



**ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ
ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ**

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ & ΤΕΧΝΟΛΟΓΙΑΣ ΥΠΟΛΟΓΙΣΤΩΝ

**ΤΟΜΕΑΣ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ & ΤΕΧΝΟΛΟΓΙΑΣ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ
ΕΡΓΑΣΤΗΡΙΟ ΕΝΣΥΡΜΑΤΗΣ ΤΗΛΕΠΙΚΟΙΝΩΝΙΑΣ**

**«Χρήση εξελικτικών νευρωνικών δικτύων και επεξεργασίας
φυσικής γλώσσας για την πρόβλεψη ισοτιμιών
συναλλάγματος»**

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΓΕΩΡΓΙΟΣ Δ. ΚΑΦΤΑΝΗΣ

ΕΠΙΒΛΕΠΩΝ: Κ. ΣΓΑΡΜΠΑΣ

ΠΑΤΡΑ, 2021

Πανεπιστήμιο Πατρών, Τμήμα Ηλεκτρολόγων Μηχανικών και Τεχνολογίας Υπολογιστών
Γεώργιος Καφτάνης.....(ονοματεπώνυμο)

© ...2021.....(έτος) – Με την επιφύλαξη παντός δικαιώματος

Το σύνολο της εργασίας αποτελεί πρωτότυπο έργο, παραχθέν από τον/την ...Γεώργιος Καφτάνης....(ονοματεπώνυμο), και δεν παραβιάζει δικαιώματα τρίτων καθ' οιονδήποτε τρόπο. Αν η εργασία περιέχει υλικό, το οποίο δεν έχει παραχθεί από την ίδιο/α, αυτό είναι ευδιάκριτο και αναφέρεται ρητώς εντός του κειμένου της εργασίας ως προϊόν εργασίας τρίτου, σημειώνοντας με παρομοίως σαφή τρόπο τα στοιχεία ταυτοποίησής του, ενώ παράλληλα βεβαιώνει πως στην περίπτωση χρήσης αυτούσιων γραφικών αναπαραστάσεων, εικόνων, γραφημάτων κλπ., έχει λάβει τη χωρίς περιορισμούς άδεια του κατόχου των πνευματικών δικαιωμάτων για την συμπερίληψη και επακόλουθη δημοσίευση του υλικού αυτού.

ΠΙΣΤΟΠΟΙΗΣΗ

Πιστοποιείται ότι η διπλωματική εργασία με θέμα:

«Χρήση εξελικτικών νευρωνικών δικτύων και επεξεργασίας φυσικής γλώσσας για την πρόβλεψη ισοτιμιών συναλλάγματος»

Του φοιτητή του τμήματος Ηλεκτρολόγων Μηχανικών και Τεχνολογίας Υπολογιστών, της Πολυτεχνικής Σχολής του Πανεπιστημίου Πατρών: **ΟΝΟΜΑΤΕΠΩΝΥΜΟ:...Γεώργιος Καφτάνης.....** με Αριθμό Μητρώου ...1046981..... παρουσιάστηκε δημοσίως στο Τμήμα Ηλεκτρολόγων Μηχανικών και Τεχνολογίας Υπολογιστών του Πανεπιστημίου Πατρών στις ...01/07/2021..... και εξετάστηκε και εγκρίθηκε από την ακόλουθη Εξεταστική Επιτροπή:

	ΥΠΟΓΡΑΦΕΣ
Κυριάκος Σγάρμπας..... (Ονοματεπώνυμο) Αναπληρωτής...Καθηγητής.....(Βαθμίδα) Πανεπιστημίου Πατρών Επιβλέπων/Επιβλέπουσα	
Ευάγγελος Δερματάς..... (Ονοματεπώνυμο) Αναπληρωτής...Καθηγητής.....(Βαθμίδα) Πανεπιστημίου Πατρών Μέλος τριμελούς εξεταστικής επιτροπής	
Κωνσταντίνος Μουστάκας... (Ονοματεπώνυμο) Καθηγητής.....(Βαθμίδα) Πανεπιστημίου Πατρών Μέλος τριμελούς εξεταστικής επιτροπής	

Πάτρα,01/07/2021.....

ΤΙΤΛΟΣ:

Χρήση εξελικτικών νευρωνικών δικτύων και επεξεργασίας φυσικής γλώσσας για την πρόβλεψη ισοτιμιών συναλλάγματος

ΣΥΓΓΡΑΦΕΑΣ:

ΓΕΩΡΓΙΟΣ ΚΑΦΤΑΝΗΣ
(Α.Μ. 1046981)



ΠΕΡΙΛΗΨΗ:

Στην παρούσα διπλωματική εργασία επιχειρείται η αξιοποίηση γενετικών αλγορίθμων και νευρωνικών δικτύων με σκοπό την επίτευξη προβλέψεων στον χώρο της οικονομίας. Συγκεκριμένα, η μεθοδολογία που προέκυψε, ενσωματώθηκε σε εφαρμογή που δραστηριοποιείται σε πρόβλεψη της ισοτιμίας EUR/USD και EUR/BTC. Η εφαρμογή δημιουργήθηκε με τρόπο τέτοιο ώστε να διασφαλίζεται η συνεχή ροή πραγματικών δεδομένων από API διεθνών ανταλλακτηρίων. Παράλληλα, ενσωματώθηκε ένας μηχανισμός παροχής πραγματικών αναρτήσεων από μέσο κοινωνικής δικτύωσης, αλλά και τίτλων ειδήσεων. Μέσω ανάλυση συναισθήματος (sentimental analysis) στα δεδομένα αυτά, αποδεικνύεται η συσχέτιση τους με τις ισοτιμίες. Με τον τρόπο αυτό παρέχεται στον χρήστη ένα επιπλέον εργαλείο πρόβλεψης.

ABSTRACT:

In the present diploma dissertation, the utilization of genetic algorithms and neural networks is utilized in order to achieve predictions in the field of economics. Specifically, the resulting methodology was incorporated into an application that is active in forecasting the EUR / USD and EUR / BTC exchange rate. The application was created in such a way as to ensure the continuous flow of real data from APIs of international exchanges. At the same time, a mechanism was provided for the provision of real posts through social media, as well as news headlines. Through sentimental analysis in these data, their correlation with the exchange rates is proved. This provides the user with an additional forecasting tool.

ΠΡΟΛΟΓΟΣ

Η επιθυμία για πρόβλεψη των μελλοντικών τιμών των ισοτιμιών των συναλλαγμάτων αποτελεί διαχρονικό πρόβλημα. Οι λόγοι προσανατολίζονται σε επενδυτικούς σκοπούς, καθώς και στην ανάγκη για προστασία των περιουσιακών κεφαλαίων μέσω αποφυγής του πληθωρισμού ή της υποτίμησης ενός μέσου συναλλαγής. Οι δυσκολίες της επίτευξης προβλέψεων με μεγάλη ακρίβεια είναι εξαιρετικά μεγάλες, κυρίως λόγω των πολλών παραγόντων που την επηρεάζουν.

Στην παρούσα διπλωματική εργασία δίνεται έμφαση στην βελτιστοποίηση νευρωνικών δικτύων με την χρήση γενετικών αλγορίθμων. Παράλληλα εφαρμόζεται επεξεργασία φυσικής γλώσσας σε αναρτήσεις και τίτλους ειδήσεων με σκοπό την εκτίμηση του «κλίματος» της αγοράς, καθώς και στην συσχέτιση του με την ισοτιμία του αντίστοιχου μέσου συναλλάγματος. Η χρήση αυτών των μεθόδων ενσαρκώνεται σε μια εφαρμογή, η οποία περιλαμβάνει διάφορες λειτουργίες:

1. Ενσωμάτωση (integration) με API ανταλλακτηρίου κρυπτονομισμάτων με σκοπό την συνεχή παροχή πραγματικών δεδομένων.
2. Ενσωμάτωση (integration) με API ανταλλακτηρίου παραδοσιακών νομισμάτων και συνεχή παροχή δεδομένων.
3. Ενσωμάτωση (integration) με API μέσου κοινωνικής δικτύωσης καθώς και με πάροχο τίτλων ειδήσεων διεθνούς εμβέλειας.
4. Σχεδιασμός γενετικού αλγορίθμου για βελτιστοποίηση παραμέτρων και αρχιτεκτονικής LSTM δικτύων.
5. Αξιολόγηση μοντέλων LSTM στην πρόβλεψη ισοτιμιών συναλλαγμάτων
6. Δυνατότητα για πρόβλεψη μελλοντικής τιμής με ήδη εκπαιδευμένο δίκτυο
7. Δυνατότητα για εξαγωγή συμπεράσματος του «κλίματος» της αγοράς μέσω ανάλυση συναισθημάτων (sentimental analysis)

Τα νομίσματα που επιλέχθηκαν για τις μετρήσεις είναι το αμερικάνικο δολάριο (USD) και το κρυπτονόμισμα Bitcoin (BTC). Ως νόμισμα αναφοράς επιλέχθηκε το ευρώ (EUR).

Περιεχόμενα

ΠΕΡΙΛΗΨΗ	5
ABSTACT	5
ΠΡΟΛΟΓΟΣ	6
ΚΕΦΑΛΑΙΟ 1	9
ΕΙΣΑΓΩΓΗ	9
1.1. Φυσικά νομίσματα	9
1.2. Κρυπτονομίσματα	10
1.3. Σημασία πρόβλεψης ισοτιμιών συναλλάγματος	11
ΚΕΦΑΛΑΙΟ 2	13
ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ	13
2.1 ΕΙΣΑΓΩΓΗ	13
2.2 Λειτουργία Τεχνητού Νευρώνα	13
2.3 Συναρτήσεις ενεργοποίησης	14
2.4 Αρχιτεκτονικές νευρωνικών δικτύων	14
2.5 Εκπαίδευση Νευρωνικών δικτύων	15
2.6 Τεχνικές βελτιστοποίησης νευρωνικών δικτύων	17
2.6.1 Τεχνική Drop out	17
2.6.2 Τεχνική Early Stopping	18
2.7 Χρήση Νευρωνικών Δικτύων για την πρόβλεψη χρονοσειρών	18
2.7.1 Η Long short-term memory αρχιτεκτονική νευρωνικών δικτύων	18
2.7.2 Αμφίδρομα Long short-term memory δίκτυα	20
2.7.3 Προετοιμασία δεδομένων για την Long short-term memory αρχιτεκτονική	21
2.7.4 Αξιολόγηση μοντέλων Long short-term memory αρχιτεκτονικής	22
ΚΕΦΑΛΑΙΟ 3	25
ΕΦΑΡΜΟΓΕΣ ΓΕΝΕΤΙΚΩΝ ΑΓΛΟΡΙΘΜΩΝ	25
3.1 Εισαγωγή	25
3.2 Συνδυασμός νευρωνικών δικτύων και γενετικών αλγορίθμων	27
3.2.1 Χρήση γενετικών αλγορίθμων για την εκπαίδευση νευρωνικών δικτύων	27
3.2.2 Χρήση γενετικών αλγορίθμων για την επιλογή αρχιτεκτονικής νευρωνικού δικτύου	28
3.2.3 Παράλληλη εκπαίδευση και επιλογή βέλτιστης αρχιτεκτονικής νευρωνικού δικτύου	28
3.2.4 Βελτιστοποίηση LSTM με την χρήση γενετικών αλγορίθμων	29

ΚΕΦΑΛΑΙΟ 4	31
ΕΠΕΞΕΡΓΑΣΙΑ ΦΥΣΙΚΗΣ ΓΛΩΣΣΑΣ	31
4.1 ΕΙΣΑΓΩΓΗ	31
4.2 Προεπεξεργασία δεδομένων	31
4.3 Αλγόριθμοι επεξεργασίας φυσικής γλώσσας	32
4.4 Εξόρυξη γνώμης (Sentiment analysis)	33
4.5 Εξόρυξη γνώμης για την πρόβλεψη ισοτιμιών συναλλάγματος	35
ΚΕΦΑΛΑΙΟ 5	37
Η ΕΦΑΡΜΟΓΗ	37
5.1 Αρχιτεκτονική	37
5.2 Σύνδεση με την βάση δεδομένων	38
5.3 Συλλογή Δεδομένων	39
5.3.1 Συλλογή δεδομένων ισοτιμιών συναλλάγματος	40
5.3.2 Συλλογή τίτλων ειδήσεων και tweets	42
5.4 Διαχείριση Δικτύων	44
5.4.1 Σχεδιασμός LSTM δικτύου και διεπαφής με τον χρήστη	44
5.4.2 Βελτιστοποίηση LSTM δικτύου με χρήση γενετικών αλγορίθμων	48
5.4.3 Ανάλυση συναισθήματος αγοράς	49
5.5 Ενδεχόμενες μελλοντικές βελτιώσεις	50
ΚΕΦΑΛΑΙΟ 6	51
ΓΕΝΙΚΕΣ ΜΕΤΡΗΣΕΙΣ	51
6.1 Πρόβλεψη ισοτιμίας συναλλαγμάτων με χρήση απλών LSTM δικτύων	51
6.1.1 Ανάλυση προβλέψεων για την ισοτιμία EUR/USD	51
6.1.2 Ανάλυση προβλέψεων για την ισοτιμία EUR/BTC	54
6.2 Πρόβλεψη ισοτιμίας συναλλαγμάτων με χρήση εξελικτικών LSTM δικτύων (GA-LSTM)	57
6.2.1 Ανάλυση προβλέψεων για την ισοτιμία EUR/USD	57
6.2.2 Ανάλυση προβλέψεων για την ισοτιμία EUR/BTC	58
6.3 Πρόβλεψη τάσης ισοτιμίας συναλλαγμάτων με χρήση επεξεργασίας φυσικής γλώσσας	59
ΒΙΒΛΙΟΓΡΑΦΙΑ	61

ΚΕΦΑΛΑΙΟ 1

ΕΙΣΑΓΩΓΗ

1.1. Φυσικά νομίσματα

Η χρήση νομισμάτων επιλέχθηκε για να δώσει λύσεις στα προβλήματα που δημιουργούσε ο αντιπραγματισμός (η οικονομία εκείνη εντός της οποίας λαμβάνει χώρα η ανταλλαγή αγαθών μεταξύ των οικονομικών μονάδων χωρίς τη διαμεσολάβηση του χρήματος). Αρχικά, η αγοραστική αξία ενός νομίσματος εξαρτιόταν από την ιστορική του αξία και/ή από την εγγενή αξία του μετάλλου του, (για παράδειγμα χρυσός ή άργυρος).

Ωστόσο, στη σύγχρονη εποχή, τα περισσότερα νομίσματα είναι κατασκευασμένα από βασικά μέταλλα και η αξία τους συναρτάται από το status τους ως πιστωτικό χρήμα (fiat money). Στις σύγχρονες κοινωνίες το νόμισμα μαζί με τα χαρτονομίσματα αποτελεί την πιο συνηθισμένη μορφή χρήματος . Αυτό σημαίνει ότι η αξία του νομίσματος μεταβάλλεται ανάλογα με την ασκούμενη νομισματική πολιτική και συνεπώς λειτουργεί περισσότερο ως συμβολικό νόμισμα με την ακριβή έννοια της λέξης .

Επιπλέον, όλες οι οικονομίες του κόσμου έχουν εμπορικές, πολιτικές, στρατιωτικές, διπλωματικές σχέσεις με χώρες του εξωτερικού. Αυτό συνεπάγεται ένα μεγάλο πλήθος εμπορικών και οικονομικών συναλλαγών μεταξύ των χωρών και δημιουργεί την ανάγκη της συγκρισιμότητας και μετατρεψιμότητας των διαφόρων νομισμάτων ώστε να είναι δυνατές οι ανταλλαγές εμπορεύσιμων αγαθών.

Η τιμή του συναλλάγματος ορίζεται όπως η τιμή προϊόντων, δηλαδή πόσες μονάδες του εγχωρίου νομίσματος απαιτούνται για την αγορά μιας μονάδας του ξένου νομίσματος. Σε καθεστώς ελεύθερης ισοτιμίας η τιμή του συναλλάγματος μπορεί να μεταβάλλεται καθημερινά, έστω και ελάχιστα και τότε η υποτίμηση ή η ανατίμηση είναι μία αργή διαδικασία. Στις χώρες που έχουν σταθερή ισοτιμία, όταν τα ελλείμματα δεν μπορούν πλέον να καλυφθούν από τα κρατικά συναλλαγματικά διαθέσιμα, η υποτίμηση γίνεται ξαφνικά. Το αρμόδιο όργανο ανακοινώνει ότι η τιμή του συναλλάγματος αυξήθηκε και συνεπώς το νόμισμα της χώρας υποτιμήθηκε. Σκοπός της υποτίμησης είναι να αυξηθούν οι εξαγωγές της χώρας διότι τα προϊόντα της θα γίνουν φθηνότερα και να μειωθούν οι εισαγωγές διότι τα ξένα προϊόντα θα γίνουν ακριβότερα.

1.2. Κρυπτονομίσματα

Το Κρυπτονόμισμα είναι μία peer-to-peer αποκεντρωμένη ηλεκτρονική μορφή χρήματος η οποία βασίζεται πάνω στις αρχές της κρυπτογραφίας για την διασφάλιση του δικτύου και την επαλήθευση των συναλλαγών. Δεν υπάρχει σε φυσική μορφή (όπως χαρτονομίσματα) και συνήθως δεν εκδίδεται από κεντρική αρχή (τράπεζα). Το δίκτυο είναι αποκεντρωμένο και διανεμημένο ισόποσα. Αυτό σημαίνει ότι δεν υπάρχει κάποιο πρόσωπο του δικτύου που να υπερέχει έναντι κάποιου άλλου προσώπου κατ' οποιονδήποτε τρόπο, οπότε υπάρχει απουσία προτεραιότητας (όποιου είδους). Κάθε κρυπτογράφηση λειτουργεί μέσω του blockchain, που χρησιμεύει ως βάση δεδομένων δημόσιας χρηματοοικονομικής συναλλαγής.

Το blockchain είναι μια νέα τεχνολογία η οποία παρουσιάζεται ως μία δημόσια, μη δυνάμενη να τροποποιηθεί ως προς το ιστορικό της, σειρά δεδομένων ομαδοποιημένων σε χρονικά αριθμημένα blocks. Κάθε αλυσίδα αποτελείται από πολλά μπλοκ και κάθε μπλοκ έχει τα παρακάτω βασικά στοιχεία:

1. Το κρυπτογραφικό αποτύπωμα του προηγούμενου στην αλυσίδα, μπλοκ.
2. Το κρυπτογραφικό αποτύπωμα του εν λόγω μπλοκ.
3. Το timestamp (χρονική του σφραγίδα – σε unix time)
4. Ένα μοναδικά δημιουργημένο αριθμητικό λέξημα, την απόδειξη μόχθου του μπλοκ (cryptographic nonce).
5. Το κρυπτογραφικό αποτύπωμα που προκύπτει από τη λειτουργία μιας συνάρτησης κατακερματισμού πάνω στο σύνολο των εισαχθέντων και ξεχωριστών συναλλαγών (transactions) στο εν λόγω μπλοκ, συμπεριλαμβανομένων και αυτών των συναλλαγών.

Το πορτοφόλι κρυπτογράφησης είναι ένα πρόγραμμα, μια υπηρεσία ή μια συσκευή που αποθηκεύει τα δημόσια και / ή ιδιωτικά κλειδιά για συναλλαγές κρυπτογράφησης. Ένα δημόσιο κλειδί επιτρέπει σε άλλους να πραγματοποιούν πληρωμές στη διεύθυνση που προέρχεται από αυτό, ενώ ένα ιδιωτικό κλειδί επιτρέπει τη δαπάνη κρυπτογράφησης από αυτήν τη διεύθυνση.

Το δημοφιλέστερο κρυπτονόμισμα είναι το Bitcoin. Εφευρέθηκε το 2008 από έναν άνθρωπο (ή από μία ομάδα) χρησιμοποιώντας το όνομα Satoshi Nakamoto. Είναι ανοιχτού κώδικα, ο σχεδιασμός του είναι δημόσιος, κανείς δεν είναι ιδιοκτήτης ούτε το ελέγχει. Η απόκτηση Bitcoins μπορεί να γίνει με τους παρακάτω τρόπους.

1. Αγοράζοντας μέσω πραγματικών χρημάτων (σε Bitcoin ανταλλακτήρια).
2. Δεχόμενος πληρωμές σε Bitcoins (πχ. πουλώντας προϊόντα ή υπηρεσίες).
3. Μπορείς να δημιουργήσεις νέα μέσω ενός υπολογιστή (mining). Οι συναλλαγές για να επιβεβαιωθούν χρειάζονται την συνεισφορά (μέσω υπολογιστικής δύναμης) τυχαίων υπολογιστών απ' όλα τον κόσμο. Νέα bitcoins εκδίδονται κάθε δέκα λεπτά τα οποία δίνονται ως ανταμοιβή σε αυτούς που συνεισφέρουν στην επιβεβαίωση των συναλλαγών.

Η τιμή ενός bitcoin καθορίζεται από την προσφορά και τη ζήτηση. Όταν η ζήτηση για bitcoins αυξάνεται, η τιμή αυξάνεται και όταν η ζήτηση πέφτει, η τιμή πέφτει. Ο μέγιστος αριθμός bitcoins είναι τα 21 εκατομμύρια. Υπάρχει μόνο ένας περιορισμένος αριθμός bitcoins στην κυκλοφορία και τα νέα bitcoins δημιουργούνται με μειωμένο και προβλέψιμο ρυθμό πράγμα που σημαίνει ότι η ζήτηση πρέπει να ακολουθεί το επίπεδο του πληθωρισμού για να κρατήσει σταθερή την τιμή. Το Bitcoin δεν είναι σχεδιασμένο να είναι ένα αποπληθωριστικό νόμισμα. Είναι πιο ακριβές να ειπωθεί ότι το Bitcoin προορίζεται να μεγεθυνθεί τα πρώτα χρόνια του (λόγω της αυξανόμενης ζήτησης) και να σταθεροποιηθεί τα επόμενα χρόνια. Η μόνη στιγμή που η ποσότητα των bitcoin σε κυκλοφορία θα μειωθεί είναι αν οι άνθρωποι απρόσεκτα χάσουν τα πορτοφόλια τους αποτυγχάνοντας να δημιουργήσουν και να ανακτήσουν αντίγραφα ασφαλείας.

Στις μέρες μας, παρατηρείται όλο και αυξανόμενη προθυμία των ανθρώπων να κάνουν χρήση των κρυπτονομισμάτων. Αυτό οδηγεί σε αυξανόμενη κεφαλαιοποίηση τους όπως φαίνεται και στο σχήμα 1.1. Το γεγονός αυτό καθιστά την μελέτη τους ιδιαίτερα σημαντική.



Σχήμα 1.1 Κεφαλαιοποίηση της αγοράς κρυπτονομισμάτων. [46]

1.3. Σημασία πρόβλεψης ισοτιμιών συναλλάγματος

Η πρόβλεψη της ισοτιμίας συναλλάγματος προσφέρει μεγάλες δυνατότητες κέρδους και αποτελεί βασικό κίνητρο για τους περισσότερους ερευνητές σε αυτόν τον τομέα. Η ακριβέστερη πρόβλεψη της μεταβολής των τιμών θα οδηγήσει σε μεγιστοποίηση των κερδών των επενδυτών. Παράλληλα, μπορεί να προστατεύσει άλλους από την απώλεια χρηματικών ποσών. Επιπλέον, τα μοντέλα μακροπρόθεσμης πρόβλεψης ισοτιμιών μπορούν να αποτελέσουν σημαντικά εργαλεία για την χάραξη κατάλληλων νομισματικών πολιτικών των νομισματικών ενώσεων ή χωρών (π.χ. αποφυγή πληθωρισμού).

