ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Τεχνικές πρόβλεψης χρηματιστηριακών τιμών αξιοποιώντας σχόλια από ιστοσελίδες και κοινωνικά μέσα δικτύωσης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Ευάγγελου Ραφτόπουλου

Επιβλέπων: Ανδρέας – Γεώργιος Σταφυλοπάτης Καθηγητής Ε.Μ.Π

Συνεπιβλέπων: Γεώργιος Σιόλας ΕΔΙΠ Ε.Μ.Π

ΕΡΓΑΣΤΗΡΙΟ ΕΥΦΥΩΝ ΣΥΣΤΗΜΑΤΩΝ

Εθνικό Μετσόβιο Πολυτεχνείο

Σχολή Ηλεκτρολόγων Μηχανικών
& Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών

Τεχνικές πρόβλεψης χρηματιστηριακών τιμών αξιοποιώντας σχόλια από ιστοσελίδες και κοινωνικά μέσα δικτύωσης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Ευάγγελου Ραφτόπουλου

Επιβλέπων: Ανδρέας – Γεώργιος Σταφυλοπάτης Καθηγητής Ε.Μ.Π

Συνεπιβλέπων: Γεώργιος Σιόλας ΕΔΙΠ Ε.Μ.Π

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 17η Δεκεμβρίου 2019

……	…….	……..
Σταφυλοπάτης Ανδρέας-Γεώργιος	**Στάμου Γεώργιος**	**Τσανάκας Παναγιώτης**
Καθηγητής ΕΜΠ	Αναπληρωτής Καθηγητής ΕΜΠ	Καθηγητής ΕΜΠ

Αθήνα Δεκέμβριος 2019

……..

Ραφτόπουλος Ευάγγελος
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικών Υπολογιστών Ε.Μ.Π

Περίληψη

Η ραγδαία ανάπτυξη της τεχνολογίας έχει συμβάλει σημαντικά σε πολλούς τομείς της ζωής. Ένας από αυτούς είναι η δυνατότητα προβλέψεων που έχει ανθήσει τα τελευταία χρόνια χάρης των συστημάτων μηχανικής μάθησης και νευρωνικών δικτύων. Η παρούσα διπλωματική εργασία πραγματεύεται το θέμα της πρόβλεψης της μελλοντικής συμπεριφοράς των χρηματιστηριακών τιμών αξιοποιώντας όχι μόνο τιμές του παρελθόντος αλλά και σχόλια από πλατφόρμες κοινωνικής δικτύωσης, με σκοπό την βελτίωση της ακρίβειας των προβλέψεων. Για την επίτευξη αυτού του στόχου κρίθηκε αναγκαία η δημιουργία ενός νευρωνικού δικτύου το οποίο θα προβαίνει σε ανάλυση συναισθήματος, δηλαδή θα αναγνωρίζει πόσο θετικό ή αρνητικό είναι το κάθε σχόλιο, και έπειτα η κατασκευή ενός μοντέλου το οποίο θα είναι υπεύθυνο για τις προβλέψεις των χρηματιστηριακών τιμών και πιο συγκεκριμένα για την τιμή κλεισίματος (close price). Το δίκτυο αυτό θα τροφοδοτηθεί την πρώτη φορά με ιστορικές χρηματιστηριακές τιμές και με τις πληροφορίες των δεδομένων κειμένου και έπειτα μόνο με τις ιστορικές χρηματιστηριακές τιμές. Τέλος θα γίνει σύγκριση μεταξύ των αποτελεσμάτων αυτών με σκοπό να αποφανθεί αν δίνοντας και τα αποτελέσματα της ανάλυσης συναισθήματος των σχολίων βελτιώθηκαν οι προβλέψεις.

Λέξεις κλειδιά

Μηχανική Μάθησης, Προβλέπεις, Χρηματιστηριακές τιμές, Συναισθηματική ανάλυση, επαναληπτικά νευρωνικά δίκτυα, διανυσματική αναπαράσταση λέξεων

Abstract

The rapid development of technology has considerably affected several aspects of life, one of which is the ability to make future predictions. This one has flourished the last years due to the mechanic systems learning and neural networks as well. The thesis in question discusses the prediction of the future behavior of stock market reclaiming not only past prices but also comments of social network platform in order to ameliorate the accuracy of predictions. Therefore, the creation of a neural network was of great necessity since it will proceed to sentiment analysis. That means the network will be able to distinguish the good or bad effect of every comment and then construct a model which will be responsible for the prediction of stock market and, particularly, it will predict the close price of a share. At first, this network will feed with historical stock prices as well as with information taken from text data and then only with the former ones. Finally, there will be a comparison between the results aiming to reach a conclusion if, when giving the results of both the sentiment analysis and the comments, the predictions were improved.

Key words

Machine learning, predictions, stock prices, sentiment analysis, recurrent neural networks, word embeddings

Ευχαριστίες

Η παρούσα διπλωματική εργασία εκπονήθηκε στα πλαίσια του προπτυχιακού κύκλου σπουδών της σχολής Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών και πραγματοποιήθηκε σε συνεργασία με το Εργαστήριο Ευφυών Συστημάτων του τομέα Τεχνολογίας Πληροφορικής και Υπολογιστών. Αρχικά, θα ήθελα να ευχαριστήσω τον καθηγητή κ. Ανδρέα Σταφυλοπάτη, Καθηγητή Ε.Μ.Π, ο οποίος ήταν ο επιβλέπων της εργασίας μου. Ακολούθως, θα ήθελα να ευχαριστήσω θερμά τον Κ. Γεώργιο Σιόλα ο οποίος μου πρόσφερε την κάθε δυνατή βοήθεια και σωστή καθοδήγηση καθ΄ όλη την διάρκεια της διπλωματικής. Τέλος, θα ήθελα να ευχαριστήσω τους φίλους μου για την υποστήριξη που μου πρόσφεραν καθ’ όλη την διάρκεια των σπουδών μου.

Περιεχόμενα

Κεφάλαιο 1

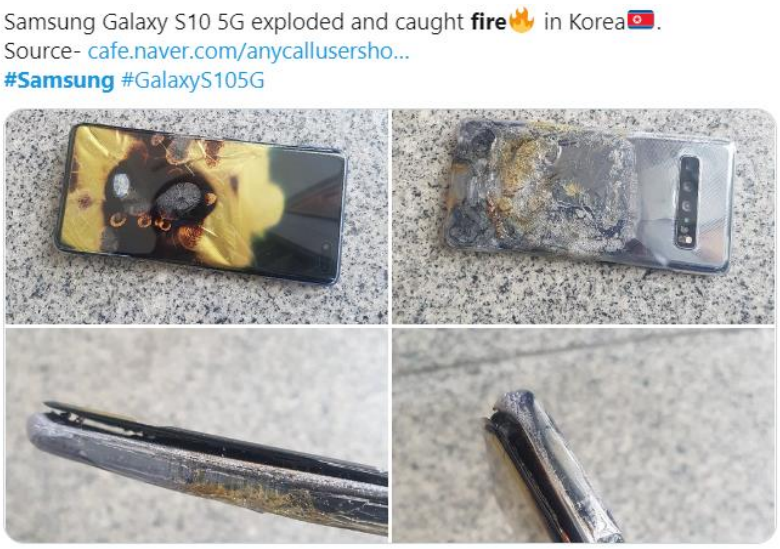
1. Εισαγωγή

Ποιος δεν θα ήθελε να προβλέψει το μέλλον ; ο άνθρωπος έχει την τάση να επιδιώκει να μάθει τι του επιφυλάσσει το μέλλον. Είτε πρόκειται για την απλή πρόβλεψη του καιρού μέχρι την πρόβλεψη του χρηματιστηρίου. Πλέον υπάρχουν οι κατάλληλες τεχνολογίες και προϋποθέσεις ώστε να προβεί σε βάσιμες εκτιμήσεις που αφορούν το άμεσο μέλλον. Σε αυτή την εργασία θα προσπαθήσουμε να προβλέψουμε τις μελλοντικές τιμές του χρηματιστηρίου και την γενική του συμπεριφορά του στο άμεσο μέλλον. Πριν κάποια χρόνια θα έλεγε κανείς ότι αυτό είναι ακατόρθωτο, αλλά την σήμερον ημέρα έχουν αναπτυχθεί τόσο η επιστήμη και η τεχνολογία που δεν φαντάζει πλέον τόσο δύσκολο. Ένας τομέας της τεχνολογίας που έχει ανθίσει τα τελευταία χρόνια είναι τα νευρωνικά δίκτυα και η τεχνητή νοημοσύνη. Έχουν κατασκευαστεί πληθώρα αλγορίθμων μηχανικής μάθησης οι οποίοι είναι σε θέση να επιλύουν πολλά προβλήματα που το ανθρώπινο μυαλό δεν μπορεί να επιλύσει κυρίως λόγω του όγκου των δεδομένων. Ένα από αυτά τα προβλήματα που οι αλγόριθμοι μηχανικής μάθησης έχουν βοηθήσει σημαντικά είναι η πρόβλεψη χρονοσειρών.

Οι άνθρωποι από το παρελθόν μέχρι σήμερα έχουν την τάση να προβαίνουν σε ενέργειες αγοροπωλησιών κινητών αξιών όπως είναι μερίδια κεφαλαίων (μετοχές), τραπεζικά ομόλογα και άλλα είδη εμπορευμάτων. Δεν είναι λίγοι εκείνοι οι οποίοι επενδύουν μεγάλα ποσά κεφαλαίων για την αγορά μετοχών κάποιων εταιριών[1]. Για αυτό τον λόγο η πρόβλεψη της συμπεριφοράς των χρηματιστηριακών μετοχών έχει προσελκύσει μεγάλο αριθμό ερευνητών. Πλέον με την εξέλιξη της τεχνολογίας και της επιστήμης ανακαλύπτονται συνέχεια νέοι μέθοδοι για την επίτευξη αυτού του στόχου. Πολλοί μέθοδοι βασίζονται σε στατιστικές και μαθηματικές μελέτες και άλλοι πιο σύγχρονοι σε τεχνικές μηχανικής μάθησης και νευρωνικά συστήματα.

Στην παρούσα εποχή τα νευρωνικά δίκτυα χρησιμοποιούνται για την αντιμετώπιση πολλών προβλημάτων. Ένα τεχνητό νευρωνικό δίκτυο μπορεί να βοηθήσει σε περιπτώσεις που υπάρχουν μη γραμμικές διαδικασίες, όπου η συσχέτιση δεν είναι γνωστή εξ΄ αρχής και άρα είναι δύσκολο να επιτευχθεί βέλτιστη προσαρμογή. Η κύρια ιδέα των νευρωνικών δικτύων είναι το φιλτράρισμα των εισόδων, που αποτελούν και τις ανεξάρτητες μεταβλητές, μέσω ενός ή περισσοτέρων κρυφών επιπέδων, πρώτου παραχθεί η τελική έξοδος. Τα νευρωνικά δίκτυα έχουν εφαρμοσθεί σε διάφορες πτυχές των προβλέψεων, από απευθείας παραγωγή προβλέψεων έως την βελτιστοποίηση συγκεκριμένων παραμέτρων[2].

Οι ερευνητές πέρα από την αξιοποίηση της συμπεριφοράς του παρελθόντος των χρηματιστηριακών μετοχών προσπαθούν να συνδυάσουν και άλλα δεδομένα στην προσπάθειά τους να παράξουν καλύτερα αποτελέσματα[1]. Μία προσπάθεια είναι με την αξιοποίηση νέων και σχολίων από ειδικούς ή και απλούς ανθρώπους. Για παράδειγμα αν διαβάσουμε το σχόλιο



Εικόνα 1: tweet που έχει σχέση με την εταιρία Samsung

τότε το πιο πιθανό σενάριο είναι οι μετοχές της εταιρίας Samsung να έχουν καθοδική πτώση στο άμεσο μέλλον. Αντίθετα, αν μεγάλος αριθμός ανθρώπων σχολιάσει ή δημοσιεύσει στα μέσα κοινωνικής δικτύωσης θετικές απόψεις για προϊόντα ή υπηρεσίες κάποιας εταιρίας τότε είναι πιθανό οι μετοχές αυτής της εταιρίας να έχουν ανοδική πορεία. Όπως έχει αποδειχθεί σε διάφορες έρευνες με την αξιοποίηση τέτοιου είδους πληροφοριών όπως είναι τα συναίσθημα και η διάθεση του κοινού, μπορούμε να πετύχουμε πιο ακριβή πρόβλεψη[1].

Σκοπός αυτής της ερευνητικής εργασίας είναι η πρόβλεψη της συμπεριφοράς και συγκεκριμένα της τιμής κλεισίματος (close) των μετοχών κάποιων γνωστών μεγάλων εταιριών όπως είναι η apple, google, microsoft κλπ αξιοποιώντας τόσο τα ιστορικά δεδομένα των χρηματιστηριακών τους τιμών όσο και σχόλια χρηστών που έγιναν κάποιες μέρες πριν την ημέρα που προσπαθούμε να προβλέψουμε.

Προκειμένου να πραγματοποιήσουμε προβλέψεις με τις ποσοτικές μεθόδους είναι αναγκαίο να έχουμε στη διάθεση μας μεγάλο όγκο πληροφοριών αλλά και να θεωρήσουμε ότι το μέλλον θα λειτουργεί όπως και το παρελθόν. Υπάρχουν δύο υποκατηγορίες ποσοτικών μεθόδων πρόβλεψης: το μοντέλο χρονοσειρών (time series model) και το αιτιοκρατικό ή επεξηγηματικό μοντέλο (causal relationship).

Το μοντέλο χρονοσειρών, είναι το πιο διαδεδομένο είδος ποσοτικού μοντέλου πρόβλεψης. Για την εφαρμογή του πρέπει να υπάρχουν ιστορικά στοιχεία για το μέγεθος που θα επιχειρηθεί να προβλεφθεί. Το μοντέλο χρονοσειρών βασίζεται στην υπόθεση ότι η μεταβολή της τιμής του μεγέθους ακολουθεί ένα συγκεκριμένο λανθάνον πρότυπο που επαναλαμβάνεται στο χρόνο και παραμένει σταθερό. Οι προβλέψεις παράγονται με την αναγνώριση αυτού του προτύπου και την προέκταση του στο μέλλον. Παράδειγμα τέτοιων μεθόδων είναι η εξομάλυνση (smoothing), η αποσύνθεση (decomposition), και οι αυτοπαλινδομικές μέθοδοι[3].

Το αιτιοκρατικό μοντέλο στηρίζεται στην βασική υπόθεση ότι υπάρχει μια σταθερή σχέση μεταξύ του υπό πρόβλεψη μεγέθους (εξαρτημένη μεταβλητή) και ορισμένων παραμέτρων (ανεξάρτητη μεταβλητή) που το επηρεάζουν. Το πιο σημαντικό πλεονέκτημα των αιτιοκρατικών μεθόδων είναι ότι προσφέρουν στον χρήστη την δυνατότητα να προβλέψει την μελλοντική τιμή κάποιου μεγέθους, για διάφορους συνδυασμούς των μεταβλητών εισόδου[3].

Στα πλαίσια αυτής της εργασίας χρησιμοποιήθηκαν μέθοδοι μηχανικής μάθησης και νευρωνικών δικτύων για την πρόβλεψη της συμπεριφοράς των χρηματιστηριακών τιμών. Κατασκευάστηκαν τρία μοντέλα εκ των οποίων τα δύο επιτελούν την ίδια λειτουργία με διαφορετικό τρόπο. Αρχικά, πρώτο βήμα ήταν η δημιουργία ενός αποτελεσματικού μοντέλου το οποίο θα πραγματοποιεί ανάλυση συναισθήματος (sentiment analysis) στα δεδομένα κειμένων. Πιο συγκεκριμένα, θα δέχεται ως είσοδο ένα σχόλιο και θα παράγει ένα ακέραιο αριθμό από το -1 έως το 1 αναλόγως το πόσο θετικό ή αρνητικό αντίστοιχα είναι το σχόλιο αυτό. Όσο πιο κοντά στο -1

βρίσκεται η τιμή αυτή τόσο πιο αρνητικό είναι το σχόλιο ενώ όσο πιο κοντά στο 1 βρίσκεται τόσο πιο θετικό είναι.

Το δεύτερο βήμα ήταν η δημιουργία ενός μοντέλου το οποίο με βάση μόνο ιστορικά δεδομένα χρηματιστηριακών τιμών όπως είναι οι τιμές close, open, high, low θα προσπαθήσει να προβλέψει τις τιμές κλεισίματος (close) στο άμεσο μέλλον. Τέλος, αυτό το μοντέλο δοκιμάσαμε να το τροφοδοτήσουμε και με το αποτέλεσμα του μοντέλου που πραγματοποιεί την ανάλυση συναισθήματος (sentiment analysis) με σκοπό να προβούμε σε καλύτερες προβλέψεις. Αν όντως βελτιωθεί η ακρίβεια των προβλέψεων τότε θα έχουμε πετύχει τον σκοπό αυτής της διπλωματικής εργασίας.

1.1 Οργάνωση Εργασίας

Στο κεφάλαιο 1 δόθηκε μια σύντομη εισαγωγή και περιγραφή του προβλήματος καθώς ορίστηκαν και οι στόχοι που προσπαθούμε να πετύχουμε.

Στο κεφάλαιο 2 περιέχονται συγγενείς εργασίες, τρόποι και μέθοδοι που χρησιμοποιούν. Είναι χωρισμένο σε τρία μέρη: αρχικά, αφορά προβλέψεις χρηματιστηριακών τιμών αξιοποιώντας ιστορικές τιμές και πληροφορίες από σχόλια, έπειτα εργασίες οι οποίες στοχεύουν στην ανάλυση συναισθήματος και τέλος εργασίες που προβλέπουν τις μελλοντικές χρηματιστηριακές τιμές χρησιμοποιώντας μόνο τις παρελθοντικές χρηματιστηριακές τιμές.

Στο κεφάλαιο 3 παρουσιάζονται και επεξηγούνται κάποιες έννοιες των Νευρωνικών Δικτύων και των Αλγορίθμων Μηχανικής Μάθησης που χρησιμοποιήθηκαν στα πλαίσια αυτής της εργασίας.

