	0.0	2.0	1.0	Another Test tweet that wasn't caught in the stream ! bitcoin
3	1132977089089556481	DetroitCrypto	J. Scardina	NaN	2019-05-27 11:49:22+00	0.0	0.0	0.0	Current Crypto Prices! \n\nBTC: 8721.99USD\nETH :266.62 USD\nLT...
4	1132977092340191232	mmursaleen72	Muhammad Mursaleen	NaN	2019-05-27 11:49:23+00	0.0	0.0	0.0	Spiv (Nosar Baz): BITCOIN Is An Asset & NOT A Currency.\n\nhtt...

Σχήμα 6.1: Δείγμα αρχικού συνόλου δεδομένων.

6.3 Βασική επεξεργασία δεδομένων

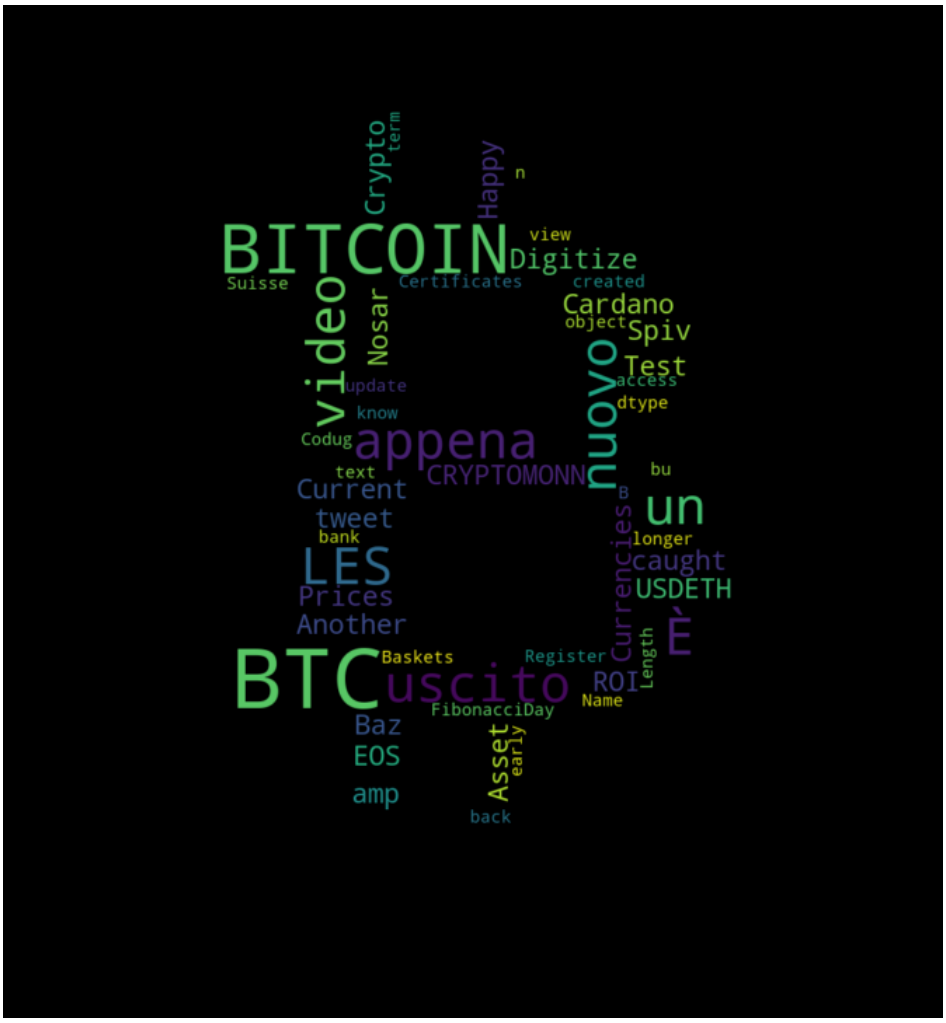
Αρχικά από το σύνολο δεδομένων αφαιρέθηκαν tweets από τα οποία έλειπαν βασικές πληροφορίες που αφορούν τα πεδία username, fullname, timestamp, replies, likes, retweets και text. Έπειτα από το κείμενο του tweet αφαιρέθηκαν σύνδεσμοι και διευθύνσεις τα οποία εντοπίστηκαν ως συμβολοσειρές που ξεκινούν με τα αναγνωριστικά "http" και "www". Ακολούθως διαγράφηκαν τα "mentions" που βρίσκονται μέσα στο κείμενο των tweets και έχουν τον αναγνωριστικό χαρακτήρα "@" ακολουθούμενο από το όνομα του χρήστη στον οποίο αναφέρεται καθώς και ο χαρακτήρας "\n" που δηλώνει αλλαγή γραμμής. Έπειτα από αυτες τις αφαιρέσεις, διαγράφηκαν και τα tweets τα οποία έμειναν κενά.

Θεωρήθηκε, σημαντικό να αφαιρεθούν tweets που δημοσιεύθηκαν από bots. Τα bots έχουν την δυνατότητα να παράξουν μεγάλο αριθμό tweets σε μικρό διάστημα κάτι το οποίο μπορεί να επηρεάσει το συνολικό συναίσθημα προς μια συγκεκριμένη κατεύθυνση και τελικά να μην ανταποκρίνεται στην πραγματικότητα. Σε μια προσπάθεια εντοπισμού τέτοιου είδους δημοσιεύσεων αφαιρέθηκαν όσα tweets περιέχουν μία από τις παρακάτω λέξεις και εκφράσεις: "win", "free", "prize", "100%", "earn" και "risk free". Τέλος αφαιρέθηκαν τα tweets που δεν περιέχουν τις λέξεις-κλειδιά "btc" και "bitcoin". Σε αυτό το στάδιο της επεξεργασίας το σύνολο των δεδομένων αποτελείται από συνολικά 10.942.567 tweets.

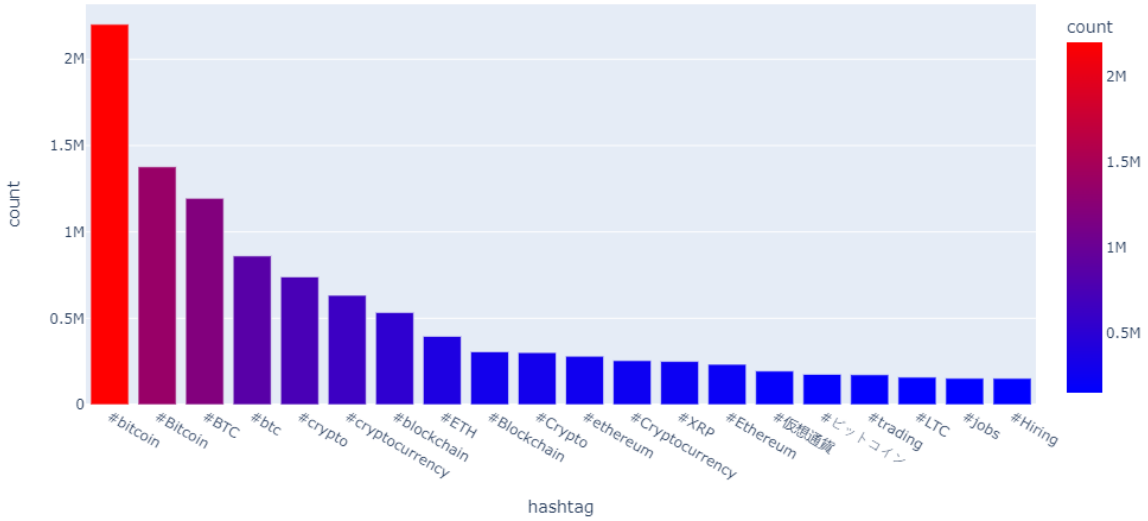
6.4 Περιγραφή των tweets (Παγκοσμίως)

Στην εικόνα 6.2 απεικονίζονται οι πιο συχνά χρησιμοποιούμενες λέξεις στα tweets. Μεγαλύτερο μέγεθος λέξης συνεπάγεται μεγαλύτερη συχνότητα χρήσης της συγκεκριμένης λέξης. Στο σχήμα 6.3 παρουσιάζονται και τα πιο δημοφιλή hashtags. Η κατανομή των tweets στο διάστημα μεταξύ 19-04-2007 και 23-11-2019 απεικονίζεται στα γραφήματα 6.4 (κατανομή ανα έτος), 6.5 (κατανομή ανα μήνα), 6.6 (κατανομή ανα ημέρα του μήνα) και 6.7 (κατανομή

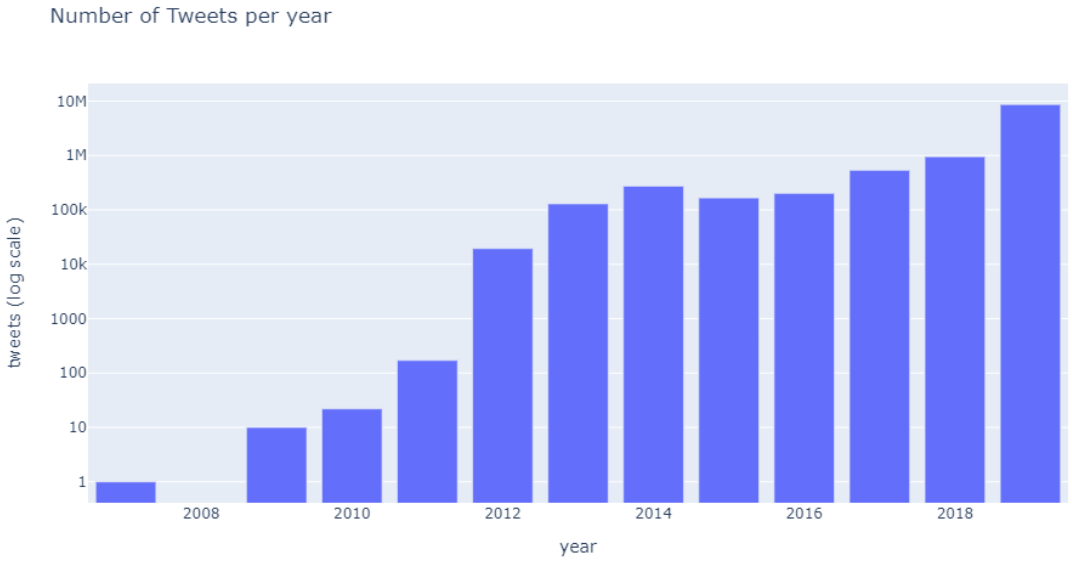
ανα ημέρα).



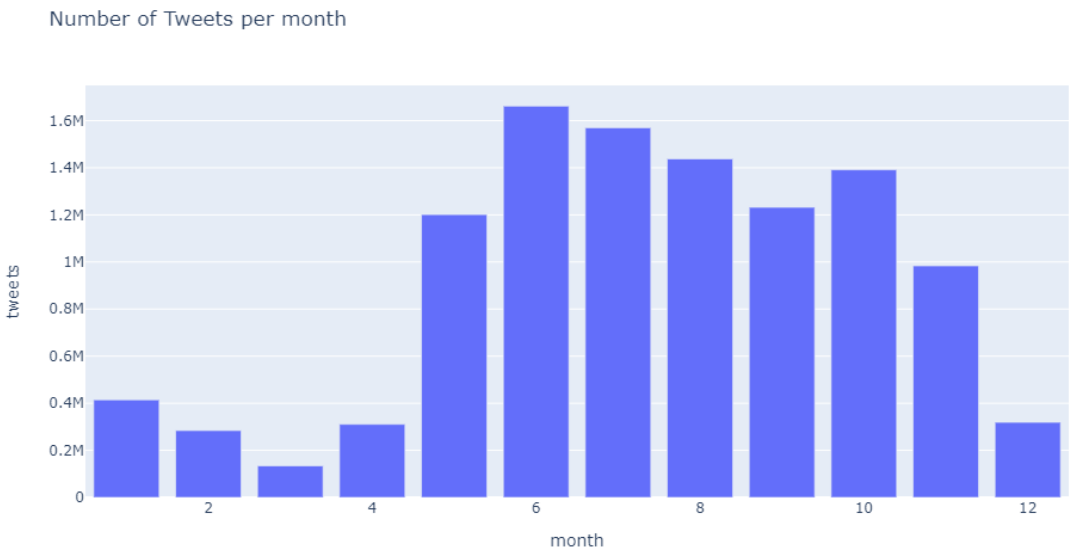
Σχήμα 6.2: Wordcloud: Συχνά χρησιμοποιούμενες λέξεις.