Είναι ένα από τα πιο απαιτητικά ζητήματα λόγω των πολλών παραγόντων που εμπλέκονται όπως τα επιτόκια, η πολιτική και η οικονομία καθώς και η γενικότερη συμπεριφορά των επενδυτών. Όσον αφορά τα νομίσματα καθοριστικό ρόλο έχει η ύπαρξη πληθωρισμού (αύξηση της προσφοράς χρήματος) ή αποπληθωρισμού (μείωση της προσφοράς χρήματος). Όσον αφορά τα κρυπτονομίσματα οι δυσκολίες εντείνονται λόγω

του μικρού χρόνου ύπαρξης αυτών, καθώς και της σχετικά μικρής ακόμα κεφαλαιοποίησης της αγοράς (market cap) που οδηγεί σε απρόβλεπτες και μεγάλες μεταβολές . Το γεγονός αυτό όμως δημιουργεί μεγαλύτερα περιθώρια κέρδους, αλλά σαφώς δεν πρέπει να αγνοούνται και οι κίνδυνοι. Παράλληλα ορισμένα κρυπτονομίσματα (π.χ. το ethereum) έχουν πληθώρα χρήσεων και λειτουργιών τα οποία μπορούν να προσεγγίσουν επιπλέον χρήστες, από τους κλασσικούς επενδυτές. Όλα αυτά συντελούν στο συμπέρασμα ότι η πρόβλεψη των ισοτιμιών αποτελεί ένα αρκετά δύσκολο , αλλά και πολύ σημαντικό εγχείρημα , το οποίο απασχολεί αρκετά τους ερευνητές.

ΚΕΦΑΛΑΙΟ 2

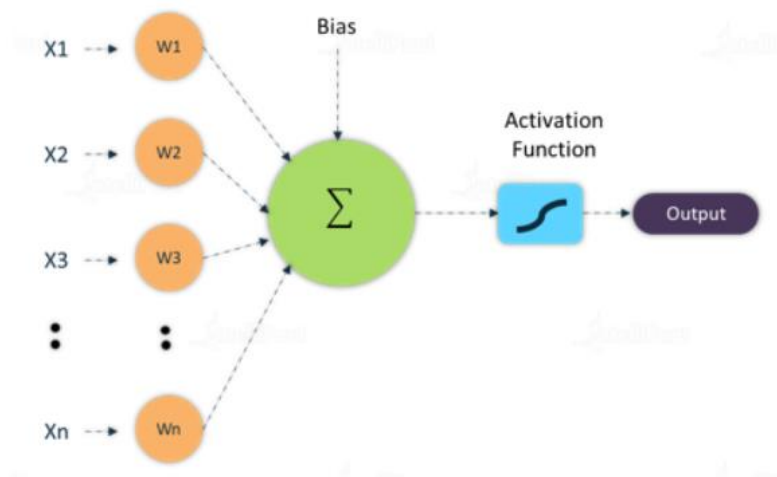
ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ

2.1 ΕΙΣΑΓΩΓΗ

Τα Νευρωνικά δίκτυα αποτελούν μια προσπάθεια προσέγγισης της λειτουργίας του ανθρώπινου εγκεφάλου από μια μηχανή. Αποτελούν ένα κύκλωμα διασυνδεδεμένων νευρώνων, προσομοιώνοντας έτσι την λειτουργία των βιολογικών νευρώνων με ένα μαθηματικό μοντέλο. Κάθε νευρώνας έχει πολλές εισόδους αλλά μόνο μια έξοδο, η οποία μπορεί να αποτελέσει είσοδο για το επόμενο επίπεδο νευρώνων. Οι συνάψεις μεταξύ των νευρώνων χαρακτηρίζονται από ένα συντελεστή βάρους, ο οποίος και προσδιορίζει και την σημαντικότητα του. Η έξοδος κάθε νευρώνα σε σχέση με τις εισόδους και τους συντελεστές βάρους καθορίζεται από την συνάρτηση μεταφοράς. Η εκπαίδευση του δικτύου γίνεται μέσω ενός αλγορίθμου εκπαίδευσης. Σκοπός αυτής της διαδικασίας είναι η μεταβολή των βαρών των συνάψεων ώστε να επιτυγχάνεται η επιθυμητή έξοδος. Θα πρέπει να αναφέρουμε ότι σε ένα νευρωνικό δίκτυο δεν είναι απαραίτητο ο αριθμός των νευρώνων που αποτελούν το κάθε επίπεδο να είναι ίδιος.

2.2 Λειτουργία Τεχνητού Νευρώνα

Ο τεχνητός νευρώνας αποτελεί δομικό στοιχείο του δικτύου. Κάθε κόμβος δέχεται ένα σύνολο αριθμητικών εισόδων μέσω των συνάψεων και παράγει μία έξοδο με βάση την συνάρτηση μεταφοράς της. Υπάρχουν τρεις τύποι νευρώνων: οι νευρώνες εισόδου, οι νευρώνες εξόδου και οι κρυμμένοι νευρώνες. Οι νευρώνες εισόδου μεσολαβούν απλώς ανάμεσα στις εισόδους του δικτύου και στους κρυφούς νευρώνες, δίχως εκτελώντας υπολογισμούς. Οι εξωτερικοί νευρώνες μας δίνουν την τελική έξοδο του δικτύου. Οι κρυφοί νευρώνες πολλαπλασιάζουν κάθε είσοδό της με το βάρος της αντίστοιχης σύναψης και υπολογίζουν το ολικό άθροισμα των γινομένων. Το άθροισμα αυτό τροφοδοτείται στη συνάρτηση ενεργοποίησης, η οποία δίνει και την τελική έξοδο του νευρώνα. Στο άθροισμα υπάρχει η δυνατότητα να προσθέτει μια σταθερά (bias), η οποία μετατοπίζει την συνάρτηση ενεργοποίησης.



Σχήμα 2.1 Δομή Τεχνητού Νευρώνα

2.3 Συναρτήσεις ενεργοποίησης

Ως συνάρτηση μεταφοράς του νευρώνα μπορούμε να χρησιμοποιήσουμε ένα πλήθος συναρτήσεων. Η επιλογή γίνεται με βάση την εφαρμογή, με στόχο την μέγιστη δυνατή αποτελεσματικότητα του νευρωνικού δικτύου. Δεν είναι απαραίτητο όλοι οι νευρώνες να υλοποιούν την ίδια συνάρτηση μεταφοράς. 3 Βασικοί τύποι συναρτήσεων είναι η Βηματική, η τμηματικά γραμμική (RELU) και η σιγμοειδής.

$$\text{Βηματική Συνάρτηση } Y(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geq 0 \end{cases}$$

$$\text{RELU } Y(x) = \begin{cases} 0, & x < 0 \\ x, & x \geq 0 \end{cases}$$

$$\text{Σιγμοειδής } y(x) = \frac{1}{1 + e^{-x}}$$

2.4 Αρχιτεκτονικές νευρωνικών δικτύων

α) Τεχνητά νευρωνικά δίκτυα πρόσθιας τροφοδότησης (Feed-Forward Neural Networks)

Αυτός είναι ο απλούστερος και συνηθέστερος τύπος νευρικού δικτύου σε πρακτικές εφαρμογές. Το πρώτο επίπεδο είναι η είσοδος και το τελευταίο επίπεδο είναι η έξοδος. Δεν υπάρχουν βρόχοι ανατροφοδότησης και το σήμα μεταφέρεται μόνο προς μία κατεύθυνση.

β) Τεχνητά νευρωνικά δίκτυα με ανατροφοδότηση (Recurrent Networks)

Στα δίκτυα με ανατροφοδότηση είναι εφικτό να έχουμε βρόχους καθώς το σήμα μπορεί να έχει 2 κατευθύνσεις. Είναι δυναμικά περίπλοκα και σαφώς δυσκολότερα να εκπαιδευτούν, αλλά προσεγγίζουν περισσότερο τα βιολογικά δίκτυα. Έχουν την ικανότητα να «συγκρατούν» πληροφορίες για μεγάλο διάστημα.

γ) Συμμετρικά συνδεδεμένα δίκτυα (Symmetrically Connected Networks)

Τα δίκτυα αυτά διαφέρουν από τα δίκτυα με ανατροφοδότηση στο γεγονός ότι οι συνδέσεις ανάμεσα στις 2 κατευθύνσεις είναι συμμετρικές (έχουν ίδιο βάρος). Είναι ευκολότερο να αναλυθούν αλλά έχουν περιορισμούς στην αποτελεσματικότητα και στις εφαρμογές.

2.5 Εκπαίδευση Νευρωνικών δικτύων

Τα Τεχνητά Νευρωνικά Δίκτυα έχουν την ικανότητα να μαθαίνουν από το περιβάλλον τους και να βελτιώνουν την απόδοσή τους μέσω της διαδικασίας της μάθησης, διαδικασία αντιστοιχεί με την μάθηση ανθρώπινη μάθηση. Η διαδικασία της μάθησης επιτυγχάνεται μέσω μιας επαναληπτικής διαδικασίας σταδιακής προσαρμογής των παραμέτρων του δικτύου σε τιμές κατάλληλες (εκπαίδευση) για να επιλυθεί το εκάστοτε πρόβλημα. Οι κυριότερες κατηγορίες που διακρίνονται τα τεχνητά νευρωνικά δίκτυα με βάση την διαδικασία εκπαίδευσης είναι οι παρακάτω:

α) Εκπαίδευση με επίβλεψη (*supervised training*)

Σ' αυτόν τον τρόπο εκπαίδευσης η εκπαίδευση βασίζεται στην εκ των προτέρων γνώση των σωστών εξόδων, που πρέπει να παράγει το δίκτυο για την συγκεκριμένη είσοδο. Επομένως, προτού ξεκινήσει η διαδικασία είναι απαραίτητο να υπάρχει διαθέσιμο ένα σύνολο γνωστών ζευγαριών εισόδων-εξόδων. Για τον λόγο αυτό συχνά διαφοροποιούμε τα δεδομένα σε δεδομένα εκπαίδευσης και δεδομένα δοκιμής (ενδεικτικά 70% δεδομένα εκπαίδευσης – 30% δεδομένα δοκιμής). Στόχος είναι να φτάσει το ΤΝΔ σε θέση να παράγει την βέλτιστη έξοδο-απόκριση προσαρμόζοντας τις παραμέτρους του (βάρη συνάψεων) ελαχιστοποιώντας το σφάλμα της εξόδου (επιθυμητική έξοδος – πραγματική έξοδος). Ως μέσο απόδοσης του δικτύου χρησιμοποιείται συχνά το μέσο τετραγωνικό σφάλμα ή το άθροισμα των μέσων τετραγωνικών σφαλμάτων πάνω στα δείγματα εκπαίδευσης. Υπάρχουν διάφορες τεχνικές εκπαίδευσης με επίβλεψη, η πιο διαδεδομένη όμως είναι η back-propagation.

Η τεχνική backpropagation υπολογίζει αποτελεσματικά την κλίση της συνάρτησης σφάλματος σε σχέση με τα βάρη του δικτύου για ένα μόνο παράδειγμα εισόδου-εξόδου. Υπολογίζει την κλίση συνάρτησης σφάλματος για ένα επίπεδο την φορά, ξεκινώντας από το ανώτερο επίπεδο (layer) και καταλήγοντας στο επίπεδο εισόδου. Η ενημέρωση των βαρών στον αλγόριθμο αυτό γίνεται σύμφωνα με τους παρακάτω τύπους.

$$w_j^r(new) = w_j^r(old) + \Delta w_j^r \quad (2.1)$$

$$\Delta w_j^r = -\mu \sum_{i=1}^N \delta_j^r(i) y^{r-1}(i) \quad (2.2)$$

$$\delta_j^L(i) = e_j(i) f'(u_j^L(i)) \quad (2.3)$$

$$\delta_j^{r-1} = e_j^{r-1}(i) f'(u_j^{r-1}(i)) \quad (2.4)$$

$$e_j^{r-1} = \sum_{k=1}^{k_r} \delta_k^r(i) w_{kj}^r \quad (2.5)$$

Η σταθερά μ στην εξίσωση (2.2) ονομάζεται σταθερά μάθησης. Η τιμή αυτή πρέπει να είναι αρκετά μικρή έτσι ώστε να εγγυάται η σύγκλιση, αλλά όχι πάρα πολύ μικρή για να επιτυγχάνεται γρήγορα.

Οι υπολογισμοί ξεκινούν από $r=L$ (επίπεδο εξόδου) και διαδίδονται προς τα πίσω για $r=L-1, r=L-2, \dots, 1$. Η εξίσωση (2.3) αφορά το επίπεδο εξόδου, ενώ η (2.4) τα υπόλοιπα επίπεδα ($r < L$). Στην (2.3) η παράγωγος f' υπολογίζεται εύκολα καθώς η εξάρτηση του σφάλματος από την τιμή u που δέχεται η συνάρτηση ενεργοποίησης είναι σαφής [43].

β) Εκπαίδευση χωρίς επίβλεψη (unsupervised training)

Στην εκπαίδευση χωρίς επίβλεψη, το δίκτυο διαθέτει τις εισόδους αλλά χωρίς τα επιθυμητά αποτελέσματα. Το ίδιο το σύστημα πρέπει στη συνέχεια να αποφασίσει ποιες δυνατότητες θα χρησιμοποιήσει για την ομαδοποίηση των δεδομένων εισόδων. Οι ελεύθερες παράμετροι του συστήματος πρέπει με την πάροδο της διαδικασίας εκπαίδευσης να βελτιστοποιούνται ως προς το μέτρο αυτό. Κάποια παραδείγματα ΤΝΔ που χρησιμοποιούν την εκπαίδευση χωρίς επίβλεψη είναι ο αυτοοργανωμένος χάρτης χαρακτηριστικών (Self-Organized feature map, SOM) και η ART (Adaptive Resonance Theory).

γ) Εκπαίδευση με ενίσχυση (reinforced training)

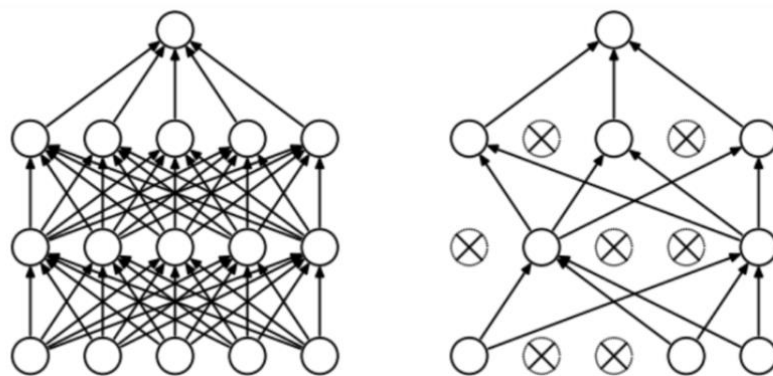
Η εκμάθηση ενίσχυσης διαφέρει από την εποπτευόμενη μάθηση στο ότι δεν χρειάζεται να παρουσιάζονται ζεύγη εισόδου / εξόδου με ετικέτα και δεν χρειάζονται να

διορθωθούν ρητά οι βέλτιστες ενέργειες. Ο υπολογιστής χρησιμοποιεί δοκιμή και σφάλμα για να βρει μια λύση στο πρόβλημα. Το δίκτυο λαμβάνει ανταμοιβές ή κυρώσεις για τις ενέργειες που εκτελεί. Στόχος του είναι να μεγιστοποιήσει τη συνολική ανταμοιβή. Παρόλο που ο προγραμματιστής καθορίζει την πολιτική ανταμοιβής, δεν δίνει σαφείς οδηγίες για την επίλυση του προβλήματος. Έτσι, το μοντέλο, αναλαμβάνει να εκτελέσει την εργασία αυτή, ξεκινώντας από εντελώς τυχαίες δοκιμές και τελειώνοντας με εξελιγμένες τακτικές και υπεράνθρωπες δεξιότητες. Η τεχνική αυτή απαιτεί σημαντική υπολογιστή ισχύ. Γνωστές τεχνικές εκπαίδευσης με ενίσχυση είναι η Monte Carlo method και η Q-learning.

2.6 Τεχνικές βελτιστοποίησης νευρωνικών δικτύων

2.6.1 Τεχνική Drop out

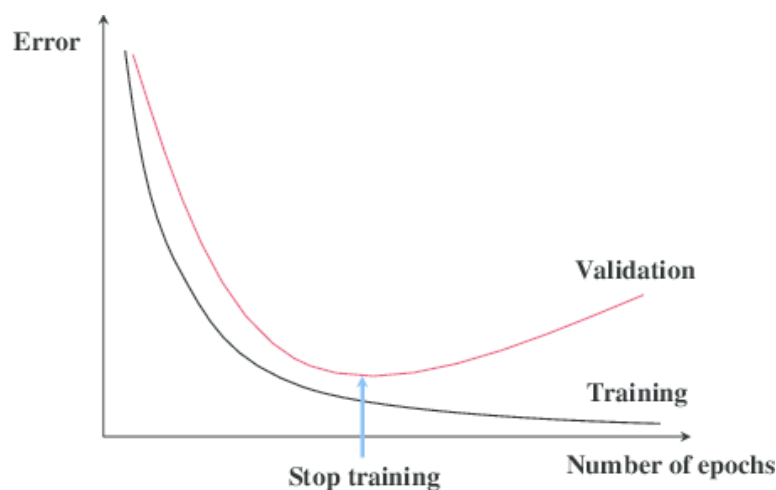
Ένα σημαντικό πρόβλημα στην εκπαίδευση νευρωνικών δικτύων αποτελεί το overfitting (η υπερβολική προσαρμογή του δικτύου στα δεδομένα εκπαίδευσης). Μια σημαντική λύση στο πρόβλημα είναι η τεχνική του Drop out, η οποία προτάθηκε από τον Srivastava το 2014 [17]. Η υλοποίηση του γίνεται με την τυχαία απενεργοποίηση μερικών νευρώνων κατά τη διάρκεια κάθε επανάληψης της εκπαίδευσης. Αυτό σημαίνει ότι τα βάρη των συνδέσεων των νευρώνων αυτών παραμένουν αμετάβλητα. Το Dropout αναγκάζει ένα νευρωνικό δίκτυο να μάθει περισσότερα ισχυρά χαρακτηριστικά που είναι χρήσιμα αξιοποιώντας πολλά διαφορετικά τυχαία υποσύνολα νευρώνων. Η τεχνική αυτή συνήθως αυξάνει την αποτελεσματικότητα του δικτύου σε άγνωστα προς αυτό δεδομένα, ενώ παράλληλα επιταχύνει την διαδικασία εκπαίδευσης καθώς η πολυπλοκότητα του δικτύου μειώνεται. Λύνοντας το πρόβλημα του overfitting σε μεγάλα δίκτυα έγιναν δυνατές μεγαλύτερες και ακριβέστερες αρχιτεκτονικές Deep Learning.



Σχήμα 2.2 Κανονικό δίκτυο (αριστερά) – Δίκτυο μετά από εφαρμογή Drop out (δεξιά)

2.6.2 Τεχνική Early Stopping

Epoch είναι ένας όρος που χρησιμοποιείται στη μηχανική μάθηση και υποδεικνύει τον αριθμό περασμάτων ολόκληρου του συνόλου δεδομένων εκπαίδευσης που έχει ολοκληρώσει ο αλγόριθμος μηχανικής μάθησης. Η επιλογή του κατάλληλου αριθμού επαναλήψεων (epochs) παίζει σημαντικό ρόλο και για την εκπαίδευση του δικτύου. Υπερβολικά μεγάλος αριθμός epochs μπορεί να οδηγήσει σε overfitting, ενώ πολύ μικρός αριθμός σε underfitting. Η τεχνική του Early stopping επιτρέπει την επιλογή μεγάλου αριθμού επαναλήψεων με την διακοπή αυτών όταν σταματά να παρατηρείται βελτίωση στο σφάλμα. Η μέτρηση του σφάλματος συνήθως γίνεται με ένα τμήμα των δεδομένων εκπαίδευσης που ονομάζεται validation dataset.



Σχήμα 2.3 Αναπαράσταση λειτουργίας Early stopping

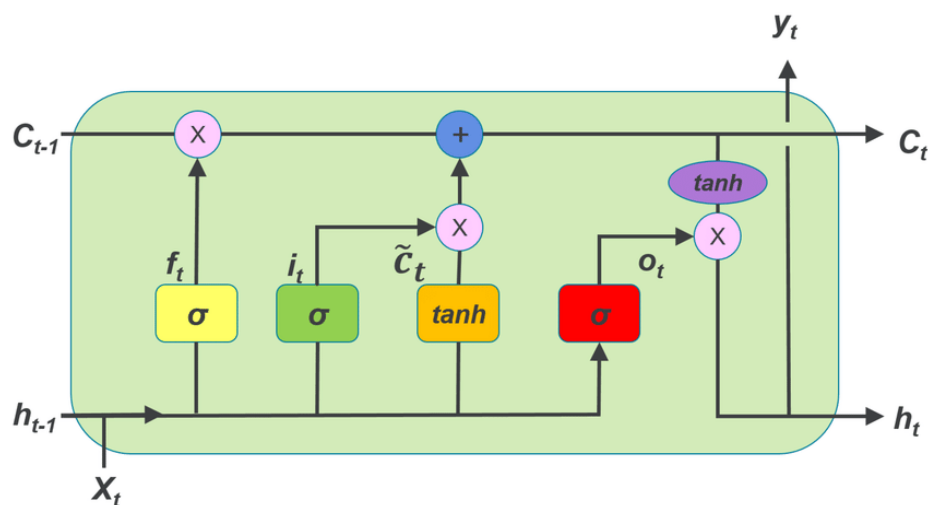
2.7 Χρήση Νευρωνικών Δικτύων για την πρόβλεψη χρονοσειρών

2.7.1 Η Long short-term memory αρχιτεκτονική νευρωνικών δικτύων

Για την πρόβλεψη χρονοσειρών στην παρούσα διπλωματική προτιμάται η Long short-term memory (LSTM) αρχιτεκτονική νευρωνικού δικτύου, η οποία είναι τύπου RNN (Recurrent neural network). Σε αντίθεση με τα τυπικά νευρικά δίκτυα feedforward, το LSTM έχει συνδέσεις ανατροφοδότησης. Ένα LSTM κελί (cell) αποτελείται από τα παρακάτω κομμάτια:

- Cell state (C_t). Αυτό αντιπροσωπεύει την εσωτερική μνήμη του κελιού που αποθηκεύει τόσο τη βραχυπρόθεσμη όσο και τη μακροπρόθεσμη μνήμη.