Στο κεφάλαιο 4 δίνονται τα νευρωνικά δίκτυα που κατασκευάστηκαν και χρησιμοποιήθηκαν για την ανάλυση συναισθήματος και την πρόβλεψη των χρηματιστηριακών τιμών. Επιπλέον, γίνεται η ανάλυση των dataset που χρησιμοποιήθηκαν.

Στο κεφάλαιο 5 παρουσιάζονται και επεξηγούνται τα αποτελέσματα που παράχθηκαν, ενώ στο κεφάλαιο 6 αναφέρονται τα συμπεράσματα που καταλήξαμε. Τέλος, στο κεφάλαιο 7 τονίζονται κάποιες πιθανές μελλοντικές επεκτάσεις της εργασίας.

Κεφάλαιο 2

2. Συγγενείς εργασίες

Στο κεφάλαιο αυτό θα παραθέσουμε συνοπτικά την ερευνητική δουλειά που έχει γίνει σε πεδία συναφή με αυτή την εργασία. Αναλυτικότερα, θα χωρίσουμε αυτό το κομμάτι σε τρία μέρη. Αρχικά, θα γίνει μια αναφορά σε ερευνητικές εργασίες οι οποίες είχαν ως στόχο να προβλέψουν την μελλοντική συμπεριφορά των τιμών του χρηματιστηρίου αξιοποιώντας τόσο τις προηγούμενες τιμές του παρελθόντος όσο και σχόλια, νέα και πληροφορίες στον τομέα αυτό. Στην συνέχεια, θα παρουσιαστούν ερευνητικά έργα που είχαν ως στόχο να πετύχουν ικανοποιητικά αποτελέσματα στην ανάλυση συναισθήματος (sentiment analysis) και οι μέθοδοι που χρησιμοποίησαν για να πετύχουν τον στόχο τους. Τέλος, θα αναφερθούν μέθοδοι που στόχευαν στην πρόβλεψη των μελλοντικών τιμών χρησιμοποιώντας μόνο ιστορικές χρηματιστηριακές τιμές.

2.1 Πρόβλεψη τιμών αξιοποιώντας ιστορικές τιμές και πληροφορίες κειμένων

Η ανάλυση και η πρόβλεψη της χρηματιστηριακής αγοράς είχε πάντα μεγάλο ενδιαφέρον από τον ακαδημαϊκό κόσμο. Έχουν προταθεί αρκετές διαφορετικές προσεγγίσεις, από την πρόβλεψη της χρονοσειράς (time series prediction) έως την ανάλυση κειμένων από ειδήσεις ακόμα και ανάλυση των κοινωνικών δικτύων. Αρχικά, οι επιστημονικές έρευνες βασίστηκαν στην Efficient Market Hypothesis (EMH) (Farma, 1965)[4], σύμφωνα με την οποία οι τιμές των μετοχών αντανακλούν όλες τις σχετικές πληροφορίες που είναι διαθέσιμες ανά πάσα στιγμή. Σε ένα τέτοιο μοντέλο αν ένας άνθρωπος επιθυμεί να επενδύσει σε μία μετοχή είτε με την μέθοδο της τεχνικής πρόβλεψης των μελλοντικών τιμών βάση της μελέτης των τιμών του παρελθόντος, είτε με τυχαία επιλογή, θα έχει τον ίδιο βαθμό κινδύνου. Με άλλα λόγια είναι αδύνατο να προβλέψουμε την μελλοντική συμπεριφορά του χρηματιστηρίου. Ωστόσο, τις τελευταίες δεκαετίες, ένα μεγάλο πλήθος εργασιών έχει απορρίψει αυτή την υπόθεση (Qian and Rasheed, 2007)[5] δείχνοντας ότι οι τιμές των μετοχών ακολουθούν τυχαίες πορειές μόνο για σύντομες χρονικές περιόδους και κατά συνέπεια υποστηρίζουν ότι γενικά θα μπορούσαν να προβλεφθούν.

Δύο κύριες προσεγγίσεις έχουν πραγματοποιηθεί για την πρόβλεψη της συμπεριφοράς των χρηματιστηριακών τιμών. Η μια χρησιμοποιεί μόνο ιστορικά δεδομένα όπως χρηματιστηριακούε δείκτες (stock index prices) (Atsalakis and Valavanis, 2009) και η άλλη χρησιμοποιεί επιπλέων και σχετικές ειδήσεις και πληροφορίες κειμένων για την πρόβλεψη των τάσεων τους (Mittermayer and Knolmayer, 2006)[6]. Άλλες έρευνες χρησιμοποιούν blog posts για την πρόβλεψη της συμπεριφοράς των χρηματιστηριακών τιμών, προσδιορίζοντας την συσχέτιση μεταξύ της δραστηριότητας σε blogs και μέσα κοινωνικής δικτύωσης και των μεταβολών στις τιμές των μετοχών αλλά και τον όγκο των συναλλαγών (Antweiler and Frank, 2004)[7]. (Gilbert and Karahalios, 2010)[8] δημιούργησαν έναν δείκτη διάθεσης (Anxiety Index) χρησιμοποιώντας πάνω από 20 εκατομμύρια δημοσιεύσεις από την ιστοσελίδα του LiveJournal, και όταν ο δείκτης αυξήθηκε σημαντικά το S&P 500 πρόβλεψε την τιμή του κλεισίματος ελάχιστα πιο χαμηλά από την αναμενόμενη.

Το twitter αποτελεί μια τεράστια βάση δεδομένων η οποία παρέχει πληροφορίες για πολλά διαφορετικά θέματα. Μπορεί να υποστηριχθεί ότι αυτή η βάση μπορεί να αποτελέσει ένδειξη για την διάθεση των ανθρώπων (public mood). Πάνω σε αυτό στηρίχθηκε η εργασία (bollen et al. 2011a)[1] όπου οι συγγραφείς της κατέληξαν στο γεγονός ότι κοινωνικά, πολιτικά, πολιτιστικά και οικονομικά γεγονότα έχουν άμεσο και σημαντικό αντίκτυπο στην δημόσια διάθεση στο twitter. Η δημόσια διάθεση έχει χρησιμοποιηθεί για την πρόβλεψη πολλών φαινομένων όπως είναι οι πωλήσεις μιας ταινίας. Η εργασία (Bollen et al.,2011b)[1] μέτρησε δείκτες διάθεσης (mood states) όπως positive, negative, calm, alert, sure, vital, kind and happy μέσω ανάλυσης συναισθήματος σε πάνω από 9 εκατομμύρια tweets που δημοσιεύτηκαν το 2008. Αυτά τα tweets φιλτραρίστηκαν από μερικές εκφράσεις όπως “I am feeling” και όχι μόνο εκείνα που σχετίζονται με το χρηματιστήριο. Ανέλυσαν τα tweets με δύο mood tools: OpinionFinder το οποίο μετράει το πόσο θετικό ή αρνητικό είναι ένα tweet

και το Google-Profile of Mood States (GPOMS) που μετράει την διάθεση σε έξι διαστάσεις. Καταλήξαν ότι το calm mood προσφέρει το καλύτερο αποτέλεσμα πρόβλεψης για το Dow Jones Industrial Average (DJIA) με ακρίβεια 86,7% σε πρόβλεψη κάθε μέρας του Δεκεμβρίου και έδειξαν ότι tweets 3ων ημερών στο παρελθόν δίνει το καλύτερο αποτέλεσμα. Μία παρόμοια έρευνα έκαναν και (Mittal and Goel, 2012)[9] όπου χρησιμοποίησαν μόνο τα συναισθήματα (calm, happy, alert, kind).

Επιπρόσθετα, τέσσερις άλλοι μέθοδοι τεχνικής μάθησης είναι οι Linear Regression, Support Vector Machine (SVMs), Logistic Regression Και SOFNN που χρησιμοποιήθηκαν για την πρόβλεψη ανόδου ή καθόδου μετοχών. Από τα παραπάνω το μοντέλο SOFNN έδωσε τα καλύτερα αποτελέσματα με ακρίβεια 76%. Οι (Oliveira et al., 2013)[10] ασχολήθηκε με κάποια διάσημα λεξικά όπως είναι τα Harvard general inquirer, Opinion Lexicon, MPQA Subjectivity Lexicon, SentiWordNet, Emoticons με σκοπό να επιτεύξουν μια καλύτερη ανάλυση συναισθήματος.

Μία άλλη προσπάθεια χρησιμοποιώντας LSTM έγινε από τους (Jiahong Li, Hui bu, Junjie Wu)[11] οι οποίοι εκμεταλλεύτηκαν τόσο post σε forums όσο και σε ιστορικές τιμές μετοχών χρηματιστηρίων για να προβλέψουν μελλοντικές τιμές. Το μοντέλο τους αρχικά τροφοδοτείται με κείμενα από φόρουμ και κάνοντας χρήση Naïve Bayes τα κατατάσσει σε τρις κατηγορίες θετικά, αρνητικά και ουδέτερα αναλόγως αν το περιεχόμενό τους είναι θετικό, αρνητικό ή ουδέτερο όσον αφορά το συναίσθημα. Έπειτα για κάθε μέρα μετράει το sentiment score τους και τροφοδοτεί σε ένα LSTM τα αποτελέσματα αυτά καθώς επίσης και ιστορικές τιμές των χρηματιστηριακών τιμών. Μετά από εκπαίδευση του μοντέλου είναι σε θέση να προβλέψει μελλοντικές τιμές κλεισίματος (close) με ακρίβεια 87.86%.

Ένα άλλο μοντέλο που χρησιμοποιείται για την πρόβλεψη μελλοντικών χρηματιστηριακών τιμών είναι το Support Vector Machine. Οι L. Yu et al. (2005, 2009)[12] χρησιμοποίησαν SVM και την μέθοδο ελαχίστων τετραγώνων και χρησιμοποιώντας γενετικούς αλγορίθμους για την επιλογή των παραμέτρων του συστήματος και κατάφερε να παράξει ικανοποιητικά αποτελέσματα.

2.2 Μέθοδοι Ανάλυση συναισθήματος

Η έννοια της Συναισθηματικής Ανάλυσης (Sentiment Analysis) συναντάται συχνά στην ξένη βιβλιογραφία και ως Εξόρυξη ΄Άποψης (Opinion Mining). Ο όρος συναισθηματική ανάλυση σύμφωνα με τον (Bing Liu 2015)[13] συναντάτε πρώτη φορά κατά πάσα πιθανότητα στην εργασία των Nasukawa and Yi (2003)[14].

Η ανάλυση συναισθήματος μπορεί να κατηγοριοποιηθεί σε δύο άξονες. Αρχικά πρωταρχικός στόχος των ερευνητών ήταν να μπορέσουν να συμπεράνουν αν το περιεχόμενο των δεδομένων κειμένου ήταν θετικό ή αρνητικό δηλαδή το πρόβλημα ήταν δυικό. Τα τελευταία χρόνια, γίνεται προσπάθεια με σκοπό να μπορούν να αποφανθούν και για άλλες διαστάσεις του συναισθήματος όπως είναι το άγχος, η ηρεμία κλπ όπως έγινε από τον Bollen[1].

Η πλειοψηφία της μέχρι στιγμής έρευνας έχει επικεντρωθεί στην πρόβλεψη δυαδικών ετικετών σε δεδομένα κειμένου (πχ περιέχει χαρά - δεν περιέχει χαρά, περιέχει φόβο - δεν περιέχει φόβο, κτλ). Το 2005 οι (Cecilia Ovesdotter Alm, Dan Roth, Richard Sproat)[15] χρησιμοποιώντας Επιβλεπόμενη Μάθηση και ακολουθώντας την αρχιτεκτονική εκμάθησης SNoW, μη έχοντας αρκετά δεδομένα προφορικού λόγου ώστε να προβλέψουν συναίσθημα στην ομιλία, όπως αρχικά σκόπευαν, ξεκίνησαν με δεδομένα γραπτού λόγου και προσπάθησαν να τα ομαδοποιήσουν με βάση τον βαθμό έντασης βασικών συναισθημάτων, χρησιμοποιώντας Bag of Words (BoW) καθώς και γλωσσικά χαρακτηριστικά. Το 2013 μία ομάδα ερευνητών από το University of Washington, Seattle έχοντας συλλέξει επί 4 χρόνια μηνύματα από chat που έχουν ανταλλαχθεί σε επιστημονικές συνεργασίες, στοχεύουν στην αυτόματη εξαγωγή μίας ετικέτας που να προβλέπει κάθε φορά το είδος των συναισθημάτων που περιγράφουν καλύτερα το εκάστοτε μήνυμα. Η διαδικασία δημιουργίας των πραγματικών ετικετών για τα δεδομένα έγινε χειροκίνητα από την ίδια την ομάδα, καταλήγοντας σε 12 συναισθήματα.

Επιπλέον, η επιστημονική κοινότητα έχει δείξει αρκετό ενδιαφέρον στην σύνεση της εξαγωγής συναισθήματος από τα Μέσα Κοινωνικής Δικτύωσης για λόγους κέρδους όπως για παράδειγμα έχει παρουσιαστεί από τους Lemmon and Portniaguina (2006)[16] υπάρχει σχέση ανάμεσα στο συναίσθημα και την αυτοπεποίθηση που έχουν οι επενδυτές, και στην αγορά μετοχών,

ενώ οι Gilbert and Karahalios (2010)[8] υλοποιώντας ένα μοντέλο με πάνω από 85% ακρίβεια δείχνουν πως η εξαγωγή συναισθημάτων από το διαδίκτυο μπορεί να οδηγήσει σε αρκετά ικανοποιητικά αποτελέσματα πρόβλεψης για τις μελλοντικές τιμές των μετοχών.

Για την πρόβλεψη των tweets χρησιμοποιήθηκαν μοντέλα όπως Naive Bayes, MaxEnt and Support Vector Machines (SVM) με τα καλύτερα αποτελέσματα για άλλη μια φορά να έρχονται, από το SVM μοντέλο.

Όσον αφορά την κατασκευή του μοντέλου που θα πραγματοποιεί ανάλυση συναισθήματος, βασιστήκαμε σε έναν διαγωνισμό "SemEval-2017 Task 5 Fine-Grained Sentiment Analysis on Financial"[17] που έλαβε μέρος το 2017. Στόχος των διαγωνιζόμενων ήταν να επιτύχουν την πιο ακριβή ανάλυση συναισθήματος με βάση ένα dataset που τους δόθηκε. Πιο συγκεκριμένα, το dataset που χρησιμοποιήθηκε για την ανάλυση συναισθήματος ήταν ένα αρχείο train.csv όπου περιείχε 1700 tweets μαζί με τα sentiment score τους που δόθηκαν στον διαγωνισμό "SemEval-2017 sentiment Analysis in Twitter". Η αξιολόγηση του μοντέλου μας έγινε με βάση ένα άλλο csv με το οποίο και αξιολογήθηκαν οι διαγωνιζόμενοι στο οποίο πάλι είχε tweets και δίπλα το label του «σωστού» sentiment score τους. Σκοπός ήταν να μεγιστοποιηθεί το cosine similarity μεταξύ των προβλέψεών μας σε σχέση με τα δοσμένα scores.

Όσον αφορά την κατασκευή του μοντέλου που θα πραγματοποιεί ανάλυση συναισθήματος, υπάρχουν αρκετοί μέθοδοι που χρησιμοποίησαν ερευνητές. Αρχικά, υπάρχουν πολλοί τρόποι κατά το στάδιο του pre-processing δηλαδή την προ-επεξεργασία του κειμένου των tweets πρώτού τροφοδοτηθούν στο νευρωνικό σύστημα. Σε αυτό το στάδιο μπορούν να αφαιρεθούν τα σημεία στίξης, τα URLs που πιθανός θα υπάρχουν σε πολλά tweets, οι αναφορές σε @username, καθώς και όλοι οι χαρακτήρες να μετατραπούν σε πεζούς. Στην συνέχεια ακολουθεί το tokenization δηλαδή το tweet που αποτελείται από μια σειρά συμβολοσειρών να χωριστεί ανά λέξη.

Υπάρχουν αρκετοί μέθοδοι κατασκευής μοντέλων που δέχονται σαν είσοδο τα δεδομένα μετά την προ-επεξεργασία και πραγματοποιούν ανάλυση συναισθήματος, δηλαδή για κάθε tweet να παράγει έναν ακέραιο αριθμό από το -1 έως το 1 που θα δείχνει πόσο θετικό ή αρνητικό είναι. Μπορούν να κατηγοριοποιηθούν σε κατηγορίες ανάλογα την τεχνική που χρησιμοποιούν. Κάποιες από αυτές είναι το machine learning (ML) η deep Learning (DL) και οι βασισμένες σε λεξικό (Lexicon-based). Ακόμα πολλοί ερευνητές χρησιμοποιούν περισσότερες από μία τεχνικές (hybrid). Η πιο ευρέως χρησιμοποιούμενη μέθοδος είναι ο συνδυασμός του Machine-learning με το Lexicon-based.
Οι πιο γνωστές Machine-learning τεχνικές που χρησιμοποιούνται κάνουν χρήση αλγορίθμων

- Artificial Neural Network (ANN) όπως έκανε ο Symeonidis et al. (2017)[18] Saleiro et al. (2017)[19],
- Random Forests όπως δοκίμασαν οι Seyeditabari et al. (2017)[20] και Saleiro et al. (2017)[19],
- support vector machine (SVM) όπως πραγματοποίησε ο Kumar et al. (2017)[21] στην ερευνά του.

Ακόμα υπάρχουν και άλλοι ευρέως διαδεδομένοι αλγόριθμοι που χρησιμοποιούνται αποτελεσματικά στην ανάλυση συναισθήματος μερικοί από τους οποίους είναι το support vector regression (SVR) η linear και logistic regression ακόμα και naive Bayes. Από τους παραπάνω αλγορίθμους αυτοί που χρησιμοποιούνται περισσότερο είναι οι Random Forest τα support vector machine και τα support vector regression.
Όσον αφορά τις Deep Learning-based τεχνικές, αυτές που χρησιμοποιούνται ευρέως είναι τα:

- Convolution Neural Network (CNN) τα οποία και χρησιμοποίησαν οι Pivovarova et al. (2017)[22], Ghosal et al. (2017)[21],
- Recurrent Neural Network (RNN) και τα
- Long short term memory (LSTM) Ghosal et al. (2017)[21] και τα Bidirectional Gated Recurrent Unit (Bi-GRU) Kar et al. (2017).