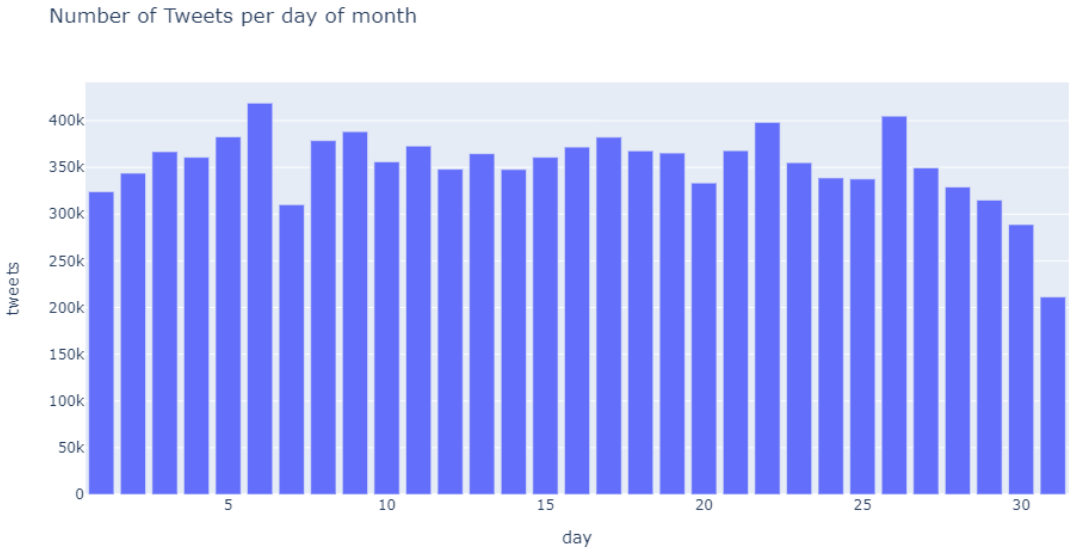
Σχήμα 6.3: Συχνά χρησιμοποιούμενα hashtags.



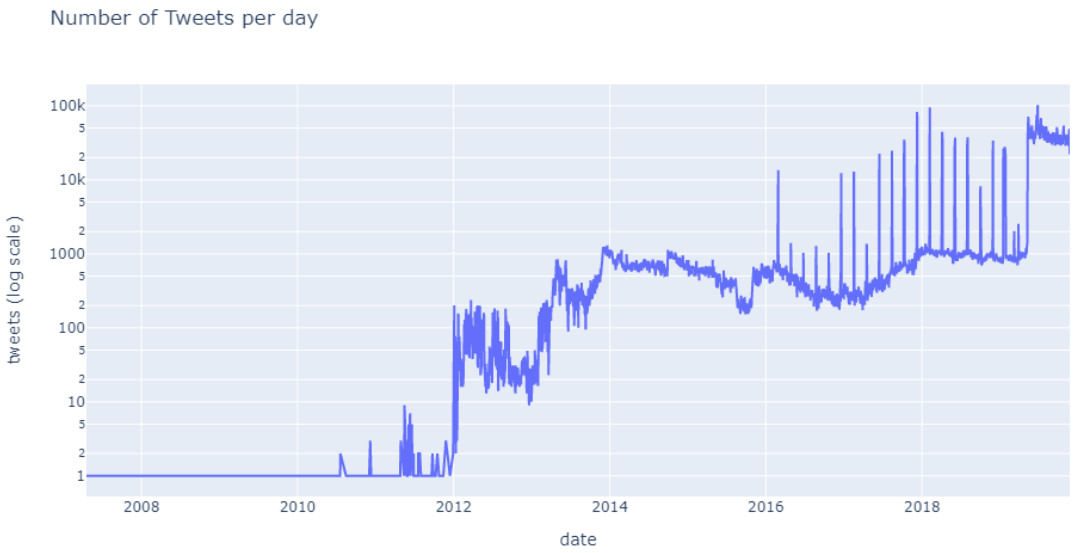
Σχήμα 6.4: Κατανομή tweets ανα έτος.



Σχήμα 6.5: Κατανομή tweets ανα μήνα.



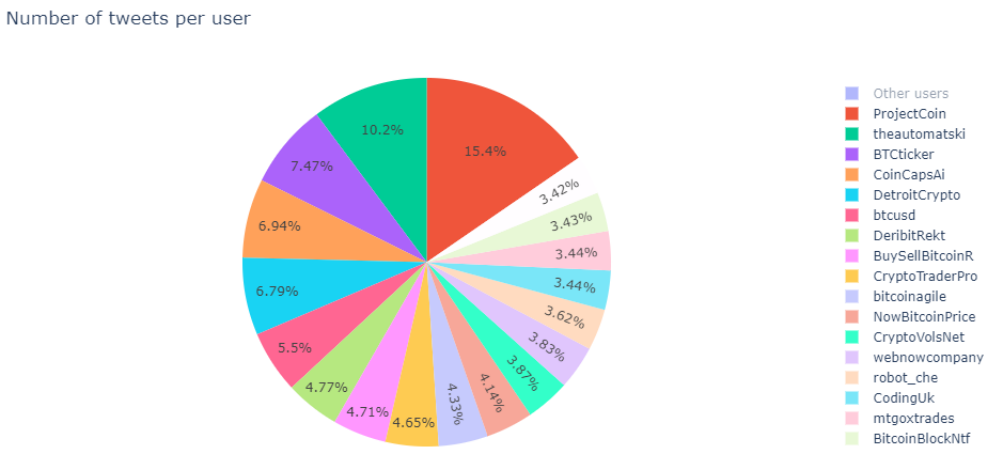
Σχήμα 6.6: Κατανομή tweets ανα ημέρα του μήνα.



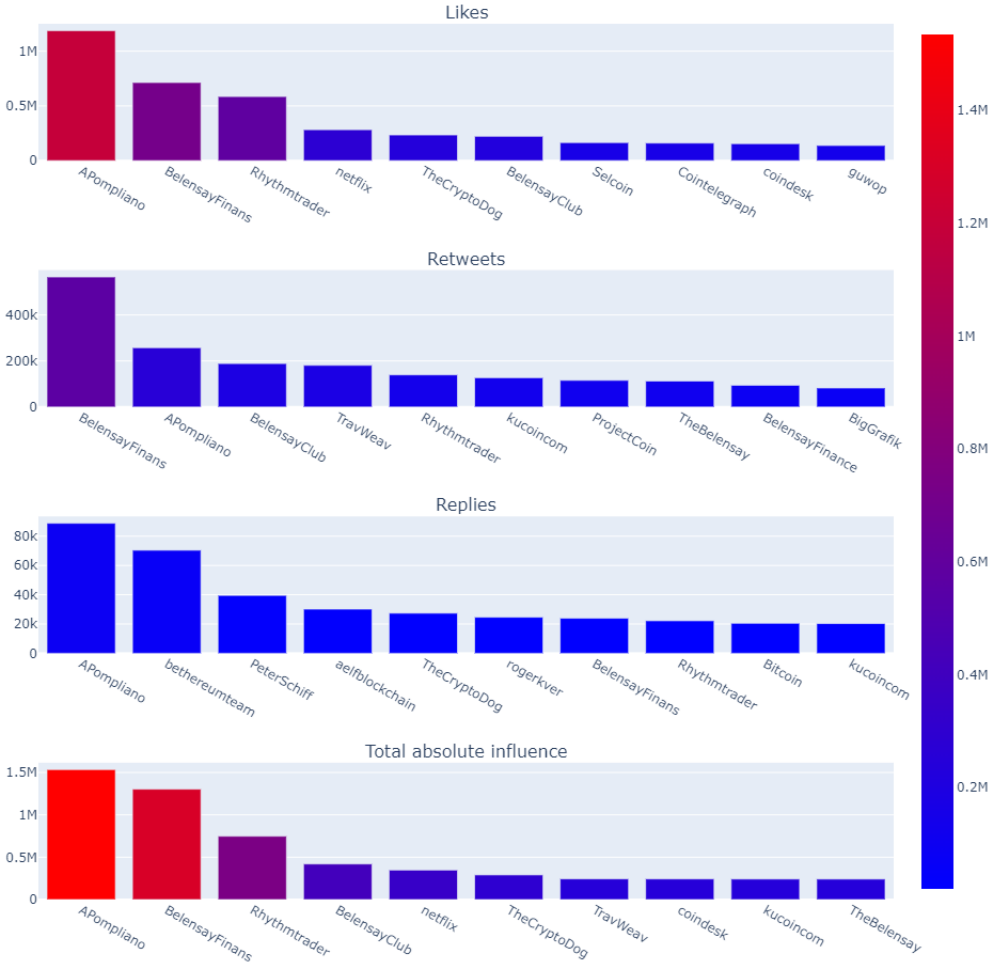
Σχήμα 6.7: Κατανομή tweets ανα ημέρα.

6.5 Περιγραφή των Χρηστών (Παγκοσμίως)

Τα παραπάνω tweets δημοσιεύθηκαν από συνολικά 920.170 χρήστες. Οι χρήστες που δημοσίευσαν τα περισσότερα tweets υπολογίστηκαν βάσει του ποσοστού των tweets που δημοσίευσαν ως προς τον συνολικό αριθμό tweets που περιλαμβάνει το σύνολο των δεδομένων και παρουσιάζονται στο σχήμα 6.8. Θεωρήθηκε χρήσιμο να υπολογισθεί η επιρροή των χρηστών αυτών στο κοινό, με βάση την απήχηση των δημοσιεύσεων τους. Ως δείκτης αυτής της επιρροής χρησιμοποιείται ο συνολικός αριθμός των likes, comments και retweets όλων των tweets του καθε χρήστη. Ο δείκτης αυτός θα χρησιμοποιηθεί και αργότερα ως συντελεστής βαρύτητας στον υπολογισμό του συνολικού συναισθήματος του κάθε tweet. Τα δεδομένα των 10 χρηστών με τους υψηλότερους δείκτες επιρροής παρουσιάζονται στα σχήματα 6.9.