- Hidden state (h_t). Είναι η τιμή της εξόδου που υπολογίζεται για τις τρέχουσες εισόδους και δίνεται ως είσοδο στα επόμενα κελιά.
- Forget gate (f_t). Είναι υπεύθυνη για την αφαίρεση πληροφοριών από την κατάσταση του κελιού (cell state). Οι πληροφορίες που δεν απαιτούνται πλέον για να καταλάβει το LSTM πράγματα ή οι πληροφορίες που είναι λιγότερο σημαντικές αφαιρούνται μέσω πολλαπλασιασμού ενός φίλτρου. Δέχεται ως εισόδους τα h_{t-1} και x_t . Η h_{t-1} είναι η κρυφή κατάσταση από το προηγούμενο κελί ή η έξοδος του προηγούμενου κελιού και η x_t είναι η είσοδος στο συγκεκριμένο βήμα χρόνου. Οι εισοδοί αυτοί, πολλαπλασιάζονται με το βάρος και έπειτα προστίθεται μια σταθερά, το bias. Στην συνέχεια το αποτέλεσμα εφαρμόζεται στην σιγμοειδή συνάρτηση, η οποία καθορίζει ποια τιμή θα κρατηθεί και ποια θα «ξεχαστεί». Εάν η τιμή της εξόδου της σιγμοειδής προσεγγίζει το «0» τότε η forget gate επιθυμεί να ξεχαστεί εντελώς αυτό το κομμάτι πληροφοριών, ενώ αν τείνει στο «1» το αντίθετο.
- Input gate (i_t). Η πύλη εισόδου είναι υπεύθυνη για την προσθήκη πληροφοριών στην κατάσταση κελιού (Cell state). Αυτή η προσθήκη πληροφοριών γίνεται με τα παρακάτω βήματα. Ρύθμιση των τιμών που πρέπει να προστεθούν στην κατάσταση κελιού με τη συμμετοχή μιας συνάρτησης σιγμοειδούς και των εισόδων h_{t-1} και x_t . Επίσης δημιουργείται ένα αντίστοιχο διάνυσμα με την συμμετοχή της tanh (που δίνει αποτέλεσμα από -1 μέχρι 1) . Τα δύο διανύσματα πολλαπλασιάζονται και το αποτέλεσμα προστίθεται στην κατάσταση κελιού.
- Output gate (o_t). Επιλέγει την πληροφορία από το τρέχον cell state που θα οδηγηθεί στην έξοδο. Δημιουργείται διάνυσμα μετά την εφαρμογή της συνάρτησης tanh στην κατάσταση κελιού, κλιμακώνοντας έτσι τις τιμές στο εύρος -1 έως +1. Δημιουργείται φίλτρο με τη συμμετοχή μιας συνάρτησης σιγμοειδούς και των εισόδων h_{t-1} και x_t . Τα δύο αυτά αποτελέσματα πολλαπλασιάζονται και έχουμε το αποτέλεσμα (y_t) και και την κρυφή κατάσταση (h_t) που εφαρμόζεται στο επόμενο κελί.



Σχήμα 2.4. Το LSTM κύτταρο

Παρακάτω ακολουθούν οι εξισώσεις που προκύπτουν:

$$f(t) = \sigma(w_{xf}^T x(t) + h(t-1) + b_f) \quad (2.6)$$

$$i(t) = \sigma(w_{xi}^T c(t) + h(t-1) + b_i) \quad (2.7)$$

$$o(t) = \sigma(w_{xo}^T x(t) + h(t-1) + b_o) \quad (2.8)$$

$$\tilde{c}(t) = g(t) = \tanh(w_{xg}^T x(t) + w_{hg}^T h(t-1) + b_g) \quad (2.9)$$

$$c(t) = f(t) \otimes c(t-1) + i(t) \otimes g(t) \quad (2.10)$$

$$h(t) = y(t) = o(t) \otimes \tanh(c(t)) \quad (2.11)$$

Η εκπαίδευση του δικτύου επιτυγχάνεται μέσω της διαδικασίας backpropagation. Αρχικά η μεταβολή του $y(t)$ υπολογίζεται ως εξής:

$$\delta_{y(t)} = \Delta_{(t)} + w_{hg} \delta_{g(t+1)} + w_{hi} \delta_{i(t+1)} + w_{hf} \delta_{f(t+1)} + w_{ho} \delta_{o(t+1)} \quad (2.12)$$

Το $\Delta_{(t)}$ είναι το διάνυσμα δέλτα το οποίο έρχεται από το παραπάνω επίπεδο. Η πύλη εξόδου υπολογίζεται ως εξής: Ορίζουμε ως $\tilde{o}(t)$ με την εξίσωση $o(t) = \sigma(\tilde{o}(t))$. Έτσι, για την output gate (πύλη εξόδου) προκύπτει:

$$\delta_{\tilde{o}(t)} = \delta_{y(t)} \otimes \tanh(c(t)) \otimes \sigma'(\tilde{o}(t)) \quad (2.13)$$

Στη συνέχεια, οι αντίστοιχες βαθμίδες για τα βάρη υπολογίζονται ως το άθροισμα των εξωτερικών γινομένων δύο διανυσμάτων. Για παράδειγμα, στην πύλη εξόδου:

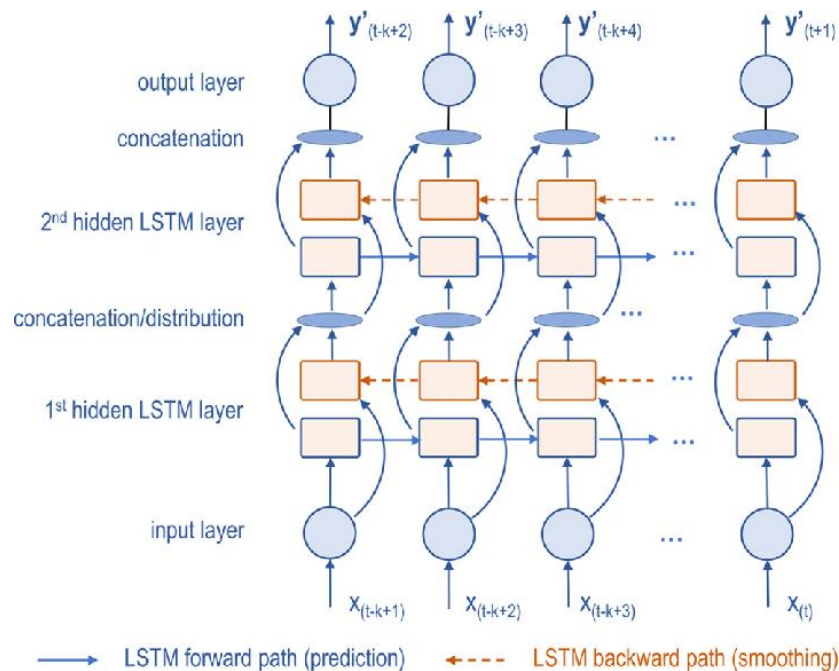
$$\delta w_{x0} = \sum_{t=0} \langle \delta_{\tilde{o}(t)}, x(t) \rangle$$

2.7.2 Αμφίδρομα Long short-term memory δίκτυα

Τα αμφίδρομα Long short-term memory δίκτυα συνδέουν δύο κρυμμένα στρώματα αντίθετων κατευθύνσεων στην ίδια έξοδο. Με αυτήν τη μορφή γενετικής βαθιάς μάθησης, το επίπεδο εξόδου μπορεί να λάβει πληροφορίες από προηγούμενες (προς τα πίσω) και μελλοντικές (προς τα εμπρός) καταστάσεις ταυτόχρονα. Εφευρέθηκαν το 1997 από τους Schuster και Paliwal και υποσχέθηκαν να αυξήσουν την ποσότητα των πληροφοριών εισαγωγής που διατίθενται στο δίκτυο [44].

Στο αμφίδρομο LSTM, τόσο η μπροστινή (forward path – prediction) όσο και η πίσω (backward path – smoothing) διαδρομή υπολογίζονται ανεξάρτητα και οι εξόδους τους

συνδυάζονται. Η τεχνική αυτή διευκολύνει ιδιαίτερα την ανακάλυψη μοτίβων. Οι εξισώσεις στο backward path προκύπτουν με αντίστοιχο τρόπο (έχουν διαφορετικά βάρη).



Σχήμα 2.5. Αμφίδρομο LSTM δίκτυο με 2 κρυφά επίπεδα [45]

2.7.3 Προετοιμασία δεδομένων για την Long short-term memory αρχιτεκτονική

Αρχικά πρέπει να γίνει η επιλογή δύο σημαντικών παραμέτρων. Ο πρώτος είναι της χρονικής διαφοράς ανάμεσα σε δύο σημεία του χρόνου. Για παράδειγμα, μπορούμε να επιλέξουμε να δειγματοληπτηθούν τα δεδομένα για κάθε πέντε λεπτά. Με τον τρόπο αυτό η πρόβλεψη που θα έχουμε θα αντιστοιχεί σε χρόνο 5 λεπτά μετά από την τελευταία είσοδο. Επιλέγοντας μεγαλύτερη συχνότητα δειγματοληψίας συνήθως αυξάνεται η ακρίβεια του μοντέλου, επιβραδύνεται όμως σημαντικά η όλη λειτουργία του δικτύου. Στην συνέχεια καλούμαστε να επιλέξουμε το μέγεθος της ακολουθίας (πλήθος χρονικών στιγμών) που θα δέχεται ως είσοδο το δίκτυο. Προκειμένου να προσφέρουμε μεγαλύτερη σταθερότητα στις εισόδους του μοντέλου μας, θα συμπεριλάβουμε προηγούμενα χρονικά βήματα ως μέρος της εισόδου. Για παράδειγμα, για μια ακολουθία μεγέθους 4 θα περιλαμβάνει το τρέχον χρονικό βήμα μαζί με τα 3 προηγούμενα χρονικά βήματα. Η διαίσθηση για αυτό είναι ότι ο θόρυβος που περιέχεται σε ένα μόνο χρονικό βήμα μπορεί να εξαλειφθεί από το πλαίσιο των άλλων χρονικών βημάτων μέσα στο παράθυρο. Επιπλέον, με τον τρόπο αυτό το δίκτυο διευκολύνεται στην αναγνώριση μοτίβων και έτσι στην αποτελεσματικότερη πρόβλεψη. Πολύ μικρή επιλογή του μεγέθους ακολουθίας δεν βοηθά το δίκτυο να «μάθει» μεγάλες εξαρτήσεις (μεγάλα μοτίβα), ενώ αντίθετα η επιλογή

πολύ μεγάλου παραθύρου μπορεί δυνητικά να οδηγήσει σε υπερβολική εφαρμογή (overfitting), και υποβάθμιση της απόδοσης του μοντέλου. Το κατάλληλο μέγεθος παραθύρου διαφέρει σίγουρα από πρόβλημα σε πρόβλημα, ανάλογα με τα δεδομένα και το μοντέλο που χρησιμοποιείται.

Στην συνέχεια, τα δεδομένα υποβάλλονται σε μια διαδικασία κανονικοποίησης. Η κανονικοποίηση των δεδομένων επιταχύνει γενικά τη μάθηση και οδηγεί σε ταχύτερη σύγκλιση. Η κανονικοποίηση είναι η επανασύνδεση δεδομένων στο αρχικό εύρος έτσι ώστε όλες οι τιμές να κυμαίνονται μεταξύ 0 και 1 (ή σε -1,1). Η διαδικασία απαιτεί να γνωρίζετε ή να είστε σε θέση να εκτιμήσετε με ακρίβεια τις ελάχιστες και μέγιστες παρατηρήσιμες τιμές. Μπορείτε να εκτιμήσετε αυτές τις τιμές από τα διαθέσιμα δεδομένα. Ενδεικτικά τα δεδομένα μπορούν να υποβληθούν στον ακόλουθο τύπο:

$$y = (x - \min) / (\max - \min)$$

2.7.4 Αξιολόγηση μοντέλων Long short-term memory αρχιτεκτονικής

Τα μοντέλα Long short-term memory αξιολογούνται ανάλογα με τον στόχο για τον οποίο έχουν δημιουργηθεί.

- Μοντέλα πρόβλεψης κατεύθυνσης τιμής. Ο κύριος στόχος είναι να προβλεφθεί η κατεύθυνση της επόμενης αλλαγής της τιμής. Τα μοντέλα αυτά αξιοποιούνται ιδιαίτερα σε περιπτώσεις μετατροπής συναλλαγμάτων ή μετοχών (trading). Στα μοντέλα αυτά δεν έχει καθοριστική σημασία η ακριβής πρόβλεψη της μελλοντικής τιμής, αλλά αν αυτή θα μειωθεί ή θα αυξηθεί. Μια πρόβλεψη μπορεί να χαρακτηριστεί ορθή αν ικανοποιείται ο παρακάτω τύπος, οποίος προτείνεται στην συγκεκριμένη διπλωματική εργασία. Σημειώνεται ότι μπορούν να υπάρξουν και διαφορετικές προσεγγίσεις.

$$((\text{πρόβλεψη τιμής μοντέλου}) - (\text{κατώφλι})) * (\text{μεταβολή πραγματικής τιμής}) \geq 0$$

$$\text{κατώφλι} = (\text{πρόβλεψη τιμής μοντέλου}) * (1 + (\text{μεταβολή στην είσοδο}))$$

$$\text{μεταβολή στην είσοδο} = \frac{\text{τρέχων τιμή στην είσοδο} - \text{προηγούμενη τιμή στην είσοδο}}{\text{προηγούμενη τιμή στην είσοδο}}$$

$$(\text{μεταβολή πραγματικής τιμής}) = (\text{νέα πραγματική τιμή}) - (\text{προηγούμενη πραγματική τιμή})$$

Κάθε πρόβλεψη μπορεί να ταξινομηθεί σε τέσσερις κατηγορίες:

1. αληθώς θετική (true positive – tp)

2. ψευδώς θετική (false positive – fp)
3. αληθώς αρνητική (true negative – tn)
4. ψευδώς αρνητική (false negative – fn)

Συνήθως η θετική πρόβλεψη αντιστοιχεί σε πρόβλεψη για άνοδο της τιμής, ενώ η αρνητική σε μείωση. Παράλληλα, η αληθής σε ορθή ενώ η ψευδή σε λάθος πρόβλεψη. Οι συνηθέστεροι δείκτες πρόβλεψης τιμών για την αξιολόγηση της απόδοσης είναι οι :

$$\text{precision} = \frac{tp}{tp + fp}$$

$$\text{recall} = \frac{tp}{tp + fn}$$

$$\text{accuracy} = \frac{tp + tn}{tp + fp + tn + fn}$$

$$F - \text{measure} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

Ο δείκτης precision ονομάζεται επίσης θετική προγνωστική τιμή , ενώ ο δείκτης recall ευαισθησία. Ο δείκτης accuracy μας δείχνει το ποσοστό των σωστών προβλέψεων, ενώ το F-measure (ή F-score) είναι το αρμονικό μέσο της του precision και του recall.

- Μοντέλα πρόβλεψης τιμής. Ο κύριος στόχος είναι να προβλεφθεί η ακριβής τιμή . Μία συνηθισμένη μέθοδος για την αξιολόγηση τέτοιων δικτύων είναι ο υπολογισμός της απόκλισης της μέσης τετραγωνικής τιμής (RMSE). Είναι ένα συχνά χρησιμοποιούμενο μέτρο των διαφορών μεταξύ των τιμών που προβλέπονται από ένα μοντέλο ή έναν εκτιμητή και τις τιμές που παρατηρούνται. Είναι φανερό ότι μικρότερη τιμή του RMSE οδηγεί σε ακριβέστερη πρόβλεψη του μοντέλου

$$RMSE = \sqrt{\sum_{i=1}^n \frac{(\hat{y}_i - y_i)^2}{n}}$$

Επιπλέον σημαντική παράμετρος υπολογισμού τέτοιων σφαλμάτων είναι το μέσο απόλυτο ποσοστό σφάλματος (Mean Absolute Percentage Error). Είναι ένα μέτρο που δείχνει πόσο ακριβής είναι ένα σύστημα πρόβλεψης. Μετρά αυτήν την ακρίβεια ως

ποσοστό . Σε σχέση με το RMSE είναι προτιμότερο όταν θέλουμε να συγκρίνουμε μοντέλα με διαφορετικά δεδομένα.

$$MAPE = \frac{100}{n} \sum_{i=1}^n \left| \frac{\hat{y}_i - y_i}{y_i} \right|$$

ΚΕΦΑΛΑΙΟ 3

ΕΦΑΡΜΟΓΕΣ ΓΕΝΕΤΙΚΩΝ ΑΛΓΟΡΙΘΜΩΝ

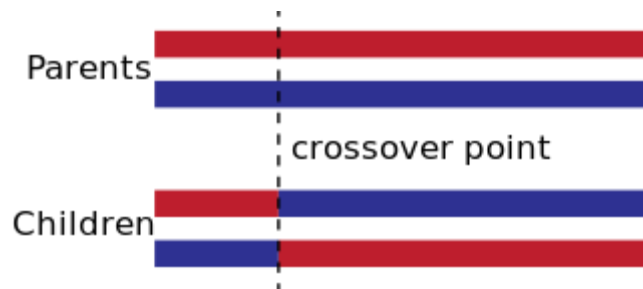
3.1 Εισαγωγή

Οι εξελικτικοί αλγόριθμοι είναι μια ευρετική (heuristic-based) προσέγγιση που βασίζεται στην επίλυση προβλημάτων που δεν μπορούν να λυθούν εύκολα σε πολυωνυμικό χρόνο . Ένας εξελικτικός αλγόριθμος χρησιμοποιεί μηχανισμούς εμπνευσμένους από τη βιολογική εξέλιξη, όπως αναπαραγωγή, μετάλλαξη, ανασυνδυασμό και επιλογή.

Η μεγαλύτερη κατηγορία εξελικτικών αλγορίθμων είναι οι γενετικοί αλγόριθμοι, οι οποίοι είναι εμπνευσμένοι από την θεωρία του Δαρβίνου. Αυτοί οι αλγόριθμοι αντικατοπτρίζουν τη διαδικασία της φυσικής επιλογής, όπου τα πιο κατάλληλα άτομα επιλέγονται για αναπαραγωγή προκειμένου να παράγουν απογόνους της επόμενης γενιάς. Αποτελείται από τα παρακάτω βήματα :

- Αρχικοποίηση πληθυσμός (Initial population) . Για να ξεκινήσουμε τον αλγόριθμό μας, πρέπει πρώτα να δημιουργήσουμε έναν αρχικό πληθυσμό λύσεων. Ο πληθυσμός θα περιέχει έναν αυθαίρετο αριθμό πιθανών λύσεων στο πρόβλημα, που συχνά ονομάζονται μέλη. Είναι σημαντικό ο πληθυσμός να περιλαμβάνει ένα ευρύ φάσμα λύσεων, διότι ουσιαστικά αντιπροσωπεύει μια ομάδα γονιδίων. Οι λύσεις ονομάζονται χρωμοσώματα.
- Επιλογή (selection). Μόλις δημιουργηθεί ένας πληθυσμός, τα μέλη του πληθυσμού πρέπει τώρα να αξιολογηθούν σύμφωνα με μια συνάρτηση (fitness function). Η συνάρτηση αυτή δίνει ένα αριθμητικό αποτέλεσμα. Με αυτόν το τρόπο υπολογίζουμε την καταλληλότητα όλων των μελών και επιλέγουμε ένα τμήμα των μελών με τη μεγαλύτερη βαθμολογία.
- Διασταύρωση (crossover). Συνδυάζει τις γενετικές πληροφορίες δύο γονέων για να δημιουργήσει νέους απογόνους. Για να ζευγαρώσουν κάθε ζευγάρι γονέων, επιλέγεται τυχαία ένα σημείο διασταύρωσης από τα γονίδια.
- Μετάλλαξη (mutation). Σε ορισμένους νέους απογόνους που σχηματίζονται, ορισμένα από τα γονίδια τους μπορούν να υποβληθούν σε μετάλλαξη με χαμηλή τυχαία πιθανότητα. Αυτό σημαίνει ότι μερικά από τα γονίδια των χρωμοσωμάτων μπορούν να αλλάξουν τιμή.

Η διαδικασία αυτή επαναλαμβάνεται μέχρις ότου να ικανοποιηθεί τουλάχιστον ένα από τα κριτήρια τερματισμού. Τα κριτήρια αυτά είναι ο μέγιστος αριθμός επαναλήψεων και ένα επιθυμητό ελάχιστο σφάλμα.



Σχήμα 3.1 Διασταύρωση (crossover)

Before Mutation

A5

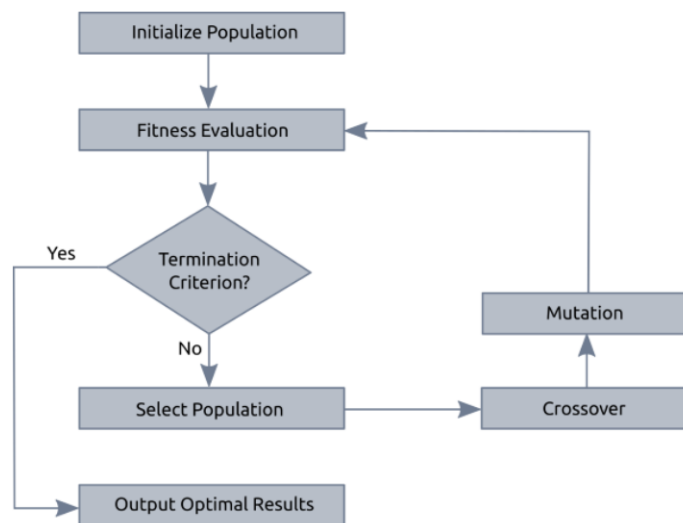
1	1	1	0	0	0
---	---	---	---	---	---

After Mutation

A5

1	1	0	1	1	0
---	---	---	---	---	---

Σχήμα 3.2 Μετάλλαξη (mutation)



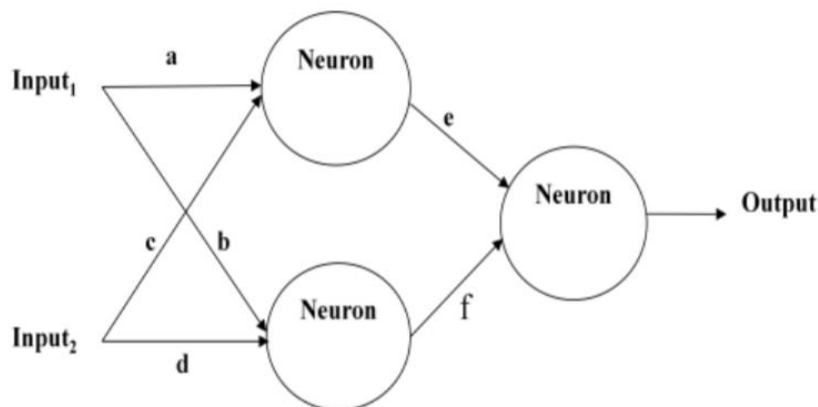
Σχήμα 3.3 Δομή Γενετικού αλγορίθμου

3.2 Συνδυασμός νευρωνικών δικτύων και γενετικών αλγορίθμων

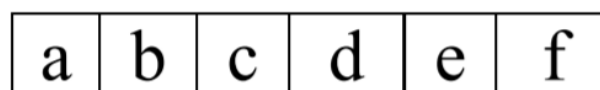
3.2.1 Χρήση γενετικών αλγορίθμων για την εκπαίδευση νευρωνικών δικτύων

Ο πιο διαδεδομένος αλγόριθμος εκπαίδευσης ΤΝΔ είναι ο αλγόριθμος Backpropagation. Ο αλγόριθμος αυτός έχει ως στόχο να ελαχιστοποιήσει το ολικό μέσο τετραγωνικό σφάλμα μεταξύ της πραγματικής και της επιθυμητής εξόδου. Στα μειονεκτήματά του όμως είναι ότι μπορεί να «πέσει» σε τοπικό ελάχιστο της συνάρτησης σφάλματος και είναι μη αποτελεσματικός σε μη παραγωγίσιμες και πολύπλοκες συναρτήσεις. Ένας αποτελεσματικός τρόπος για την παράκαμψη αυτών των προβλημάτων είναι η χρήση γενετικών αλγορίθμων.