Δεν είναι λίγα τα λεξικά που μπορούν να χρησιμοποιηθούν για την πραγματοποίηση της ανάλυσης συναισθήματος. Τα πιο χρησιμοποιούμενα λεξικά τα οποία μπορούν να χρησιμοποιηθούν είτε μόνα τους είτε ως συνδυασμό με άλλες μεθόδους είναι τα εξής:

• «Loughran and McDonald Sentiment Word Lists» (Loughran and McDonald, 2011b), και χρησιμοποιήθηκε από τους Seyeditabari et al. (2017)[20], Saleiro et al. (2017)[19], Ghosal et al. (2017)[21]
• SentiWordNet, και χρησιμοποιήθηκε από τους Cabanski et al. (2017), Chen et al. (2017); Jiang et al. (2017)
• VADER (Hutto and Gilbert, 2014)[8] το οποίο χρησιμοποιήθηκε από Cabanski et al. (2017)
• Opinion Lexicon (Hu and Liu, 2004)[13] χρησιμοποιήθηκε από Cabanski et al. (2017); Kumar et al. (2017);
• MPQA Subjectivity Lexicon (Wilson et al., 2009) που χρησιμοποιήθηκε από Kumar et al. (2017)[21] Jiang et al. (2017); Saleiro et al. (2017)[19]; Ghosal et al. (2017)[21]

Εκείνα τα οποία χρησιμοποιούνται περισσότερο είναι τα Loughran and McDonald Sentiment Word Lists, Opinion Lexicon και το MPQA Subjectivity Lexicon.

Τέλος, δεν είναι λίγοι εκείνοι οι οποίοι συνδύασαν περισσότερες από μια μεθόδους. Ο Cabanski et al. (2017) σύγκρινε κάποιες hybrid τεχνικές με την hybrid(DL, lex) για να πετυχαίνει τα καλύτερα αποτελέσματα στο dataset που χρησιμοποιούσε. Από την άλλη πλευρά Kumar et al. (2017) συνέδεσε Support Vector machine και logistic Regression.

2.3 Πρόβλεψη τιμών χρησιμοποιώντας μόνο ιστορικά δεδομένα

Σε αυτό το σημείο θα αναφερθούν μέθοδοι τους οποίους χρησιμοποίησαν ερευνητές προκειμένου να προβούν σε αποτελεσματικές προβλέψεις συμπεριφοράς των χρηματιστηριακών τιμών αξιοποιώντας μόνο ιστορικούς δείκτες και τιμές χωρίς να αντλούν πληροφορίες από κειμενικά δεδομένα όπως είδαμε προηγουμένως.

Κατά τη μελέτη κάποιου φαινομένου, η ανάπτυξη ενός μαθηματικού μοντέλου για την προσομοίωση των μη γραμμικών σχέσεων μεταξύ της εισόδου και της εξόδου είναι ένα δύσκολο έργο λόγω της πολύπλοκης φύσης αυτών των φαινομένων. Τα συστήματα τεχνητής νοημοσύνης όπως τα τεχνητά νευρωνικά δίκτυα (ANN), το σύστημα Fuzzy Inference (FIS) και το προσαρμοστικό σύστημα νεύρο-ασαφούς συμπερασμού (ANFIS) εφαρμόστηκαν για να μοντελοποιήσουν ένα ευρύ φάσμα δύσκολων προβλημάτων στην επιστήμη και στη μηχανική[23]. Ένα μοντέλο AAN εμφανίζει καλύτερες επιδόσεις στην πρόβλεψη της χρεοκοπίας σε σχέση με μοντέλα που χρησιμοποιούν στατιστικές μεθόδους όπως διακριτική ανάλυση (discriminant analysis) και ανάλυση στατιστικών δεδομένων (logistic regression)[24]. Έρευνες έδειξαν ότι τα ANN έχουν μεγαλύτερη ακρίβεια πρόβλεψης έναντι των στατιστικών μεθόδων εξαιτίας της πολύπλοκης σχέσης μεταξύ των οικονομικών και άλλων δεδομένων εισόδου[25].

Οι Guresen, Kayakutlu, and Daim (2011) [26] ερεύνησαν την απόδοση των πολυστρωματικών νευρωνικών δικτύων (MLP), των δυναμικών ANN και των υβριδικών ANN στην πρόβλεψη των χρηματιστηριακών τιμών. Οι Chen, Leung, and Daouk (2003) [27] χρησιμοποίησαν πιθανοτικά νευρωνικά δίκτυα (PNN) με σκοπό να προβλέψουν την κατεύθυνση των χρηματιστηριακών δεικτών του Taiwan και κατέληξαν στο συμπέρασμα ότι τα πιθανοτικά PNN έχουν υψηλότερη απόδοση σε σύγκριση με άλλες μεθόδους όπως είναι οι Kalman Filter και τα μοντέλα τυχαίων περιπάτων (random forecasting models). Οι Kuo, Chen, and Hwang (2001) [28] ανάπτυξαν ένα σύστημα υποστήριξης λήψης αποφάσεων (decision support system) χρησιμοποιώντας έναν γενετικό αλγόριθμο που βασιζόταν σε ένα ασαφή νευρωνικό δίκτυο (GFNN) και σε ένα νευρωνικό δίκτυο ANN για την πρόβλεψη ανόδου ή καθόδου των χρηματιστηριακών τιμών. Στην συνέχεια το ίδιο προσπάθησαν να καταφέρουν και οι Qiu, Liu, and Wang (2012) [9] οι οποίοι ανάπτυξαν ένα μοντέλο βασιζόμενο σε ασαφή σειρές (fuzzy time series) και σε δέντρα αποφάσεων (C-fuzzy decision tree). Τέλος οι Atsalakis and Valavanis (2009) [29] κατασκεύασαν ένα μοντέλο νευροασαφών δικτύων με σκοπό να προβλέψουν την τιμή κλεισίματος των μετοχών της επόμενης μέρας.

Η πιο κοντινή ερευνητική εργασία με αυτή που ακολουθήθηκε σε αυτή την διατριβή ήταν των Murtaza Roondiwala Harshal Patel Shraddha Varma [30] οι οποίοι ανέπτυξαν ένα μοντέλο Long short-Term Memory (LSTM) με σκοπό την πρόβλεψη των δεκτών του χρηματιστηρίου NIFTY αξιοποιώντας δεδομένα 5 χρόνων.

Κεφάλαιο 3

3. Θεωρητικό Υπόβαθρο

Παρακάτω θα γίνει επεξήγηση κάποιων βασικών εννοιών οι οποίοι χρησιμοποιήθηκαν στο πλαίσιο αυτής της εργασίας.

3.1 Μηχανική Μάθηση

Αυτή η εργασία θα βασιστεί σε τεχνικές μηχανικής μάθησης και νευρωνικών δικτύων. Η μηχανική μάθηση (machine learning) αποτελεί ένα πεδίο της επιστήμης των υπολογιστών το οποίο χρησιμοποιεί τεχνικές στατιστικής ώστε να δώσει σε υπολογιστικά συστήματα την δυνατότητα να μάθουν από δεδομένα χωρίς να χρησιμοποιούν κάποιον ντετερμινιστικό αλγόριθμο. Ουσιαστικά πρόκειται για την εκμάθηση υπολογιστικών μηχανών ώστε να προβαίνουν σε αυτόματη λήψη αποφάσεων και προβλέψεων. Οι αλγόριθμοι της μηχανικής μάθησης χωρίζονται σε τρεις κατηγορίες: μάθηση με επίβλεψη (supervised Learning), μάθηση χωρίς επίβλεψη (unsupervised Learning) και Ενισχυμένη μάθηση (Reinforcement Learning) [31]

- Επιβλεπόμενη Μάθηση (Supervised Learning) είναι η διαδικασία όπου ο αλγόριθμος κατασκευάζει μια συνάρτηση που απεικονίζει δεδομένες εισόδους (σύνολο εκπαίδευσης) σε γνωστές επιθυμητές εξόδους, με απώτερο στόχο τη γενίκευση της συνάρτησης αυτής και για εισόδους με άγνωστη έξοδο. Χρησιμοποιείται σε προβλήματα ταξινόμησης (classification), πρόγνωσης (prediction), παλινδρόμησης (regression) [31].

- Μάθηση χωρίς επίβλεψη: το σύστημα καλείται μόνο του να ανακαλύψει συσχετίσεις ή ομάδες σε ένα σύνολο δεδομένων, δημιουργώντας πρότυπα, χωρίς να είναι γνωστό αν υπάρχουν, πόσα και ποια είναι. Χρησιμοποιείται σε προβλήματα ανάλυσης Συσχετισμών (association Analysis) και ομαδοποίησης (clustering) [31].

- Ενισχυτική Μάθηση (Reinforcement Learning), όπου ο αλγόριθμος μαθαίνει μια στρατηγική ενεργειών μέσα από άμεση αλληλεπίδραση με το περιβάλλον. Χρησιμοποιείται κυρίως σε προβλήματα Σχεδιασμού (Planning), όπως για παράδειγμα ο έλεγχος κίνησης ρομπότ και η βελτιστοποίηση εργασιών σε εργοστασιακούς χώρου[31].

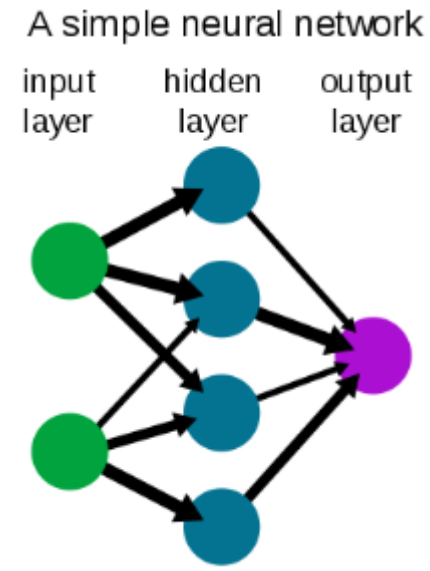
Για κάθε πρόβλημα προς επίλυση στο χώρο της Μηχανικής Μάθησης υπάρχει ένας κατάλληλος τρόπος μάθησης και για κάθε τρόπο μάθησης υπάρχει τουλάχιστον ένας κατάλληλος αλγόριθμος που μπορεί να χρησιμοποιηθεί[31].

3.2 Νευρωνικό δίκτυο

Το νευρωνικό δίκτυο είναι ένα δίκτυο από απλούς υπολογιστικούς κόμβους (νευρώνες), διασυνδεδεμένους μεταξύ τους. Είναι εμπνευσμένο από το Κεντρικό Νευρικό Σύστημα (ΚΝΣ), το οποίο προσπαθεί να προσομοιώσει[32].

Οι νευρώνες είναι τα δομικά στοιχεία του δικτύου. Κάθε τέτοιος κόμβος δέχεται ένα σύνολο αριθμητικών εισόδων από διαφορετικές πηγές (είτε από άλλους νευρώνες, είτε από το περιβάλλον), επιτελεί έναν υπολογισμό με βάση αυτές τις εισόδους και παράγει μία έξοδο. Η εν λόγω έξοδος είτε κατευθύνεται στο περιβάλλον, είτε τροφοδοτείται ως είσοδος σε άλλους νευρώνες του δικτύου. Υπάρχουν τρεις τύποι νευρώνων: οι νευρώνες εισόδου, οι νευρώνες εξόδου και οι υπολογιστικοί νευρώνες ή κρυμμένοι νευρώνες. Οι νευρώνες εισόδου δεν επιτελούν κανέναν υπολογισμό, μεσολαβούν απλώς ανάμεσα στις περιβαλλοντικές εισόδους του δικτύου και στους υπολογιστικούς νευρώνες. Οι νευρώνες εξόδου διοχετεύουν στο περιβάλλον τις τελικές αριθμητικές εξόδους του δικτύου. Οι υπολογιστικοί νευρώνες πολλαπλασιάζουν κάθε είσοδό τους με το αντίστοιχο συναπτικό βάρος και υπολογίζουν το ολικό άθροισμα των γινομένων. Το άθροισμα αυτό τροφοδοτείται ως

όρισμα στη συνάρτηση ενεργοποίησης, την οποία υλοποιεί εσωτερικά κάθε κόμβος. Η τιμή που λαμβάνει η συνάρτηση για το εν λόγω όρισμα είναι και η έξοδος του νευρώνα για τις τρέχουσες εισόδους και βάρη[32].



Εικόνα 2:Παράδειγμα τεχνητού νευρωνικού δικτύου

3.3 Χρονοσειρές

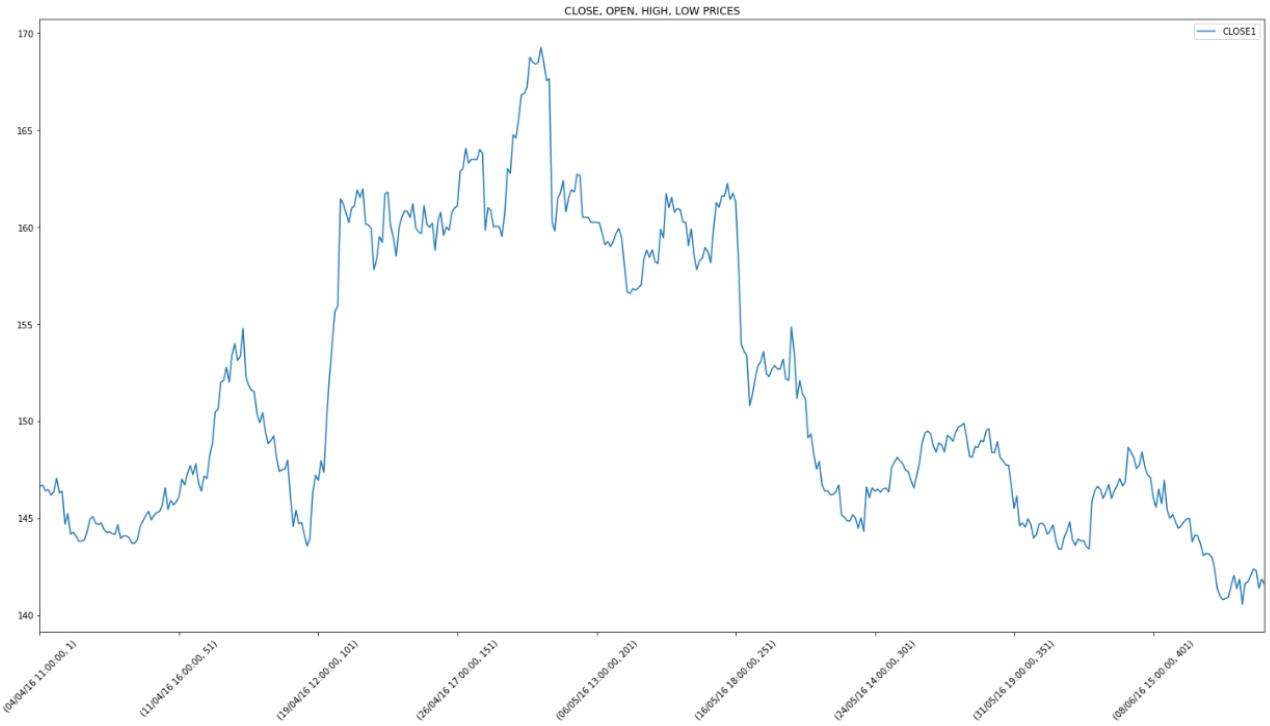
Το σύνολο των δεδομένων, τα οποία συλλέγονται διαχρονικά και εκφράζουν την εξέλιξη των τιμών μιας μεταβλητής κατά τη διάρκεια ίσων διαδοχικών χρονικών περιόδων ονομάζεται χρονοσειρά (ή χρονολογική σειρά, time series). Ειδικότερα, η χρονοσειρά αποτελείται από ένα σύνολο παρατηρήσεων, οι τιμές της οποίας λαμβάνονται σε ίσες χρονικές στιγμές ή περιόδους, π.χ. έτος, τρίμηνο, μήνας κ.ά.[33]. Ουσιαστικά, πρόκειται για μια στοχαστική διαδικασία, αφού οι τιμές του μεγέθους επηρεάζονται από τυχαίους παράγοντες, ενώ η τιμή κάθε χρονικής στιγμής συνιστά και μια ξεχωριστή τυχαία μεταβλητή.

Οι χρονοσειρές απαιτούν μόνο τις παρελθοντικές τιμές της μεταβλητής των διαδοχικών καταστάσεων στο χρόνο. Έτσι, μπορούν να αναλυθούν ώστε να εξαχθούν συμπεράσματα για την συμπεριφορά της μεταβλητής. Με βάση την πληροφορία από το παρελθόν, μας επιτρέπεται να προβλέψουμε τις τιμές της στο μέλλον.

Μαθηματικά, χρονοσειρά είναι ένα σύνολο παρατηρήσεων $y_1, y_2, \ldots, y_T$ όπου ο δείκτης Τ παριστάνει ισαπέχοντα χρονικά σημεία ή διαστήματα. Οι παρατηρήσεις $y_1, y_2, \ldots, y_T$ είναι συγκεκριμένες τιμές των τυχαίων μεταβλητών $Y_1, Y_2, \ldots, Y_T$ και είναι μέρος μόνο μιας άπειρης ακολουθίας τυχαίων μεταβλητών και συμβολίζεται με $\{Y_T\}$ [33].