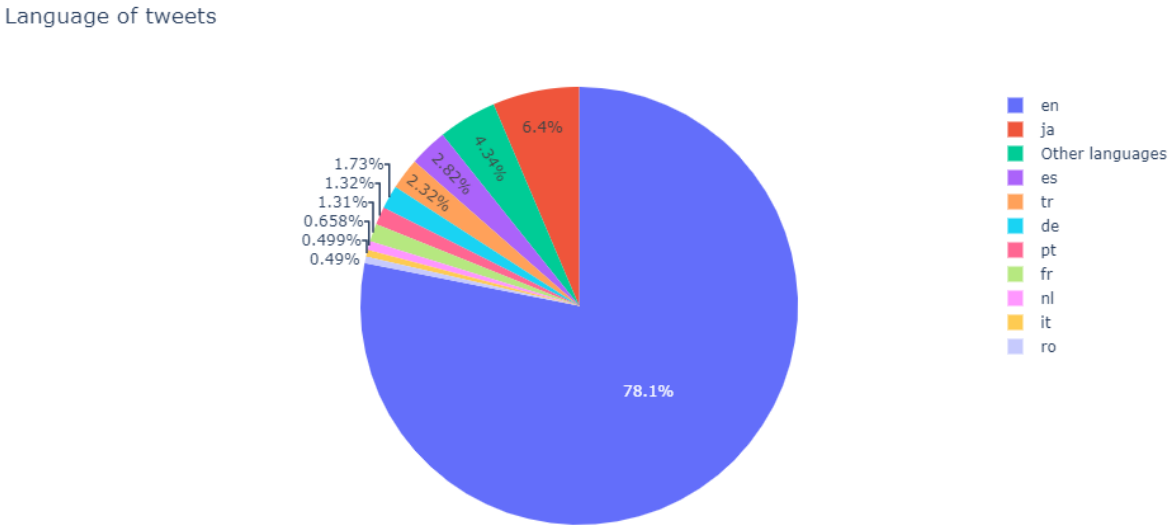
Σχήμα 6.8: tweets ανα χρήστη.



Σχήμα 6.9: Επιρροή των χρηστών.

6.6 Γλώσσα των tweets

Η γλώσσα του κειμένου είναι βασική παράμετρος της ανάλυσης του συναισθήματος οπότε έπρεπε να γίνει είτε μετάφραση όλων των tweets στα αγγλικά είτε αναγνώριση της γλώσσας και έπειτα αναγνώριση του συναισθήματος με βάση την κάθε γλώσσα. Δεδομένου του πολύ μεγάλου όγκου των δεδομένων, και οι δύο αυτές διαδικασίες απαιτούν αρκετό χρόνο και υπολογιστική ισχύ τα οποία δεν ήταν διαθέσιμα, δίνουν όμως έρεισμα για μελλοντική επέκταση της παρούσας εργασίας. Η κατανομή των tweets με βάση την γλώσσα τους παρουσιάζεται στο σχήμα 6.10. Όπως φαίνεται στο σχήμα τα σχόλια που αναγνωρίστηκαν ως αγγλικά αφορούν το 78.1% του συνόλου δεδομένων δηλαδή 8.548.785 από τα 10.942.567 tweets, ποσοστό το οποίο μπορεί να θεωρηθεί αντιπροσωπευτικό του παρόντος συνόλου.

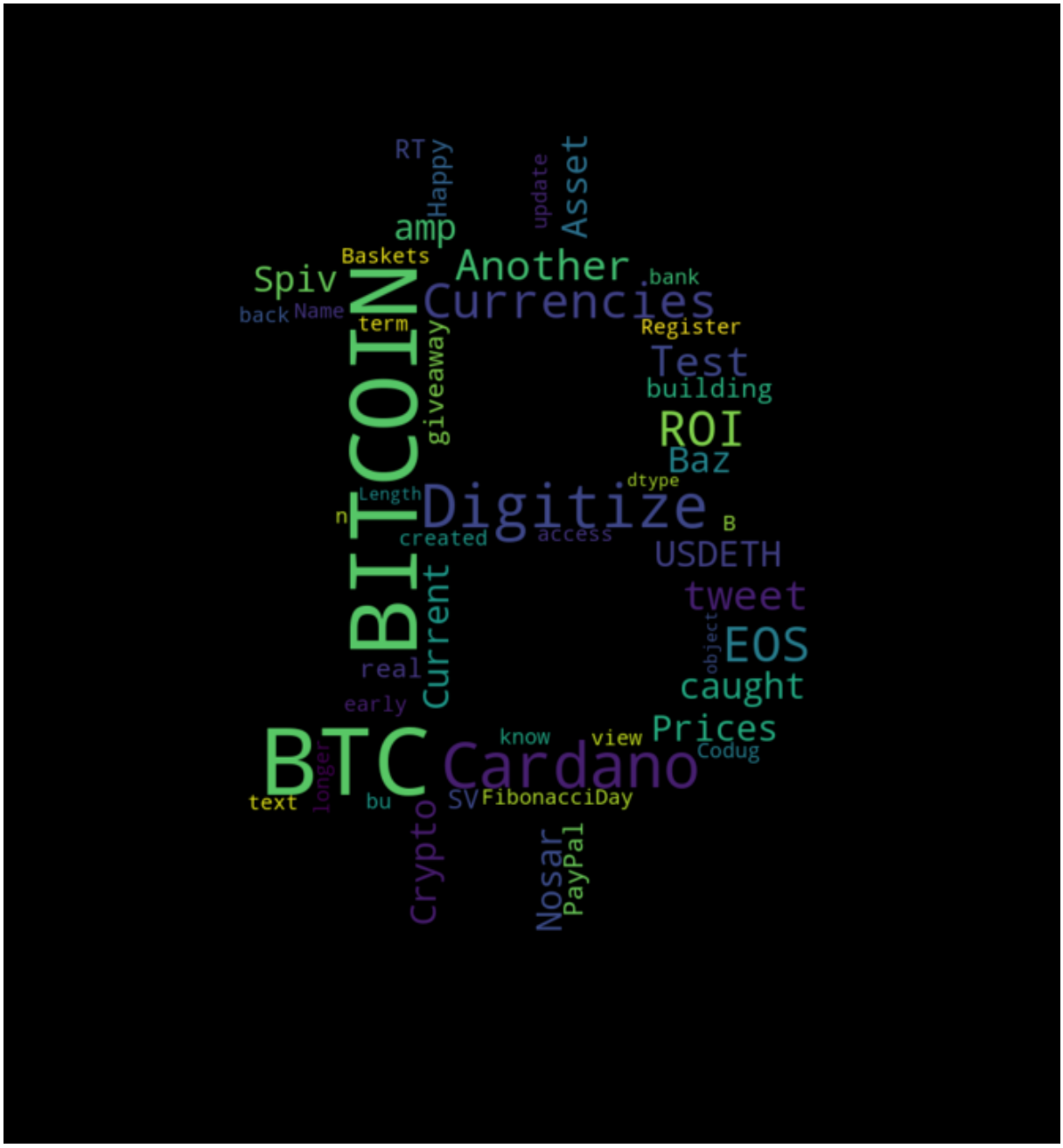


Σχήμα 6.10: Κατανομή των tweets με βάση την γλώσσα.

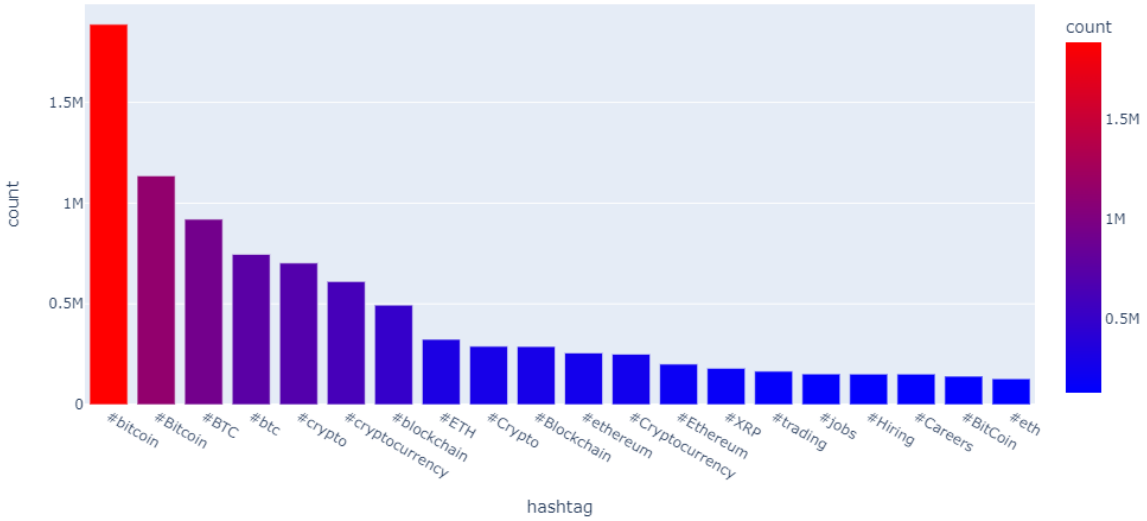
6.7 Περιγραφή τελικού συνόλου δεδομένων

Το σύνολο των δεδομένων το οποίο τελικά θα χρησιμοποιηθεί για την αναγνώριση του συναισθήματος περιέχει 8.548.785 tweets δημοσιευμένα από 581.614 χρήστες στο διάστημα μεταξύ 2013-04-29 και 2019-11-23, ώστε να συμπίπτει με το διάστημα των διαθέσιμων δεδομένων της τιμής του bitcoin και αφορούν μόνο τα tweets των οποίων ως βασική γλώσσα αναγνωρίστηκε η αγγλική.

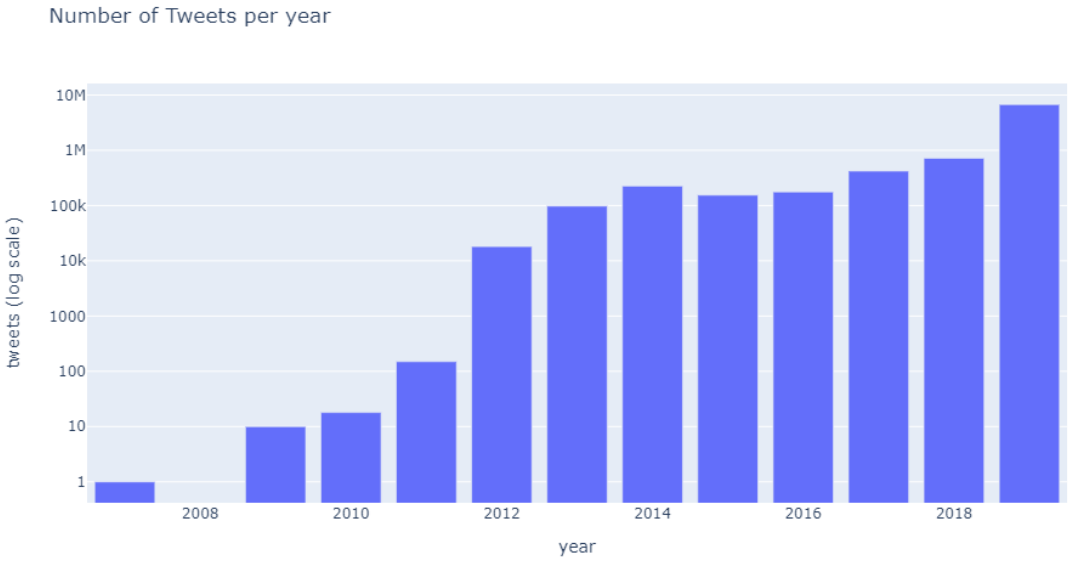
Στην εικόνα 6.11 απεικονίζονται οι πιο συχνά χρησιμοποιούμενες λέξεις στα tweets και στο σχήμα 6.12 παρουσιάζονται και τα πιό δημοφιλή hashtags. Η κατανομή των tweets στο διάστημα μεταξύ 19-04-2007 και 23-11-2019 απεικονίζεται στα γραφήματα 6.13 (κατανομή ανα έτος), 6.14 (κατανομή ανα μήνα), 6.15 (κατανομή ανα ημέρα του μήνα), 6.16 (κατανομή ανα ημέρα). Στο σχήμα 6.17 βρίσκονται οι χρήστες με τα περισσότερα tweets και το σχήμα 6.18 δείχνει τους χρήστες με την μεγαλύτερη επιρροή στα tweets που αφορούν το Bitcoin.