Η βασική ιδέα της χρήσης γενετικών αλγορίθμων για την εκπαίδευση νευρωνικών δικτύων είναι να χρησιμοποιηθούν για την εύρεση των βέλτιστων βαρών των συνάψεων. Ένα παράδειγμα δημιουργίας χρωμοσώματος από τα νευρωνικά δίκτυα φαίνεται στα Σχήματα 3.3 και 3.4. Όλα τα βάρη των συνάψεων με συγκεκριμένη σειρά δημιουργούν ένα σύνολο χαρακτήρων. Δημιουργώντας διάφορα χρωμοσώματα στην ουσία δημιουργούμε ένα πλήθος νευρωνικών δικτύων. Για κάθε χρωμόσωμα υπολογίζουμε την συνάρτηση Αξιολόγησης (Fitness). Η συνάρτηση αυτή εξαρτάται συνήθως από το σφάλμα της πραγματικής εξόδου από την επιθυμητή του νευρωνικού δικτύου (π.χ. $\text{fitness} = \frac{1}{\text{error}}$). Μικρότερο σφάλμα οδηγεί σε μεγαλύτερη καταλληλότητα (fitness).



Σχήμα 3.3 Απλό νευρωνικό δίκτυο



3.4 Δημιουργία χρωμοσώματος από το ΤΝΔ του σχήματος 3.3

3.2.2 Χρήση γενετικών αλγορίθμων για την επιλογή αρχιτεκτονικής νευρωνικού δικτύου

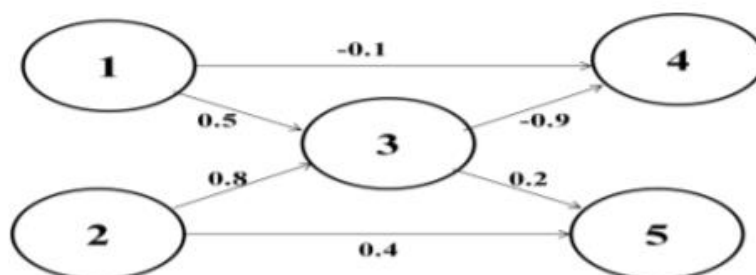
Η αρχιτεκτονική ενός Νευρωνικού δικτύου έχει σημαντικό αντίκτυπο στην αποτελεσματικότητα . Είναι εφικτή η χρήση γενετικών αλγορίθμων για την εύρεση της κατάλληλης αρχιτεκτονικής του νευρωνικού δικτύου. Ένας τρόπος για την υλοποίηση αυτού είναι η αναπαράσταση για κάθε νευρώνα με 1 της σύνδεσης της και με 0 της μη σύνδεσης της με τους υπόλοιπους αντίστοιχους νευρώνες. Για παράδειγμα για την αρχιτεκτονική του δικτύου του Σχήματος 3.3 θα έχουμε την παρακάτω (Σχήμα 3.5) αναπαράσταση πινάκων. Η αρίθμηση των νευρώνων ορίστηκε με Νευρώνας Α για τον πάνω νευρώνα και στην συνέχεια σύμφωνα με την φορά του ρολογιού. Η θέση του πίνακα (i,i) αναπαριστά την σύνδεση του νευρώνα με τον εαυτό του , οπότε στο συγκεκριμένο παράδειγμα είναι 0.

Νευρώνας Α	0 1 0
Νευρώνας Β	1 0 1
Νευρώνας Γ	0 1 0

Σχήμα 3.5 Αναπαράσταση πίνακα

3.2.3 Παράλληλη εκπαίδευση και επιλογή βέλτιστης αρχιτεκτονικής νευρωνικού δικτύου

Συνδυάζοντας τις 2 παραπάνω μεθόδους μπορούμε να επιταχύνουμε την διαδικασία. Για να γίνει αυτού επιλέγεται ξανά για 0 η αναπαράσταση της μη σύνδεσης 2 νευρώνων i,j στο κελί του πίνακα $M(i,j)$, αλλά αυτή την φορά η σύνδεση των 2 αντιστοιχεί στο βάρος της αντίστοιχης σύναψης. Με τα σχήματα 3.6 και 3.7 παρουσιάζεται ένα παράδειγμα αυτής της υλοποίησης. Ο πίνακας που προκύπτει αντιστοιχεί στο αλφαριθμητικό (χρωμόσωμα) 0 0 0 0.5 0.1 0 0 0 0 0.8 0 0.4 0 0 0 0 -0.9 0.2 0 0 0 0 0 0 0 0 0 0 0.



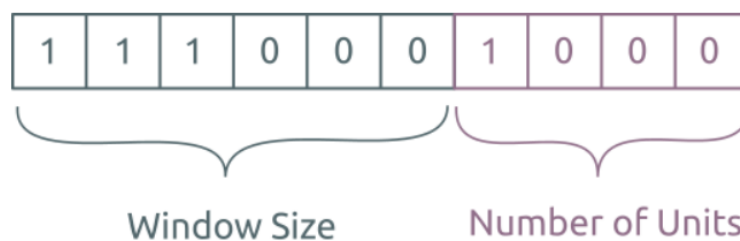
Σχήμα 3.6 Παράδειγμα Νευρωνικού Δικτύου

+0.0 +0.0 +0.0 +0.5 - 0.1 +0.0
 +0.0 +0.0 +0.0 +0.8 +0.0 +0.4
 +0.0 +0.0 +0.0 +0.0 - 0.9 +0.2
 +0.0 +0.0 +0.0 +0.0 +0.0 +0.0
 +0.0 +0.0 +0.0 +0.0 +0.0 +0.0

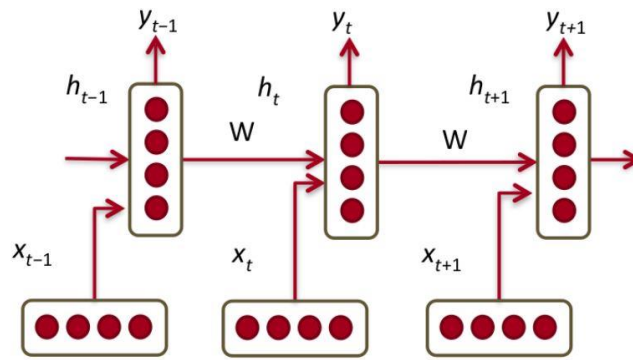
Σχήμα 3.7 Αναπαράσταση πίνακα

3.2.4 Βελτιστοποίηση LSTM με την χρήση γενετικών αλγορίθμων

Για την αναπαράσταση και την δημιουργία του γενετικού αλγορίθμου απαιτείται η αναπαράσταση λύσης (ορισμός χρωμοσώματος) και υλοποίηση της συνάρτησης αξιολόγησης (fitness). Σε ότι αναφορά την χρήση τους σε LSTM νευρωνικά δίκτυα, ένα παράδειγμα χρωμοσώματος είναι το Σχήμα 3.7. Το window size σ ένα τέτοιο δίκτυο αφορά τον αριθμό των προηγούμενων χρονικών στιγμών που συνοδεύουν μια χρονική στιγμή ώστε να σχηματίσουν την ακολουθία εισόδου του δικτύου. Κάθε κρυφό επίπεδο έχει κρυφά κελιά (hidden cells), όσο και ο αριθμός των βημάτων χρόνου (window size). Και επιπλέον, κάθε κρυφό κελί αποτελείται από πολλές κρυφές μονάδες (hidden units), όπως φαίνεται και στο σχήμα 3.8. Οι κρυφές μονάδες αντιστοιχούν στην μεταβλητή Number of Units του σχήματος 3.7 και επιπλέον αντιστοιχούν στην διάσταση του χώρου εξόδου. Ως συνάρτηση αξιολόγησης fitness χρησιμοποιείται $\frac{1}{RMSE}$ (RMSE = Root-Mean-Square Error).



Σχήμα 3.7 Αναπαράσταση χρωμοσώματος για LSTM δίκτυα



Σχήμα 3.8 LSTM διάγραμμα

Μία ενδεδειγμένη βελτιστοποίηση των LSTM με την χρήση γενετικών αλγορίθμων συνοψίζεται με τα εξής βήματα:

- Δημιουργία πολλών LSTM νευρωνικών δικτύων (population) με διαφορετικούς αριθμούς βημάτων χρόνου (window size) και κρυφές μονάδες (number of units).
- Αξιολόγηση κάθε δικτύου με την συνάρτηση αξιολόγησης fitness. Επιλογή των καταλληλότερων δικτύων. Διασταύρωση και μετάλλαξη (με μια μικρή πιθανότητα) των χρωμοσωμάτων τους και δημιουργία νέων δικτύων.
- Επανάληψη της προηγούμενης διαδικασίας για έναν αριθμό επαναλήψεων ή μέχρι να επιτευχθεί κάποιο κριτήριο σφάλματος.
- Το δίκτυο με το μικρότερο σφάλμα που προκύπτει μας δίνει τις τιμές των window size και number of units που θα χρησιμοποιήσουμε στο δίκτυο.

ΚΕΦΑΛΑΙΟ 4

ΕΠΕΞΕΡΓΑΣΙΑ ΦΥΣΙΚΗΣ ΓΛΩΣΣΑΣ

4.1 ΕΙΣΑΓΩΓΗ

Η επεξεργασία της φυσικής γλώσσας (NLP), είναι ένας κλάδος της τεχνητής νοημοσύνης που ασχολείται με την αλληλεπίδραση μεταξύ υπολογιστών και ανθρώπων χρησιμοποιώντας τη φυσική γλώσσα. Ο απώτερος στόχος του είναι η ανάγνωση, η αποκρυπτογράφηση και η κατανόηση της φυσικής γλώσσας από ένα υπολογιστικό σύστημα.

Η επεξεργασία της φυσικής γλώσσας μπορεί να μας δώσει λύσεις σε διάφορα προβλήματα όπως:

- Εξαγωγή συναισθημάτων (sentimental analysis) από μία πρόταση ή κείμενο.
- Αυτόματη εξαγωγή περίληψης από ένα κείμενο.
- Αναγνώριση ονομαστικής οντότητας (Named-entity recognition) , δηλαδή η ταξινόμηση των οντοτήτων που αναφέρονται σε μη δομημένο κείμενο σε προκαθορισμένες κατηγορίες όπως ονόματα προσώπων, οργανώσεις, τοποθεσίες, νομισματικές αξίες, ποσοστά κ.λπ.
- Αναγνώριση φωνής.
- Διεξαγωγή μιας συνομιλίας μέσω διαδικτύου μέσω κειμένου ή κειμένου σε ομιλία, χωρίς την παρουσία άμεσης επαφής με έναν άνθρωπο (chatbot).
- Αυτόματη μετάφραση κειμένου.
- Προβολή κατάλληλων διαφημίσεων σύμφωνα με τα ανθρώπινα ενδιαφέροντα.
- Εξαγωγή πληροφοριών από το διαδίκτυο.
- Ορθογραφικός και γραμματικός έλεγχος.

4.2 Προεπεξεργασία δεδομένων

Υπάρχουν δύο κύριες φάσεις για την επεξεργασία φυσικής γλώσσας: προεπεξεργασία δεδομένων και ανάπτυξη αλγορίθμων. Η προεπεξεργασία δεδομένων αποσκοπεί στην προετοιμασία των δεδομένων ώστε να γίνουν λειτουργικά για τον αλγόριθμο που θα τα χρησιμοποιήσει. Μπορεί να περιλαμβάνει τις παρακάτω μεθόδους:

- Tokenization . Κατά την διαδικασία αυτή χωρίζεται μια φράση, πρόταση, παράγραφος ή ολόκληρο το κείμενο σε μικρότερες ενότητες, όπως

μεμονωμένες λέξεις ή όρους. Κάθε μία από αυτές τις μικρότερες μονάδες ονομάζεται token.

- Stop word removal. Διαγράφονται λέξεις οι οποίες είναι οι πιο συνηθισμένες λέξεις σε οποιαδήποτε γλώσσα (όπως άρθρα, προθέσεις, αντωνυμίες, συνδέσεις κ.λπ.) καθώς δεν προσδίδουν επιπλέον πληροφορίες.
- Lemmatization and stemming. Στόχος τους είναι να μειωθούν οι παραμορφωτικές μορφές και μερικές φορές οι παράγωγες μορφές μιας λέξης σε μια κοινή μορφή βάσης. Ο Lemmatization λαμβάνει υπόψη τη μορφολογική ανάλυση των λέξεων. Για να γίνει αυτό, είναι απαραίτητο να υπάρχουν λεπτομερή λεξικά τα οποία μπορεί να αναζητήσει ο αλγόριθμος για να συνδέσει τη φόρμα με το λήμμα της. Οι αλγόριθμοι Stemming λειτουργούν κόβοντας το τέλος ή την αρχή της λέξης, λαμβάνοντας υπόψη μια λίστα κοινών προθημάτων και επιθημάτων που μπορούν να βρεθούν σε λέξη.
- Part-of-speech tagging. Είναι μια διαδικασία μετατροπής μιας πρότασης σε μία λίστα πλειάδων (όπου κάθε πλειάδα είναι της μορφής (λέξη, ετικέτα)). Η ετικέτα υποδηλώνει εάν η λέξη είναι ουσιαστικό, επίθετο, ρήμα και ούτω καθεξής

4.3 Αλγόριθμοι επεξεργασίας φυσικής γλώσσας

Μόλις τα δεδομένα έχουν υποστεί προεπεξεργασία, αναπτύσσεται ένας αλγόριθμος για την επεξεργασία τους. Οι κατηγορίες αλγόριθμων είναι οι εξής:

- Σύστημα βασισμένους σε κανόνες (Rules-based system). Αυτό το σύστημα χρησιμοποιεί προσεκτικά σχεδιασμένους γλωσσικούς κανόνες. Αυτή η προσέγγιση χρησιμοποιήθηκε νωρίς στην ανάπτυξη της επεξεργασίας φυσικής γλώσσας και εξακολουθεί να χρησιμοποιείται. Εξακολουθούν να χρησιμοποιούνται επειδή είναι δοκιμασμένα και έχουν αποδειχθεί ότι λειτουργούν σωστά. Συνήθως έχουν υψηλή απόδοση σε συγκεκριμένες περιπτώσεις χρήσης, αλλά συχνά υποφέρουν από υποβάθμιση απόδοσης όταν γενικεύονται.
- Αλγόριθμοι «παραδοσιακής» μηχανικής μάθησης. Περιλαμβάνουν πιθανοτικά μοντέλα και γραμμικούς ταξινομητές. Επιπλέον, σ αυτά δεν ανήκουν τα νευρωνικά δίκτυα. Οι αλγόριθμοι αυτοί αποτελούνται από συγκεκριμένα βήματα. Αρχικά, η προεπεξεργασία δεδομένων με την αφαίρεση ανεπιθύμητων δεδομένων. Στην συνέχεια, η τεχνική χαρακτηριστικών (feature engineering) χρησιμοποιείται για τον μετασχηματισμό δεδομένων ακατέργαστου κειμένου σε ελκυστική αριθμητική μορφή, έτσι ώστε ένα μοντέλο να μπορεί να εκπαιδευτεί σε αυτά τα δεδομένα. Οι συνηθέστερες τεχνικές feature engineering είναι η Bag-of-words και η n-gram. Η Bag-of-words είναι μια τεχνική τεχνικών χαρακτηριστικών που δημιουργεί αναπαραστάσεις χαρακτηριστικών με βάση τη συχνότητα εμφάνισης λέξεων. Κατά την διαδικασία αυτή, δημιουργείται μια λίστα με όλες τις χρησιμοποιούμενες

λέξεις. Στην συνέχεια, για κάθε πρόταση δημιουργείται ένα διάνυσμα ίδιας διάστασης με το λεξικό. Σε κάθε θέση υπάρχει ένας αριθμός που υποδεικνύει την συχνότητα εμφάνισης αυτής της λέξης στην συγκεκριμένη πρόταση. Η τεχνική n-gram διασπά επίσης το κείμενο σε μικρότερα κομμάτια. Για παράδειγμα, η 2 gram θα σπάσουν το κείμενο σε οντότητες δύο γραμμάτων (ή δύο λέξεων).

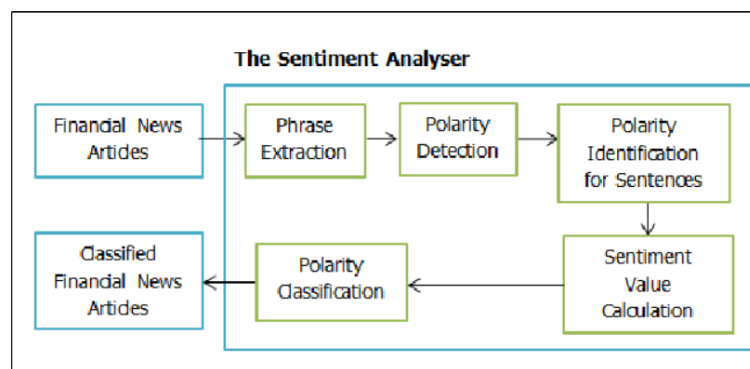
- Νευρωνικά δίκτυα. Σημαντικό πλεονέκτημα εδώ είναι ότι η feature engineering παραλείπεται. Ωστόσο, απαιτούνται αρκετά δεδομένα εκπαίδευσης. Τα συνηθέστερα νευρωνικά δίκτυα χρήσης στο NLP περιλαμβάνουν recurrent neural networks (RNNs) και convolutional neural networks (CNN).

4.4 Εξόρυξη γνώμης (Sentiment analysis)

Η εξόρυξη γνώμης ή ανάλυση συναισθημάτων είναι η διαδικασία προσδιορισμού του κατά πόσον ένα κομμάτι γραφής είναι θετικό, αρνητικό ή ουδέτερο. Η διαδικασία είναι ιδιαίτερα χρήσιμη σε πρακτικές εφαρμογές. Η ανάλυση συναισθημάτων βοηθά τους αναλυτές δεδομένων σε μεγάλες επιχειρήσεις να εκτιμήσουν την κοινή γνώμη για διάφορους σκοπούς, όπως η παρακολούθηση της φήμης της επωνυμίας και του προϊόντος και η κατανόηση των εμπειριών των πελατών. Παράλληλα μπορεί να χρησιμοποιηθεί για την εκτίμηση μεταβολών τιμών μετοχών και συναλλαγμάτων. Μέσω των αλγορίθμων αυτών μπορούν να εντοπιστούν συγκεκριμένες εταιρείες ή συναλλάγματα που παρουσιάζουν θετικό συναίσθημα σε άρθρα ειδήσεων ή στο ευρύ κοινών (μέσω των μέσων κοινωνικής δικτύωσης). Τα δεδομένα αυτά μπορούν να αντληθούν από διάφορες πλατφόρμες των μέσων κοινωνικής δικτύωσης ή των μεγαλύτερων πρατηρίων ειδήσεων μέσω API. Μία ακόμα αξιοσημείωτη χρήση είναι η εκτίμηση εκλογικών αποτελεσμάτων και γενικότερα ψηφοφοριών.

Η ανάλυση συναισθημάτων μπορεί επίσης να κατηγοριοποιηθεί σε δύο είδη, με βάση τον τύπο της παραγωγής που παράγει η ανάλυση.

- Κατηγορηματικό (Categorical/Polarity). Σ' αυτό το είδος το κείμενο κατηγοριοποιείται σε «θετικό», «αρνητικό» ή «ουδέτερο».



Σχήμα 4.1 Παράδειγμα διαδικασίας κατηγοριοποίησης κειμένου [46]

- Βαθμωτό Scalar/Degree . Το κείμενο αντιστοιχεί σε μια τιμή από -1,1. Οι αρνητικές τιμές δηλώνουν αρνητικό περιεχόμενο , ενώ οι θετικές θετικό περιεχόμενο, ενώ παράλληλα τιμές που τείνουν στο 0 ουδέτερο περιεχόμενο. Στην κατηγορία αυτή γίνεται σαφή διάκριση των «έντονα» θετικών και των «έντονα» αρνητικών γνωμών-συναισθημάτων που καταλαμβάνουν τιμές κοντά στο 1 και στο -1 αντίστοιχα.

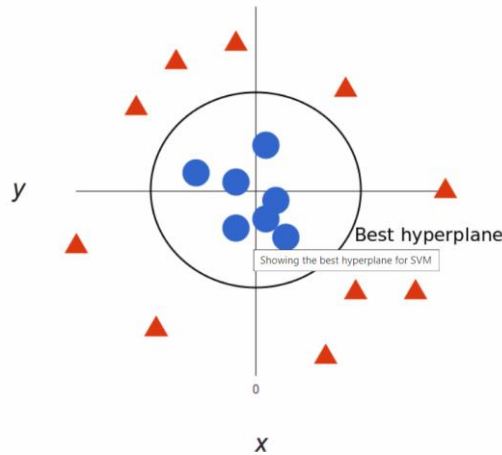
Η ανάλυση συναισθημάτων μπορεί να υλοποιηθεί με διάφορους αλγορίθμους:

- Σύστημα βασισμένους σε κανόνες (Rules-based system). Οι αλγόριθμοι αυτοί επιτυγχάνουν πολύ καλά αποτελέσματα όταν ειδικεύονται. Ο αλγόριθμος υπολογίζει τη βαθμολογία συναισθημάτων από ένα σύνολο κανόνων που δημιουργήθηκαν με μη αυτόματο τρόπο. Για παράδειγμα, μπορείτε να μετρήσετε τον αριθμό που κάποιος χρησιμοποίησε λέξεις όπως «εξαιρετικός», «τέλεια» κλπ. στην κριτική του και να υπολογίσετε το εκτιμώμενο συναίσθημα. Παράλληλα, σε περιπτώσεις όπως των αναρτήσεων σε μέσα κοινωνικής δικτύωσης συνηθίζεται να ακολουθείται παρόμοια δομή σύνταξης των κειμένων και λεξιλόγιο από τους χρήστες. Στις περιπτώσεις αυτές , ένα εξειδικευμένο σύστημα βασισμένο σε κανόνες μπορεί να λειτουργήσει με πολύ καλή επιτυχία.
- Naïve Bayes. Αποτελούν μια αρκετά απλή ομάδα πιθανολογικών αλγορίθμων, για την ταξινόμηση ανάλυσης συναισθημάτων, που αποδίδει μια πιθανότητα ότι μια δεδομένη λέξη ή φράση θα πρέπει να θεωρείται θετική ή αρνητική.

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

Σχήμα 4.2 Ο κανόνας του Bayes

- Γραμμική παλινδρόμηση. Η γραμμική παλινδρόμηση είναι μια προσέγγιση για τη μοντελοποίηση της σχέσης μεταξύ μιας βαθμωτής εξαρτημένης μεταβλητής Y και μία ή περισσότερες επεξηγηματικές μεταβλητές (ή ανεξάρτητη μεταβλητή) συμβολίζεται X.
- Μηχανή διανυσματικής υποστήριξης (Support vector machine). Αποτελούν παρόμοια, αλλά πιο προηγμένη τεχνική από την τεχνική παλινδρόμηση. Η σχέση εισόδου (λέξη) και εξόδου (Κατηγορία) δεν περιορίζεται σε μια γραμμική σχέση.



Σχήμα 4.3 Διαχωρισμός δεδομένων με SVM

- Δέντρο απόφασης (Decision tree). Τα δέντρα απόφασης έχουν συχνά καλή ακρίβεια και έχουν την πρόσθετη ιδιότητα να παράγουν εύκολα επεξηγούμενα αποτελέσματα με τη μορφή δέντρων.
- Deep learning. Συνήθως χρησιμοποιούνται Recurrent Neural Networks και συγκεκριμένα τα Long short-term memory που εξηγήθηκαν εκτεταμένα στο κεφάλαιο 3.