Οι χρονοσειρές διακρίνονται σε συνεχείς χρονοσειρές και σε διακριτές. Συνεχείς χρονοσειρές είναι αυτές που η τιμή του φαινομένου παρατηρείται συνεχώς. Παράδειγμα συνεχών χρονοσειρών είναι η συνεχόμενη καταγραφή της θερμοκρασίας του αέρα ή η συνεχής παρακολούθηση των σεισμών. Διακριτές χρονοσειρές είναι αυτές όπου η τιμή του φαινομένου καταγράφεται σε ορισμένα χρονικά διαστήματα. Παράδειγμα διακριτών χρονοσειρών είναι η τιμή μιας μετοχής ανά ημέρα ή ο αριθμός των ηλιακών κηλίδων ανά έτος, όπου υπάρχουν τιμές σε συγκεκριμένα χρονικά διαστήματα.

Οι χρονοσειρές βρίσκουν εφαρμογές σε διάφορα πεδία, όπως στα Οικονομικά, την Ιατρική, την Περιβαλλοντολογία κ.ά.[34]. Παραδείγματα χρονοσειρών είναι η ημερήσια τιμή κλεισίματος μιας μετοχής στο Χρηματιστήριο, το ετήσιο ακαθάριστο εθνικό προϊόν μιας χώρας, οι μηνιαίες πωλήσεις ενός προϊόντος, οι ημερήσιες θερμοκρασίες μιας πόλης, η εγκεφαλική λειτουργία ανά δευτερόλεπτο και άλλα πολλά. Στο παράδειγμα που ακολουθεί, παρουσιάζεται μια χρονοσειρά που αναφέρεται στις τιμές κλεισίματος της μετοχής Apple (AAPL) κατά την διάρκεια 2 μηνών.



Εικόνα 3: Διάγραμμα χρονοσειράς της τιμής κλεισίματος της μετοχής AAPL

Για να γίνει η σωστή μελέτη μιας χρονοσειράς πρέπει κανείς να ξεκινήσει με την επισκόπηση του γραφήματός της στο πεδίο του χρόνου, από το οποίο μπορούν να ανιχνευθούν τα βασικά χαρακτηριστικά της: η τάση, η κυκλικότητα, η εποχικότητα και οι ακραίες τιμές [33], [35].

3.3.1 Τάση χρονοσειράς

Η τάση (trend) θα μπορούσε να ορισθεί ως μια μακροπρόθεσμη μεταβολή του μέσου επιπέδου των τιμών μιας χρονοσειράς. Έτσι, η τάση των τιμών μπορεί να είναι ανοδική, πτωτική ή σταθερή σε ένα συγκεκριμένο χρονικό διάστημα [36]. Συχνά, μπορεί να εκτιμηθεί από διάφορες οικογένειες καμπυλών, όπως μια ευθεία γραμμή ή μια εκθετική καμπύλη. Για να είναι ασφαλή τα συμπεράσματα που θα εξαχθούν για το αν μια σειρά παρουσιάζει τάση ή όχι θα πρέπει να έχουμε ένα ικανό αριθμό παρατηρήσεων και να εκτιμηθεί ένα κατάλληλο χρονικό διάστημα.

3.3.2 Κυκλικότητα χρονοσειράς

Η κυκλικότητα (cyclic) αντιπροσωπεύει μια μεταβολή που εμφανίζεται λόγω εξωγενών παραγόντων κατά μεγάλες περιόδους. Οι περίοδοι αυτοί είναι μεγαλύτερες του έτους και συνήθως της τάξεως της πενταετίας και δεκαετίας, χωρίς όμως αυτό να σημαίνει ότι είναι σταθερού μήκους [36]. Στις γραφικές παραστάσεις των χρονοσειρών παρουσιάζεται ως μια κυματοειδής γραμμή που κινείται ανάμεσα στην υψηλότερη και χαμηλότερη στάθμη.

3.3.3 Εποχικότητα χρονοσειράς

Η εποχικότητα (seasonal) μπορεί να εκφραστεί σαν μια περιοδική διακύμανση η οποία έχει σταθερό και μικρότερο ή ίσο μήκος ενός έτους. Η διακύμανση αυτή είναι άμεσα κατανοητή και προβλέψιμη, διότι τα δεδομένα ορισμένων χρονοσειρών επαναλαμβάνονται με τον ίδιο περίπου τρόπο σε σχέση με το χρόνο. Συνίσταται σε χρονοσειρές, όπως η ποσότητα κατανάλωσης του πετρελαίου θέρμανσης, η οποία είναι μεγαλύτερη κατά τους χειμερινούς μήνες κάθε έτους και όπως η μηνιαία κατανάλωση παγωτού η οποία είναι μεγαλύτερη κατά την καλοκαιρινή περίοδο σε σχέση με την χειμερινή. Εφόσον, η εποχική διακύμανση παρουσιάζεται με συστηματικό τρόπο, είναι ένα χαρακτηριστικό εύκολα οπτικά αναγνωρίσιμο που μπορεί να μετρηθεί και να απομονωθεί, ώστε να μην επηρεάζει τα δεδομένα μας. Η νέα χρονοσειρά που προκύπτει ονομάζεται αποεποχικοποιημένη χρονοσειρά.

3.3.4 Ακραίες τιμές χρονοσειράς

Οι ακραίες τιμές (outliers) είναι οι απομονωμένες παρατηρήσεις που εμφανίζονται στο γράφημα κάποιας χρονοσειράς ως απότομες αλλαγές στο πρότυπο συμπεριφοράς της. Οι ακραίες τιμές είναι μη προβλέψιμες και η επίδρασή τους στην χρονοσειρά έχει μικρή χρονική διάρκεια. Η ερμηνεία τέτοιων παρατηρήσεων χρειάζεται ιδιαίτερη προσοχή, διότι απαιτείται θεωρητική γνώση, κριτική ικανότητα και κοινή λογική. Ένα outlier μπορεί να αντιπροσωπεύει μια ασυνήθιστη παρατήρηση που οφείλεται σε κάποιο απρόβλεπτο γεγονός. Για παράδειγμα, μια απεργία μπορεί να προκαλέσει μεγάλη πτώση στην παραγωγή μιας βιοτεχνίας [35].

3.3.5 Στατιστικά μεγέθη χρονοσειράς

Μέση τιμή: Η μέση τιμή ή αναμενόμενη τιμή μιας χρονοσειράς Y δίνεται από την σχέση:

$$\mu_t = E(Y_t) = \int_{-\infty}^{+\infty} y_t f_{y_t}(y_t) dy_t$$

Η μέση τιμή μ_t σχετίζεται άμεσα με την έννοια της τάσης της χρονοσειράς, εφόσον εκφράζεται ως συνάρτηση της χρονικής στιγμής t της παρατήρησης Y_t. Συγκεκριμένα, αν μια χρονοσειρά παρουσιάζει αυξητική ή πτωτική τάση αντίστοιχα σε ένα χρονικό διάστημα, αυτό θα αποτυπώνεται και στη μέση τιμή ως συνάρτηση του χρόνου.

Αυτοσυνδιακύμανση: Υποθέτουμε ότι έχουμε δύο τυχαίες μεταβλητές X και W. Η συνδιακύμανση (covariance) των εν λόγω τυχαίων μεταβλητών δίνεται από την σχέση:

$$Cov(X,W) = E(X-\mu_x)(W-\mu_w)$$

Αυτοσυσχέτιση: Ο συντελεστής αυτοσυσχέτισης είναι ένας στατιστικός δείκτης ο οποίος χρησιμοποιείται στην ανάλυση χρονοσειρών για τον καθορισμό της τυχαιότητας ή μη της χρονοσειράς. Η αυτοσυσχέτιση (autocorrelation) j-οστής τάξης P_{jt} της τυχαίας μεταβλητής Y_t με μια καθυστερημένη εκδοχή της Y_{t-j} ορίζεται ως εξής:

$$P_{jt} = \frac{E(Y_t-\mu_t)(Y_{t-j}-\mu_{t-j})}{E(Y_t-\mu_t)^2}$$

Μια χρονοσειρά είναι στάσιμη αν θεωρήσουμε ότι οι στατιστικές της ιδιότητες παραμένουν σταθερές στο χρόνο, δηλαδή όταν δεν υπάρχει συστηματική αλλαγή του μέσου όρου και της διασποράς της στο χρόνο [36]. Αυτή είναι μια υπόθεση που δύσκολα μπορεί να υιοθετηθεί σε πολλά πραγματικά προβλήματα, αλλά μπορεί να χρησιμοποιηθεί ως υπόθεση εργασίας για την εξαγωγή χρήσιμων συμπερασμάτων.

Η μεγαλύτερη πρόκληση στην ανάλυση χρονοσειρών είναι η πρόβλεψη, δηλαδή πως η ακολουθία των παρατηρήσεων θα συνεχιστεί στο μέλλον. Το ζητούμενο είναι να ακολουθεί μια διαδικασία που θα εξασφαλίσει ότι θα παραχθούν όσο το δυνατόν πιο ακριβείς προβλέψεις, αξιοποιώντας στο έπακρο όλη την διαθέσιμη ιστορική πληροφορία [36].

3.4 Διαδικτυακή πλατφόρμα κοινωνικής δικτύωσης Twitter

Στην σύγχρονη εποχή το διαδίκτυο κατέχει κυρίαρχο ρόλο στην ζωή του ανθρώπου. Ο καθένας έχει την δυνατότητα να ενημερώνεται για τα τρέχοντα νέα αλλά ταυτόχρονα μπορεί να εκφράσει εύκολα την άποψή του για αυτά και γενικά για ό,τι άλλο επιθυμεί. Αυτή την δυνατότητα παρέχουν και τα μέσα κοινωνικής δικτύωσης. Πρόκειται για πλατφόρμες όπου οι χρήστες μπορούν να κοινοποιούν, να επικοινωνούν και να μοιράζονται υλικό μεταξύ τους. Υπάρχουν πολλά μέσα κοινωνικής δικτύωσης όπως είναι το Facebook, twitter, LinkedIn και το καθένα προσφέρει διαφορετικές δυνατότητες [37].

Το twitter, από το οποίο θα αντλήσουμε πληροφορίες σε αυτή την εργασία, είναι ένα μέσο κοινωνικής δικτύωσης που επιτρέπει στους χρήστες του να στέλνουν και να διαβάζουν σύντομα μηνύματα (μέχρι 280 χαρακτήρες), τα οποία ονομάζονται tweets. Τα μηνύματα μπορούν να αναγνωστούν και από μη συνδεδεμένους χρήστες, αλλά μόνο οι συνδεδεμένοι μπορούν να δημοσιεύσουν κείμενα. Πλέον υπάρχουν περισσότεροι από 300 εκατομμύρια άνθρωποι που χρησιμοποιούν το twitter. Μερικά από τα χαρακτηριστικά του είναι η δυνατότητα του λογαριασμού να αποκτά άτομα που ακολουθούν τις δραστηριότητές του (followers) αλλά και να ακολουθεί άλλους χρήστες. Επιπρόσθετα δίνεται η δυνατότητα στον χρήστη να κάνει Like σε σχόλια που του αρέσουν ή ακόμα και να αναπαραγάγει το σχόλιο αυτό προς τα άτομα που τον ακολουθούν γνωστό ως (Re-tweet) [37].

Όμως, το βασικό στοιχείο του twitter είναι τα hashtag τα οποία είναι κάτι σαν λέξεις-κλειδιά. Η λέξη hashtag προέρχεται από την ένωση των λέξεων hash και tag, δηλαδή του συμβόλου # και μιας ετικέτας (λέξης). Για παράδειγμα, #AAPL. Αυτά χρησιμοποιούνται για την διευκόλυνση της ομαδοποίησης των tweets σε κατηγορίες. Για παράδειγμα το #AAPL δηλώνει ότι ο χρήστης θα κάνει κάποιο σχόλιο για την εταιρία της Apple. Έτσι όταν κάποιος πληκτρολογήσει στην αναζήτηση το hashtag #AAPL θα του εμφανιστούν όλα τα σχόλια που αφορούν την εταιρία της Apple. επίσης το twitter API δίνει την δυνατότητα στους χρήστες να κατεβάζουν tweets με κάποιο συγκεκριμένο hashtag αν και πλέον έχει θέσει αυστηρούς περιορισμούς όσον αφορά το πλήθος των tweets που μπορεί κάποιος να κατεβάσει για κάποια χρονική διάρκεια [37].

Ένα ερώτημα που καλούμαστε να εξετάσουμε είναι αν τα αποτυπώματα που αφήνουν οι χρήστες των μέσων κοινωνικής δικτύωσης παρέχουν αρκετή πληροφορία για να βοηθήσουν στην πρόβλεψη κάποιων γεγονότων και πιο συγκεκριμένα στην εργασία αυτή αν μπορούν να αξιοποιηθούν αποτελεσματικά ώστε να προβλεφθεί η μελλοντική συμπεριφορά των χρηματιστηριακών τιμών μια μετοχής.

3.5 Ανάλυση συναισθήματος

Η ανάλυση συναισθήματος (sentiment analysis) ή αλλιώς εξόρυξη γνώμης (opinion mining) βασίζεται στην φυσική επεξεργασία και στην ανάλυση κειμένου, ώστε συστηματικά να εντοπίζει, να εξάγει, να ποσοτικοποιεί και να μελετά συναισθηματικές καταστάσεις και υποκειμενικές πληροφορίες. Στην βιομηχανία, η ανάλυση συναισθήματος έχει βρει ευρεία εφαρμογή σε περιοχές όπως είναι οι κριτικές προϊόντων, οι απαντήσεις ερωτηματολογίων, τα κοινωνικά δίκτυα, ενώ οι εφαρμογές είναι πολλές και πολύ διαφορετικές μεταξύ τους, από την εξωστρεφή προώθηση προϊόντων μέχρι λ.χ. την ανάλυση της άποψης των καταναλωτών [38].

Αποτελεί ένα πεδίο έρευνας, που προσελκύει έντονο ενδιαφέρον τα τελευταία χρόνια εξαιτίας της μεγάλης επιρροής των κοινωνικών δικτύων στην καθημερινότητά μας, του αυτοματοποιημένου τρόπου που παρέχει για την ανάλυση της γραπτής πληροφορίας που αφθονεί σε διαδικτυακές πηγές αλλά και της σημαντικής προόδου που σημειώνεται τελευταία στα πεδία της μηχανικής μάθησης, της τεχνητής νοημοσύνης και της βαθιάς μάθησης [39].

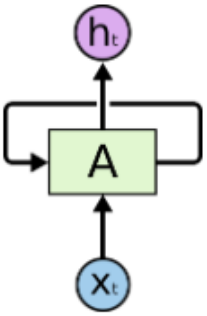
Στην παρούσα διπλωματική εργασία θα χρησιμοποιήσουμε tweets προκειμένου να προβούμε στην ανάλυση συναισθήματος. Πρόκεινται για κείμενα μικρού μήκους αποτελούμενα συνήθως από μία με δύο προτάσεις. Αυτό, όμως δυσχεραίνει την ανάλυση συναισθήματος μιας και

είναι πιο δύσκολο να συμπεράνει το σύστημα αν είναι θετικό ή αρνητικό το σχόλιο αυτό σε τόσο μικρή έκταση του σχολίου.

3.6 Επαναληπτικά νευρωνικά δίκτυα

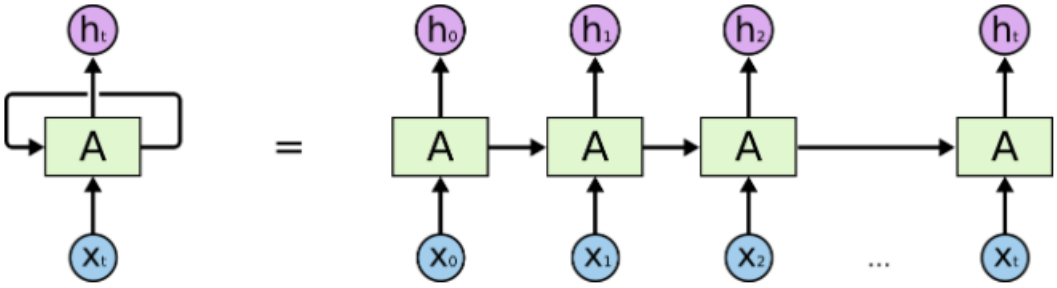
Οι άνθρωποι δεν ξεκινούν τον συλλογισμό της σκέψης τους κάθε δευτερόλεπτο. Καθώς διαβάζουν ένα κείμενο κατανοούν την σημασία κάθε λέξης με βάση την κατανόηση των προηγούμενων λέξεων και δεν διαγράφουν από την μνήμη τους ότι έχουν διαβάσει προηγουμένως. Τα παραδοσιακά νευρωνικά δίκτυα δεν μπορούν να το καταφέρουν αυτό και αποτελεί ένα από τα μεγάλα τους μειονεκτήματα. Για παράδειγμα, ας φανταστούμε ότι θέλουμε να κατανοήσουμε το κάθε γεγονός που συμβαίνει σε κάθε σημείο μιας ταινίας. Δεν είναι σαφές το πώς ένα παραδοσιακό νευρωνικό δίκτυο θα μπορούσε να χρησιμοποιηθεί ώστε να κρατάει στην μνήμη του τα προηγούμενα χρονικά γεγονότα της ταινίας ώστε να τα χρησιμοποιήσει αργότερα για τα επόμενα[40].

Τα recurrent neural networks αντιμετωπίζουν αυτό το ζήτημα. Πρόκειται για δίκτυα με βρόγχους (loops in them), και επιτρέπουν την διατήρηση πληροφοριών του παρελθόντος στο μέλλον.