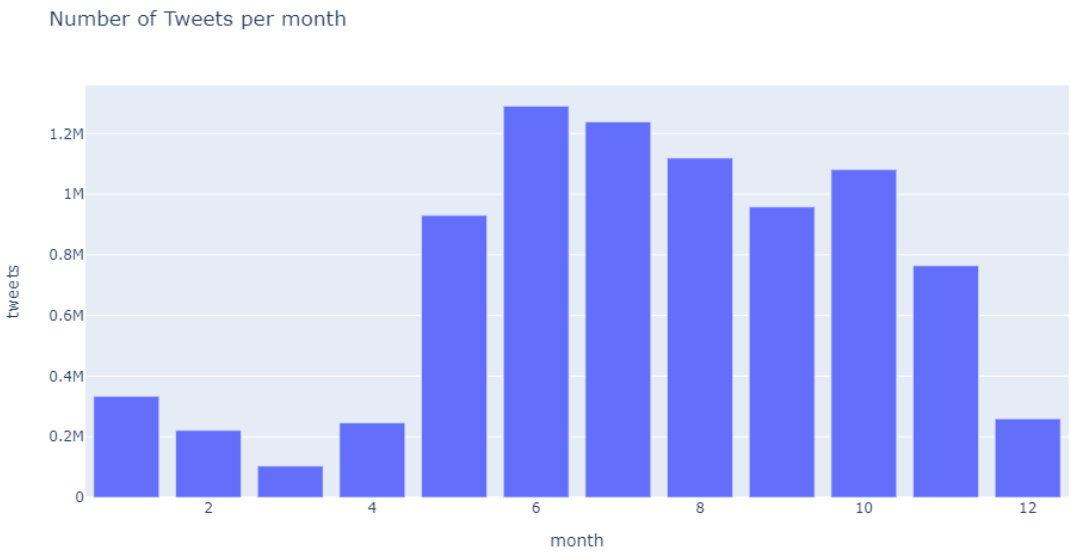
Σχήμα 6.11: Wordcloud: Συχνά χρησιμοποιούμενες λέξεις (αγγλικά tweets).



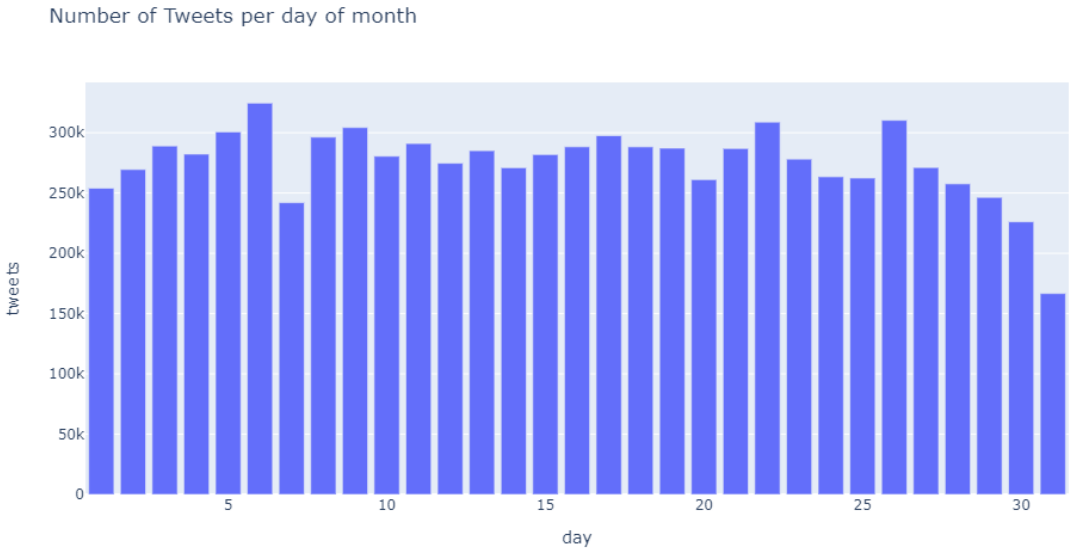
Σχήμα 6.12: Συχνά χρησιμοποιούμενα hashtags (αγγλικά tweets).



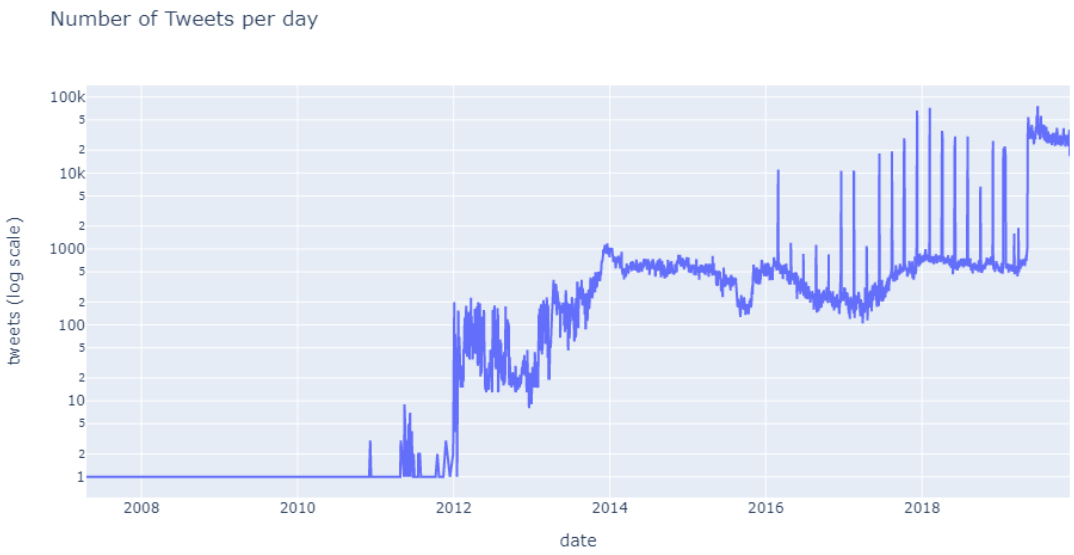
Σχήμα 6.13: Κατανομή tweets ανα έτος (αγγλικά tweets).



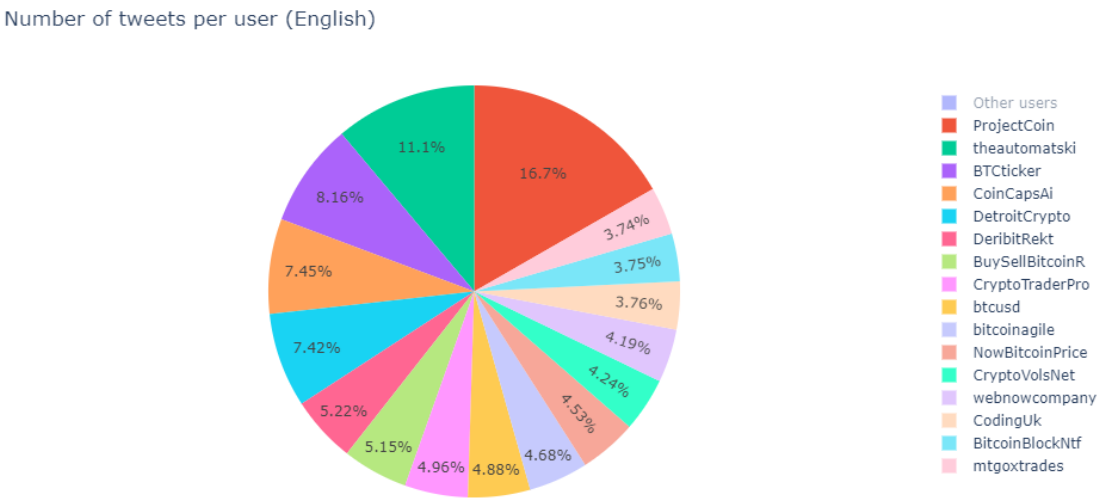
Σχήμα 6.14: Κατανομή tweets ανα μήνα (αγγλικά tweets).



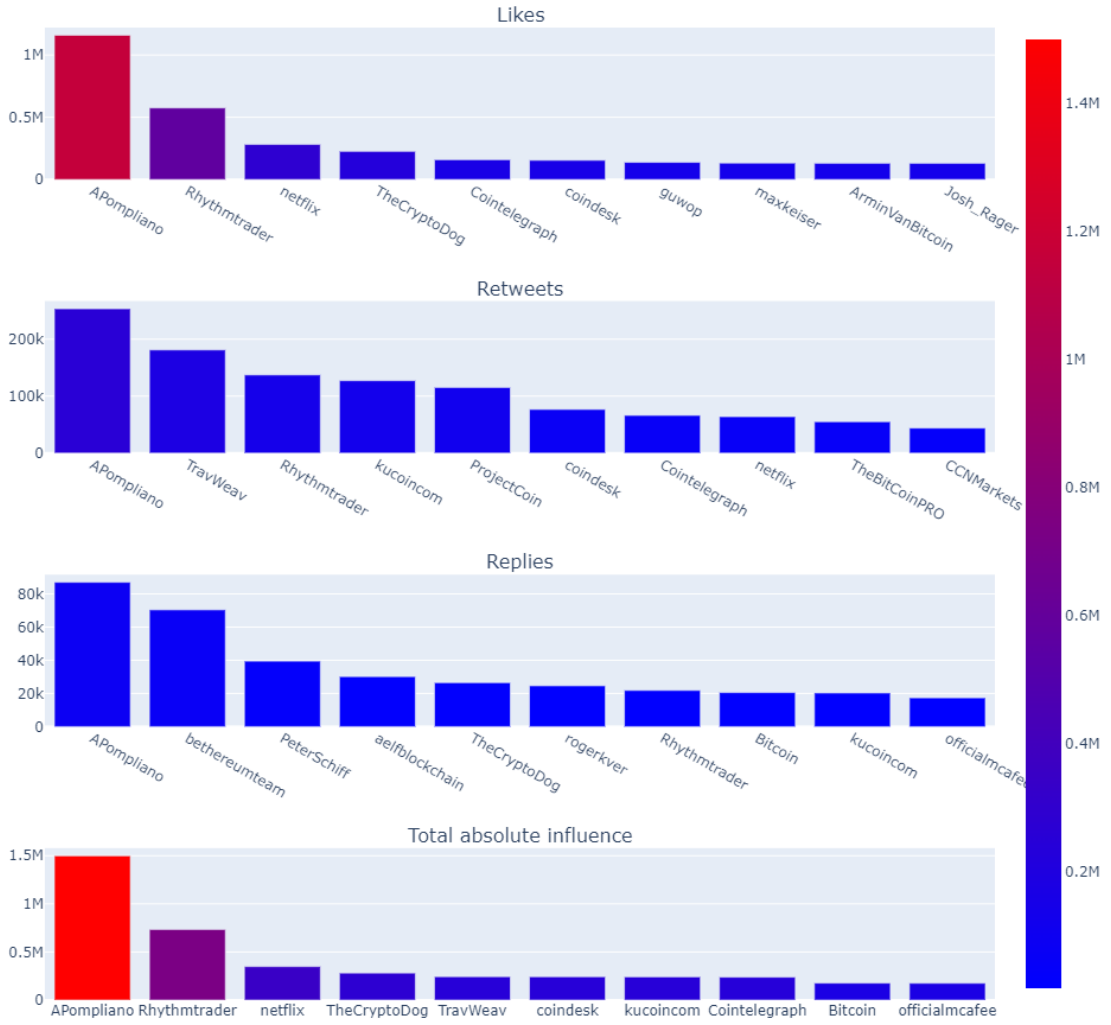
Σχήμα 6.15: Κατανομή tweets ανα ημέρα του μήνα (αγγλικά tweets).