4.5 Εξόρυξη γνώμης για την πρόβλεψη ισοτιμιών συναλλάγματος

Στον σύγχρονο κόσμο, οι άνθρωποι κρίνουν τον κόσμο γύρω τους, κάνοντας θετικές και αρνητικές κριτικές για άτομα, προϊόντα, μέρη και εκδηλώσεις. Οι κριτικές αυτές πέρα από σκέψεις και ιδιωτικές συζητήσεις μεταξύ των ατόμων συνηθίζεται να εκφράζονται μέσω αναρτήσεων στα μέσα κοινωνικής δικτύωσης. Το ίδιο φαίνεται να ισχύει και για αντικείμενα όπως η άνοδος ή η πτώση μετοχών και αξίας συναλλαγμάτων. Με τον τρόπο αυτό στα μέσα κοινωνικής δικτύωσης συσσωρεύεται σωρεία πληροφοριών και η εκμετάλλευση της μπορεί να αντικατοπτρίσει σε ικανοποιητικό βαθμό την κοινή γνώμη. Παράλληλα, σημαντική είναι και η συλλογή κατάλληλων τίτλων ειδήσεων, καθώς μπορούν να εκτιμήσουν βραχυχρόνιες ή μακροχρόνιες μεταβολές, αλλά και να επηρεάσουν την κοινή γνώμη.

Μία μέθοδος εκτίμησης των μεταβολών αυτών είναι υπολογίζοντας την θετικότητα των αναρτήσεων σε μια συγκεκριμένη χρονική στιγμή. Για βραχυπρόθεσμες προβλέψεις έχει παρατηρηθεί ότι ο ενθουσιασμός και τα γενικότερα θετικά μηνύματα προκαλούν άνοδο της τιμής, ενώ αρνητικά συναισθήματα όπως ο φόβος και ο πανικός οδηγούν σε απότομη πτώση. Αντίθετα, σε πιο μακροπρόθεσμες προβλέψεις μπορούμε να ισχυριστούμε ότι ο παρατεταμένος φόβος στην αγορά σε συνδυασμό με πτώση της τιμής

μπορεί να οδηγεί σε επενδυτική ευκαιρία, ενώ παράλληλα παρατεταμένη άνοδος σε συνδυασμό με απληστία συνήθως οδηγεί σε πτώση.

ΚΕΦΑΛΑΙΟ 5

Η ΕΦΑΡΜΟΓΗ

Για τους σκοπούς της τρέχουσας διπλωματικής εργασίας υλοποιήθηκε ο σχεδιασμός και η συγγραφή πηγαίου κώδικα , με σκοπό την δημιουργία εφαρμογής , η οποία θα δραστηριοποιείται γύρω από την πρόβλεψη ισοτιμιών συναλλάγματος. Στην εφαρμογή δόθηκε το όνομα Market Predictor. Ο κύριος σκοπός της είναι η πρόβλεψη της ακριβής τιμής των ισοτιμιών (EUR-USD και EUR-BTC) , αλλά και της μεταβολής αυτών στον χρόνο. Παράλληλα μέσω της συσχέτισης των τίτλων ειδήσεων και αναρτήσεων σε μέσα κοινωνικής δικτύωσης με τις ισοτιμίες και τις μεταβολές αυτών παρέχεται μια επιπλέον πληροφορία προς εκμετάλλευση προς τον χρήστη. Υπάρχει δυνατότητα για εύκολη προσθήκη νομισμάτων. Για την πραγματοποίηση αυτού αντλήθηκαν και αξιοποιήθηκαν πραγματικά δεδομένα ισοτιμιών και ειδήσεων σε πραγματικό χρόνο. Παράλληλα σχεδιάστηκαν LSTM (Long short-term memory) δίκτυα, τα οποία βελτιστοποιήθηκαν με την χρήση γενετικών αλγορίθμων και τέλος πραγματοποιήθηκε ανάλυση συναισθημάτων (sentimental analysis).

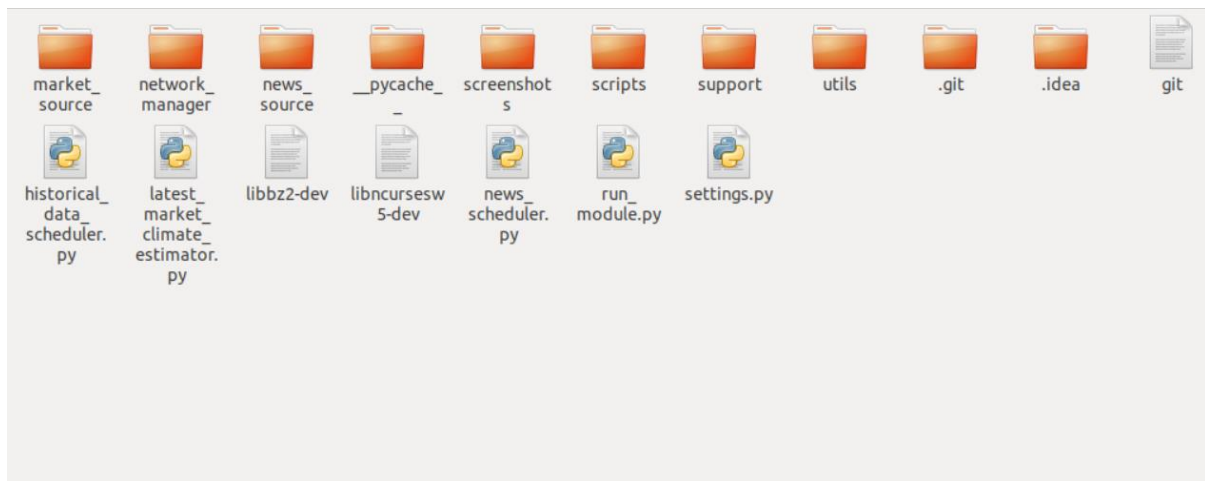
5.1 Αρχιτεκτονική

Η εφαρμογή σχεδιάστηκε σε γλώσσα προγραμματισμού Python. Η έκδοση που δοκιμάστηκε και για την οποία υλοποιήθηκε είναι η 3.6.9. Η Python είναι διερμηνευόμενη (interpreted), γενικού σκοπού (general-purpose) και υψηλού επιπέδου, γλώσσα προγραμματισμού. Ο λόγος που επιλέχθηκε είναι ότι διαθέτει ευρεία γκάμα βιβλιοθηκών τόσο για την εκτέλεση δικτυακών λειτουργιών, αλλά κυρίως στον τομέα της Μηχανικής Μάθησης. Παράλληλα χαρακτηριστικό της είναι η αναγνωσιμότητα του κώδικά της και η ευκολία χρήσης της.

Η βάση δεδομένων που επιλέχθηκε στην παρούσα εφαρμογή είναι η ArangoDB [40]. Η ArangoDB είναι ένα δωρεάν και ανοιχτού κώδικα εγγενές σύστημα βάσεων δεδομένων πολλαπλών μοντέλων που αναπτύχθηκε από την ArangoDB GmbH. Είναι ένα σύστημα βάσης δεδομένων NoSQL. Η γλώσσα ερωτήματος (query language) που περιλαμβάνει ονομάζεται AQL (ArangoDB Query Language) είναι δηλωτική και επιτρέπει τον συνδυασμό διαφορετικών προτύπων πρόσβασης δεδομένων σε ένα μόνο ερώτημα.

Η εφαρμογή αποτελείται από τέσσερα μέρη (components). Τα δύο πρώτα μπορούν να λειτουργήσουν εντελώς αυτόνομα από τα υπόλοιπα και αφορούν την τροφοδοσία της βάσης δεδομένων με πραγματικά δεδομένα για τις ισοτιμίες συναλλαγμάτων και τους τίτλους ειδήσεων και αναρτήσεων αντίστοιχα. Τα υπόλοιπα δύο προϋποθέτουν μόνο την ύπαρξη δεδομένων στην βάση δεδομένων για την λειτουργία τους. Πραγματοποιούν προβλέψεις για μελλοντικές ισοτιμίες των επιθυμητών συναλλαγμάτων (για την επιλεγμένη

χρονική στιγμή - μετά από πέντε , δέκα λεπτά.) και ανάλυση του «κλίματος» της αγοράς για την δεδομένη χρονική στιγμή αντίστοιχα .



Σχήμα 5.1 Στιγμιότυπο από το την εφαρμογή

Στον φάκελο scripts βρίσκεται ένα μικρό script σε bash το οποίο κατά την εκτέλεση του σε Unix λειτουργικό σύστημα εγκαθιστά τις περισσότερες βιβλιοθήκες που είναι απαραίτητες για την εκτέλεση της εφαρμογής. Σημειώνεται εδώ ότι για την εκτέλεση των LSTM μοντέλων είναι επίσης απαραίτητη η εγκατάσταση της NVIDIA CUDA® Deep Neural Network library (cuDNN), η οποία δεν γίνεται μέσω του script. Ιδιαίτερη προσοχή πρέπει να δοθεί στην συμβατότητα της TanserFlow βιβλιοθήκης με την cuDNN.

Παράλληλα, πολλά τμήματα στον κώδικα ακολουθούν την λογική του ασύγχρονου προγραμματισμού. Για τον λόγω αυτό, για να εκτελέσουμε κάποιο αρχείο (.py) που δεν βρίσκεται στον κύριο φάκελο (ίδιο επίπεδο με το run_module.py) πρέπει να το εκτελέσουμε μέσω της run_module.py. Εκτελώντας κάποιο αρχείο μέσω της run_module.py εκτελείται η συνάρτηση async_main που βρίσκεται μέσα σε αυτό.

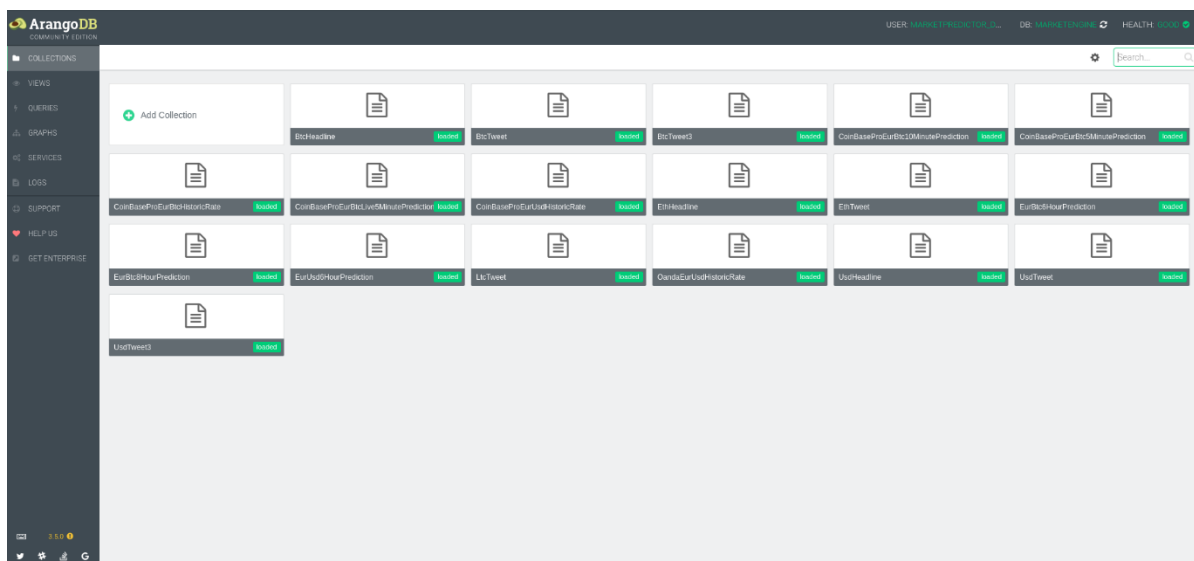
5.2 Σύνδεση με την βάση δεδομένων

Στο αρχείο settings.py υπάρχουν οι παράμετροι arango_nodes, arango_db_name, arango_username και arango_password, οι οποίοι είναι απαραίτητοι για την σύνδεση και την επικοινωνία με την βάση δεδομένων. Η arango_nodes αφορά το μονοπάτι που βρίσκεται η επιθυμητή βάση και arango_db_name το όνομα της βάσης δεδομένων. Η arango_username και η arango_password αφορούν το όνομα χρήστη και ο κωδικός του αντίστοιχα.

Στο αρχείο connect_with_db.py υλοποιήθηκαν όλες οι συναρτήσεις και οι κλάσεις που αφορούν την σύνδεση της εφαρμογής με την βάση δεδομένων. Η ConnectwithDB

δημιουργήθηκε με σκοπό να αποτελεί υπερκλάση όλων των κλάσεων που τα αντικείμενα τους δύνανται να αποθηκευτούν στην βάση δεδομένων. Διαθέτει διάφορες συναρτήσεις όπως η `db_write` και `db_read` οι οποίες διευκολύνουν την γρήγορη εγγραφή και ανάγνωση εγγράφων από την βάση. Μέσω της συνάρτησης `do_aql` δίνεται η δυνατότητα για απευθείας εκτέλεση ερωτημάτων με αξιοποίηση της ArangoDB Query Language. Με την βοήθεια αυτής υλοποιήθηκαν οι παρακάτω συναρτήσεις:

- `receive_last_timestamp` . Επιστρέφει το πιο πρόσφατο timestamp (σφραγίδα χρόνου) ενός collection της βάσης.
- `receive_first_timestamp` . Επιστρέφει το παλαιότερο (πρώτο) timestamp (σφραγίδα χρόνου) ενός collection της βάσης.
- `extract_sorted_collection`. Επιστρέφει ταξινομημένα τα έγγραφα μια συλλογής της βάσης με φορά ταξινόμησης καθώς και ανώτερα και κατώτερα χρονικά όρια που δίνονται ως παράμετροι. Επίσης δίνονται ως παράμετροι η ονομασία του χαρακτηριστικού που αφορά τον χρόνο , καθώς και το όριο των εγγράφων που θέλουμε να επιστραφούν.
- `count_collection_entries_with_timestamp`. Επιστρέφει τον αριθμό των εγγράφων που βρίσκονται σε μια συλλογή της βάσης και είναι νεότερα μιας χρονικής τιμής που δίνεται ως παράμετρο.



Σχήμα 5.2 Το dashboard της βάσης δεδομένων

5.3 Συλλογή Δεδομένων

Για την συλλογή δεδομένων έγινε υλοποίηση κώδικα σύνδεσης (integration) με τα API 2 γνωστών διεθνών ανταλλακτηρίων. Αυτά είναι το Coinbase pro, μέσω του οποίου τροφοδοτούμε την βάση δεδομένων με ιστορικά αλλά και νέα δεδομένα σχετικά με ισοτιμίες του ευρώ με τα δημοφιλέστερα κρυπτονομίσματα (Bitcoin, Ethereum, Litecoin

κλπ), καθώς και του Oanda για την ισοτιμία του ευρώ και του αμερικάνικου δολαρίου (EUR/USD). Παράλληλα, υλοποιήθηκε κώδικας για την συνεχή άντληση τίτλων ειδήσεων από το newsapi.org, το οποίο περιλαμβάνει τίτλους ειδήσεων από παγκόσμιας εμβέλειας πρακτορεία ειδήσεων. Τέλος, μέσω κώδικα σύνδεσης με το API του twitter γίνεται εφικτή η τροφοδότηση του συστήματος με tweets που αφορούν το εκάστοτε ενδιαφερόμενο νόμισμα.

5.3.1 Συλλογή δεδομένων ισοτιμιών συναλλάγματος

Για κάθε ένα από τα integrations (ενσωματώσεις) , τα οποία θα αφορούν συλλογή οικονομικών και συναλλαγματικών δεδομένων θα δημιουργείτε ξεχωριστό αρχείο το οποίο θα περιλαμβάνει αντίστοιχη κλάση. Όλες οι κλάσεις αυτές αποτελούν υποκλάσεις της MarketTrackerBase , η οποία βρίσκεται στο αρχείο market_tracker_base.py . Η κλάση αυτή έχει ως σκοπό να δώσει κοινή αρχιτεκτονική σ όλες τις υποκλάσεις οι οποίες αποτελούν μηχανισμοί συλλογής δεδομένων ισοτιμίας συναλλάγματος (trackers), που έχουν ή θα δημιουργηθούν στο μέλλον, καθιστώντας τους υποχρεωτική την υλοποίηση 2 συγκεκριμένων συναρτήσεων. Αυτές είναι η receive_live_market_rate που αφορά την συλλογή των τρεχουσών τιμών και της receive_historical_data με αρμοδιότητα την συλλογή τιμών από μια συγκεκριμένη τιμή του χρόνου που υπάρχει ως παράμετρος στο αρχείο settings.py μέχρι την τωρινή χρονική στιγμή.

Η υλοποίηση της σύνδεσης με το coinbase pro έγινε μέσω του αρχείου coinbase_pro_tracker.py και της κλάσης CoinbaseProTracker. Η δημιουργία αντίστοιχου αντικειμένου απαιτεί τον ορισμό ενός νομίσματος ως νόμισμα αναφοράς. Αυτό γίνεται θέτοντας το ως παράμετρο στην συνάρτηση αρχικοποίησης της κλάσης (__init__). Στο σημείο αυτό αξιοποιήθηκε η επίσημη βιβλιοθήκη coinbasepro-python της εταιρείας για την ταχύτερη και αξιόπιστη επικοινωνία με το API. Υπάρχει δυνατότητα για την εύκολη προσθήκη ή αφαίρεση νομισμάτων που ακολουθούνται (εφόσον υποστηρίζονται από το ανταλλακτήριο) μέσω αντίστοιχης παραμέτρου στο settings.py. Αξίζει να αναφερθεί ότι ελήφθησαν υπόψιν όλοι οι περιορισμοί σε αριθμούς , συχνότητα και μεγέθους αιτημάτων (requests) προς το API. Για την δημιουργία του αντικειμένου που αντιστοιχεί στην αναλυτική αναφορά του όγκου των συναλλαγών , της ενδιαμέσης υψηλής και χαμηλής τιμής καθώς και της τιμής ανοίγματος και κλεισίματος μιας συγκεκριμένης χρονικής περιόδου δημιουργήθηκε η συνάρτηση get_historic_market_currency_rate_class στο αρχείο exchange_rate.py. Η συνάρτηση αυτή δέχεται ως παραμέτρους 2 νομίσματα (που αποτελούν την ενδιαφερόμενη ισοτιμία) , ένα αλφαριθμητικό που αντιστοιχεί στην πηγή δεδομένων και έπειτα δημιουργεί αντίστοιχη κλάση (υποκλάση της ExchangeRate στο ίδιο αρχείο). Για παράδειγμα αν δεχτεί ως παραμέτρους EUR, BTC και CoinbasePro δημιουργεί την κλάση CoinbaseProEurBtcHistoricRate. Στο σχήμα 5.3 δείχνεται η δημιουργία αντικειμένου με τα αντίστοιχα χαρακτηριστικά και η αποθήκευση του στην βάση ως Document του αντίστοιχου Collection στην βάση δεδομένων. Το πρωτεύον κλειδί του

Document είναι το timestamp της ενδιαφερόμενης χρονικής στιγμής. Το timestamp μετريέται σε unix time (ισούται με τα δευτερόλεπτα που έχουν περάσει από την 1/1/1970 και 00:00:00 UTC).

```
1. history_entry=get_historic_market_currency_rate_class(  
2.     self.base_currency,  
3.     market_currency,  
4.     market_source)()  
5. self.base_currency,market_currency, market_source)()  
6. history_entry.set_timestamp(time_stamp)  
7. history_entry.set_trade_volume(volume)  
8. history_entry.set_interval_high(high)  
9. history_entry.set_interval_low(low)  
10. history_entry.set_open_rate(open)  
11. history_entry.set_close_rate(close)  
12. history_entry.db_write()
```

Σχήμα 5.3 Απόσπασμα από το αρχείο *coinbase_pro_tracker.py*

Για την συλλογή πληροφοριών σχετικά με την τιμή του αμερικάνικου δολαρίου (USD) χρησιμοποιήθηκε το API του ανταλλακτηρίου Oanda. Για τον λόγο αυτό δημιουργήθηκε το αρχείο *oanda_tracker.py* το οποίο περιλαμβάνει την κλάση *OandaTracker* υποκλάση της *MarketTrackerBase*, με αντίστοιχο τρόπο όπως και η *CoinbaseProTracker*. Και σ' αυτή την κλάση ακολουθείται η ίδια αρχιτεκτονική, λογική και συναρτήσεις (με ίδια ονόματα) και παραμέτρους. Για την επικοινωνία με το API δεν αξιοποιήθηκε κάποια εξειδικευμένη βιβλιοθήκη, αλλά τα αιτήματα γίνονται μέσω της βιβλιοθήκης *requests*. Στην συνάρτηση αρχικοποίησης της κλάσης δημιουργούνται τα *headers* (επικεφαλίδα) του αιτήματος. Μέσω αυτού γίνεται το *Authorization* (εξουσιοδότηση του χρήστη) το οποίο περιλαμβάνει το *token* μέσω παραμέτρου που βρίσκεται στο αρχείο *settings.py*.

Είναι φανερό ότι απαιτείται ένα ανώτερο επίπεδο συγχρονισμού και χειρισμού των *trackers* αυτών. Το επίπεδο αυτό θα διευκολύνει την χρήση τους από τον χρήστη, ενώ παράλληλα θα δώσει την δυνατότητα για αδιάκοπη λειτουργία τους. Για τον λόγο αυτό δημιουργήθηκε το αρχείο *historical_data_scheduler.py* το οποίο περιλαμβάνει την κλάση με το όνομα *Scheduler* καθώς και πλήρη αξιοποίηση αυτής με σκοπό εκμετάλευσης ολόκληρου του μηχανισμού άντλησης τιμών απλώς εκτελώντας το αρχείο αυτό. Μέσω τις παραμέτρου *active_market_tracker* του *settings.py* έχουμε την δυνατότητα να απενεργοποιούμε ή να προσθέτουμε κάθε φορά κάποιον από τους *trackers*, καθώς και να ρυθμίζουμε την επιθυμητή χρονική απόσταση δεδομένων (πχ. Αν επιθυμούμε να έχουμε την τιμή για κάθε 1 λεπτό κλπ.) . Παράλληλα, μέσω της «*market_source*»_*historical_freq_download* έχουμε την δυνατότητα να ρυθμίζουμε την χρονική απόσταση που θα χωρίζει δύο αιτήματα στο API του tracker. Ο χρόνος αυτός είναι σε δευτερόλεπτα. Επιπλέον μέσω της λίστας «*market_source*»_*base_currencies* δηλώνεται για κάθε tracker στο αρχείο *settings.py* ποιο νόμισμα (υπάρχει δυνατότητα για παραπάνω

από ένα) θα αποτελέσει το νόμισμα αναφοράς. Ο Scheduler ακολουθεί την λογική του ασύγχρονου προγραμματισμού , μέσω χρήσης της βιβλιοθήκης `asyncio`. Παράλληλα μέσω της βιβλιοθήκης `multiprocessing` διασφαλίζεται η εκτέλεση κάθε tracker σε ξεχωριστή διεργασία (Process), καθώς και η παράλληλη εκτέλεση αυτών. Ο διαχωρισμός σε ξεχωριστές διεργασίας επιπλέον μας διευκολύνει να αποφύγουμε ενδεχόμενα `memory leaks`. Εξίσου σημαντικό είναι το γεγονός ότι διασφαλίζεται ότι ενδεχόμενο bug ή αστοχία σε έναν tracker δεν θα επηρεάσει την ομαλή λειτουργία των υπολοίπων. Κατά την εκτέλεση του scheduler ο χρήστης ενημερώνεται συνεχώς με μηνύματα στο τερματικό για εξέλιξη και την πορεία της συλλογής δεδομένων όπως φαίνεται στην εικόνα 5.2.

```
enabling scheduler for: coinbase_pro
enabling scheduler for: oanda
2021-06-08T23:08:04.156637+03:00 executing operations for: <class 'market_source.plu
2021-06-08T23:08:04.183552+03:00 executing operations for: <class 'market_source.plu
2021-06-08 23:08:06 BTC-EUR LAST CLOSE PRICE: 35690.28 TIMESTAMP: 1621268700
2021-06-08 23:08:07 EUR-USD LAST CLOSE PRICE: 1.221 TIMESTAMP: 1621521240
2021-06-08 23:08:13 BTC-EUR LAST CLOSE PRICE: 36216.40 TIMESTAMP: 1621280400
2021-06-08 23:08:15 EUR-USD LAST CLOSE PRICE: 1.223 TIMESTAMP: 1621550700
2021-06-08 23:08:21 BTC-EUR LAST CLOSE PRICE: 35168.58 TIMESTAMP: 1621292100
2021-06-08 23:08:23 EUR-USD LAST CLOSE PRICE: 1.223 TIMESTAMP: 1621580160
2021-06-08 23:08:28 BTC-EUR LAST CLOSE PRICE: 36814.64 TIMESTAMP: 1621303740
2021-06-08 23:08:31 EUR-USD LAST CLOSE PRICE: 1.218 TIMESTAMP: 1621609680
2021-06-08 23:08:36 BTC-EUR LAST CLOSE PRICE: 37336.96 TIMESTAMP: 1621315440
```

Σχήμα 5.4 Τα μηνύματα αναφοράς του `historical_data_scheduler.py`

5.3.2 Συλλογή τίτλων ειδήσεων και tweets

Η συλλογή τίτλων ειδήσεων και tweets προσανατολίζεται γύρω από την κλάση `NewsTrackerBase` η οποία βρίσκεται στο αρχείο `news_tracker_base.py` (στον φάκελο `news_source`). Η κλάση αυτή καθιστά υποχρεωτική την υλοποίηση των συναρτήσεων `receive_live_market_rate` και `receive_historical_data` για κάθε υποκλάση της. Η συνάρτηση `receive_new_data` αφορά την τροφοδοσία της βάσης με δεδομένα πραγματικού χρόνου . Αξίζει να αναφερθεί ότι λόγω περιορισμών των ελεύθερων API που χρησιμοποιήθηκαν δεν δόθηκε η δυνατότητα για πλήρη αξιοποίηση ενδεχόμενων δυνατοτήτων και των σκοπών της `receive_historical_data`. Παρόλα αυτά για να τηρηθεί ο γενικότερος σχεδιασμός και αρχιτεκτονική του συστήματος, στις περιπτώσεις αυτές υλοποιείται η συνάρτηση αυτή ως κενή συνάρτηση.