Εικόνα 4: Βρόγχος επαναληπτικού νευρωνικού δικτύου

Στο διπλανό σχήμα, το οποίο αποτελεί ένα κομμάτι νευρωνικού δικτύου, το A δέχεται κάποια δεδομένα εισόδου X_t και παράγει μία έξοδο h_t. Ο βρόγχος (loop) επιτρέπει τη μετάδοση πληροφοριών από το ένα βήμα του δικτύου στο επόμενο. Αυτοί οι βρόγχοι κάνουν τα επαναληπτικά νευρωνικά δίκτυα (recurrent neural networks) να φαίνονται κάπως μυστηριώδη. Ωστόσο, εάν συλλογιστούμε λίγο παραπάνω, καταλαβαίνουμε ότι δεν είναι εντελώς διαφορετικό από ένα κανονικό νευρωνικό δίκτυο. Ένα επαναλαμβανόμενο νευρωνικό δίκτυο (recurrent neural network) μπορεί να θεωρηθεί ως πολλαπλά αντίγραφα του ίδιου δικτύου, το καθένα από τα οποία «μεταφέρει» πληροφορία στον διάδοχο του. Ας δούμε τι θα συμβεί εάν ξετυλίξουμε τον βρόγχο:



Εικόνα 5: Οι βρόγχοι ενός επαναληπτικού νευρωνικού δικτύου

Αυτή η μορφή που μοιάζει με αλυσίδα αποκαλύπτει ότι τα επαναληπτικά νευρωνικά δίκτυα σχετίζονται στενά με ακολουθίες. Αποτελεί την κλασσική αρχιτεκτονική του νευρωνικού δικτύου που χρησιμοποιείται για τέτοιου είδους δεδομένα[40].

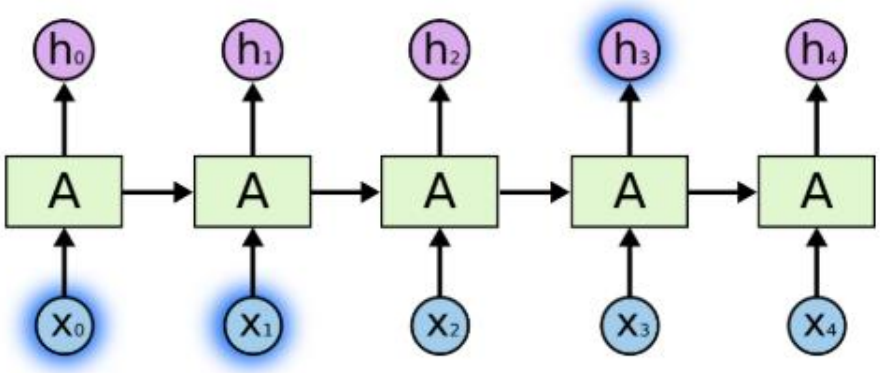
Είναι ευρέως χρησιμοποιούμενα νευρωνικά δίκτυα και τα τελευταία χρόνια υπάρχει μεγάλη επιτυχία στην ανάπτυξη και στην εφαρμογή των RNN σε πληθώρα είδη προβλημάτων όπως είναι αναγνώριση φωνής (speech recognition), γλωσσικά μοντέλα (language modeling), μετάφραση (translation), επεξεργασία εικόνας (image captioning).

Βασικός παράγοντας για την μεγάλη επιτυχία που έχουν τα RNN στις μέρες μας είναι η χρήση των LSTMs που αποτελούν ένα πολύ ιδιαίτερο είδος RNN το οποίο σε πολλά προβλήματα λειτουργεί πολύ καλύτερα από την απλή μορφή των RNN.

3.7 Το πρόβλημα της μακροχρόνιας εξάρτησης

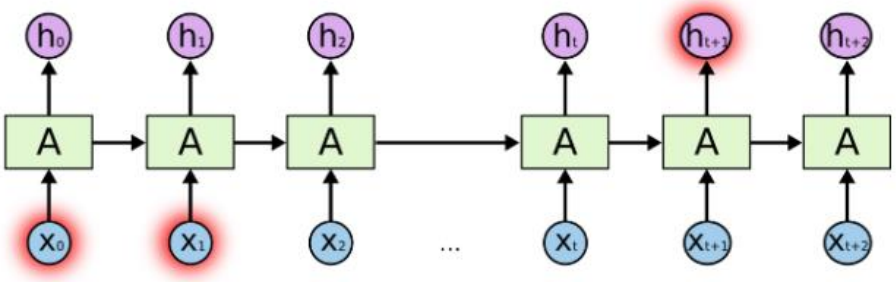
Ένας από τους βασικούς λόγους που θα θέλαμε να χρησιμοποιήσουμε δίκτυα RNN είναι το γεγονός ότι μπορούν να συνδυάσουν προηγούμενες πληροφορίες με το παρόν, όπως για παράδειγμα σε ένα video frame που μπορούν να «κατανοήσουν» το παρόν frame έχοντας «αποθηκεύσει» τα frames του παρελθόντος. Όμως έχουν ένα πρόβλημα το οποίο θα αναλύσουμε παρακάτω.

Μερικές φορές, για να παράξουμε την επιθυμητή έξοδο πρέπει να εξετάσουμε μόνο τις πρόσφατες πληροφορίες. Αυτό γίνεται κατανοητό αν σκεφτούμε για παράδειγμα ένα γλωσσολογικό μοντέλο το οποίο προσπαθεί να προβλέψει την επόμενη λέξη σε μία πρόταση με βάση τις προηγούμενες λέξεις. Εάν προσπαθήσουμε να προβλέψουμε την τελευταία λέξη της πρότασης «the clouds are in the sky" δεν χρειαζόμαστε καμία επιπλέον πληροφορία αφού είναι προφανές ότι η επόμενη λέξη θα είναι το "sky". Σε αυτές τις περιπτώσεις στις οποίες το κενό μεταξύ των σχετικών πληροφοριών είναι μικρό τα δίκτυα RNN μπορούν να χρησιμοποιηθούν αποδοτικά και αποτελεσματικά αφού μπορούν να μάθουν να χρησιμοποιούν τις προηγούμενες πληροφορίες.



Εικόνα 6: Μνήμη του δικτύου

Υπάρχουν όμως και πολλές περιπτώσεις στις οποίες για να παράξουμε το επιθυμητό αποτέλεσμα χρειαζόμαστε περισσότερο περιεχόμενο πληροφορίας. Έστω ότι τώρα επιθυμούμε να προβλέψουμε την τελευταία λέξη στο κείμενο "I grew up in France…I speak fluent French". Οι τελευταίες πληροφορίες δείχνουν ότι η επόμενη λέξη θα είναι το όνομα μιας γλώσσας. Όμως αν θέλουμε να προσδιορίσουμε και το όνομα αυτό τότε χρειαζόμαστε την πρόταση "I grew up in France" η οποία όμως μπορεί να βρίσκεται αρκετά πιο πριν της πρότασης «I speak fluent French" την οποία εξετάζουμε αφού μεσολαβούν αρκετές άλλες προτάσεις και πληροφορίες. Δυστυχώς, όταν το κενό μεταξύ των επιθυμητών πληροφοριών είναι αρκετά μεγάλο τα κλασσικά δίκτυα RNN δεν μπορούν να χρησιμοποιηθούν ώστε να συνδυάσουν τις πληροφορίες του παρελθόντος με το μέλλον.



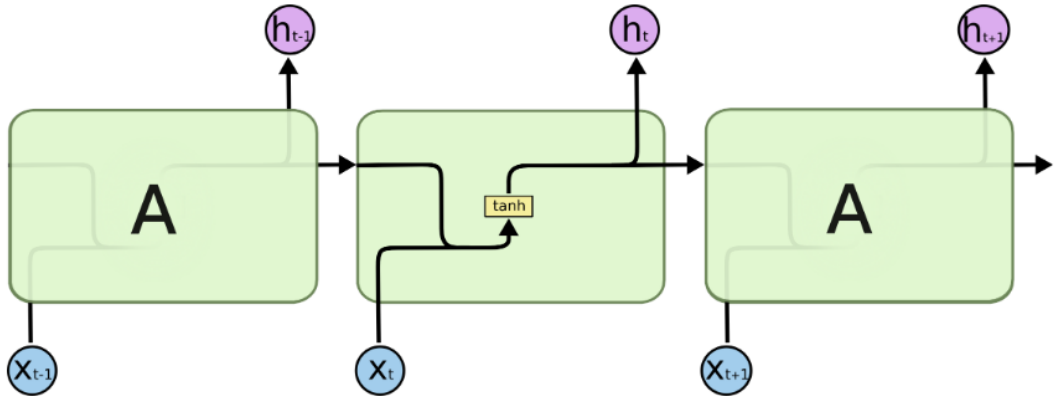
Εικόνα 7: Το πρόβλημα της μακροχρόνιας εξάρτησης

Θεωρητικά τα δίκτυα RNN είναι ικανά να χειριστούν τέτοιες «μακροχρόνιες εξαρτήσεις». Αν επιλεγούν προσεκτικά οι κατάλληλοι παράμετροι, τότε θα καταφέρουν να επιλύουν αυτά τα προβλήματα. Όμως στην πράξη όπως απέδειξε ο Hochreiter το 1991 αυτό είναι πρακτικά αδύνατο. Αυτού του είδους προβλήματα μπορούν να αντιμετωπίσουν με επιτυχία τα νευρωνικά δίκτυα LSTMs

3.8 Δίκτυα Μακράς Βραχυπρόθεσμης Μνήμης

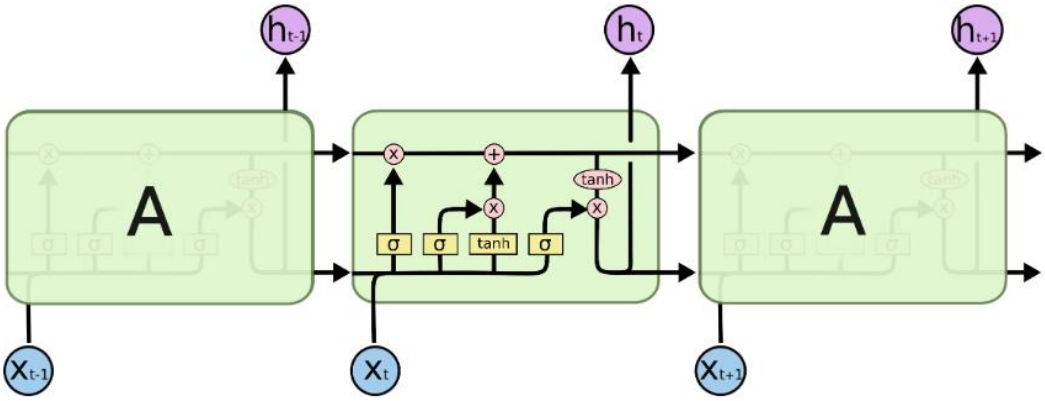
Τα Δίκτυα Μακράς Βραχυπρόθεσμης Μνήμης (Long Short Memory Networks) που συνήθως αναφέρονται ως "LSTMs" για συντομία, είναι ένα συγκεκριμένο είδος δικτύων RNN, τα οποία είναι ικανά να αντιμετωπίζουν τα προβλήματα μακροχρόνιων εξαρτήσεων. Πρώτος τα μελέτησε ο Hochreiter & Schmidhuber το 1997 [41] και έπειτα ασχολήθηκαν με αυτά και πολλοί άλλοι ερευνητές. Χρησιμοποιούνται ευρέως στις μέρες μας αφού είναι ικανά να επιλύουν αποτελεσματικά μεγάλη ποικιλία προβλημάτων.

Τα LSTMs είναι σχεδιασμένα με τρόπο τέτοιο ώστε να μπορούν να αντιμετωπίσουν τα προβλήματα μακροχρόνιας εξάρτησης. Έχουν την ικανότητα να αποθηκεύουν πληροφορίες από μεγάλες χρονικές περιόδους πράγμα που τα απλά RNN δεν είχαν. Όλα τα RNN δίκτυα έχουν την μορφή μιας αλυσίδας επαναλαμβανόμενων δομών (modules) του νευρωνικού δικτύου. Στα απλά RNN αυτή η επαναλαμβανόμενη δομή (Module) έχει απλή αρχιτεκτονική, όπως μόνο ένα στρώμα υπερεφαπτομένης (tanh).



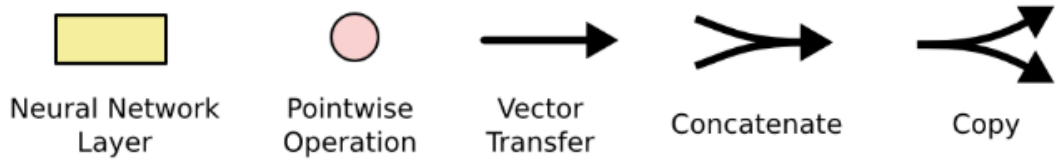
Εικόνα 8: Η επαναλαμβανόμενη δομή σε ένα τυπικό RNN περιέχει μόνο το στρώμα της υπερεφαπτομένης

Τα LSTMs έχουν επίσης αυτή την αλυσιδωτή μορφή αλλά η επαναλαμβανόμενη δομή έχει διαφορετική αρχιτεκτονική. Αντί να έχει ένα μοναδικό νευρωνικό επίπεδο, έχει τέσσερα τα οποία αλληλεπιδρούν με ένα πολύ συγκεκριμένο τρόπο. Η βασική τους μορφή φαίνεται στο ακόλουθα σχήμα



Εικόνα 9: Η επαναλαμβανόμενη δομή ενός LSTM αποτελείται από τέσσερα στρώματα

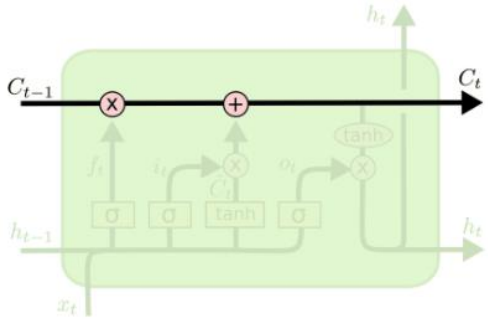
Θα ακολουθήσει μια πιο λεπτομερής ανάλυση για την δομή και την λειτουργία των νευρωνικών LSTM. Αρχικά σημαντικό είναι να γίνει η επεξήγηση κάποιων συμβόλων που θα χρειαστούν στην συνέχεια:



Εικόνα 10: Διαγράμματα για την επεξήγηση των LSTM

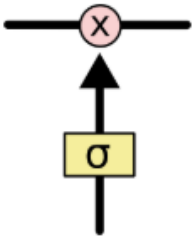
Στο παραπάνω διάγραμμα κάθε γραμμή μεταφέρει ένα διάνυσμα από την έξοδο ενός κόμβου στις εισόδους των επόμενων. Οι μωβ κύκλοι αντιπροσωπεύουν λειτουργίες όπως είναι η πρόσθεση, πολλαπλασιασμός διανυσμάτων και τα κίτρινα πλαίσια είναι μαθηματικά επίπεδα του νευρωνικού δικτύου. Οι γραμμές συγχώνευσης υποδηλώνουν αλληλεσύνδεση ενώ μια γραμμή αντιγραφής υποδηλώνει ότι το περιεχόμενο της αντιγράφεται και τα αντίγραφα ακολουθούν διαφορετικές διαδρομές.

Η σημαντική δομή για τα LSTM είναι η κυτταρική κατάσταση (cell state), η οριζόντια γραμμή που υπάρχει στην κορυφή του διαγράμματος που φαίνεται στην εικόνα 11. Αυτή πρόκειται για γραμμή μεταφοράς και μεταφέρονται οι πληροφορίες με κάποιες μικρές γραμμικές αλληλεπιδράσεις. Με αυτή, η μεταφορά της πληροφορίας γίνεται πολύ εύκολα.



Εικόνα 11: Η κυτταρική κατάσταση ενός LSTM

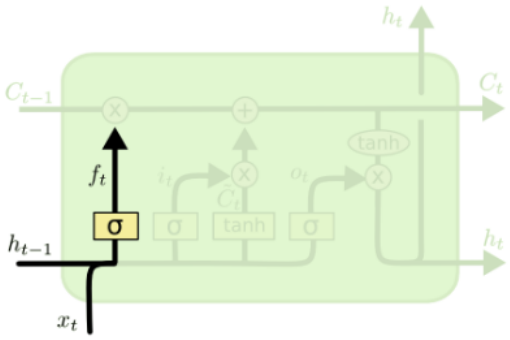
Τα LSTMs έχουν την δυνατότητα να αφαιρούν ή να προσθέτουν πληροφορίες στην κυτταρική κατάσταση (cell state). Για αυτές τις ενέργειες υπεύθυνοι είναι κάποιες άλλες δομές που λέγονται πύλες (gates). Αποτελούνται από στρώμα σιγμοειδούς νευρωνικού, και από μία ενέργεια πολλαπλασιασμού.



Το στρώμα σιγμοειδούς εξάγει αριθμούς μεταξύ του 0 και του 1 ανάλογα με το πόσο πολύ πληροφορία θα πρέπει να μεταφερθεί. Αν η τιμή είναι 0 τότε αυτό σημαίνει ότι όλη η πληροφορία «διαγράφεται» ενώ αν είναι 1 τότε όλη η πληροφορία μεταφέρεται. Κάθε νευρωνικό δίκτυο LSTM περιέχει τρεις τέτοιες πύλες για να ελέγχει την κυτταρική κατάσταση (cell state).

Εικόνα 12: Στρώμα Σιγμοειδούς Συνάρτησης

Ας εξετάσουμε αναλυτικά τον τρόπο με τον οποίο ένα LSTM μεταφέρει την πληροφορία. Το πρώτο βήμα είναι να αποφασίσει ποιες πληροφορίες θα ξεχαστούν-πεταχτούν μέσω της κυτταρικής κατάστασης (cell state). Η απόφαση αυτή γίνεται από το στρώμα σιγμοειδούς (sigmoid) που ονομάζεται πύλη λήθης ή αλλιώς“forget gate layer”. Αυτό τροφοδοτείται με το h_{t-1} και με την είσοδο x_t και δίνει ως έξοδο ένα αριθμό μεταξύ των τιμών 0 και 1. Το 1 αντιπροσωπεύει την απόφαση να κρατηθεί όλη η πληροφορία ενώ το 0 σημαίνει να μην κρατηθεί τίποτα από την πληροφορία, δηλαδή να την ξεχάσει.