Σχήμα 6.16: Κατανομή tweets ανα ημέρα (αγγλικά tweets).



Σχήμα 6.17: tweets ανα χρήστη (αγγλικά tweets).



Σχήμα 6.18: Επιρροή των χρηστών (αγγλικά tweets).

6.8 Συναισθηματική ανάλυση

Η ανάλυση του συναισθήματος του κάθε tweet έγινε με τη χρήση του εργαλείου VADER (Valence Aware Dictionary for sEntiment Reasoning) [42], εργαλείο σχεδιασμένο ειδικά για την ανάλυση συναισθήματος σε κείμενο που έχει αναρτηθεί στα κοινωνικά δίκτυα. Σε κάθε λέξη του κειμένου, το VADER, μέσω της βιβλιοθήκης του, αποδίδει έναν αριθμό μεταξύ 0 και 1 σε κάθε μια από τις κατηγορίες θετικό, αρνητικό και ουδέτερο. Έπειτα για κάθε tweet υπολογίζεται, μέσω αθροίσματος των αποτελεσμάτων κάθε κατηγορίας των λέξεων που το αποτελούν, το συνολικό αποτέλεσμα της πρότασης για κάθε μια από τις κατηγορίες θετικό αρνητικό και ουδέτερο. Ο τελικός δείκτης, ο οποίος χρησιμοποιείται στην συνέχεια, είναι η σύνθετη πολικότητα (compound polarity) η οποία προκύπτει από το άθροισμα των τιμών των επι μέρους κατηγοριών, προσαρμοσμένο σύμφωνα με ειδικούς κανόνες και κανονικοποιημένο μεταξύ -1 (το πιο ακραίο αρνητικό) και 1 (το πιο ακραίο θετικό). Το σχήμα 6.19 περιέχει κάποια χαρακτηριστικά παραδείγματα:

$$\text{"}VADER\ is\ smart,\ handsome,\ and\ funny.\text{"}$$
$$\{'pos': 0.746,\ 'compound': 0.8316,\ 'neu': 0.254,\ 'neg': 0.0\}$$

$$\text{"}VADER\ is\ very\ smart,\ handsome,\ and\ funny.\text{"}$$
$$\{'pos': 0.701,\ 'compound': 0.8545,\ 'neu': 0.299,\ 'neg': 0.0\}$$

$$\text{"}VADER\ is\ not\ smart,\ handsome,\ nor\ funny.\text{"}$$
$$\{'pos': 0.0,\ 'compound': -0.7424,\ 'neu': 0.354,\ 'neg': 0.646\}$$

Σχήμα 6.19: Παραδείγματα προσδιορισμού σύνθετης πολικότητας

Ανάλογα λοιπόν με την σύνθετη πολικότητα κάθε tweet ανήκει σε μια από τις 5 κατηγορίες σύμφωνα με τους κανόνες του σχήματος 6.20.

$$compound\ score = \begin{cases} [-1,-0.5) & , sentiment = Strongly\ Negative \\ [-0.5,0) & , sentiment = Weakly\ Negative \\ 0 & , sentiment = Neutral \\ (0,0.5] & , sentiment = Weakly\ Positive \\ (0.5,1] & , sentiment = Strongly\ Positive \end{cases}$$

Σχήμα 6.20: Απόφαση συναισθήματος βάσει της σύνθετης πολικότητας

Πειραματικά υπολογίζεται και ένας συντελεστής βαρύτητας για το κάθε tweet ο οποίος

προκύπτει από τα στοιχεία του tweet όπως ορίζει το σχήμα 6.21.

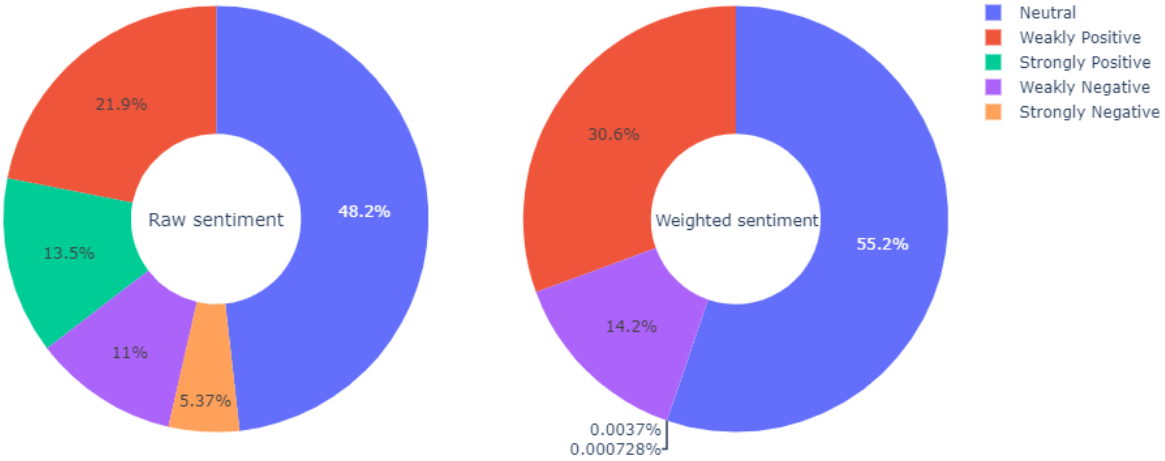
δείκτης επιρροής του *tweet* = *sum*(αριθμός *likes* του *tweet*,
αριθμός *comments* του *tweet*,
αριθμός *retweets* του *tweet*)

δείκτης επιρροής χρήστη = *sum*(αριθμός *likes* όλων των *tweets* του χρήστη ,
αριθμός *comments* όλων των *tweets* του χρήστη,
αριθμός *retweets* όλων των *tweets* του χρήστη)

συνολικός δείκτης επιρροής του *tweet* = *mean*(δείκτης επιρροής του *tweet*,
δείκτης επιρροής χρήστη)

Σχήμα 6.21: Τρόπος υπολογισμού του δείκτη βαρύτητας.

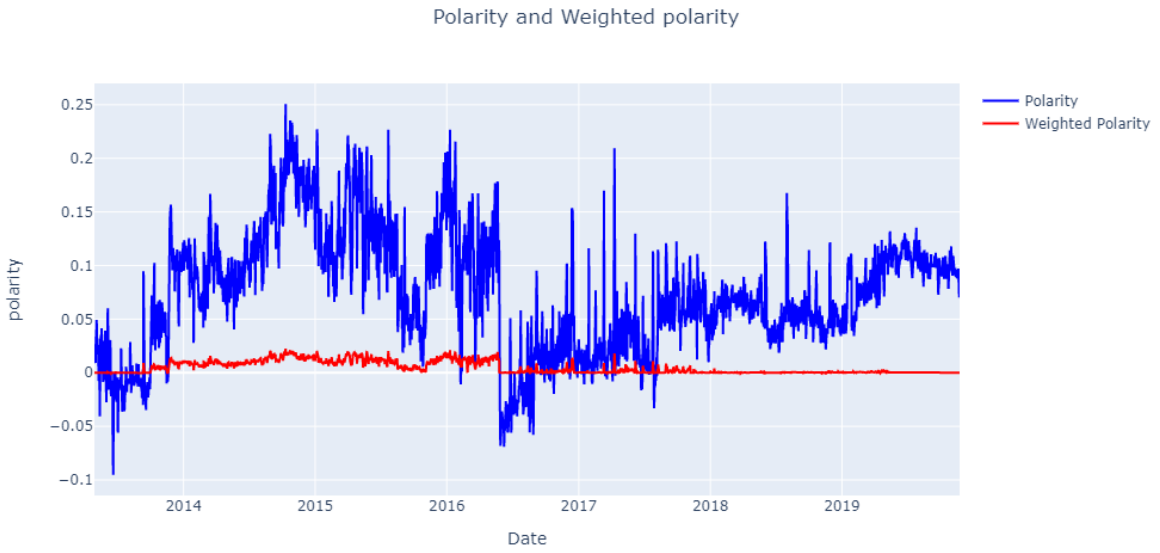
Χρησιμοποιώντας τον συντελεστή βαρύτητας στη σύνθετη πολικότητα, προκύπτει και η σταθμισμένη σύνθετη πολικότητα (weighted compound polarity). Οπότε εφαρμόζοντας τους ίδιους κανόνες (του σχήματος 6.20) και για τη νέα σταθμισμένη σύνθετη πολικότητα, η κατανομή των tweets στις 5 κατηγορίες συναισθήματος βάσει της σύνθετης πολικότητας και της σταθμισμένης σύνθετης πολικότητας παρουσιάζεται στο σχήμα 6.22.



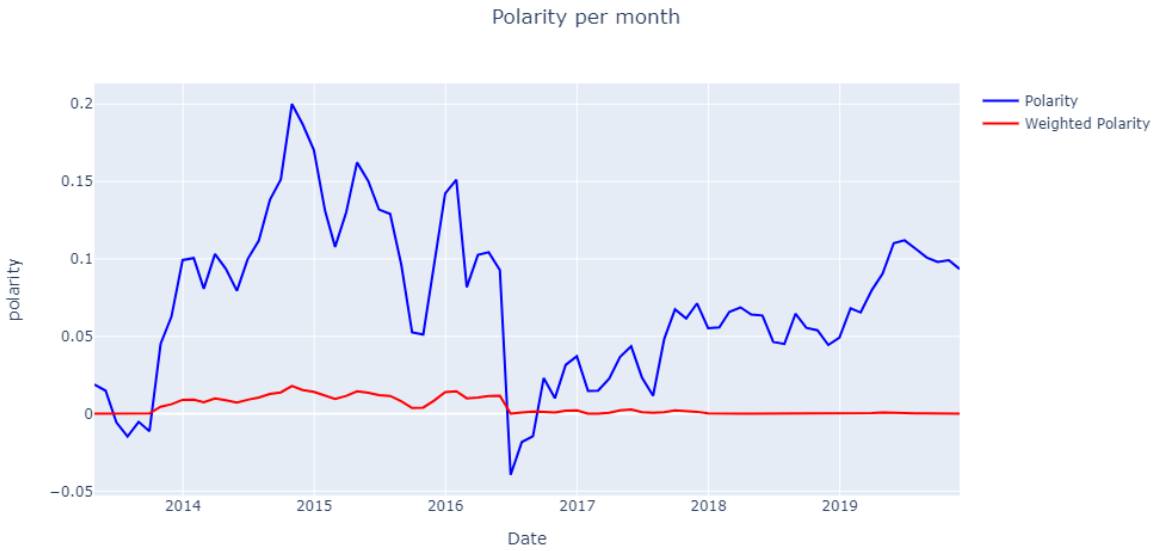
Σχήμα 6.22: Κατανομή των tweets στις 5 κατηγορίες συναισθήματος

Τέλος υπολογίστηκε το συνολικό ημερήσιο συναίσθημα ως ο μέσος όρος της σύνθετης πολικότητας και της σταθμισμένης σύνθετης πολικότητας των tweets που δημοσιεύθηκαν την συγκεκριμένη μέρα. Στα σχήματα 6.23, 6.24 και 6.25 παρουσιάζεται αντίστοιχα η ημερήσια, μηνιαία και ετήσια χρονική εξέλιξη της σύνθετης πολικότητας και της σταθμισμένης σύνθετης πολικότητας. Οι δείκτες αυτοί, της σύνθετης πολικότητας, της σταθμισμένης σύνθετης πολικότητας και του αριθμού των tweets θα χρησιμοποιηθούν στο επόμενο κεφάλαιο