Για τον σκοπό της συλλογής τίτλων ειδήσεων διεθνούς εμβέλειας πρακτορείων ειδήσεων αξιοποιείται η δωρεάν έκδοση του API του `newsapi.org`. Το API αυτό μας παρέχει τίτλους ειδήσεων ανάλογα με την κατηγορία που ζητούμε (στην προκειμένη περίπτωση νόμισμα ή κρυπτονόμισμα). Για τον σκοπό αυτό δημιουργήθηκε η κλάση `GoogleNewsTracker` η οποία αποτελεί υποκλάση της `NewsTrackerBase` που αναφέρθηκε

προηγουμένως. Η GoogleNewsTracker βρίσκεται στο αρχείο googlenews_tracker.py . Δέχεται στην συνάρτηση αρχικοποίησης (__init__) ως παράμετρο το ενδιαφερόμενο νόμισμα . Τα αιτήματα στο API γίνονται και εδώ μέσω της βιβλιοθήκης requests καθώς δεν υπάρχει εξιδεικευμένη βιβλιοθήκη. Το googlenews_api_id είναι απαραίτητο για την πραγματοποίηση των αιτημάτων αυτών και βρίσκεται ως παράμετρος στο settings.py . Για την αποθήκευση των δεδομένων αυτών ως document σε συγκεκριμένο collection της βάσης δεδομένων αξιοποιείται η συνάρτηση get_headline_currency_class που έχει δημιουργηθεί στο αρχείο headline_entry.py. Η συνάρτηση αυτή δέχεται ως παράμετρο το νόμισμα για το οποίο γίνεται η αναζήτηση τίτλων ειδήσεων και έτσι δημιουργείται η ενδιαφερόμενη κλάση. Για παράδειγμα , αν αναζητούμε ειδήσεις για το Bitcoin θα δώσουμε ως παράμετρο «BTC» και θα δημιουργηθεί η κλάση με όνομα BtcHeadline , το οποίο αντιστοιχεί σε collection στην βάση με το ίδιο όνομα. Στο Σχήμα 5.3 δείχνεται η δημιουργία των headline_entries και η εγγραφή τους στην βάση.

```

1. for article in articles:
2.     headline_entry = get_headline_currency_class(
3.                                     self.base_currency
4.                                     )()
5.     source = article[self.source_str]
6.     source_id=source['id']
7.     try:
8.         author = article[self.author_str]
9.         title = article[self.title_str]
10.        title = title.encode('latin1').decode('utf8')
11.        description=article[self.description_str]
12.        publishedAt=article[self.publishedAt_str]
13.        publishedAt_timestamp = int(time.mktime(
14.                                    time.strptime(
15.                                        publishedAt,
16.                                        self.pattern)
17.                                    ))
18.        content=article[self.content_str]
19.    except:
20.        continue
21.    headline_entry.set_basic(
22.        publishedAt_timestamp,
23.        source_id,
24.        Search
25.    )
26.    headline_entry.set_title(title)
27.    headline_entry.set_author(author)
28.    headline_entry.set_content(content)
29.    headline_entry.set_timestamp(timestamp_current)
30.    headline_entry.db_write()

```

Σχήμα 5.5 Απόσπασμα από το αρχείο googlenews_tracker.py

Το αρχείο twitter_tracker.py αφορά την κλάση TwitterNewsTracker η οποία ακολουθεί και αυτή την αρχιτεκτονική της NewsTrackerBase. Μέσω της

TwitterNewsTracker αξιοποιούνται οι δυνατότητες του επίσημου API του Twitter χρησιμοποιώντας και την βιβλιοθήκη της python «tweepry». Η δημιουργία αντίστοιχου αντικειμένου προϋποθέτει την επιλογή νομίσματος ή κρυπτονομίσματος παρακολούθησης. Στο αρχείο settings.py είναι αναγκαίο να δοθούν οι παράμετροι twitter_api_key, twitter_api_secret_key, twitter_access_token twitter_access_secret τα οποία αποτελούν αναγκαία προϋπόθεση για την χρήση του συγκεκριμένου API. Για την αποφυγή tweets ψεύτικων λογαριασμών δίνεται η παράμετρος «filter:verified» στα αιτήματα. Επίσης, δίνεται παράμετρος να δεχόμαστε δεδομένα μόνο στην αγγλική γλώσσα. Παράλληλα, συλλέγονται και δευτερεύουσες πληροφορίες όπως τα Likes, τα retweets οι followers του χρήστη που ανέβασε την ανάρτηση. Η αντίστοιχη συνάρτηση εδώ είναι η get_tweet_currency_class. Για παράδειγμα αν δεχθεί ως παράμετρο το «BTC» θα μας επιστραφεί κλάση BtcTweet (που αντιστοιχεί στην βάση δεδομένων σε collection με το ίδιο όνομα).

Η επιθυμία για εύκολη αξιοποίηση των δυνατοτήτων των παραπάνω κλάσεων οδήγησε στην δημιουργία του news_scheduler.py η οποία ακολουθεί υλοποίηση αντίστοιχη με αυτή του historical_data_scheduler.py που εξηγήθηκε στην ενότητα 5.3.1. Η παράμετρος εδώ που μας επιτρέπει να απενεργοποιούμε και να ενεργοποιούμε plugins που αφορούν υλοποιήσεις άντλησης δεδομένων είναι η active_news_tracker. Εκτελώντας το αρχείο αυτό έχουμε συνεχή παροχή δεδομένων τίτλου ειδήσεων και tweets για όλα τα νομίσματα και κρυπτονομίσματα που παρακολουθούνται στο σύστημα μας.

```
['BTC', 'USD']
2021-06-08T23:10:14.786950+03:00 executing operations for: <class 'news_source.plugins.googlenews_tracker.GoogleNewsTracker'>
2021-06-08T23:10:16.387789+03:00 executing operations for: <class 'news_source.plugins.twitter_tracker.TwitterNewsTracker'>
2021-06-08T23:10:37.472484+03:00 executing operations for: <class 'news_source.plugins.googlenews_tracker.GoogleNewsTracker'>
2021-06-08T23:10:38.688056+03:00 executing operations for: <class 'news_source.plugins.twitter_tracker.TwitterNewsTracker'>
```

Σχήμα 5.5 Τα μηνύματα αναφοράς του news_scheduler.py

5.4 Διαχείριση Δικτύων

5.4.1 Σχεδιασμός LSTM δικτύου και διεπαφής με τον χρήστη

Αρχικά, δημιουργήθηκε ένα δίκτυο LSTM με συγκεκριμένες παραμέτρους που επιλέχθηκαν εμπειρικά και μετά από ορισμένες δοκιμές. Το δίκτυο αυτό δημιουργήθηκε στο αρχείο network_manager.py και στην κλάση NetworkManager. Οι παράμετροι στην συνάρτηση αρχικοποίησης είναι οι εξής:

- **policy_class.** Δέχεται μια κλάση που περιλαμβάνει παραμέτρους που σχετίζονται με το ανταλλακτήριο από το οποίο λαμβάνονται τα δεδομένα.

- **market_currency.** Δέχεται ένα αλφαριθμητικό, το οποίο υποδεικνύει το νόμισμα για το οποίο θα γίνει η πρόβλεψη.
- **hybrid_network=False.** Δέχεται μια λογική τιμή, η οποία δηλώνει αν θα χρησιμοποιεί η διαδικασία βελτιστοποίησης των παραμέτρων του νευρωνικού δικτύου μέσω γενετικού αλγορίθμου.
- **timepoint_minutes_duration=5.** Δέχεται μια ακέραια αριθμητική τιμή, που δηλώνει την χρονική απόσταση ανάμεσα σε δύο διαδοχικές τιμές ισοτιμίας.
- **future_timepoints_predict=0.** Σε περίπτωση που η τιμή αυτή είναι μεγαλύτερη του μηδενός, τότε πραγματοποιείται πρόβλεψη για μελλοντικές χρονικές στιγμές (για τις οποίες δεν υπάρχουν δεδομένα) ίσες με την τιμή αυτή.
- **network_profile='presentation'.** Δηλώνει το «κλειδί» του χρησιμοποιημένου δικτύου και χρησιμεύει στην χρησιμοποίηση και δοκιμή διαφόρων διαφορετικών δικτύων.

Στις περιπτώσεις όπου μετά την παράμετρο υπάρχει το σύμβολο '=' και μια τιμή, η τιμή αυτή θεωρείται ως η προκαθορισμένη (default) τιμή της παραμέτρου.

Ο βασικός αλγόριθμος χρησιμοποιεί Bidirectional (αμφίδρομα) LSTM. Η χρησιμοποιούμενη βιβλιοθήκη είναι η TensorFlow. Η υλοποίηση LSTM που επιλέχθηκε είναι η CuDNNLSTM, η οποία δίνει την δυνατότητα ταχύτερης εκτέλεσης των μοντέλων σε κάρτες γραφικών. Το μοντέλο διαθέτει 3 επίπεδα (layers) LSTM. Παράλληλα, χρησιμοποιήθηκε η τεχνική του Dropout και σε κάθε επίπεδο δόθηκε η τιμή 0.15 (15%). Ως windows size δόθηκε η τιμή 128.

```

1. DROPOUT = 0.15
2. number_of_units = 128
3. model = keras.Sequential()
4. model.add(Bidirectional(CuDNNLSTM(number_of_units * 2, return_sequences=True),
5.                               input_shape=(1, SEQ_LEN)))
6. model.add(Dropout(rate=DROPOUT))
7. model.add(Bidirectional(CuDNNLSTM((number_of_units), return_sequences=True)))
8. model.add(Dropout(rate=DROPOUT))
9. model.add(Bidirectional(CuDNNLSTM(number_of_units, return_sequences=False)))
10. model.add(Dense(units=1))
11. model.add(Activation('linear'))

```

Σχήμα 5.5 Δημιουργία βασικού LSTM δικτύου

Βασική συνάρτηση της κλάσης αυτής, μέσω της οποίας ρυθμίζονται σημαντικές παράμετροι είναι η network_execution. Η συνάρτηση αυτή δέχεται ως παραμέτρους τις εξής:

- **real_time_prediction=False.** Δέχεται λογική τιμή. Η τιμή True διαφοροποιεί την διαδικασία εκπαίδευσης και αξιολόγησης. Σκοπός της τεχνικής αυτής είναι να προσομοιωθεί μια ρεαλιστική διαδικασία πρόβλεψης με εκπαίδευση νέων δεδομένων σε πραγματικό χρόνο και προσπάθεια πρόβλεψης μια τιμής. Για παράδειγμα, επιλέγεται σε ένα ήδη εκπαιδευμένο με κάποια δεδομένα δίκτυο, για

κάθε 17 νέες τιμές (στο πέρασμα του χρόνου) που προκύπτουν να γίνεται επανεκπαίδευση του δικτύου με τις 16 και να προβλέπεται η 17ή . Η διαδικασία αυτή επαναλαμβάνεται για αρκετές φορές με σκοπό την εκτίμηση του σφάλματος.

- **train_percentage=0.8.** Η τιμή αυτή δηλώνει το ποσοστό των δεδομένων τα οποία θα χρησιμοποιηθούν για εκπαίδευση. Η τιμή της έχει νόημα μόνο όταν `real_time_prediction=False`.
- **timestamp_points=8000.** Δηλώνει το όριο στις χρονικές στιγμές που θα χρησιμοποιηθούν σ αυτήν την εκτέλεση του δικτύου.
- **loaded_model_path=None.** Η παράμετρος αυτή χρησιμοποιείται σε περίπτωση που θέλουμε να χρησιμοποιηθεί συγκεκριμένο και ήδη αποθηκευμένο δίκτυο.
- **loaded_last_training_timestamp=None.** Η τιμή αυτή φανερώνει την χρονική στιγμή που σταμάτησε την εκπαίδευση το προηγούμενο δίκτυο.
- **dataframe_data=None.** Χρησιμοποιείται για την διαδικασία `real_time_prediction`. Ουσιαστικά μέσω αυτής δίνεται η δυνατότητα για απευθείας ανάθεση των χρονικών στιγμών (μέσω ενός dataframe) που θα χρησιμοποιήσει το δίκτυο.
- **store_statistic_entries=False.** Η τιμή `True` οδηγεί στην αποθήκευση των αποτελεσμάτων των προβλέψεων στην βάση δεδομένων.
- **return_queue=None.** Χρησιμοποιείται για την διαδικασία `real_time_prediction` και στην βελτιστοποίηση του δικτύου μέσω γενετικών αλγορίθμων.
- **model_parameters_list=None.** Χρησιμοποιείται στις περιπτώσεις εκτέλεσης γενετικού αλγορίθμου, αλλά και όταν θέλουμε να θέσουμε εμείς τις παραμέτρους του δικτύου. Δέχεται μια λίστα της μορφής `["window_size", "number_of_units", "extra_layer", "epochs", "batch_size"]`.

Σε «κλασική» λειτουργία του δικτύου (πχ διαχωρισμός των δεδομένων σε 80% εκπαίδευσης και 20% για δοκιμή) οι παράμετροι `loaded_model_path`, `loaded_last_training_timestamp`, `dataframe_data` και `return_queue` έχουν τις default `None` τιμές και η `real_time_prediction` έχει τιμή `False`.

```
1. plugin = NetworkManager(policy_class = CoinbaseProPolicy,
2.                          market_currency = 'BTC',
3.                          hybrid_network = False,
4.                          timepoint_minutes_duration = 5,
5.                          network_profile = 'presentation5')
6.
7. train_percentage=0.8
8.
9. plugin.network_execution(real_time_prediction = False,
10.                         train_percentage = train_percentage,
11.                         timestamp_points = 17000,
12.                         store_statistic_entries = True
13.                         )
```

Σχήμα 5.6 Παράδειγμα εκτέλεσης εφαρμογής με διαχωρισμό δεδομένων (80% για εκπαίδευση – 20% για δοκιμή)

Σε λειτουργία προσομοίωσης «online learning» του δικτύου, η `network_execution` εκτελείται διαδοχικά μέσω της συνάρτησης `train_and_extract_movement_accuracy` της κλάσης `NetworkManager`. Η μόνη παράμετρος που δέχεται αυτή η συνάρτηση είναι ο αριθμός των επαναλήψεων που θέλουμε να εκτελεστούν. Η συνάρτηση χρησιμοποιεί σε ένα ήδη εκπαιδευμένο δίκτυο νέα δεδομένα σε κομμάτια των 17. Σημειώνεται εδώ ότι κάθε «δεδομένο» αποτελεί μια ακολουθία χρονικών στιγμών μεγέθους που καθορίζεται από την παράμετρο `window size`, σύμφωνα με την λειτουργία των LSTM δικτύων. Οι 16 ακολουθίες τιμών χρησιμοποιούνται από το δίκτυο για εκπαίδευση και έπειτα εκτιμάται η τιμή της 17^{ης}. Το ποσοστό ακρίβειας καθορίζεται από το σύνολο των προβλέψεων.

```
1. plugin = NetworkManager(policy_class = CoinbaseProPolicy,  
2.     market_currency = 'BTC',  
3.     hybrid_network = False,  
4.     timepoint_minutes_duration = 5,  
5.     network_profile = 'presentation5')  
6.  
7. plugin.train_and_extract_movement_accuracy(1500)
```

Σχήμα 5.7 Παράδειγμα εκτέλεσης εφαρμογής για τον υπολογισμό της ακρίβειας του σε «πραγματική» λειτουργία

Τέλος, δίνεται η δυνατότητα για πρόβλεψη μελλοντικών χρονικών στιγμών (που δεν είναι γνωστή η τιμή τους για την βάση δεδομένων) μέσω της παραμέτρου `future_timepoints_predict`. Όπως εξηγήθηκε και προηγουμένως η παράμετρος αυτή δείχνει τον αριθμό των προβλέψεων στο μέλλον που επιθυμούμε να γίνουν.

```
14. plugin = NetworkManager(policy_class = CoinbaseProPolicy,  
15.     market_currency = 'BTC',  
16.     hybrid_network = False,  
17.     timepoint_minutes_duration = 5,  
18.     future_timepoints_predict = 1,  
19.     network_profile = 'presentation5')  
20.  
21. train_percentage=None  
22.  
23. plugin.network_execution(real_time_prediction = False,  
24.     train_percentage = train_percentage,  
25.     timestamp_points = 5000  
26. )
```

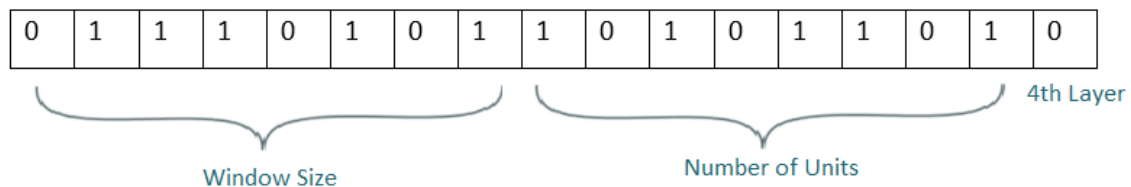
Σχήμα 5.8 Παράδειγμα εκτέλεσης εφαρμογής για τον πρόβλεψη μελλοντικής τιμής

5.4.2 Βελτιστοποίηση LSTM δικτύου με χρήση γενετικών αλγορίθμων

Η ανάγκη για εύρεση των βέλτιστων παραμέτρων με σκοπό την βελτίωση της ακρίβειας πρόβλεψης του δικτύου οδήγησε στην δημιουργία της συνάρτησης `genetic_algorithm` της κλάσης `NetworkManager`. Η συνάρτηση αυτή υλοποιεί γενετικό αλγόριθμο με σκοπό την εύρεση βέλτιστου συνδυασμού παραμέτρων και αρχιτεκτονικής. Στόχος είναι η μεγιστοποίηση του ποσοστού επιτυχημένων προβλέψεων που αφορά την κατεύθυνση της μεταβολής της τιμής.

Ο **πληθυσμός** αντιστοιχεί σε διάφορα νευρωνικά δίκτυα με διαφορετικές παραμέτρους και αρχιτεκτονική.

Ως **χρωμόσωμα** επιλέχθηκε μια λέξη 17 bits. Τα πρώτα 8 bits αντιστοιχούν στην παράμετρο `window size` του LSTM δικτύου. Τα επόμενα 8 bits συνδέονται με τον αριθμό των κρυφών κελιών σε κάθε επίπεδο του LSTM δικτύου. Το 17^ο bit αφορά την ύπαρξη ή όχι 4^{ου} επιπέδου στο LSTM δίκτυο.



Σχήμα 5.9 Αναπαράσταση χρωμοσώματος

Η συνάρτηση αξιολόγησης (**fitness**) υπολογίζει το ποσοστό των σωστών προβλέψεων που κάνει το δίκτυο σε επίπεδο πρόβλεψης κατεύθυνσης της τιμής του συναλλάγματος. Μεγαλύτερο ποσοστό σωστών προβλέψεων αντιστοιχεί και σε καλύτερο δίκτυο. Παράλληλα, υπολογίζει και το μέσο ποσοστιαίο σφάλμα. Το `fitness score` προκύπτει από την διαφορά του ποσοστού επιτυχημένων προβλέψεων με το μισό του μέσου ποσοστιαίου σφάλματος (MAPE). Με τον τρόπο αυτό το δίκτυο μέσω εξέλιξης αποκτά χαρακτηριστικά κατάλληλα για την αύξηση των επιδόσεων του σχετικά με την πρόβλεψη της κατεύθυνσης της τιμής χωρίς να υστερεί σε ακρίβεια πρόβλεψης τιμής.

Οι παράμετροι του γενετικού αλγορίθμου που επιλέχθηκαν είναι :

- Μέγεθος πληθυσμού = 16
- Αριθμός γενιών = 6
- Μέγεθος χρωμοσώματος = 17
- Πιθανότητα μετάλλαξης 0.1


```

1. plugin = NetworkManager(policy_class = CoinbaseProPolicy,
2.                          market_currency = 'BTC',
3.                          hybrid_network = True,
4.                          timepoint_minutes_duration = 5
5.                          network_profile = 'presentation5')
6.
7. plugin.genetic_algorithm(1500)
8.

```

Σχήμα 5.7 Παράδειγμα εκτέλεσης εφαρμογής για τον εύρεση βέλτιστης αρχιτεκτονικής του LSTM δικτύου μέσω γενετικού αλγορίθμου

Αξίζει να σημειωθεί ότι τα αρχεία του δικτύου μπορούν να αποθηκευτούν σε 3 φακέλους. Αν έχουμε απλή εκτέλεση του δικτύου (`hybrid_network=False` και `real_time_prediction=False`) τότε το δίκτυο αποθηκεύεται στον φάκελο `network_manager/network_files/simple_lstm`. Στην περίπτωση που θέλουμε να εκτελέσουμε την υβριδική μέθοδο (`hybrid_network=True`) τότε το δίκτυο αποθηκεύεται στον φάκελο `network_manager/network_files/hybrid_lstm`. Τέλος όταν θέσουμε `real_time_prediction=True` τότε το δίκτυο αποθηκεύεται στον φάκελο `network_manager/network_files/online_lstm`. Κατά την εκτέλεση της εφαρμογής ελέγχεται η ύπαρξη δικτύων (με το ίδιο `network_profile`). Σε περίπτωση εύρεσης κάποιου δικτύου αντίστοιχης λειτουργίας, τότε το δίκτυο διαβάζεται και η εκτέλεση μιας πιθανής εκπαίδευσης συνεχίζεται με το ήδη υπάρχον δίκτυο.