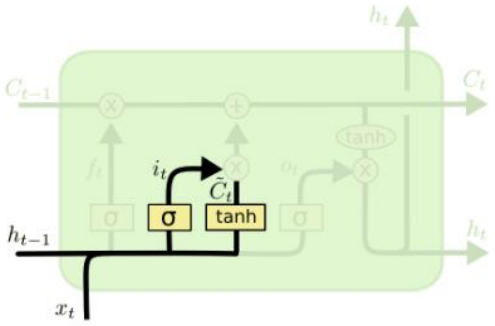


$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

Εικόνα 13: Forget Gate Layer

Στο παράδειγμα με την πρόβλεψη της επόμενης λέξης η κυτταρική κατάσταση (cell state) μπορεί να χρησιμοποιηθεί για να κρατάει το γένος-φύλλο του παρόντος υποκειμένου, έτσι ώστε να χρησιμοποιούνται οι σωστές αντωνυμίες από το νευρωνικό σύστημα. Όταν όμως ανιχνεύσει ένα νέο υποκείμενο τότε πρέπει να ξεχάσει το γένος του παλιού υποκειμένου.

Στο επόμενο βήμα το LSTM πρέπει να αποφασίσει ποια νέα πληροφορία πρέπει να αποθηκευτεί στην κυτταρική κατάσταση (cell state). Αυτό πραγματοποιείται από δύο στρώματα. Καταρχάς ένα στρώμα σιγμοειδούς (sigmoid) το οποίο ονομάζεται πύλη εισόδου ή "input gate layer" και αποφασίζει ποιες τιμές βαρών θα ενημερωθούν. Το δεύτερο μέρος αποτελείται από ένα στρώμα υπερεφαπτομένης (tanh) το οποίο δημιουργεί ένα διάνυσμα νέων υποψήφιων τιμών βαρών C_t, οι οποίες θα μπορούσαν να προστεθούν στην κυτταρική κατάσταση (state). Έπειτα αυτά τα δύο στρώματα θα συνδυαστούν ώστε να δημιουργηθεί η ενημέρωση για την επόμενη κατάσταση (state).

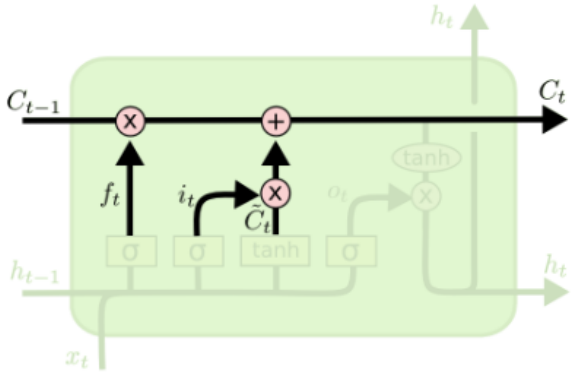


$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$

Εικόνα 14: Στρώμα εισόδου και στρώμα υπερεφαπτωμένης

Στο παράδειγμα με το γλωσσικό μοντέλο, πρέπει να προστεθεί το φύλο του νέου υποκειμένου στην κυτταρική κατάσταση (cell state) για να αντικατασταθεί το παλιό που θα ξεχαστεί από το «forget gate layer».

Τώρα είναι καιρός να ενημερωθεί η κατάσταση της «παλιάς» κυτταρικής κατάστασης (cell state) C_{t-1} στην νέα C_t. Στα προηγούμενα βήματα έχει ήδη αποφασιστεί τι πρέπει να γίνει, και απλά μένει να ολοκληρωθεί. Αυτό συμβαίνει με τον πολλαπλασιασμό της παλιάς κατάστασης με το f_t, ώστε να ξεχαστούν οι πληροφορίες που αποφασίστηκαν να απορριφθούν στο "forget state layer". Έπειτα προστίθεται το $i_t * C_t$. Αυτές είναι οι νέες υποψήφιες τιμές οι οποίες είναι κλιμακωμένες από το πόσο αποφασιστικέ να ενημερωθεί η κάθε κατάσταση.

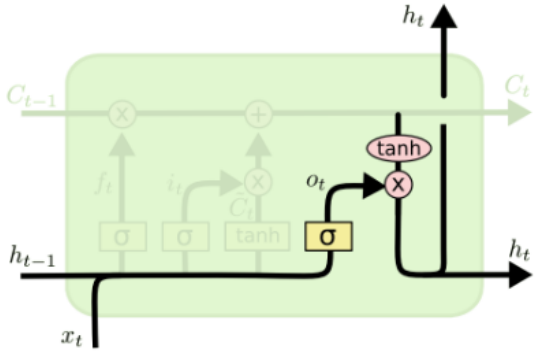


$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

Εικόνα 15: Ενημέρωση της κυτταρικής κατάστασης

Στο παράδειγμα με το γλωσσικό μοντέλο, είμαστε στο σημείο που πραγματικά πετάμε το γένος του παλιού υποκειμένου και προσθέτουμε την πληροφορία για το γένος του νέου υποκειμένου, όπως αποφασίστηκε στα προηγούμενα βήματα.

Τέλος, πρέπει να αποφασιστεί ποια θα είναι η έξοδος. Η έξοδος θα βασιστεί στην κυτταρική κατάσταση (cell state), αλλά πρώτα θα φιλτραριστεί. Αρχικά, θα περαστεί από ένα στρώμα σιγμοειδούς (sigmoid layer) το οποίο θα αποφασίσει ποια μέρη της κυτταρικής κατάστασης (cell state) θα «προχωρήσουν» στην έξοδο. Έπειτα η τρέχον κυτταρική κατάσταση (cell state) C_t θα περάσει από ένα στρώμα υπερεφαπτομένης (tanh), ώστε οι τιμές να είναι μεταξύ του -1 και του 1 και αυτό πολλαπλασιάζεται με την έξοδο της σιγμοειδούς πύλης, έτσι ώστε να δωθούν ως έξοδος μόνο οι πληροφορίες που αποφασίστηκαν. Αυτό φαίνεται στο επόμενο σχήμα



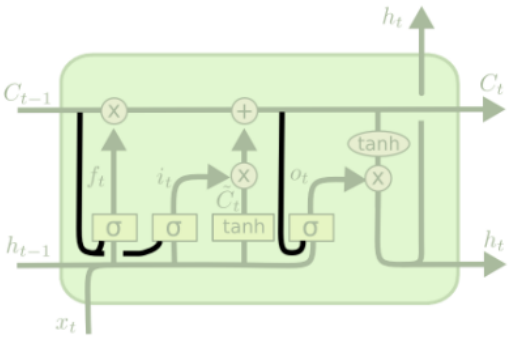
$$o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right)$$
$$h_t = o_t * \tanh\left(C_t\right)$$

Εικόνα 16: Έξοδος του LSTM

Στο παράδειγμα του γλωσσικού μοντέλου, μόλις αντιληφθεί ένα νέο υποκείμενο, θα δώσει ως έξοδο πληροφορίες σχετικές με το ρήμα, για την περίπτωση όπου αυτή είναι η λέξη που πρέπει να προβλέψει. Πιο συγκεκριμένα, θα μπορούσε να δώσει ως έξοδο αν το υποκείμενο είναι σε ενικό ή πληθυντικό αριθμό, ώστε να γνωρίζει σε ποια μορφή θα πρέπει να εμφανιστεί το ρήμα αν αυτή είναι η προβλεπόμενη λέξη.

Σε αυτό το σημείο πρέπει να τονιστεί ότι υπάρχουν και πολλές παραλλαγές του LSTM που εξετάσαμε παραπάνω, κάποιες από αυτές είναι οι εξής:

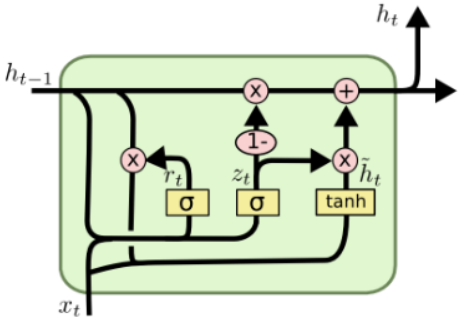
- Να έχει “peepholes connections” δηλαδή να επιτρέπεται στα στρώματα των gates να βλέπουν την κυτταρική κατάσταση (cell state).



$$f_t = \sigma\left(W_f \cdot [\boldsymbol{C_{t-1}}, h_{t-1}, x_t] \; + \; b_f\right)$$
$$i_t = \sigma\left(W_i \cdot [\boldsymbol{C_{t-1}}, h_{t-1}, x_t] \; + \; b_i\right)$$
$$o_t = \sigma\left(W_o \cdot [\boldsymbol{C_t}, h_{t-1}, x_t] \; + \; b_o\right)$$

Εικόνα 17: Peepholes connections LSTM

- Gated Recurrent Unit (GRU). Αυτά συνδυάζουν την forget και την input gate σε μία ενιαία πύλη ενημέρωσης «update gate”. Επίσης συγχωνεύει την κυτταρική κατάσταση (cell state) και η κρυφή πύλη κάνει κάποιες επιπλέων αλλαγές. Το αποτέλεσμα είναι ένα μοντέλο απλούστερο από το κλασσικό LSTM και έχει αναπτυχθεί και έχει γίνει πολύ γνωστό.

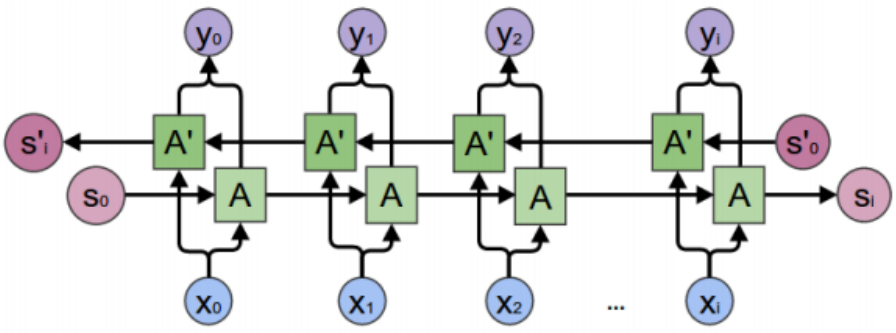


$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$
$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$
$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$
$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

Εικόνα 18: GRU, μια παραλλαγή του LSTM

3.9 Αμφίδρομα Επαναληπτικά δίκτυα

Μέχρι τώρα εξετάσαμε RNN νευρωνικά των οποίων η ροή ήταν προς μια μόνο κατεύθυνση. όμως αρκετές φορές είναι χρήσιμο, αντί να βλέπουμε μελλοντικές καταστάσεις με βάση τις προηγούμενες εισόδους, να προβλέπουμε προγενέστερες καταστάσεις χρησιμοποιώντας τις μελλοντικές ανατρέχοντας τα δεδομένα προς τα πίσω. Δημιουργήθηκε λοιπόν η ιδέα του νευρώνα ο οποίος θα λαμβάνει υπόψιν του και τις δύο πιθανές ροές της πληροφορίας στα δεδομένα, ευθέως και αντίστροφα. Οι νευρώνες αυτοί ονομάζονται Αμφίδρομα Επαναλαμβανόμενα Νευρωνικά δίκτυα (Bidirectional Recurrent Neural Networks)[42]. Παρακάτω φαίνεται η λειτουργία τους



Εικόνα 19: Αμφίδρομο Επαναλπτικό Νευρωνικό δίκτυο

Στην ουσία τα Αμφίδρομα Επαναλαμβανόμενα Νευρωνικά Δίκτυα αποτελούν συνδυασμό δύο ξεχωριστών RNN που επεξεργάζονται τα δεδομένα με αντίθετη φορά όπως φαίνεται και στην παραπάνω εικόνα

3.10 Εκπαίδευση Νευρωνικού Δικτύου

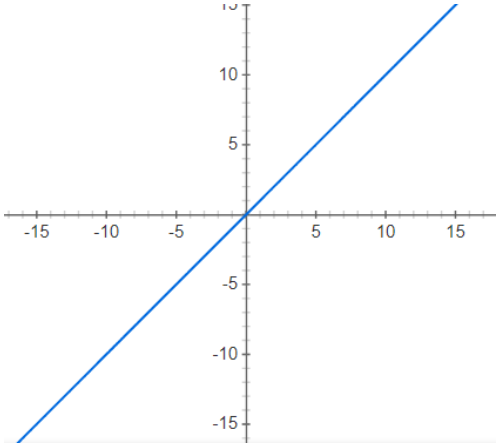
Ένα νευρωνικό δίκτυο για να παράξει αποτελέσματα πρέπει να έχει υποβληθεί σε μια διαδικασία εκπαίδευσης σε λογικά πλαίσια χρόνου και με βάση την υπολογιστική ισχύ που διαθέτουμε. Κατά την διαδικασία της εκπαίδευσης μας δίνεται η επιλογή να επιλέξουμε παραμέτρους που θα εκπαιδευτεί το μοντέλο. Για να γίνει αυτό πρέπει να επιλεχθούν οι κατάλληλες παράμετροι και μέθοδοι που σε κάποιες περιπτώσεις μπορεί εκ των προτέρων να γνωρίζουμε ποιες είναι αυτές και απλώς να χρειάζεται να υλοποιήσουμε την αρχιτεκτονική του νευρωνικού μας, ενώ σε άλλες να πρέπει να εργαστούμε με εξαντλητικές μεθόδους ή τεχνικές δοκιμής - σφάλματος (trial and error), για να αποφασίσουμε ποιες υπερπαράμετροι είναι πιο αποτελεσματικές. Αν επιλεγούν λανθασμένες υπερπαράμετροι υπάρχει ο κίνδυνος να υπάρξει υπερεκπαίδευση (overfitting) πάνω στα δεδομένα εκπαίδευσης. Το νευρωνικό δίκτυο μπαίνει σε μια επαναληπτική διαδικασία όπου σε κάθε επανάληψη προσπαθεί να βελτιώσει την απόδοσή του. Η κάθε επανάληψη ονομάζεται εποχή (epoch) και το πλήθος τους καθορίζεται από τον προγραμματιστή. Ο προγραμματιστής επίσης δίνει και άλλες παραμέτρους όπως είναι το Batch size του οποίου η επεξήγηση θα γίνει παρακάτω. Το νευρωνικό δίκτυο προσπαθεί να βελτιώσει μία συνάρτηση βελτιστοποίησης ή να μειώσει μία συνάρτηση κόστους τις οποίες επιλέγει ο προγραμματιστής ανάλογα με το μοντέλο και το πρόβλημα που επιθυμεί να επιλύσει.

3.11 Συναρτήσεις Ενεργοποίησης

Οι Συναρτήσεις Ενεργοποίησης (Activation Functions) αποτελούν αναπόσπαστο κομμάτι των νευρωνικών δικτύων. Πρόκειται για συναρτήσεις οι οποίες είναι αυτές που εισάγουν μη γραμμικές ιδιότητες στο δίκτυο. Η έξοδος κάθε νευρώνα, ύστερα από την επεξεργασία της εισόδου με τα αντίστοιχα βάρη και την πόλωση, φιλτράρεται μέσω μιας συνάρτησης ενεργοποίησης και τροφοδοτείται στο επόμενο επίπεδο του νευρωνικό δικτύου. Χωρίς την ύπαρξη των συναρτήσεων ενεργοποίησης, το σήμα της εξόδου θα ήταν μια γραμμική συνάρτηση, περιορίζοντας το σύστημά μας στα πλαίσια των γραμμικών αταξινόμητών, αδυνατώντας να εξάγει συμπεράσματα για μη γραμμικά προβλήματα. Σε πολλά στρώματα ο προγραμματιστής είναι εκείνος ο οποίος επιλέγει ή ορίζει την συνάρτηση ενεργοποίησης του στρώματος. Παρακάτω παρατίθενται μερικές ευρέως χρησιμοποιούμενες συναρτήσεις ενεργοποίησης, μερικές από τις οποίες χρησιμοποιήσαμε και στα μοντέλα μας[43].

3.11.1 Γραμμική Συνάρτηση

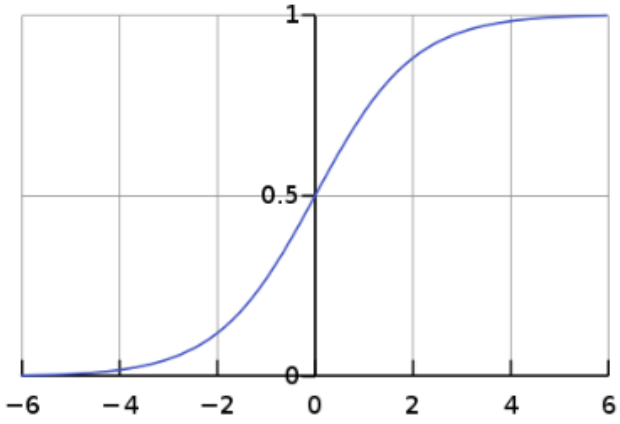
Είναι η πιο απλή συνάρτηση ενεργοποίησης και σε πολλά μοντέλα η προεπιλεγμένη επιλογή. Η συνάρτησή της είναι η f(x)=x δηλαδή ότι τιμή της δοθεί ως είσοδο αυτή θα είναι και η τιμή της εξόδου. Παρακάτω φαίνεται η γραφική της παράσταση[43].