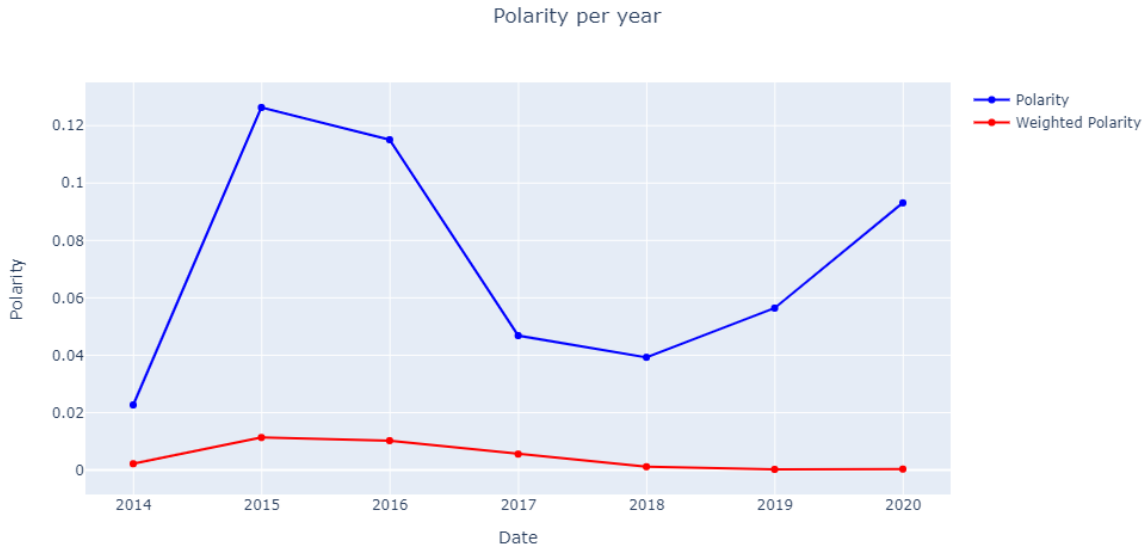
με στόχο την βελτίωση της πρόβλεψης.



Σχήμα 6.23: Ημερήσια χρονική εξέλιξη της σύνθετης πολικότητας και της σταθμισμένης σύνθετης πολικότητας



Σχήμα 6.24: Μηνιαία χρονική εξέλιξη της σύνθετης πολικότητας και της σταθμισμένης σύνθετης πολικότητας



Σχήμα 6.25: Ετήσια χρονική εξέλιξη της σύνθετης πολικότητας και της σταθμισμένης σύνθετης πολικότητας

Κεφάλαιο 7

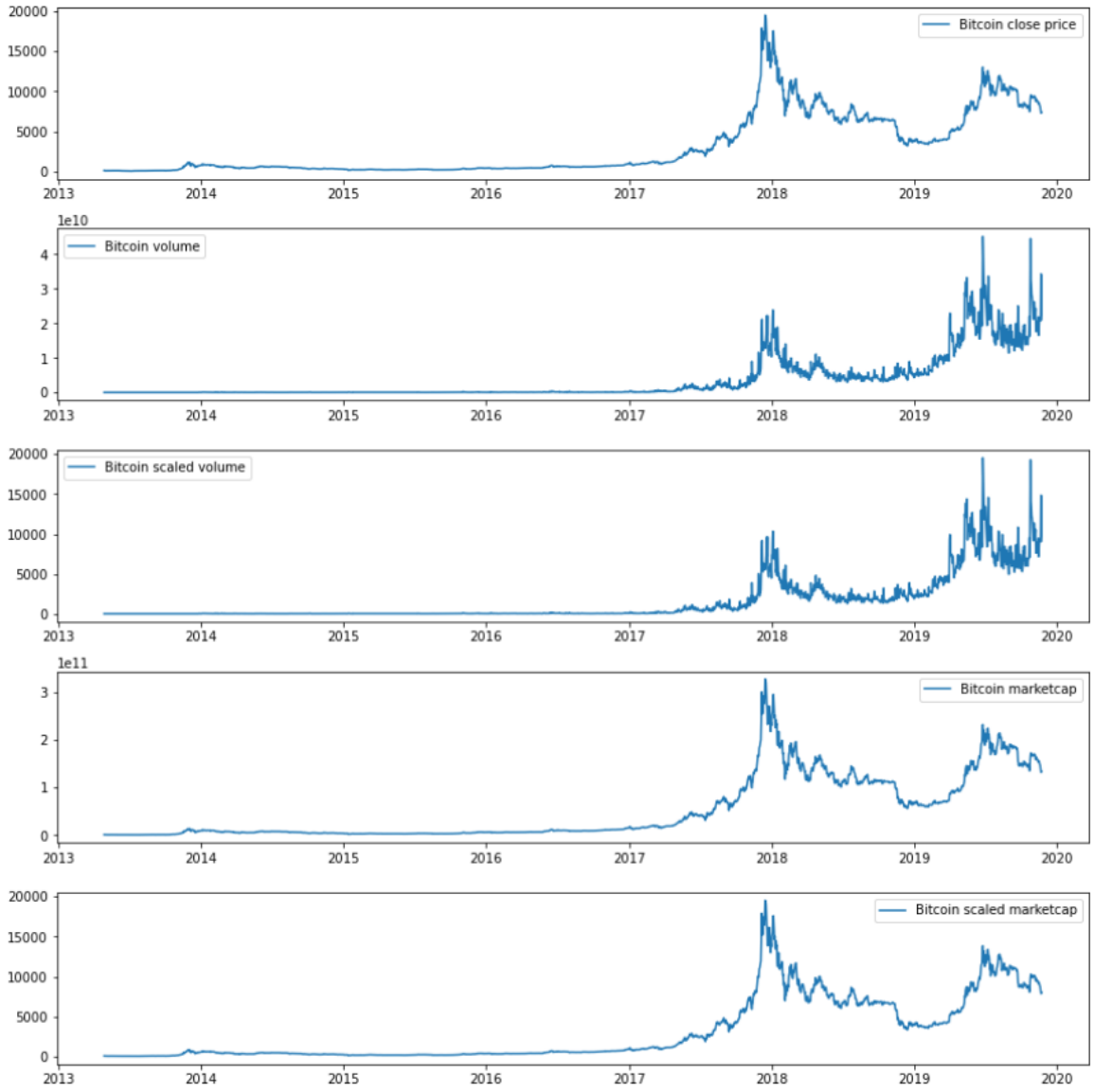
Πρόβλεψη με περισσότερες παραμέτρους

7.1 Εισαγωγή

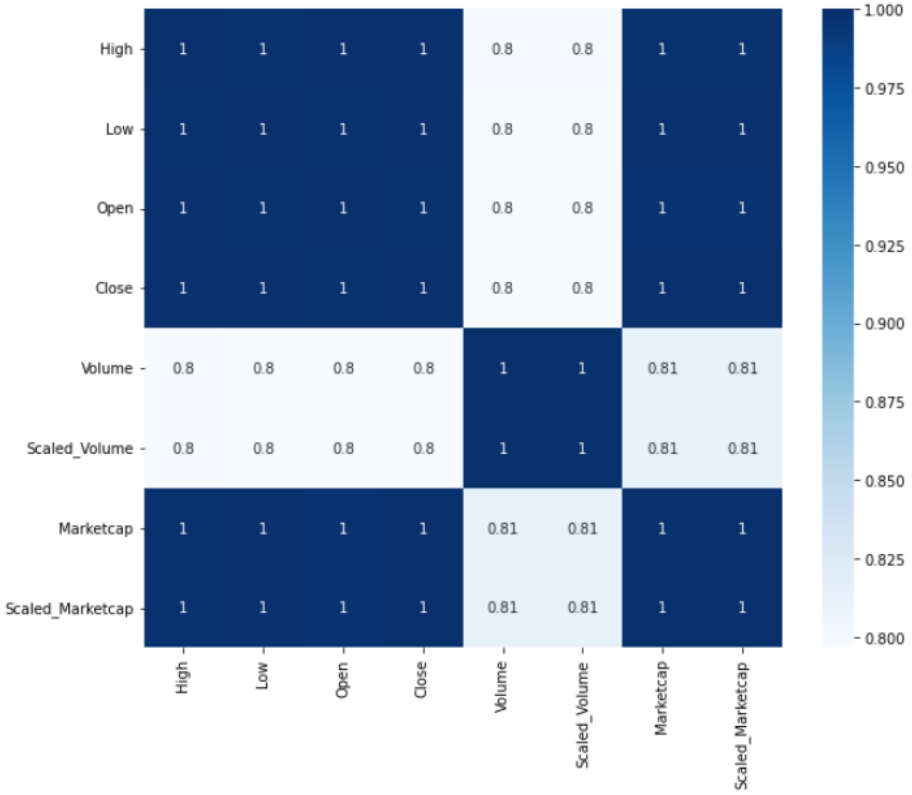
Στο κεφάλαιο 5 αναλύθηκε η χρονοσειρά της τιμής κλεισίματος του Bitcoin. Στο παρόν κεφάλαιο χρησιμοποιούνται οι δείκτες που προέκυψαν από την ανάλυση των δεδομένων του Twitter (κεφάλαιο 6) και γίνεται πειραματική πρόβλεψη για την τιμή του κρυπτονομίσματος μέσω όλων των διαθέσιμων μεταβλητών.

7.2 Προσθήκη οικονομικών δεδομένων του Bitcoin στην πρόβλεψη

Στο Κεφάλαιο 4 οι προσπάθειες πρόβλεψης έγιναν με βάση τις παρελθοντικές τιμές κλεισίματος του Bitcoin. Με σκοπό την βελτιστοποίηση του σφάλματος εισάγονται στον αλγόριθμο και άλλες χρονοσειρές που προκύπτουν από τις οικονομικές παραμέτρους του κρυπτονομίσματος όπως οι τιμές ανοίγματος, οι μέγιστες - ελάχιστες τιμές, ο όγκος καθώς και η κεφαλαιοποίηση της αγοράς σε συγκεκριμένα χρονικά διαστήματα. Προκειμένου να επιλεγούν οι κατάλληλες παράμετροι ελέγχεται η συσχέτιση τους με την τιμή κλεισίματος. Η εξέλιξη των οικονομικών δεδομένων παρουσιάζεται στο σχήμα 7.1 ενώ στο σχήμα 7.2 παρουσιάζεται ο πίνακας συσχέτισης τους.