5.4.3 Ανάλυση συναισθήματος αγοράς

Για τους σκοπούς της εκτίμησης του «κλίματος» της αγοράς δημιουργήθηκε η κλάση `MarketClimateEstimator` στο αρχείο `market_climate_estimator.py`. Η κλάση αυτή δέχεται στην συνάρτηση αρχικοποίησης 3 παραμέτρους:

1. Το νόμισμα για το οποίο θέλουμε να κάνουμε την ανάλυση.
2. Την περίοδο του χρόνου για την οποία θέλουμε να βγαίνει κάθε συμπέρασμα.
3. Το νόμισμα αναφοράς.

Αρχικά τα δεδομένα φυσικής γλώσσας υποβάλλονται σε μια προεπεξεργασία. Από αυτά αφαιρούνται ιστοσελίδες, σημεία στίξης, σύμβολα, `hashtags` με το ενδιαφερόμενο νόμισμα και διάφοροι μεταχαρακτήρες (όπως το `/s`). Τα δεδομένα συλλέγονται με αύξουσα χρονική ταξινόμηση.

Για την αξιολόγηση του μοντέλου της ανάλυσης του συναισθήματος της αγοράς και την συσχέτιση του με την εκάστοτε τιμή του νομίσματος δημιουργήθηκε η `historical_sentimental_analysis`. Η συνάρτηση αυτή διαβάζει τα δεδομένα φυσικής γλώσσας από την βάση δεδομένων, τα υποβάλλει σε προεπεξεργασία και στην συνέχεια τα χωρίζει σε κομμάτια διάρκειας χρόνου που ο χρήστης ορίζει. Στην συνέχεια κάθε κομμάτι

υποβάλλεται ξεχωριστά σε ανάλυση συναισθήματος . Στην συνέχεια υπολογίζεται μέσω αυτού το ποσοστό θετικότητας των γνώμων ή συναισθημάτων. Το αποτέλεσμα που προκύπτει αναπαρίσταται σε μια γραφική παράσταση συγχρόνως με την τιμή του νομίσματος. Με τον τρόπο αυτό φανερώνεται η σωστή εκτίμηση του «κλίματος» της αγοράς, καθώς και η άμεση αλληλεξάρτηση αυτής με την τιμή του νομίσματος.

Το μοντέλο εκτίμησης της αγοράς του συστήματος υλοποιήθηκε με την χρήση της βιβλιοθήκης nltk της python. Το εργαλείο sentimental analysis που επιλέχθηκε είναι το VADER sentiment analysis [41]. Είναι ένα ruled-based μοντέλο , το οποίο εξειδικεύεται στην ανάλυση συναισθήματος σε μέσα κοινωνικής δικτύωσης. Δημιουργοί του είναι οι Hutto, C.J. & Gilbert, E.E. (2014). Το μονοπάτι της κλάσης που επιλέχθηκε είναι το nltk.sentiment.vader.SentimentIntensityAnalyzer, το οποίο αναλαμβάνει να επιστρέψει το ποσοστό θετικών, αρνητικών και ουδέτερων «συναισθημάτων» που υπάρχουν σ' ένα κείμενο.

Τέλος, προς χρήση και ευκολία του χρήστη δημιουργήθηκε το αρχείο sentimental_analysis_scheduler.py . Το αρχείο όταν εκτελεστεί από τον χρήστη ελέγχει για ύπαρξη δεδομένων φυσικής γλώσσας στην βάση δεδομένων από την τωρινή στιγμή μέχρι χρονικής στιγμής στο παρελθόν που ρυθμίζεται από το settings.py. Εφόσον υπάρχουν, υλοποιεί ανάλυση συναισθήματος και τα αποτελέσματα εκτυπώνονται στο τερματικό. Η συγκεκριμένη λειτουργία ενδεχομένως να βοηθήσει ανθρώπους που επιθυμούν να πραγματοποιήσουν μετατροπές νομισμάτων παρέχοντας τους συμβουλευτικό χαρακτήρα.

5.5 Ενδεχόμενες μελλοντικές βελτιώσεις

Η παρούσα εφαρμογή θα μπορούσε να δεχτεί αρκετές βελτιώσεις, οι οποίες όμως απαιτούν σημαντικό χρόνο και ξεπερνούν τα πλαίσια αυτής της διπλωματικής εργασίας. Η αρχιτεκτονική όμως που έχει επιλεγεί διευκολύνει τις περισσότερες από τις πιθανές προσθήκες και βελτιώσεις. Οι προσθήκες θα μπορούσε να είναι οι εξής:

- Δημιουργία dashboard για την εκτέλεση όλων των λειτουργιών μέσω αυτού.
- Ενσωμάτωση νέων ανταλλακτηρίων με σκοπό την παρακολούθηση περισσότερων νομισμάτων.
- Παρακολούθηση περισσότερων νομισμάτων από τα ήδη υπάρχοντα ανταλλακτήρια
- Δημιουργία και νέων νευρωνικών δικτύων . Παράλληλη λειτουργία αυτών και συμψηφισμών των αποφάσεων τους (απαιτείται τεράστια επεξεργαστική ισχύ) με σκοπό την επίτευξη ακόμα μεγαλύτερης ακρίβειας στην πρόβλεψης κατεύθυνσης τιμής.
- Δημιουργία κώδικα για την υλοποίηση ενδεχόμενων συναλλαγών (σε demo account και για ακαδημαϊκούς σκοπούς) μέσω των API των ήδη ενσωματωμένων ανταλλακτηρίων

ΚΕΦΑΛΑΙΟ 6

ΓΕΝΙΚΕΣ ΜΕΤΡΗΣΕΙΣ

6.1 Πρόβλεψη ισοτιμίας συναλλαγμάτων με χρήση απλών LSTM δικτύων

6.1.1 Ανάλυση προβλέψεων για την ισοτιμία EUR/USD

Αρχικά, χρησιμοποιήθηκαν τα δεδομένα ισοτιμίας από τις 1 Ιανουαρίου 2015 και ώρα 8 και 5 μετά μεσημβρίας μέχρι 27 Μαΐου του 2020 και ώρα 7 και 50 μετά μεσημβρίας. Οι ώρες αναφέρονται σε GMT. Οι ισοτιμίες σ' αυτήν την χρονική περίοδο έχουν δειγματοληπτηθεί σε πεντάλεπτα. Τα δεδομένα αυτά χωρίστηκαν σε 80% εκπαίδευσης και 20% δοκιμής με τον εξής τρόπο:

- Το σύνολο των δεδομένων από 1 Ιανουαρίου, 2015, 8:05 μ.μ. GMT έως 16 Απριλίου, 2019, 8:25 π.μ. GMT. χρησιμοποιήθηκαν ως δεδομένα εκπαίδευσης (training dataset)
- Το σύνολο των δεδομένων από 16 Απριλίου, 2019, 8:30 π.μ. GMT. έως 27, Μαΐου, 2020, 7:50 μ.μ. GMT αποτελεί το σύνολο δοκιμής (test dataset)

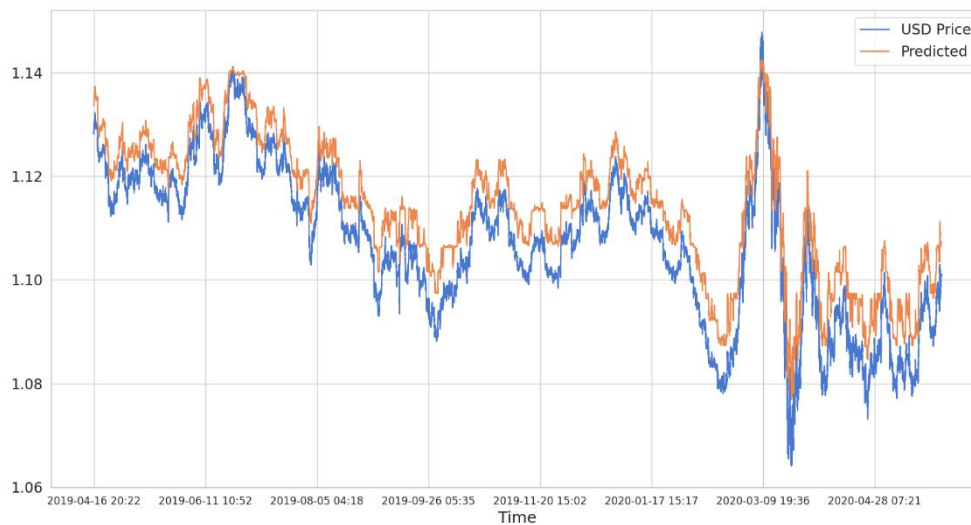
Στα δεδομένα δοκιμής γίνεται προσπάθεια πρόβλεψης της πραγματικής τιμής μέσω ελαχιστοποίησης του δείκτη RMSE (root-mean-square error). Παράλληλα όμως, γίνεται προσπάθεια πρόβλεψης της κατεύθυνσης της ισοτιμίας (άνοδος ή κάθοδος) σε κάθε χρονικό βήμα. Για να γίνει αυτό εφικτό ορίστηκε ένα κατώφλι (threshold) . Αν η προβλεπόμενη τιμή στο επόμενο βήμα ξεπερνά αυτό το κατώφλι τότε θεωρούμε ότι το μοντέλο έκανε πρόβλεψη για άνοδο της τιμής. Στην αντίθετη περίπτωση εκτιμούμε πτώση της τιμής. Όταν έχουμε εξαιρετικά μικρή μετατόπιση (σχεδόν μηδενική) θεωρούμε ότι το μοντέλο μας δεν κάνει πρόβλεψη. Το κατώφλι που επιλέξαμε καθορίζεται από την παρακάτω εξίσωση:

$$\text{threshold} = \frac{(\text{προηγούμενη τιμή πρόβλεψης}) * (1 + \frac{(\text{τελευταία πραγματική τιμή}) - (\text{προτελευταία πραγματική τιμή})}{\text{προτελευταία πραγματική τιμή}})}{2}$$

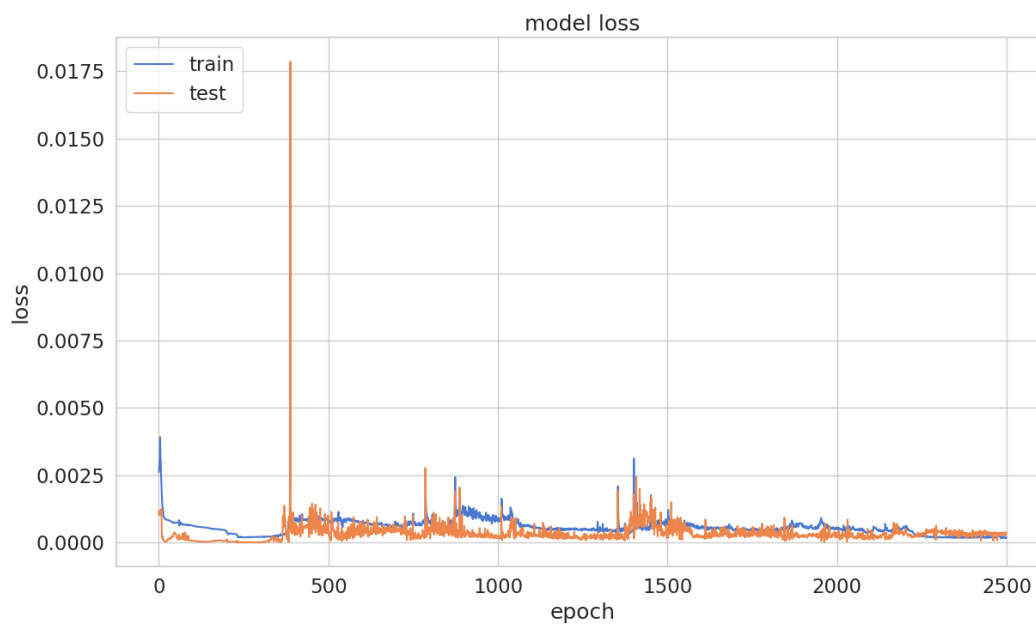
Τα αποτελέσματα των προβλέψεων των δεδομένων δοκιμής ακολουθούν στον επόμενο πίνακα.

RMSE	0.007
True Positive	18636
False Positive	17910
False Negative	18073
True Negative	18765
Precision	50,99%
Recall	50,77%
Accuracy	50,97%
F-measure	50,88%

Σχήμα 6.1 Αποτελέσματα πρόβλεψης ισοτιμίας EUR/USD (80% training-20% testing)



Σχήμα 6.2 Πραγματική τιμή και πρόβλεψη ισοτιμίας EUR/USD



Σχήμα 6.3 Το σφάλμα κατά την διαδικασία της εκπαίδευσης

Στην περίπτωση που θεωρήσουμε ότι το δίκτυο δεν είναι υποχρεωμένο να πραγματοποιεί σε κάθε χρονική στιγμή πρόβλεψη για την άνοδο ή την κάθοδο της τιμής, αλλά μόνο όταν αυτό έχει κάποιας μορφής «αυτοπεποίθηση» τα αποτελέσματα της προηγούμενης δοκιμής διαφοροποιούνται σύμφωνα με τον παρακάτω πίνακα.

Σημειώνουμε ότι ως κριτήριο «αυτοπεποίθησης» θεωρήσαμε η προβλεπόμενη τιμή να απέχει (σε απόλυτη τιμή) από το threshold τουλάχιστον 0,02% της τιμής του threshold.

Τα αποτελέσματα των προβλέψεων των δεδομένων δοκιμής ακολουθούν στον επόμενο πίνακα.

RMSE	0.007
True Positive	4574
False Positive	4344
False Negative	4337
True Negative	4589
Precision	51,29%
Recall	51,33%
Accuracy	51,35%
F-measure	51,31%

Σχήμα 6.4 Αποτελέσματα πρόβλεψης ισοτιμίας EUR/USD (80% training-20% testing , πρόβλεψη μόνο με «αυτοπεποίθηση»

Στην συνέχεια, το υπάρχον εκπαιδευμένο δίκτυο υποβλήθηκε σε διαδικασία online training. Δηλαδή, το δίκτυο που έχει εκπαιδευτεί με δεδομένα από 1 Ιανουαρίου, 2015, 8:05 μ.μ. GMT έως 16 Απριλίου, 2019, 8:25 π.μ. GMT συνεχίζει την εκπαίδευση του με προσομοίωση «πραγματικής κατάστασης». Σε αυτό λαμβάνονται νέα δεδομένα σε ομάδες των 17 ακολουθιών . Οι 16 χρησιμοποιούνται για επανεκπαίδευση του δικτύου και η 17^η για δοκιμή. Η διαδικασία αυτή επαναλαμβάνεται για 700 φορές. Από το σύνολο των προβλέψεων εκτιμάται η ακρίβεια του μοντέλου. Στον πίνακα του σχήματος 6.5 βρίσκονται τα σχετικά αποτελέσματα. Οι ισοτιμίες σ' αυτήν την χρονική περίοδο έχουν δειγματοληπτηθεί σε πεντάλεπτα.

RMSE	0,0074
True Positive	180
False Positive	148
False Negative	178
True Negative	161
Precision	54,88%
Recall	50,28%
Accuracy	51,13%
F-measure	52,48%

Σχήμα 6.5 Αποτελέσματα πρόβλεψης ισοτιμίας EUR/USD (με διαδικασία online training)

6.1.2 Ανάλυση προβλέψεων για την ισοτιμία EUR/BTC

Η πρόβλεψη της τιμής του Bitcoin αποτελεί σαφώς δυσκολότερη διαδικασία, κυρίως λόγω των συχνών ακραίων μεταβολών του. Αυτό συμβαίνει κυρίως εξαιτίας της σχετικά μικρής κεφαλαιοποίησης του σε σχέση με το αμερικανικό δολάριο ή τον χρυσό, καθώς και λόγω της αβεβαιότητας που επικρατεί για τον μέλλον και την εξέλιξη αυτού. Για τους σκοπούς της υπάρχουσας διπλωματικής, χρησιμοποιήθηκαν δεδομένα από τις 1 Ιανουαρίου 2018 και ώρα 8 και 5 μετά μεσημβρίας μέχρι 10 Αυγούστου του 2020 και ώρα 3 και 4 προ μεσημβρίας. Οι ώρες αναφέρονται σε GMT. Ως window size επιλέχθηκε η τιμή 100 και ως τιμή της μεταβλητής num_of_units το 120. Οι ισοτιμίες σ' αυτήν την χρονική περίοδο έχουν δειγματοληπτηθεί σε πεντάλεπτα. Τα δεδομένα αυτά χωρίστηκαν σε 80% εκπαίδευσης και 20% δοκιμής με τον εξής τρόπο:

- Το σύνολο των δεδομένων από 1 Ιανουαρίου, 2018, 12:04 π.μ. GMT έως 2 Νοεμβρίου, 2019, 10:04 μ.μ. GMT. χρησιμοποιήθηκαν ως δεδομένα εκπαίδευσης (training dataset)
- Το σύνολο των δεδομένων από 2 Νοεμβρίου, 2019, 10:09 μ.μ. GMT. έως 10, Αυγούστου, 2020, 3:04 π.μ. GMT αποτελεί το σύνολο δοκιμής (test dataset)

Για την διαδικασία πρόβλεψης κατεύθυνσης της τιμής επιλέχθηκε το ίδιο threshold με την παράγραφο 6.1.1. Ο παρακάτω πίνακας περιλαμβάνει τα αποτελέσματα που προέκυψαν.

RMSE	244,35
True Positive	13837
False Positive	12170
False Negative	12286
True Negative	13463
Precision	53,20%
Recall	52,69%
Accuracy	52,75%
F-measure	53,81%

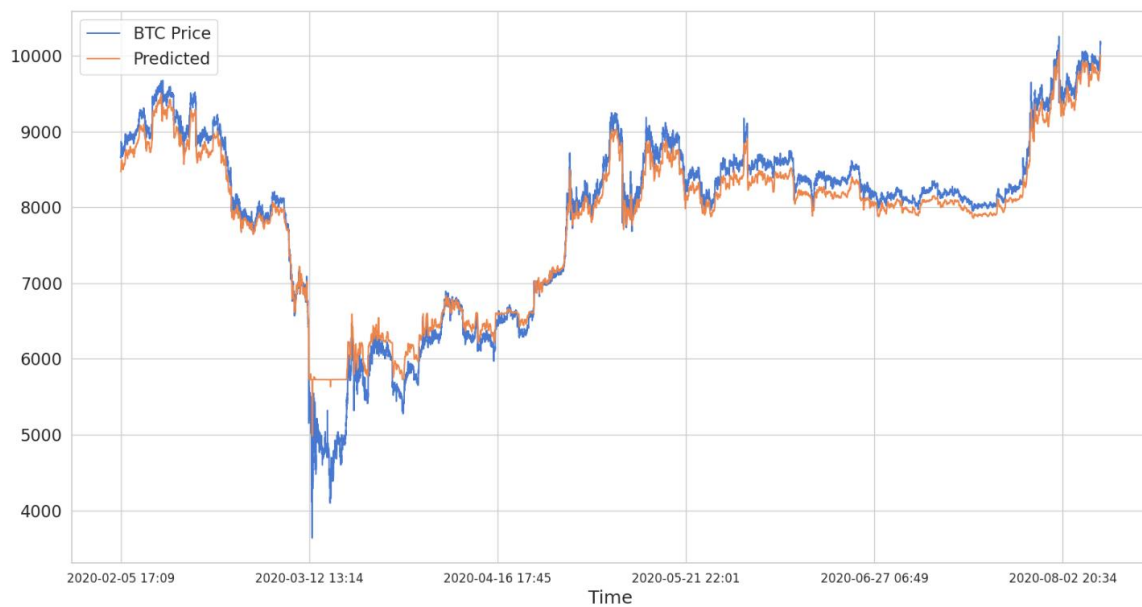
Σχήμα 6.6 Αποτελέσματα πρόβλεψης ισοτιμίας EUR/BTC (80% training-20% testing)

Στην περίπτωση που προσμετρήσουμε μόνο τις περιπτώσεις που το μοντέλο πραγματοποιεί ξεκάθαρη πρόβλεψη (ικανοποιείται το κριτήριο της απόστασης από το threshold) όπως στην ενότητα 6.1.1, με ελάχιστη απόσταση 0,04% του threshold, τα αποτελέσματα που προκύπτουν ακολουθούν στον παρακάτω πίνακα.

True Positive	9806
False Positive	8269
False Negative	8509
True Negative	9590
Precision	54,25%
Recall	53,54%
Accuracy	53,62%
F-measure	53,89%

Σχήμα 6.7 Αποτελέσματα πρόβλεψης ισοτιμίας EUR/BTC (80% training-20% testing , πρόβλεψη μόνο με «αυτοπεποίθηση»

s



Σχήμα 6.8 Πραγματική τιμή και πρόβλεψη ισοτιμίας EUR/BTC



Σχήμα 6.9 Το σφάλμα κατά την διαδικασία της εκπαίδευσης

Στην συνέχεια, το υπάρχον εκπαιδευμένο δίκτυο υποβλήθηκε σε διαδικασία online training, όπως εξηγήθηκε στην προηγούμενη ενότητα. Οι επαναλήψεις και επομένως οι προβλέψεις που έγιναν ήταν 700. Ο παρακάτω πίνακας περιλαμβάνει τις μετρήσεις που προέκυψαν.

True Positive	209
False Positive	156
False Negative	160
True Negative	173
Precision	57,26%
Recall	56,64%
Accuracy	54,72%
F-measure	56,94%

Σχήμα 6.10 Αποτελέσματα πρόβλεψης ισοτιμίας EUR/BTC (με διαδικασία online training)

6.2 Πρόβλεψη ισοτιμίας συναλλαγμάτων με χρήση εξελικτικών LSTM δικτύων (GA-LSTM)

6.2.1 Ανάλυση προβλέψεων για την ισοτιμία EUR/USD

Στην συνέχεια, έγινε προσπάθεια βελτιστοποίησης της αρχιτεκτονικής και των παραμέτρων των LSTM δικτύων, με σκοπό την επίτευξη καλύτερων αποτελεσμάτων όσον αφορά την πρόβλεψη κατεύθυνση τιμής, αλλά και την ελαχιστοποίηση του μέσου ποσοστιαίου σφάλματος. Για τον λόγο αυτό, πραγματοποιήθηκε παράλληλη εκπαίδευση και αξιολόγηση δικτύων χρησιμοποιώντας γενετικό αλγόριθμο, με τον τρόπο που εξηγήθηκε στην παράγραφο 5.4.2. Χρησιμοποιήθηκαν τα ίδια δεδομένα με την ενότητα 6.1.2. Τα δεδομένα χωρίστηκαν σε 80% δεδομένα εκπαίδευσης και 20% δεδομένα δοκιμής. Η αρχιτεκτονική που προέκυψε έχει τα εξής χαρακτηριστικά :

- Window size : 90
- Unit size : 149
- Δεν διαθέτει επιπλέον επίπεδο (layer).