Εικόνα 20: Γραφική παράσταση της γραμμικής συνάρτησης

3.11.2 Σιγμοειδής Συνάρτηση

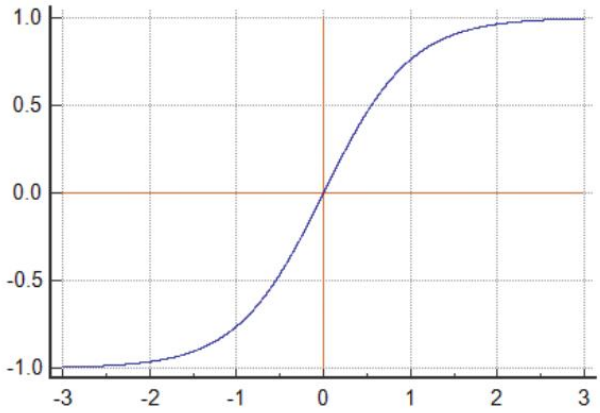
Η σιγμοειδής συνάρτηση είναι μια από τις πιο ευρέως χρησιμοποιούμενες γιατί αντιστοιχεί την τιμή της εισόδου σε τιμές του ανοικτού διαστήματος (0,1) χωρίς να μπορεί να πάρει τις ακραίες τιμές 0 και 1. Ένας επιπλέον λόγος που είναι πολύ γνωστή είναι το γεγονός ότι πραγματοποιεί κανονικοποίηση των τιμών στο διάστημα (0,1). Όμως το αρνητικό της συνάρτησης αυτής είναι ότι η κανονικοποίηση για τιμές αρκετά μεγάλες ή αρκετά μικρές δεν δίνουν αισθητή διαφοροποίηση στην έξοδο, φαινόμενο της εξαφανιζόμενης κλίσης (Vanishing gradient)[43]. Ο τύπος της σιγμοειδούς συνάρτησης είναι ο $S(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x+1}$ και η γραφική της παράσταση φαίνεται παρακάτω



Εικόνα 21: Γραφική παράσταση της σιγμοειδούς συνάρτησης

3.11.3 Συνάρτηση Υπερβολικής Εφαπτομένης

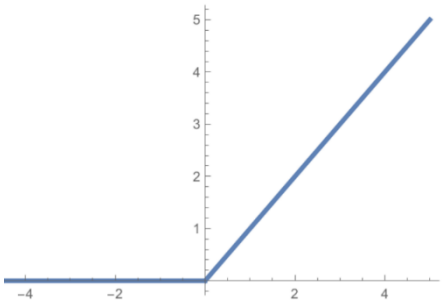
Η συνάρτηση αυτή έχει πολλές ομοιότητες με την σιγμοειδή, με την διαφορά ότι η κατανομή των τιμών γίνεται στο ανοικτό διάστημα (-1,1) χωρίς να παίρνει τις ακραίες τιμές -1 και 1. Ο τύπος της είναι f(x)=tanh(x)= $\frac{e^x-e^{-x}}{e^x+e^{-x}}$ Και η γραφική της παράσταση είναι η εξής:



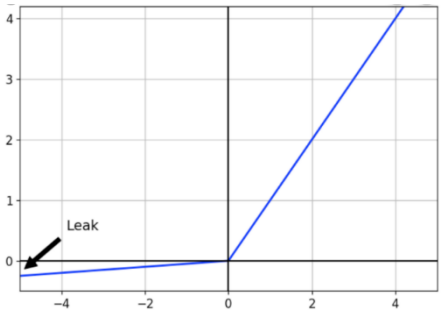
Εικόνα 22: Γραφική παράσταση της υπερβολικής εφαπτομένης

3.11.4 Ανορθωμένη Γραμμική Μονάδα

Η ανορθωμένη γραμμική μονάδα (ReLU) είναι από τις πιο διαδεδομένες συναρτήσεις ενεργοποίησης και χρησιμοποιείται σε πολλά νευρωνικά μοντέλα και κυρίως στα convolutional neural networks (CNN). Για τιμές εισόδου μεγαλύτερες του μηδενός λειτουργεί ακριβώς όπως η γραμμική συνάρτηση, ενώ μηδενίζει όλες τις αρνητικές τιμές. Κάποια από τα πλεονεκτήματά της είναι ότι οδηγεί σε γρήγορη σύγκλιση άρα είναι υπολογιστικά αποδοτική, έχει μη σταθερή παράγωγο και γιαυτό μπορεί να χρησιμοποιηθεί για την τεχνική του back Propagation και σε αντίθεση με την σιγμοειδή και την tanh δεν εμφανίζει το φαινόμενο της εξαφανιζόμενης κλίσης (vanishing gradient). Όμως εμφανίζει ένα αρνητικό φαινόμενο που ονομάζεται the dying ReLU και πρόκειται για το γεγονός ότι η εκμηδένιση όλων των αρνητικών τιμών οδηγεί στην θανάτωση όλων των νευρώνων που παίρνουν οποιαδήποτε στιγμή αρνητική τιμή και έτσι αυτοί οι νευρώνες δεν μπορούν να επηρεάσουν στην συνέχεια το νευρωνικό δίκτυο. Υπάρχει μια παραλλαγή που αντιμετωπίζει κατά κάποιο τρόπο αυτό το φαινόμενο που ονομάζεται Leaky ReLU κατά το οποίο οι αρνητικές τιμές πολλαπλασιάζονται με μια μικρή σταθερά c, και δεν μηδενίζονται. Ο τύπος της είναι $f(x) = x^+ = \max(0,x)$ και η γραφική της παράσταση είναι η παρακάτω[43].



Εικόνα 23: Γραφική παράσταση της ανορθωμένης Γραμμικής μονάδας



Εικόνα 24: Εικόνα 23: Γραφική παράσταση της ανορθωμένης Γραμμικής μονάδας με Leak

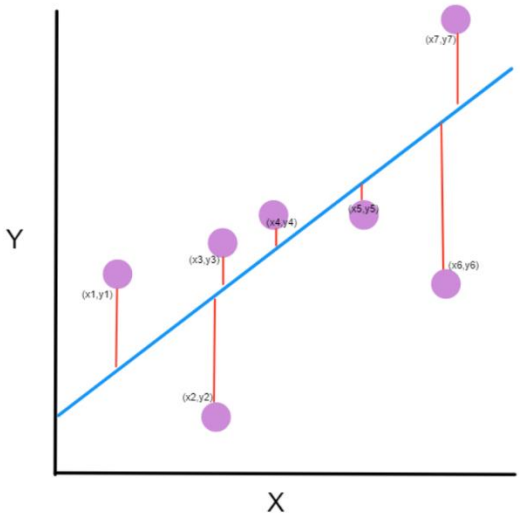
3.12 Συνάρτηση Κόστους

Το νευρωνικό δίκτυο είναι μια επαναλαμβανόμενη διαδικασία η οποία εκπαιδεύεται με ένα πλήθος επαναλήψεων (epoch). Σε κάθε επανάληψη αλλάζει τα βάρη του ώστε να βελτιώνεται η απόδοσή του. Οπότε πρέπει να υπάρχει ένα μέτρο που να μετρά το σφάλμα και να ενημερώνει το νευρωνικό ώστε να διορθώσει την συμπεριφορά του και να ελαχιστοποιήσει το σφάλμα. Γενικά με τον όρο σφάλμα εννοούμε τη διαφορά μεταξύ επιθυμητής εξόδου, δηλαδή της πραγματικής, από την έξοδο που προέβλεψε το σύστημά μας. Μια συνάρτηση σφάλματος επεξεργάζεται αυτό το σφάλμα και ουσιαστικά δείχνει πόσο ορθά έχει λειτουργήσει το μοντέλο μας. Αυτό το μέτρο ονομάζεται Συνάρτηση κόστους (Cost Function or Loss Function)[44].

Επίσης τα μοντέλα των νευρωνικών δίνουν την δυνατότητα στον προγραμματιστή να ορίσει μια μετρική αξιολόγησης (Evaluation Function) ώστε να μπορεί να δείχνει πόσο καλά πάει η βελτίωση του κατά το στάδιο της εκπαίδευσης. Επίσης, δίνεται η δυνατότητα να συγκρίνουμε την απόδοση διαφορετικών μοντέλων με βάση τις μετρικές αξιολόγησης. Υπάρχουν πολλές τέτοιες μετρικές που χρησιμοποιούνται όπως είναι το accuracy, F1-score[45]. Παρακάτω θα αναλυθούν κάποιες συναρτήσεις οι οποίες μπορούν να χρησιμοποιηθούν τόσο ως συναρτήσεις κόστους όσο και ως μετρικές συναρτήσεις και οι οποίες χρησιμοποιήθηκαν στα μοντέλα μας.

3.12.1 Μέσο Τετραγωνικό Σφάλμα

Βρίσκει μεγάλη χρήση στα επαναληπτικά μοντέλα και ουσιαστικά πρόκειται για τις αποστάσεις των σημείων (που έχουν δοθεί ως προβλέψεις από την έξοδο του νευρωνικού σε κάθε εποχή) από μία ευθεία (πραγματικές τιμές). Όσο πιο μικρή είναι η τιμή του Μέσου Τετραγωνικού Σφάλματος (Mean Square Error-MSE) τόσο πιο αποτελεσματικό είναι το νευρωνικό δίκτυο. Η σημαντική διαφοροποίηση σε σχέση με το MAE είναι το γεγονός ότι δίνει πολύ μεγαλύτερο βάρος στα μεγάλα σφάλματα και μικρότερο βάρος στα μικρά σφάλματα (λόγω τετραγωνισμού). Ο τύπος είναι ο εξής: $\text{MSE}=\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$



Εικόνα 25: Παράδειγμα Μέσο τετραγωνικό σφάλμα

3.12.2 Ριζικό Μέσο Τετραγωνικό Σφάλμα

Πρόκειται για την ίδια συνάρτηση με την MSE με την διαφορά ότι το αποτέλεσμα είναι υψωμένο στην τετραγωνική ρίζα.

$$RMSE = \sqrt{MSE}$$

3.12.3 Απόλυτο Μέσο Τετραγωνικό Σφάλμα

Πάλι μοιάζει με την MSE με την διαφορά ότι στο άθροισμα αντί για ύψωμα στο τετράγωνο έχει απόλυτη τιμή.

$$\text{MSE}=\frac{1}{n}\sum_{j=1}^{n}|y_i-\hat{y}_i|$$

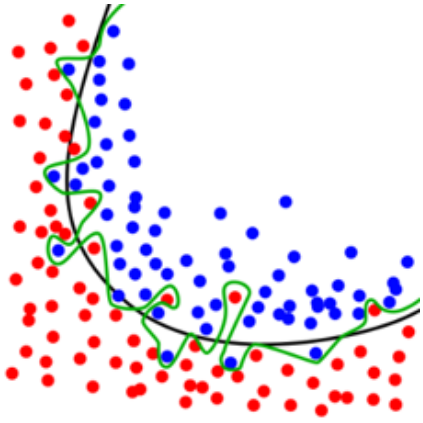
3.12.4 Ομοιότητα Συνημιτόνου

Το Cosine similarity ή cosine proximity πρόκειται για μία συνάρτηση η οποία μπορεί να χρησιμοποιηθεί είτε ως συνάρτηση κόστους είτε ως μετρική και μετράει την ομοιότητα μεταξύ δύο μη μηδενικών διανυσμάτων υπολογίζοντας το συνημίτονο της γωνίας που σχηματίζουν αυτά τα διανύσματα. Λαμβάνει τιμές από το 0 έως το 1 και όσο πιο κοντά στο 1 είναι τόσο πιο όμοια είναι τα διανύσματα που συγκρίνουμε. Ο τύπος της φαίνεται παρακάτω

$$similarity=\cos(\theta)=\frac{A*B}{||A||\,||B|}=\frac{\sum_{i=1}^{n}A_iB_i}{\sqrt{\sum_{i=1}^{n}{A_i}^2}\sqrt{\sum_{i=1}^{n}{B_i}^2}}$$

3.13 Υπερεκπαίδευση

Η γενίκευση (generalization) στην Μηχανική Μάθηση αναφέρεται στο πόσο καλά μπορούν να αποδίδουν τα δίκτυα σε παραδείγματα τα οποία τα οποία δεν περιέχονταν στο εκπαιδευτικό σύνολο δεδομένων. Ο στόχος των περισσοτέρων μοντέλων Μηχανικής Μάθησης είναι η "καλή" γενίκευση από το εκπαιδευτικό σύνολο δεδομένων, προκειμένου να κάνουν σωστές προβλέψεις στο μέλλον για δεδομένα που δεν είχαν ξαναδεί προηγουμένως. Υπερεκπαίδευση συμβαίνει όταν ένα μοντέλο μαθαίνει πολλές λεπτομερείς και θόρυβο των δεδομένων εκπαίδευσης και αυτό έχει αρνητικό αντίκτυπο στην απόδοση του μοντέλου στα νέα δεδομένα που θα χρησιμοποιήσει. Επίσης το μοντέλο πιθανόν να μην έχει γενικευτεί και να παράγει λανθασμένα αποτελέσματα στα νέα δεδομένα. Πολλές φορές παρουσιάζεται overfitting όταν χρησιμοποιείται μικρό μέγεθος δεδομένων εκπαίδευσης. Ένας τρόπος για να αποφευχθεί το φαινόμενο του overfitting είναι να χρησιμοποιηθούν νευρώνες Dropout [46].



Εικόνα 26: παράδειγμα υπερεκπαίδευσης

3.14 Συνελικτικό Επίπεδο

Το συνελικτικό επίπεδο είναι η βασική μονάδα κατασκευής ενός συνελικτικού Δικτύου, το οποίο εκτελεί και τους πιο απαιτητικούς υπολογισμούς. Ο κύριος σκοπός του επιπέδου αυτού είναι η εξαγωγή χαρακτηριστικών από την εικόνα εισόδου αν πρόκειται για πρόβλημα αναγνώρισης εικόνας. Οι παράμετροι του επιπέδου αυτού αποτελούνται από ένα σύνολο από εκπαιδεύσιμα φίλτρα. Κάθε φίλτρο είναι χωρικά μικρό (ως προς το ύψος και το πλάτος), αλλά εκτείνεται σε όλο το βάθος του όγκου της εισόδου. Ας θεωρήσουμε για παράδειγμα μια εικόνα, ένα τυπικό φίλτρο στο πρώτο επίπεδο ενός συνελικτικού δικτύου μπορεί να έχει μέγεθος 5x5x3 (5 pixels για το πλάτος, 5 pixels για το ύψος και 3 για τον αριθμό των καναλιών μίας έγχρωμης εικόνας RGB). Κατά το προωθητικό πέρασμα συνελίσσουμε κάθε φίλτρο σε όλο τον όγκο της εισόδου και υπολογίζουμε τα εσωτερικά γινόμενα μεταξύ των τιμών του φίλτρου και των τιμών της εισόδου σε οποιαδήποτε θέση. Καθώς περνάμε το φίλτρο κατά ύψος και κατά πλάτος του πίνακα εισόδου, παράγεται ένας δισδιάστατος πίνακας ενεργοποίησης ο οποίος αποδίδει τις τιμές απόκρισης του φίλτρου σε κάθε χωρική θέση. Διαισθητικά,

το δίκτυο θα εκπαιδευθεί σε φίλτρα τα οποία ενεργοποιούνται όταν βλέπουν κάποιον τύπο οπτικών χαρακτηριστικών. Έτσι έχουμε, πλέον, αποκτήσει ένα ολόκληρο σύνολο από φίλτρα σε κάθε συνελικτικό επίπεδο, κάθε ένα από τα οποία θα παράγει έναν δισδιάστατο πίνακα ενεργοποίησης. Θα στοιβάξουμε αυτούς τους πίνακες ενεργοποίησης κατά την τρίτη διάσταση (βάθος) και εν τέλει θα αποκτήσουμε την τρισδιάστατη έξοδο.

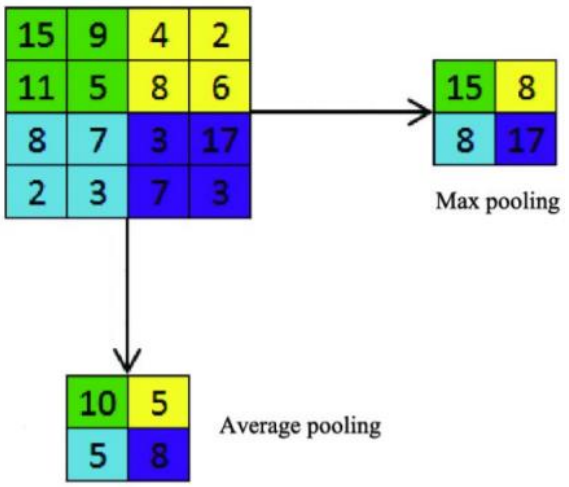
Κάνοντας παραλληλοποίηση με την λειτουργία του εγκεφάλου, μπορούμε να φανταστούμε ότι κάθε τιμή του τρισδιάστατου πίνακα εξόδου μπορεί να μεταφραστεί σαν μία έξοδο ενός νευρώνα ο οποίος κοιτάζει μόνο μία μικρή περιοχή της εισόδου και μοιράζεται τις παραμέτρους του με όλους τους νευρώνες που βρίσκονται δεξιά και αριστερά του [47].

3.15 Στρώμα Εμφύτευσης

Το στρώμα εμφίτευσης (embedding layer) χρησιμοποιείται με σκοπό να σχηματίσει διανύσματα λέξεων για τις λέξεις της εισόδου. Συνήθως συναντάται μεταξύ της εισόδου και του στρώματος LSTM δηλαδή η έξοδος του embedding layer είναι η είσοδος στο LSTM. Τα βάρη του Embedding layer μπορούν είτε να σχηματιστούν σε αυτό το στρώμα είτε να τροφοδοτηθούν σε αυτό αφού πρώτα έχουν δημιουργηθεί με κάποιον άλλο τρόπο όπως είναι το word2vec ή το GloVe των οποίων η επεξήγηση θα γίνει παρακάτω [48].

3.16 Συγκεντρωτικό Επίπεδο

Το Συγκεντρωτικό Επίπεδο (Pooling Layer) είναι ένα επίπεδο το οποίο, συνήθως, εισάγεται μεταξύ διαδοχικών συνελικτικών επιπέδων σε μια αρχιτεκτονική ενός συνελικτικού Δικτύου. Η λειτουργία του έγκειται στην προοδευτική μείωση του χωρικού μεγέθους της αναπαράστασης, στην μείωση των παραμέτρων και υπολογισμών στο δίκτυο και, συνεπώς, στον έλεγχο της υπερεκπαίδευση (overfitting) . Παρά τις όποιες χωρικές μειώσεις, το επίπεδο αυτό είναι σε θέση να διατηρεί τις πιο σημαντικές πληροφορίες της εισόδου. Στο επίπεδο αυτό, ορίζουμε μια χωρική “γειτονιά” (για παράδειγμα, ένα παράθυρο 2x2) και επιλέγουμε να διατηρήσουμε μόνο το μεγαλύτερο στοιχείο από το διαμορφωμένο πίνακα μέσα στο παράθυρο. Αντί να επιλέξουμε το μεγαλύτερο στοιχείο, θα μπορούσαμε επίσης να επιλέξουμε την μέση τιμή των στοιχείων (Average Pooling) ή το άθροισμα όλων των στοιχείων μέσα στο παράθυρο. Στην πράξη έχει αποδειχθεί ότι αποτελεσματικότερα αποδίδει το Max Pooling [47].