Σχήμα 7.1: Εξέλιξη οικονομικών δεδομένων



Σχήμα 7.2: Πίνακας συσχέτισης μεταξύ των οικονομικών δεδομένων του Bitcoin

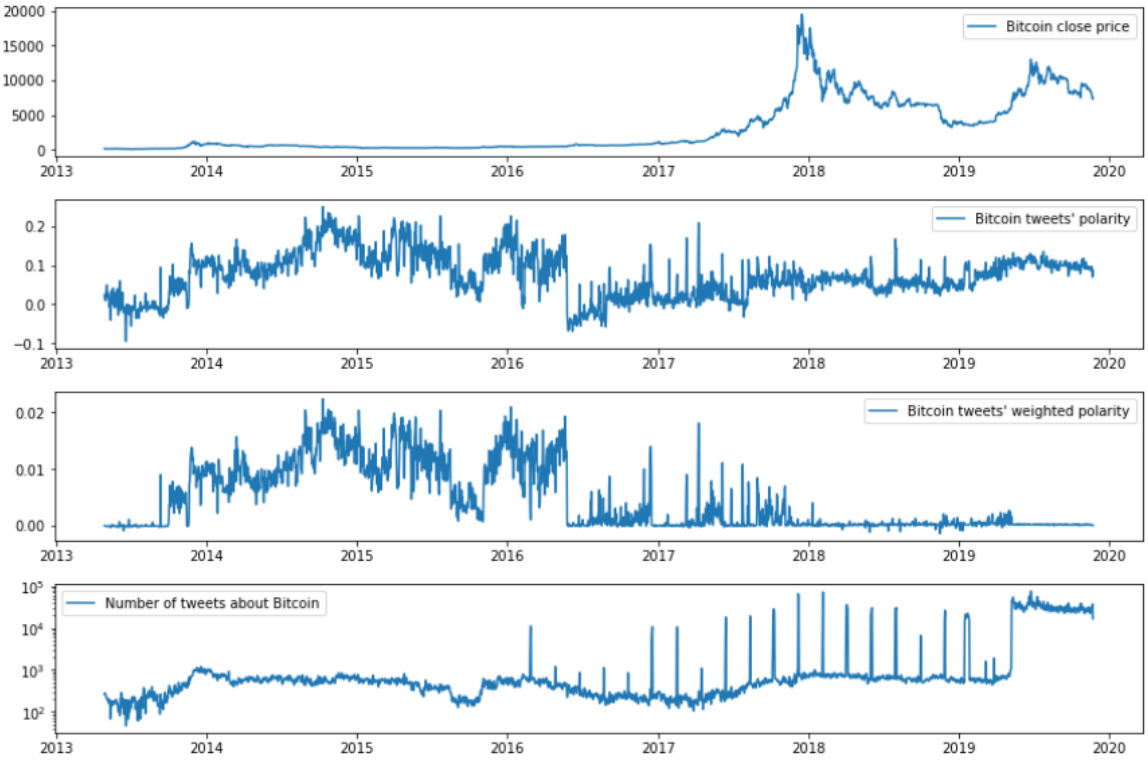
Η συσχέτιση μεταξύ των δεδομένων είναι πολύ μεγάλη οπότε η συμπερίληψή τους στην πρόβλεψη, πιθανότατα δεν θα επιφέρει κάποια βελτίωση. Στον πίνακα 7.1 παρουσιάζονται τα αποτελέσματα των εκτελέσεων της πρόβλεψης λαμβάνοντας υπόψιν κάθε φορά μια διαφορετική παράμετρο, όπου πράγματι δεν παρατηρείται κάποια σημαντική βελτίωση του σφάλματος.

Πίνακας 7.1: Αποτελέσματα πρόβλεψης από συνδυασμό παραμέτρων οικονομικών δεδομένων

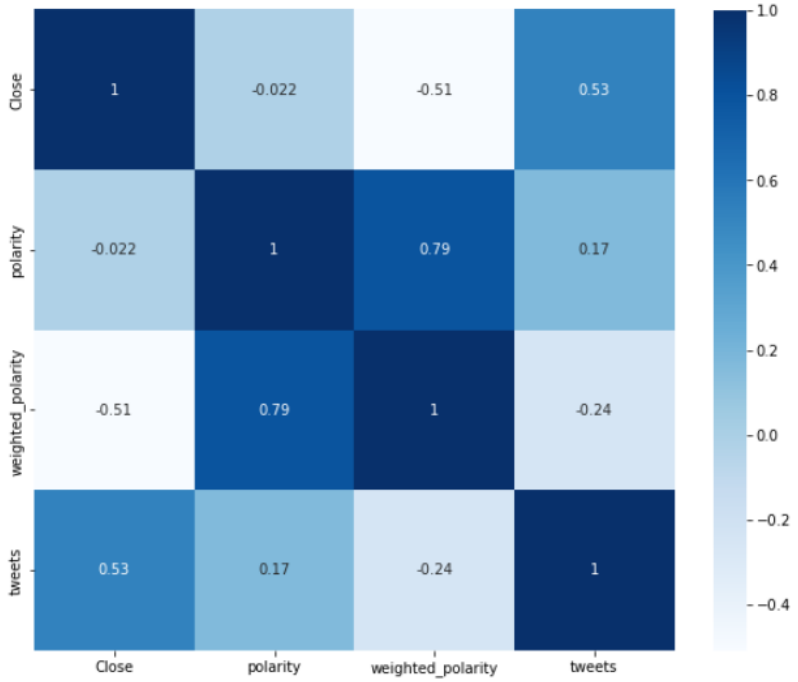
Παράμετροι	RMSE	Relative RMSE	MAPE
Close and High value	297.38	11.79%	47.79%
Close and Low value	325.06	12.89%	47%
Close and Open value	417.24	16.54%	45.97%
Close value	312.09	12.37%	47.27%
Close and Volume value	582.95	23.11%	50.85%
Close and Scaled Volume value	433.75	17.19%	46.16%
Close and Marketcap value	401.62	15.92%	49.71%
Close and Scaled Marketcap value	431.17	17.09%	49.46%

7.3 Προσθήκη δεδομένων της ανάλυσης των Tweets στην πρόβλεψη

Στο Κεφάλαιο 6 αναλύθηκε το συναίσθημα των Tweets που αφορούν το Bitcoin και προέκυψαν τρεις νέες παράμετροι που μπορούν να χρησιμοποιηθούν στην πρόβλεψη, η σύνθετη πολικότητα, η σταθμισμένη σύνθετη πολικότητα καθώς και ο αριθμός των tweets που δημοσιεύθηκαν ανα ημέρα. Προκειμένου να αξιολογηθούν οι παράμετροι ως προς την συμβολή τους στην πρόβλεψη, ελέγχεται η συσχέτιση τους με την τιμή κλεισίματος. Η εξέλιξη των δεδομένων που προέκυψαν από το Twitter παρουσιάζονται στο σχήμα 7.3 και ο πίνακας συσχέτισης μεταξύ τους στο σχήμα 7.4.



Σχήμα 7.3: Εξέλιξη των δεδομένων του Twitter



Σχήμα 7.4: Πίνακας συσχέτισης μεταξύ των δεδομένων του Twitter για το Bitcoin

Η συσχέτιση μεταξύ των δεδομένων είναι αρκετά μικρή οπότε η συμπερίληψή τους στην πρόβλεψη, θα μπορούσε να βελτιώσει την πρόβλεψη. Στον πίνακα 7.2 παρουσιάζονται τα αποτελέσματα των εκτελέσεων της πρόβλεψης λαμβάνοντας υπόψιν κάθε φορά μια διαφορετική παράμετρο. Όπως προκύπτει από τα σφάλματα, παρόλο που η συσχέτιση των δεδομένων είναι μικρή, ο συνδυασμός των παραμέτρων δεν καταφέρνει να βελτιώσει σημαντικά το σφάλμα της πρόβλεψης.

Πίνακας 7.2: Αποτελέσματα πρόβλεψης από συνδυασμό παραμέτρων των δεδομένων του Twitter

Παράμετροι	RMSE	Relative RMSE	MAPE
Close and Polarity	298.36	11.83%	47.64%
Close and Weighted Polarity	316.52	12.55%	47.04%
Close and Number of Tweets	335.94	13.32%	46.9%

7.4 Περίληψη κεφαλαίου

Η προσθήκη κάποιων μεταβλητών βελτίωσε την πρόβλεψη συνεισφέροντας σε μία μικρή μείωση του σφάλματος. Οι δείκτες οι οποίοι φαίνεται να συμβάλλουν στην καλύτερη πρόβλεψη είναι η μέγιστη τιμή του Bitcoin και η σύνθετη πολικότητα, παράγοντες οι οποίοι μειώνουν το απόλυτο RMSE κατα 14.71 και 13.73 μονάδες αντίστοιχα και το σχετικό RMSE κατα 0.58% και 0.54% αντίστοιχα.

Κεφάλαιο 8

Επίλογος

8.1 Σύνοψη και συμπεράσματα

Στην παρούσα πτυχιακή πραγματοποιήθηκε μια ανάλυση της εξέλιξης τόσο της τιμής του δημοφιλούς κρυπτονομίσματος Bitcoin, όσο και της επιρροής του στο κοινωνικό δίκτυο Twitter μελετώντας σχετικά δεδομένα. Έπειτα από τις αναλύσεις και τα πειράματα που διεξήχθησαν, προκύπτουν ορισμένα συμπεράσματα που αφορούν την προβλεψιμότητα του κρυπτονομίσματος και το κατα πόσο επηρεάζεται η τάση της τιμής του από το συνολικό συναίσθημα που εκφράζεται γι' αυτό στα social media.

Στο κεφάλαιο 5 γίνεται η πρώτη προσπάθεια πρόβλεψης με τη χρήση του μοντέλου Prophet όμως τόσο τα αποτελέσματα όσο και το εύρος του πιθανού σφάλματος δεν είναι ικανοποιητικά. Ακολουθούν μετασχηματισμοί με στόχο τη βελτίωση της σταθερότητας της χρονοσειράς και και κατ΄ επέκταση της απόδοσης των εργαλείων ARIMA και SARIMAX. Το μοντέλο SARIMAX που επιλέχθηκε ανταποκρίθηκε καλύτερα στις μεταβολές του Bitcoin από οτι το βέλτιστο μοντέλο ARIMA μιας και οι αυξομειωτικές τάσεις προβλέφθηκαν με ικανοποιητική ακρίβεια. Παρόλα αυτά η τιμή πρόβλεψης απείχε αρκετά από τις πραγματικές τιμές, γεγονός που επέφερε και μεγάλο σφάλμα RMSE. Η καλύτερη προσπάθεια πρόβλεψης έγινε με την χρήση ενος επαναλαμβανόμενου νευρωνικού δικτύου δυο στρωμάτων (GRU-LSTM) όπου προβλέφθηκαν με μεγάλη ακρίβεια οι αυξομειώσεις της τάσης, γεγονός που επιβεβαίωνουν και οι τιμές του συνολικού σφάλματος RMSE = 312.09.

Στο κεφάλαιο 6 γίνεται ανάλυση των δεδομένων του Twitter με απώτερο στόχο την εξαγωγή δεικτών που θα χρησιμοποιηθούν ως πρόσθετοι παράγοντες πρόβλεψης στο βέλτιστο νευρωνικό δίκτυο του κεφαλαίου 5. Αρχικά περιγράφονται τα δεδομένα και έπειτα γίνεται

η βασική προεπεξεργασία καθαρισμού των tweets. Ακολουθεί η ανάλυση των χρηστών και της εξέλιξης των Tweets σε παγκόσμιο επίπεδο. Αναγνωρίστηκε η γλώσσα στην οποία ήταν γραμμένα τα tweets και στη συνέχεια, στα Αγγλικά tweets εφαρμόστηκε συναισθηματική ανάλυση. Από την ανάλυση αυτή προέκυψαν οι δείκτες της πολικότητας και της σύνθετης πολικότητας στον προσδιορισμό της οποίας χρησιμοποιούνται βάρη που υπολογίστηκαν βάσει της επιρροής των tweets. Τα αποτελέσματα της συναισθηματικής ανάλυσης έδειξαν οτι το 5.37% των σχολίων εκφράζουν καθαρά αρνητική άποψη, το 11% σχετικά αρνητική, το 48.2% ουδέτερη, το 21.9% σχετικά θετική και το 13.5% καθαρά θετική άποψη.

Τέλος στο κεφάλαιο 7 γίνεται χρήση των δεικτών της πολικότητας, της σύνθετης πολικότητας, του αριθμού των tweets και των οικονομικών δεικτών του Bitcoin (μέγιστη-ελάχιστη τιμή, τιμή ανοίγματος-κλεισίματος, όγκος και κεφαλαιοποίηση αγοράς), στο βέλτιστο νευρωνικό δίκτυο της προηγούμενης ενότητας, ως πρόσθετες μεταβλητές πρόβλεψης. Έπειτα από σύγκριση των σφαλμάτων που παράγει ο κάθε συνδυασμός μεταβλητών, συμπεραίνεται οτι οι δείκτες που βελτιώνουν την πρόβλεψη είναι η μέγιστη τιμή του κρυπτονομίσματος, που δίνει RMSE = 297.38 και η πολικότητα που δίνει RMSE = 298.36, μειώνοντας το RMSE κατα 14.71 και 13.73 μονάδες αντίστοιχα.