Τα αποτελέσματα που προέκυψαν για τα δεδομένα δοκιμής περιλαμβάνονται στον παρακάτω πίνακα:

RMSE	0.0043
True Positive	18697
False Positive	17818
False Negative	18051
True Negative	18742
Precision	51,20%
Recall	50,88%
Accuracy	51,07%
F-measure	51,04%

Σχήμα 6.11 Αποτελέσματα πρόβλεψης ισοτιμίας EUR/USD (80% training-20% testing)

6.2.2 Ανάλυση προβλέψεων για την ισοτιμία EUR/BTC

Για την βελτιστοποίηση της αρχιτεκτονικής υλοποιήθηκε γενετικός αλγόριθμος με τρόπο παρόμοιο με την ενότητα 6.2.1. Τα δεδομένα που χρησιμοποιήθηκαν είναι τα ίδια με την ενότητα 6.1.2. Η αρχιτεκτονική που προέκυψε έχει τα εξής χαρακτηριστικά :

- Window size : 92
- Unit size : 177
- Δεν διαθέτει επιπλέον επίπεδο (layer).

Τα αποτελέσματα που προέκυψαν για τα δεδομένα δοκιμής περιλαμβάνονται στον παρακάτω πίνακα:

RMSE	79.0999
True Positive	13815
False Positive	12247
False Negative	12308
True Negative	13386
Precision	53,00%
Recall	52,88%
Accuracy	52,56%
F-measure	53,21%

Σχήμα 6.12 Αποτελέσματα πρόβλεψης ισοτιμίας EUR/BTC (80% training-20% testing)

Στην περίπτωση που προσμετρήσουμε μόνο τις περιπτώσεις που το μοντέλο πραγματοποιεί ξεκάθαρη πρόβλεψη (ικανοποιείται το κριτήριο της απόστασης από το threshold – απόσταση μεγαλύτερη από 0,04% του threshold) όπως στην ενότητα 6.1.1 τα αποτελέσματα που προκύπτουν ακολουθούν στον παρακάτω πίνακα.

True Positive	9879
False Positive	8249
False Negative	5856
True Negative	9668
Precision	54,50%
Recall	53,59%
Accuracy	53,80%
F-measure	54,04%

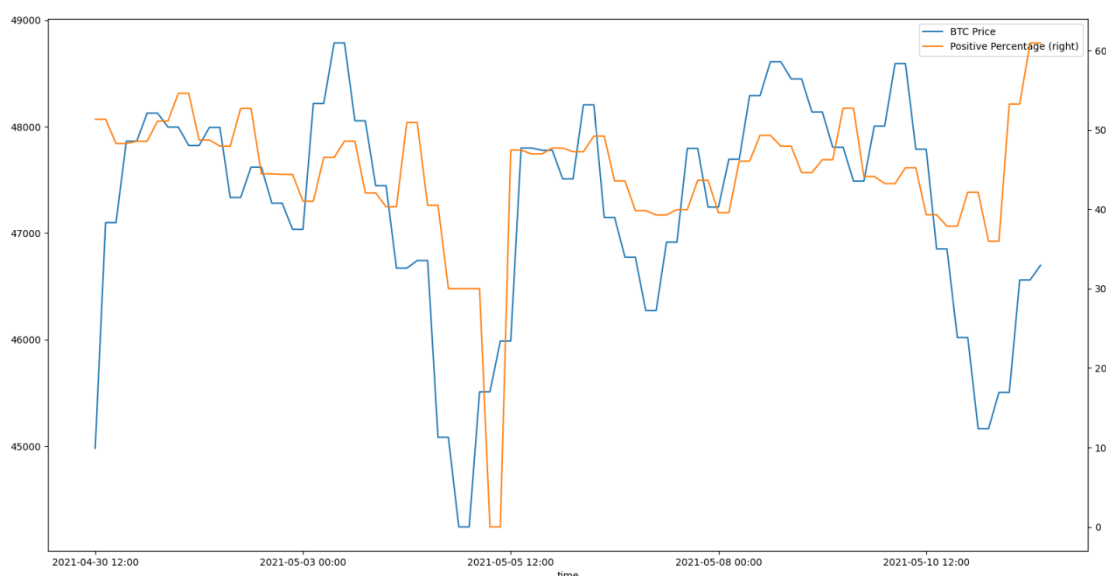
Σχήμα 6.13 Αποτελέσματα πρόβλεψης ισοτιμίας EUR/BTC (80% training-20% testing , πρόβλεψη μόνο με «αυτοπεποίθηση»)

6.3 Πρόβλεψη τάσης ισοτιμίας συναλλαγμάτων με χρήση επεξεργασίας φυσικής γλώσσας

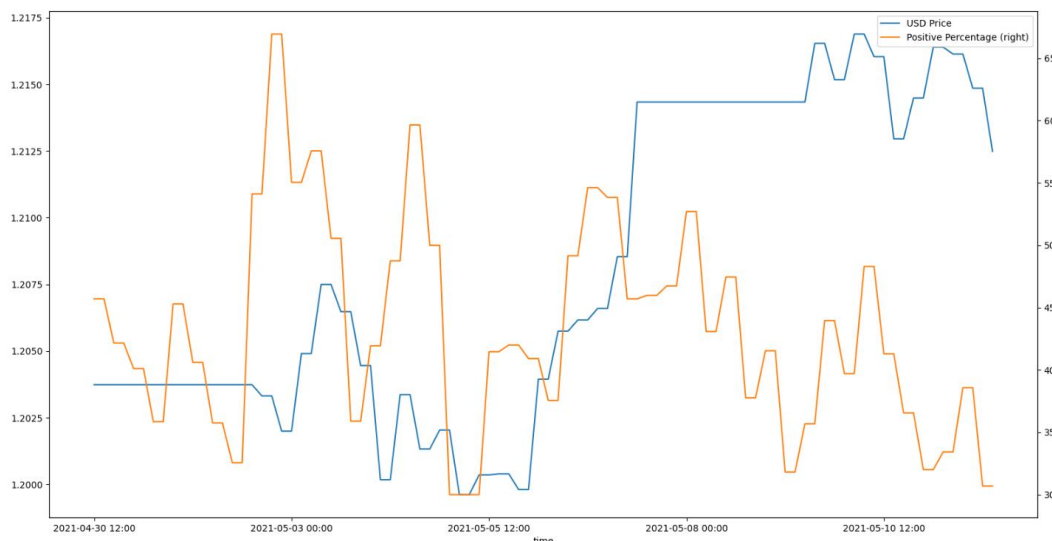
Για την χρήση επεξεργασίας φυσικής γλώσσας, όπως αναφέρθηκε εκτενώς στο κεφάλαιο 5 πραγματοποιήθηκε ανάλυση γνώμης (sentimental analysis) σε δεδομένα τίτλων ειδήσεων και tweets σχετικά με το ενδιαφερόμενο νόμισμα. Στην συνέχεια έγινε προσπάθεια απόδειξης συσχέτισης αυτών με τις μεταβολές του εκάστοτε νομίσματος.

Αρχικά, χρησιμοποιήθηκαν δεδομένα επεξεργασίας κειμένου από το newsapi.org και από το twitter από 30 Απριλίου του 2021 μέχρι 11 Μαΐου του 2021. Η συχνότητα λήψης δεδομένων από το newsapi.org επιλέχθηκε ως ένα αίτημα ανά μία ώρα, ενώ στο twitter επιλέχθηκε συχνότητα λήψης δεδομένων 1 αίτημα ανά 20 λεπτά. Τα tweets που επιστρέφονται σε κάθε αίτημα του API του twitter περιορίζεται στα 200, λόγω περιορισμών του API. Παράλληλα, έγινε προσπάθεια να επιστραφούν αξιόπιστα και αντιπροσωπευτικά tweets μέσω διάφορων παραμέτρων που χρησιμοποιήθηκαν στο API. Τα δεδομένα επιλέχθηκαν να χωριστούν σε ομάδες, με την κάθε ομάδα να καταλαμβάνει 6 ώρες. Για κάθε ομάδα υπολογίζεται το ποσοστό θετικότητας των θετικών αναρτήσεων.

Τα νομίσματα που χρησιμοποιήθηκαν στις μετρήσεις αυτές είναι το Bitcoin (κρυπτονόμισμα) και το αμερικάνικο δολάριο (USD). Στις παρακάτω γραφικές αποτυπώνεται με μπλε χρώμα η τιμή του ενδιαφερόμενου νομίσματος και με πορτοκαλί το ποσοστό θετικότητας των αναρτήσεων αυτών στο ίδιο χρονικό διάστημα σύμφωνα με την διαδικασία που εξηγήθηκε προηγουμένως. Η τιμή του νομίσματος αντιστοιχεί στην τιμή του γ άξονα που βρίσκεται αριστερά και η τιμή της θετικότητας στην τιμή των αναρτήσεων στην τιμή του γ άξονα που βρίσκεται δεξιά.



Σχήμα 6.4 Συσχέτιση θετικότητας ειδήσεων και τιμής του Bitcoin



Σχήμα 6.5 Συσχέτιση θετικότητας ειδήσεων και τιμής του USD

Από τις παραπάνω γραφικές παραστάσεις γίνεται αντιληπτό ότι η τιμή του Bitcoin έχει αρκετά μεγάλη συσχέτιση με την θετικότητα των απόψεων των ειδήσεων και των αναρτήσεων στο twitter. Αυτό συμβαίνει γιατί ως μέσο συναλλάγματος έχει σαφώς μεγαλύτερες διακυμάνσεις από ένα παραδοσιακό και σχετικά σταθερό νόμισμα όπως το αμερικάνικο δολάριο. Παράλληλα, στο διάστημα αυτό που έγιναν οι μετρήσεις αποτελεί επίκεντρο προσοχής πολλών μικρών και μεγάλων επενδυτών λόγω της ραγδαίας αύξησης της τιμής του τον τελευταίο χρόνο. Το γεγονός αυτό διευκολύνει την εύρεση επαρκών αναρτήσεων σχετικά μ' αυτό και επομένως την αποτελεσματικότερη πρόβλεψη. Παρ' όλα αυτά, υπάρχει σαφή συσχέτιση και της θετικότητας των αναρτήσεων του αμερικανικού δολαρίου με την τιμή του. Οι μικρές του μεταβολές όμως και το σαφώς μικρότερο ενδιαφέρον για την τιμή αυτού (στα μέσα κοινωνικής δικτύωσης) καθιστούν τον υπάρχον μηχανισμό συσχέτισης του τιμών συναλλάγματος με το ποσοστό θετικότητας των αναρτήσεων σχετικά μ' αυτό, σαφώς καταλληλότερο για το Bitcoin.

ΒΙΒΛΙΟΓΡΑΦΙΑ

- [1] Νόμισμα (κέρμα) , Βικιπαίδεια . Διαθέσιμο στον δικτυακό τόπο:
[https://el.wikipedia.org/wiki/Νόμισμα_\(κέρμα\)](https://el.wikipedia.org/wiki/Νόμισμα_(κέρμα)) (πρόσβαση 9 Απριλίου 2021).
- [2] Λιανός & Ψειρίδου | Οικονομική Ανάλυση & Πολιτική - Μακροοικονομική. Διαθέσιμο στον δικτυακό τόπο:
https://repository.kallipos.gr/bitstream/11419/1973/1/15305_02_MACRO_chapter_19.pdf (πρόσβαση 9 Απριλίου 2021).
- [3] Blockchain, Βικιπαίδεια . Διαθέσιμο στον δικτυακό τόπο:
<https://el.wikipedia.org/wiki/Blockchain> (πρόσβαση 9 Απριλίου 2021).
- [4] Cryptocurrency, Βικιπαίδεια . Διαθέσιμο στον δικτυακό τόπο:
<https://en.wikipedia.org/wiki/Cryptocurrency> (πρόσβαση 9 Απριλίου 2021).
- [5] Bicoín πληροφορίες ,bitcoin.org . Διαθέσιμο στον δικτυακό τόπο: <https://bitcoin.org/el/> (πρόσβαση 9 Απριλίου 2021).
- [6] Guide: What is Bitcoin and how does it work? , BBC UK . Διαθέσιμο στον δικτυακό τόπο:
<https://www.bbc.co.uk/newsround/25622442> (πρόσβαση 10 Απριλίου 2021).
- [7] Γενετικοί Αλγόριθμοι, Βικιπαίδεια . Διαθέσιμο στον δικτυακό τόπο:
https://el.wikipedia.org/wiki/Γενετικοί_Αλγόριθμοι (πρόσβαση 10 Απριλίου 2021).
- [8] Activation function, Βικιπαίδεια . Διαθέσιμο στον δικτυακό τόπο:
https://en.wikipedia.org/wiki/Activation_function (πρόσβαση 10 Απριλίου 2021).
- [9] Αυτόματη Ταξινόμηση μελωδίας σε μουσικά είδη με τη χρήση νευρωνικών δικτύων, Βιδάλη Ροδάνθη, Μεταπτυχιακή διατριβή – Πανεπιστήμιο Πειραιώς . Διαθέσιμο στον δικτυακό τόπο: <https://dione.lib.unipi.gr/xmlui/bitstream/handle/unipi/4822/Vidali.pdf> (πρόσβαση 10 Απριλίου 2021).
- [9] Αρχιτεκτονική και εκπαίδευση νευρωνικών δικτύων με γενετικούς αλγορίθμους στην πρόγνωση οικονομικών δεδομένων, Δημήτριος Α. Τσορτανίδης, Διπλωματική εργασία – Πανεπιστήμιο Πατρών 2005 . Διαθέσιμο στον δικτυακό τόπο:
https://nemertes.lis.upatras.gr/jspui/bitstream/10889/4489/1/Nimertis_Tsortanidis%28ma%29.pdf (πρόσβαση 15 Απριλίου 2021).
- [10] The 8 Neural Network Architectures Machine Learning Researchers Need to Learn , James Le. Διαθέσιμο στον δικτυακό τόπο: <https://www.kdnuggets.com/2018/02/8-neural-network-architectures-machine-learning-researchers-need-learn.html> (πρόσβαση 15 Απριλίου 2021).
- [11] Recurrent Neural Networks cheatsheet, Afshine Amidi και Shervine Amidi. Διαθέσιμο στον δικτυακό τόπο: <https://stanford.edu/~shervine/teaching/cs-230/cheatsheet-recurrent-neural-networks> (πρόσβαση 15 Απριλίου 2021).

[12] 2.5 Training an Artificial Neural Network, University of Toronto. Διαθέσιμο στον δικτυακό τόπο: <http://www2.psych.utoronto.ca/users/reingold/courses/ai/cache/neural3.html> (πρόσβαση 10 Μαΐου 2021).

[13] Backpropagation, Βικιπαίδεια . Διαθέσιμο στον δικτυακό τόπο: <https://en.wikipedia.org/wiki/Backpropagation> (12 Μαΐου 2021).

[14] What is reinforcement learning? The complete guide, Błażej Osipiński και Konrad Budek - deepsense.ai .Διαθέσιμο στον δικτυακό τόπο: <https://deepsense.ai/what-is-reinforcement-learning-the-complete-guide/> (πρόσβαση 12 Μαΐου 2021).

[15] Stock Market Predictions with LSTM in Python, Thushan Ganegedara - datacamp.com. Διαθέσιμο στον δικτυακό τόπο: <https://www.datacamp.com/community/tutorials/lstm-python-stock-market> (πρόσβαση 12 Μαΐου 2021).

[16] Understanding LSTM Networks. Διαθέσιμο στον δικτυακό τόπο: <http://colah.github.io/posts/2015-08-Understanding-LSTMs/> (πρόσβαση 12 Μαΐου 2021).

[17] Essentials of Deep Learning : Introduction to Long Short Term Memory, . PRANJAL SRIVASTAVA. Διαθέσιμο στον δικτυακό τόπο: <https://www.analyticsvidhya.com/blog/2017/12/fundamentals-of-deep-learning-introduction-to-lstm> (πρόσβαση 12 Μαΐου 2021).

[18] Introduction to Genetic Algorithms — Including Example Code, Vijini Mallawaarachchi. Διαθέσιμο στον δικτυακό τόπο: <https://towardsdatascience.com/introduction-to-genetic-algorithms-including-example-code-e396e98d8bf3> (πρόσβαση 12 Μαΐου 2021).

[19] Evolutionary algorithm, Βικιπαίδεια . Διαθέσιμο στον δικτυακό τόπο: https://en.wikipedia.org/wiki/Evolutionary_algorithm (12 Μαΐου 2021).

[20] Γενετικοί Αλγόριθμοι και Εφαρμογές , Σπυρίδων Λυκοθανάσης. Διαθέσιμο στον δικτυακό τόπο: <https://psifiakoskosmos.files.wordpress.com/2009/12/likothanasis1.pdf> (12 Μαΐου 2021).

[21] Using Genetic Algorithm for Optimizing Recurrent Neural Networks , Aaqib Saeed, University of Twente. Διαθέσιμο στον δικτυακό τόπο: <https://www.kdnuggets.com/2018/01/genetic-algorithm-optimizing-recurrent-neural-network.html> (πρόσβαση 12 Μαΐου 2021).

[22] Εξελικτικά Τεχνητά Νευρωνικά Δίκτυα, Σταμάτιος-Άγγελος Ν. Αλεξανδρόπουλος , Διπλωματική εργασία. Διαθέσιμο στον δικτυακό τόπο: <https://nemertes.lis.upatras.gr/jspui/bitstream/10889/9620/1/THESIS.PDF> (πρόσβαση 12 Μαΐου 2021).

[23] Combining Genetic Algorithms and Neural Networks: The Encoding Problem. Philipp Koehn, University of Tennessee, Knoxville. Διαθέσιμο στον δικτυακό τόπο: <https://homepages.inf.ed.ac.uk/pkoehn/publications/gann94.pdf> (πρόσβαση 12 Μαΐου 2021).

- [24] Neural Network Weight Selection Using Genetic Algorithms. David J. Montana. Διαθέσιμο στον δικτυακό τόπο: <http://davidmontana.net/papers/hybrid.pdf> (πρόσβαση 12 Μαΐου 2021).
- [25] Genetic Algorithm-Optimized Long Short-Term Memory Network for Stock Market . Hyejung Chung και Kyung-shik Shin.
- [26] Χρήση γενετικού αλγορίθμου για βελτιστοποίηση δομής παραμέτρων τεχνητών νευρωνικών δικτύων και εφαρμογή της υβριδικής μεθόδου σε προβλήματα από το χώρο της οικονομίας . Θωμάς Αμοργιανιώτης , Διπλωματική Εργασία – Πανεπιστήμιο Πατρών.
- [27] Analysis of Stock Market using Text Mining and Natural Language Processing. Sheikh Shaugat Abdullah, Mohammad Saiedur Rahaman and Mohammad Saidur Rahman - University of Bangladesh. Διαθέσιμο στον δικτυακό τόπο: https://www.researchgate.net/publication/261212654_Analysis_of_stock_market_using_text_mining_and_natural_language_processing (πρόσβαση 12 Μαΐου 2021).
- [28] What is the difference between stemming and lemmatization? Διαθέσιμο στον δικτυακό τόπο: <https://blog.bitext.com/what-is-the-difference-between-stemming-and-lemmatization/> (πρόσβαση 13 Μαΐου 2021).
- [29] Machine Learning (ML) for Natural Language Processing (NLP) . Διαθέσιμο στον δικτυακό τόπο: <https://www.lexalytics.com/lexablog/machine-learning-natural-language-processing> (πρόσβαση 13 Μαΐου 2021).
- [30] Dropout Regularization in Deep Learning Models With Keras. Jason Brownlee, machinelearningmastery.com . Διαθέσιμο στον δικτυακό τόπο: <https://machinelearningmastery.com/dropout-regularization-deep-learning-models-keras/> (πρόσβαση 13 Μαΐου 2021).
- [31] Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Yoshua Bengio, University of Toronto. Διαθέσιμο στον δικτυακό τόπο: <https://machinelearningmastery.com/dropout-regularization-deep-learning-models-keras/> (πρόσβαση 13 Μαΐου 2021).
- [32] Use Early Stopping to Halt the Training of Neural Networks At the Right Time. Jason Brownlee . Διαθέσιμο στον δικτυακό τόπο: <https://machinelearningmastery.com/how-to-stop-training-deep-neural-networks-at-the-right-time-using-early-stopping/> (πρόσβαση 13 Μαΐου 2021).
- [33] Input Window Size for Deep Recurrent Reinforcement Learning , Eric Muccino. Διαθέσιμο στον δικτυακό τόπο: <https://medium.com/mindboard/input-window-size-for-deep-recurrent-reinforcement-learning-4e5d2960756b> (πρόσβαση 13 Μαΐου 2021).
- [34] Why Data should be Normalized before Training a Neural Network, Timo Stöttner. Διαθέσιμο στον δικτυακό τόπο: <https://towardsdatascience.com/why-data-should-be-normalized-before-training-a-neural-network-c626b7f66c7d> (πρόσβαση 13 Μαΐου 2021).

- [35] Predicting High-Frequency Stock Market by Neural Networks, Timo Stöttner, Imperial College London. Διαθέσιμο στον δικτυακό τόπο:
https://www.imperial.ac.uk/media/imperial-college/faculty-of-natural-sciences/department-of-mathematics/math-finance/Tu_Yuchen_01219050.pdf (πρόσβαση 13 Μαΐου 2021).
- [36] The Complete Guide to Sentiment Analysis. Διαθέσιμο στον δικτυακό τόπο:
<https://getthematic.com/insights/sentiment-analysis/> (πρόσβαση 13 Μαΐου 2021).
- [37] Sentiment Analysis Explained. Διαθέσιμο στον δικτυακό τόπο:
<https://www.lexalytics.com/technology/sentiment-analysis> (πρόσβαση 5 Ιουνίου 2021).
- [38] Sentiment Analysis: A Definitive Guide. Διαθέσιμο στον δικτυακό τόπο:
<https://monkeylearn.com/sentiment-analysis/> (πρόσβαση 5 Ιουνίου 2021).
- [39] Basic Sentiment Analysis using NLTK. Διαθέσιμο στον δικτυακό τόπο:
<https://towardsdatascience.com/basic-binary-sentiment-analysis-using-nltk-c94ba17ae386> (πρόσβαση 5 Ιουνίου 2021).
- [40] ArangoDB, Βικιπαίδεια . Διαθέσιμο στον δικτυακό τόπο:
<https://en.wikipedia.org/wiki/ArangoDB> (5 Ιουνίου 2021).
- [41] Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.
- [42] Sentiment Analysis with Machine Learning: Process & Tutorial. Διαθέσιμο στον δικτυακό τόπο: <https://monkeylearn.com/blog/sentiment-analysis-machine-learning/> (πρόσβαση 5 Ιουνίου 2021).
- [43] Βιβλίο: Αναγνώριση Προτύπων. S.Theodoridis, K.Koutroumbas . Εκδόσεις Πασχαλίδης
- [44] Bidirectional recurrent neural networks, Βικιπαίδεια . Διαθέσιμο στον δικτυακό τόπο:
https://en.wikipedia.org/wiki/Bidirectional_recurrent_neural_networks (10 Ιουνίου 2021).
- [45] Long short-term memory for machine remaining life prediction, Department of Mechanical and Aerospace Engineering, Case Western Reserve University, Cleveland. Jianjing Zhanga, Peng Wanga, Ruqiang Yanb, Robert X. Gao. Διαθέσιμο στον δικτυακό τόπο:
https://www.researchgate.net/publication/325564535_Long_short-term_memory_for_machine_remaining_life_prediction (πρόσβαση 10 Ιουνίου 2021).
- [45] Rule-Based Sentiment Analysis for Financial News . Tan Li Im, Phang Wai San, P. Anthony . Published 2015 Computer Science 2015 IEEE . Διαθέσιμο στον δικτυακό τόπο:
<https://www.semanticscholar.org/paper/Rule-Based-Sentiment-Analysis-for-Financial-News-Im-San/854225e6b089af2b5281801ec83ac9763bf51b01> (πρόσβαση 10 Ιουνίου 2021).
- [46] Γράφημα κεφαλαιοποίησης αγοράς κρυπτονομισμάτων. Διαθέσιμο στον δικτυακό τόπο: <https://coinmarketcap.com/>