8.2 Μελλοντικές επεκτάσεις

Με σκοπό την βελτίωση των αποτελεσμάτων και την εξαγωγή περισσότερων συμπερασμάτων, δίνεται έρεισμα για περαιτέρω έρευνα στους παρακάτω τομείς της εργασίας:

- Όσον αφορά την μελέτη της γλώσσας των tweets που συλλέχθηκαν, με την χρήση διαφορετικών λεξικών μπορεί να γίνει συναισθηματική ανάλυση σε μεγαλύτερο αριθμό tweets που θα αφορούν διαφορετικές γλώσσες από την αγγλική, με στόχο μια πιο σφαιρική ανάλυση της παγκόσμιας εικόνας σχετικά με το κρυπτονόμισμα (ενότητα 6.6).

- Τα διαθέσιμα δεδομένα του Twitter που αφορούν το Bitcoin καλύπτουν το χρονικό διάστημα μέχρι τις 23-11-2019. Με την χρήση του Twitter API ή κάποιας άλλης μεθόδου εξόρυξης δεδομένων από τα social media, μπορούν να συλλεχθούν πιο σύγχρονα δεδομένα, ακόμη και τρέχοντα, με στόχο την πρόβλεψη της τιμής στο άμεσο μέλλον (ενότητα 6.2).

- Το φιλτράρισμα των tweets αναφορικά με τα δεδομένα που παράγονται από bots μπο-

ρεί να βελτιωθεί κάνοντας χρήση εξειδικευμένων μεθόδων ώστε ο θόρυβος που παράγεται από αυτά να περιοριστεί (ενότητα 6.3).

- Η βασική ιδέα της πρόβλεψης τις τιμής του Bitcoin μπορεί να επεκταθεί και σε άλλα κρυπτονομίσματα με την συλλογή σχετικών δεδομένων από τα social media ώστε να βελτιωθεί η επίδοση των νευρωνικών δικτύων κάνοντας χρήση των παραμέτρων που μελετήθηκαν στην ενότητα 5.7.

Βιβλιογραφία

[1] Sik-Ho Tsang. Empirical evaluation of gated recurrent neural networks on sequence modeling. *Sik-Ho Tsang*, 2014.

[2] Işil Yenidoğan, Aykut Çayir, Ozan Kozan, Tuğçe Dağ, and Çiğdem Arslan. Bitcoin forecasting using arima and prophet. In *2018 3rd International Conference on Computer Science and Engineering (UBMK)*, pages 621–624. IEEE, 2018.

[3] Takaomi Hirata, Takashi Kuremoto, Masanao Obayashi, Shingo Mabu, and Kunikazu Kobayashi. Time series prediction using dbn and arima. In *2015 International Conference on Computer Application Technologies*, pages 24–29, 2015.

[4] S.L Ho, M Xie, and T.N Goh. A comparative study of neural network and box-jenkins arima modeling in time series prediction. *Computers & Industrial Engineering*, 42(2):371–375, 2002.

[5] Suhwan Ji, Jongmin Kim, and Hyeonseung Im. A comparative study of bitcoin price prediction using deep learning. *Mathematics*, 7(10):898, 2019.

[6] Aniruddha Dutta, Saket Kumar, and Meheli Basu. A gated recurrent unit approach to bitcoin price prediction. *Journal of Risk and Financial Management*, 13(2):23, 2020.

[7] Jerome T Connor, R Douglas Martin, and Les E Atlas. Recurrent neural networks and robust time series prediction. *IEEE transactions on neural networks*, 5(2):240–254, 1994.

[8] Vasily Derbentsev, Andriy Matviychuk, and Vladimir N Soloviev. Forecasting of cryptocurrency prices using machine learning. In *Advanced Studies of Financial Technologies and Cryptocurrency Markets*, pages 211–231. Springer, 2020.

[9] Μουλοσιώτης Κωνσταντίνος. Πρόβλεψη Εκλογικού Αποτελέσματος με Χρήση Συναισθηματικής Ανάλυσης στα Δεδομένα του twitter. Πτυχιακή εργασία, Πανεπιστήμιο Θεσσαλίας, Σεπτ. 2021.

[10] Shihab Elbagir and Jing Yang. Twitter sentiment analysis using natural language toolkit and vader sentiment. In *Proceedings of the international multiconference of engineers and computer scientists*, volume 122, page 16, 2019.

[11] Τζαβέλλος Φοίβος-Ιωάννης. Πρόβλεψη τιμής κρυπτονομίσματος χρησιμοποιώντας ανάλυση συναισθημάτων. Πτυχιακή εργασία, Πανεπιστήμιο Θεσσαλίας, Σεπτ. 2021.

[12] Πλέσσας-Αυγητίδης Νίκος. Μελέτη Ιατρικών Και Κοινωνιολογικών Δεδομένων Που Αφορούν Τον Νέο Κορονοϊό covid-19 Με Χρήση Αλγορίθμων Εξόρυξης Δεδομένων. Πτυχιακή εργασία, Πανεπιστήμιο Θεσσαλίας, Φεβρ. 2021.

[13] Anshul Mittal and Arpit Goel. Stock prediction using twitter sentiment analysis. *Standford University, CS229 (2011 http://cs229. stanford. edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis. pdf)*, 15:2352, 2012.

[14] Tushar Rao, Saket Srivastava, et al. Analyzing stock market movements using twitter sentiment analysis. *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, 2012.

[15] Yanwei Bao, Changqin Quan, Lijuan Wang, and Fuji Ren. The role of pre-processing in twitter sentiment analysis. In *International conference on intelligent computing*, pages 615–624. Springer, 2014.

[16] Fajri Koto and Mirna Adriani. A comparative study on twitter sentiment analysis: Which features are good? In *International Conference on Applications of natural language to information systems*, pages 453–457. Springer, 2015.

[17] Xujuan Zhou, Xiaohui Tao, Jianming Yong, and Zhenyu Yang. Sentiment analysis on tweets for social events. In *Proceedings of the 2013 IEEE 17th international conference on computer supported cooperative work in design (CSCWD)*, pages 557–562. IEEE, 2013.

[18] Cryptocurrencies. `https://www.investopedia.com/`. Ημερομηνία πρόσβασης: 28-01-2022.

[19] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

[20] Vineet Chaoji, Rajeev Rastogi, and Gourav Roy. Machine learning in the real world. *Proceedings of the VLDB Endowment*, 9(13):1597–1600, 2016.

[21] F.S. Rohman, S. Abdul Sata, and N. Aziz. Application of derivative - free estimator for semi batch autocatalytic esterification reactor: Comparison study of unscented kalman filter, divided difference kalman filter and cubature kalman filter. In Krist V. Gernaey, Jakob K. Huusom, and Rafiqul Gani, editors, *12th International Symposium on Process Systems Engineering and 25th European Symposium on Computer Aided Process Engineering*, volume 37 of *Computer Aided Chemical Engineering*, pages 329–334. Elsevier, 2015.

[22] Laila Ouazzani Chahidi, Marco Fossa, Antonella Priarone, and Abdellah Mechaqrane. Evaluation of supervised learning models in predicting greenhouse energy demand and production for intelligent and sustainable operations. *Energies*, 14(19), 2021.

[23] Ummul Khair, Hasanul Fahmi, Sarudin Al Hakim, and Robbi Rahim. Forecasting error calculation with mean absolute deviation and mean absolute percentage error. In *Journal of Physics: Conference Series*, volume 930, page 012002. IOP Publishing, 2017.

[24] Bing Liu. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167, 2012.

[25] Anastasia Giachanou and Fabio Crestani. Like it or not: A survey of twitter sentiment analysis methods. *ACM Comput. Surv.*, 49(2), jun 2016.

[26] Bhumika Gupta, Monika Negi, Kanika Vishwakarma, Goldi Rawat, Priyanka Badhani, and B Tech. Study of twitter sentiment analysis using machine learning algorithms on python. *International Journal of Computer Applications*, 165(9):29–34, 2017.

[27] Sebastian Raschka. *Python machine learning*. Packt publishing ltd, 2015.

[28] Zar Zar Oo and Sabai Phyu. Time series prediction based on facebook prophet: A case study, temperature forecasting in myintkyina. *International Journal of Applied Mathematics Electronics and Computers*, pages 263 – 267, 2020.

[29] Sean J Taylor and Benjamin Letham. Forecasting at scale. *The American Statistician*, 72(1):37–45, 2018.

[30] G Peter Zhang. Time series forecasting using a hybrid arima and neural network model. *Neurocomputing*, 50:159–175, 2003.

[31] Hirotogu Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike*, pages 199–213. Springer, 1998.

[32] Akaike's information criterion (aic). `https://www.mathworks.com/help/ident/ref/idgrey.aic.html#buy66ft-5`. Ημερομηνία πρόσβασης: 31-01-2022.

[33] Agostino Tarsitano and Ilaria L Amerise. Short-term load forecasting using a two-stage sarimax model. *Energy*, 133:108–114, 2017.

[34] Richard ID Harris. Testing for unit roots using the augmented dickey-fuller test: Some issues relating to the size, power and the lag structure of the test. *Economics letters*, 38(4):381–386, 1992.

[35] Cryptocurrency historical prices. `https://www.kaggle.com/sudalairajkumar/cryptocurrencypricehistory`. Ημερομηνία πρόσβασης: 9-10-2021.

[36] Closing price. `https://www.investor.gov/introduction-investing/investing-basics/glossary/closing-price`. Ημερομηνία πρόσβασης: 21-12-2021.

[37] What is volume? `https://coinmarketcap.com/alexandria/glossary/volume`. Ημερομηνία πρόσβασης: 21-12-2021.

[38] What is market cap? `https://www.coinbase.com/learn/crypto-basics/what-is-market-cap`. Ημερομηνία πρόσβασης: 21-12-2021.

[39] SJ Taylor and B. Letham. Forecasting at scale. *PeerJ Preprints 5:e3190v2*, 2017. https://doi.org/10.7287/peerj.preprints.3190v2.

[40] Mohil Maheshkumar Patel, Sudeep Tanwar, Rajesh Gupta, and Neeraj Kumar. A deep learning-based cryptocurrency price prediction scheme for financial institutions. *Journal of Information Security and Applications*, 55:102583, 2020.

[41] Bitcoin tweets-16m tweets. `https://www.kaggle.com/alaix14/bitcoin-tweets-20160101-to-20190329`. Ημερομηνία πρόσβασης: 9-11-2021.

[42] C.J. Hutto and E.E. Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Eighth International Conference on Weblogs and Social Media (ICWSM-14)*, June 2014.

Παράρτημα

Αρχεία Κώδικα και Σύνολα Δεδομένων

Τα αρχεία κώδικα, απο τα οποία προκύπτουν τα αποτελέσματα της εργασίας, βρίσκονται ανεβασμένα σε αποθετήριο στην πλατφόρμα GitHub (σύνδεσμος: `https://github.com/atsaklidis/Thesis_ECE_UTH`). Για την εκτέλεση του κώδικα απαιτούνται τα σύνολα δεδομένων που αφορούν τις τιμές των κρυπτονομισμάτων (σύνδεσμος: `https://www.kaggle.com/sudalairajkumar/cryptocurrencypricehistory`) και τα δεδομένα του Twitter (σύνδεσμος: `https://www.kaggle.com/alaix14/bitcoin-tweets-20160101-to-20190329`) που είναι διαθέσιμα μέσω της πλατφόρμας Kaggle.