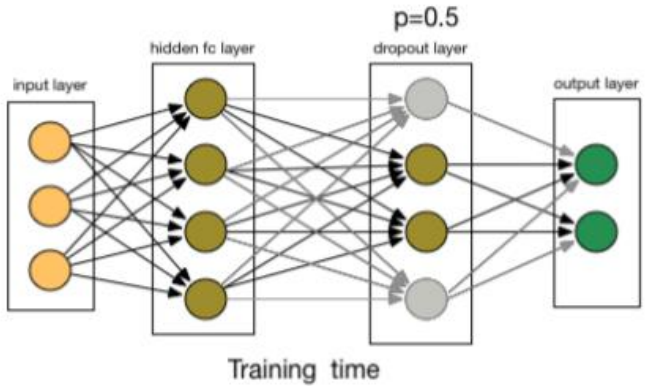
Εικόνα 27: Συγκεντρωτικό επίπεδο με max και average pooling

3.17 Πλήρως Συνδεδεμένο Επίπεδο

Όπως μπορούμε να καταλάβουμε από τον όρο Πλήρως Συνδεδεμένο Επίπεδο (Fully-Connected Layer) πρόκειται για ένα επίπεδο όπου κάθε νευρώνας σε ένα προηγούμενο επίπεδο συνδέεται με όλους τους νευρώνες του επόμενου επιπέδου. Πρόκειται για μία παραδοσιακή αρχιτεκτονική πολλών επιπέδων με νευρώνες, η οποία χρησιμοποιεί μια συνάρτηση ενεργοποίησης (συνήθως την softmax) στην έξοδό της [47].

3.18 Περιορισμός Ενεργοποίησης

Η μέθοδος ενεργοποίησης (Dropout) πρόκειται για μια προσέγγιση στην αντιμετώπιση του προβλήματος της υπερεκπαίδευσης περιλαμβάνει τον περιορισμό ενεργοποίησης (dropout). Στο επίπεδο αυτό, σε κάθε επανάληψη της εκπαίδευσης, απενεργοποιούνται τυχαία κάποιοι νευρώνες του δικτύου μαζί με όλες τις εισερχόμενες και εξερχόμενες συνδέσεις. Στο Σχήμα φαίνεται το επίπεδο περιορισμού ενεργοποίησης, όπου ο κάθε νευρώνας έχει πιθανότητα 50% να απενεργοποιηθεί (dropout=0.5) [49].



Εικόνα 28: Λειτουργία Dropout

3.19 Πυκνό στρώμα

Πρόκειται για ένα στρώμα (dense layer) το οποίο δέχεται στην είσοδό του τις εξόδους των προηγούμενων στρωμάτων και δίνει σαν έξοδο το πλήθος τιμών που επιθυμούμε να έχουμε σαν τελική έξοδο από το νευρωνικό. Στα νευρωνικά που θα κατασκευάσουμε σε αυτή την εργασία η τιμή του dense θα είναι 1 αφού επιδιώκουμε μία τιμή σαν έξοδο [49].

3.20 Διασταυρούμενη επικύρωση

Η διασταυρούμενη επικύρωση (cross validation) είναι ένας ακέραιος αριθμός ανάμεσα στο 0 και το 1 και πρόκειται για το ποσοστό των δεδομένων που θα χρησιμοποιηθούν για την επαλήθευση των δεδομένων. Θα χωρίσει σε δεδομένα εκπαίδευσης και δεδομένα αξιολόγησης τα οποία δεν θα συμμετέχουν κατά την διαδικασία της εκπαίδευσης αλλά πάνω σε αυτά θα αξιολογηθεί το μοντέλο στο τέλος της κάθε εποχής δίνοντας τις τιμές της συνάρτησης κόστους και μετρικών συναρτήσεων.

3.21 Μέγεθος Δέσμης

Το Μέγεθος δέσμης (Batch size) πρόκειται για τον αριθμό των δειγμάτων που θα χρησιμοποιηθούν από το δίκτυο πρώτού ενημερωθούν τα βάρη των νευρώνων. Για παράδειγμα αν υπάρχουν 1000 training samples και ορίσουμε batch size = 100 τότε το νευρωνικό θα ξεκινήσει με τα πρώτα 100 δείγματα από το 1 έως το 100 του training data και θα προπονήσει το δίκτυο και έπειτα θα ενημερώσει τις τιμές των βαρών. Έπειτα θα χρησιμοποιήσει τα επόμενα 100 δείγματα (101-200) και θα δουλέψει ομοίως. Ομοίως θα συνεχίσει έως ότου χρησιμοποιηθούν όλα τα training data. Ένα μικρό

batch size μειώνει την ταχύτητα της εκπαίδευσης του δικτύου, ενώ ένα μεγάλο batch size μειώνει την δυνατότητα γενίκευσης του μοντέλου σε διαφορετικά δεδομένα [50].

3.22 Αλγόριθμοι Βελτιστοποίησης

Η βελτιστοποίηση (Optimization) είναι η τεχνική που ακολουθείται προκειμένου να ελαχιστοποιηθεί (minimize) ή να μεγιστοποιηθεί maximize μια "συνάρτηση – στόχος" (objective function). Ο αλγόριθμος βελτιστοποίησης (ή ο βελτιστοποιητής) είναι η βασική προσέγγιση που χρησιμοποιείται για την εκπαίδευση ενός μοντέλου μηχανικής μάθησης για να ελαχιστοποιηθεί το σφάλμα. Δύο είναι οι βασικές μετρικές που καθορίζουν την αποτελεσματικότητα ενός αλγορίθμου βελτιστοποίησης. Αυτές είναι η ταχύτητα της σύγκλισης (δηλαδή η διαδικασία εύρεσης του ελαχίστου) και η δυνατότητα γενίκευσης (δηλαδή την ανταπόκριση του μοντέλου σε νέα δεδομένα). Δύο πολύ γνωστοί αλγόριθμοι είναι ο Adaptive Moment Estimation (Adam) και ο αλγόριθμος κατάβασης κλίσης (Stochastic Gradient Descent-SGD). Σε μία δημοσίευση είχε αναφερθεί ότι ένα μοντέλο ιδανικά θα εκπαιδευόταν «τόσο γρήγορα όσο ο Adam και τόσο καλά όσο ο SGD". Ο SGD αλγόριθμος ήταν ο πιο γνωστός τρόπος για την εκπαίδευση των βαθέων νευρωνικών δικτύων. Προτάθηκε το 1950 και ενημερώνει τις παραμέτρους του μοντέλου παρατηρώντας τις αλλαγές που πραγματοποιούνται στην συνάρτηση κόστους με σκοπό να ελαχιστοποιηθεί το σφάλμα. Πρόκειται για μία παραλλαγή της κατάβαση κλίσης (variant gradient descend) και αντί να πραγματοποιεί υπολογισμούς σε όλα τα δεδομένα εκπαίδευσης, το οποίο είναι περιττό και μη αποδοτικό, ο SGD πραγματοποιεί υπολογισμούς σε μικρότερα υποσύνολα των δεδομένων εκπαίδευσης. Είναι πολύ αποτελεσματικός αλγόριθμος γιατί πραγματοποιεί την ίδια λειτουργία με την κανονική κατάβαση κλίσης (regular gradient descent) όταν ο ρυθμός εκμάθησης είναι χαμηλός.

Ωστόσο, τα τελευταία χρόνια μεγάλος αριθμός αλγορίθμων βελτιστοποίησης έχουν προταθεί ώστε να αντιμετωπίσουν τις περιπτώσεις όπου οι μέθοδοι κατάβασης κλίσης (gradient descend) δεν είναι αποτελεσματικοί. Ένας από τους πιο γνωστούς βελτιστοποιητές που είναι ευρέως χρησιμοποιούμενος στην βαθιά μάθηση είναι ο Adam. Ουσιαστικά πρόκειται για έναν αλγόριθμο βαλτιστοποίησης των στοχαστικών συναρτήσεων κόστους που συνδυάζει τα πλεονεκτήματα των SGD και του Root Mean Square Propagation (RMSProp) και υπολογίζει τους επιμέρους ρυθμούς εκμάθησης για διαφορετικές παραμέτρους. Το σημαντικό πλεονέκτημα του αλγορίθμου αυτού είναι το γεγονός ότι η σύγκλιση επιτυγχάνεται πιο γρήγορα όμως υπάρχει ο κίνδυνος να μην καταφέρει να φτάσει το ολικό ελάχιστο της συνάρτησης κόστους.

Στην εργασία αυτή δοκιμάστηκαν και οι δύο αυτοί βελτιστοποιητές και τα αποτελέσματά τους ήταν πολύ όμοια. Για αυτό τον λόγο επιλέχθηκε να χρησιμοποιηθεί ο Adam εξαιτίας του γεγονότος ότι καταλήγει σε σύγκλιση συγκριτικά πιο γρήγορα [51].

3.23 Διανυσματική Αναπαράσταση Λέξεων

Όπως είναι γνωστό για να μπορέσει ένας υπολογιστής να επιλύσει κάποιο πρόβλημα πρέπει να του δοθούν τα δεδομένα σε μορφή κατανοητή για αυτόν. Όταν θέλουμε ένα νευρωνικό δίκτυο να επεξεργαστεί κείμενο, τότε πρέπει να το τροφοδοτήσουμε με είσοδο που μπορεί να «καταλάβει» και όχι με απλό πέρασμα των λέξεων από τις οποίες αποτελείται. Μία μορφή κατανοητή για το νευρωνικό δίκτυο είναι οι ακέραιοι αριθμοί. Για αυτό τον λόγο πρέπει να μετατρέψουμε την κάθε λέξη του κειμένου σε ένα ακέραιο αριθμό. Υπάρχουν πολλοί μέθοδοι με τους οποίους μπορούμε να το καταφέρουμε αυτό.

Ένας απλός τρόπος είναι να αντιστοιχίσουμε την κάθε λέξη του κειμένου με έναν μοναδικό αριθμό. Για παράδειγμα στην πρόταση «The cat sat on the mat" θα αντιστοιχούσαμε την λέξη cat με το 1 την λέξη mat με το 2 κλπ. Έτσι θα σχηματίζονταν το διάνυσμα [5,1,4,3,5,2]. Όμως αυτή η τεχνική είναι μη αποδοτική γιατί η αντιστοίχιση είναι αυθαίρετη και δεν εκμεταλλεύεται καμία σχέση μεταξύ των λέξεων.

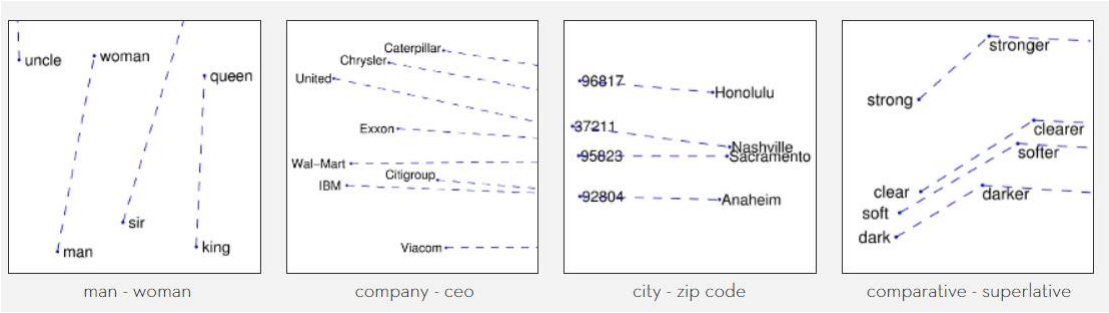
3.23.1 Εμφύτευση One-hot

Ένας άλλος απλός τρόπος είναι με την μέθοδο “one hot embedding”. Για να αναπαραστήσουμε κάθε λέξη θα δημιουργήσουμε ένα μηδενικό διάνυσμα με μήκος ίσο με το λεξιλόγιο και έπειτα θα βάλουμε «1» στο index που αντιστοιχεί στην λέξη. Για παράδειγμα στην πρόταση «the sky is blue», θα σχηματιστούν τα εξής διανύσματα: the = [1,0,0,0], sky=[0,1,0,0], is = [0,0,1,0], blue=[0,0,0,1] [52].

	the	sky	Is	Blue
The	1	0	0	0
Sky	0	1	0	0
Is	0	0	1	0
Blue	0	0	0	1

Όπως καταλαβαίνουμε το πλήθος των διανυσμάτων θα ισούται με το πλήθος των ξεχωριστών λέξεων του κειμένου που θέλουμε να τροφοδοτήσουμε στο νευρωνικό σύστημα και μάλιστα η πλειοψηφία των διανυσμάτων θα περιέχουν περισσότερα «0». Όμως αυτή η τεχνική δεν είναι αποδοτική σε κείμενα μεγάλου μήκους γιατί δημιουργείται υπερβολικά μεγάλος αριθμός διαστάσεων (όσες είναι και οι ξεχωριστές λέξεις του vocabulary) και διανυσμάτων και αυτό έχει ως συνέπεια να είναι δύσκολη και χρονοβόρα η επεξεργασία του από ένα σύστημα τεχνικής μάθησης.

Εύκολα μπορεί κάποιος να διαπιστώσει ότι οι λέξεις good και great έχουν παρόμοια σημασία και συχνά χρησιμοποιούνται μαζί. Αυτό το γεγονός εκμεταλλεύονται οι άλλοι τρόποι για την δημιουργία πιο αποδοτικών διανυσμάτων για τις λέξεις ενός κειμένου, δηλαδή προσπαθούν να μειώσουν τις διαστάσεις των ξεχωριστών διανυσμάτων που δημιουργούνται από ένα κείμενο, ώστε να είναι πιο «εύκολο» από το νευρωνικό σύστημα να επεξεργαστεί αυτά τα δεδομένα. Οι πιο γνωστές μέθοδοι είναι το GloVe και το word2vector. Και οι δύο αυτοί τρόποι βασίζονται στο γεγονός ότι σημασιολογικά παρόμοιες λέξεις τοποθετούνται σε «κοντινά» διανύσματα λέξεων, ενώ λέξεις με διαφορετική σημασία απωθούνται.



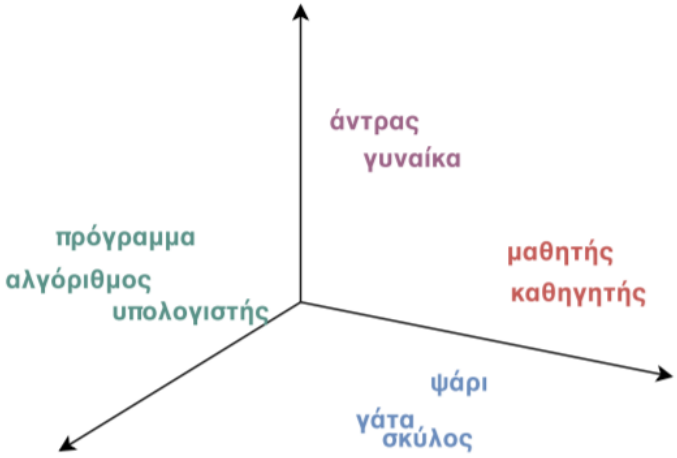
Εικόνα 29: Σημασιολογικά παρόμοιες λέξεις τοποθετούνται σε κοντινά διανύσματα λέξεων

3.23.2 Εμφύτευση Λέξης

Η εμφύτευση λέξης (word embedding**)** πρόκειται για αναπαράσταση λέξης σε διάνυσμα πραγματικών αριθμών και λίγων διαστάσεων συνήθως μερικών εκατοντάδων (100-500) και είναι μια μέθοδος που έχει αναπτυχθεί πολύ την τελευταία δεκαετία. Η λογική βασίζεται στην προσπάθεια όπου παρόμοιες σημασιολογικά λέξεις πρέπει να χαρτογραφούνται κοντά στον ίδιο διανυσματικό χώρο. Έτσι μία λέξη αναπαρίσταται πλέον όχι μονοδιάστατα αλλά ως συσχέτιση με άλλες λέξεις. Αναπαριστώντας μία λέξη ως ένα διάνυσμα και όχι μονοσήμαντα όπως με την one-hot αναπαράσταση, μπορούμε να συγκρατήσουμε τις αλληλεπιδράσεις της με τις υπόλοιπες λέξεις. Κάθε διάσταση στον διανυσματικό χώρο των λέξεων, αντιστοιχεί κατά κάποιον τρόπο σε μία αφηρημένη έννοια και η τιμή που έχει κάθε λέξη σε μία διάσταση, αντικατοπτρίζει το βαθμό στον οποίο σχετίζεται μαζί της. Το θετικό με την μέθοδο αυτή είναι ότι επειδή τα διανύσματα έχουν δημιουργηθεί από εκπαίδευση πολύ

μεγάλου dataset, έχουμε συσχετίσεις δεδομένων που μπορεί στα Δεδομένα Εκπαίδευσής μας να μην υπάρχουν[53].

Η ιδέα της αναπαράστασης μίας λέξης ως διάνυσμα είναι αρκετά παλιά. Επιπλέον, υπάρχουν αρκετά παραδείγματα λεξικών (Mohammad κ.ά. 2013; Staiano κ.ά. 2014; Cambria κ.ά. 2016) στα οποία λέξεις συσχετίζονται, με χειροκίνητο τρόπο από ειδικούς (Γλωσσολόγους, Ψυχολόγους κλπ.), με ορισμένα συναισθήματα ή συναισθηματικές καταστάσεις (“φόβος”, “χαρά”, “θυμός”, …). Στην παρακάτω εικόνα φαίνεται μια αναπαράσταση λέξεων ως διανύσματα.

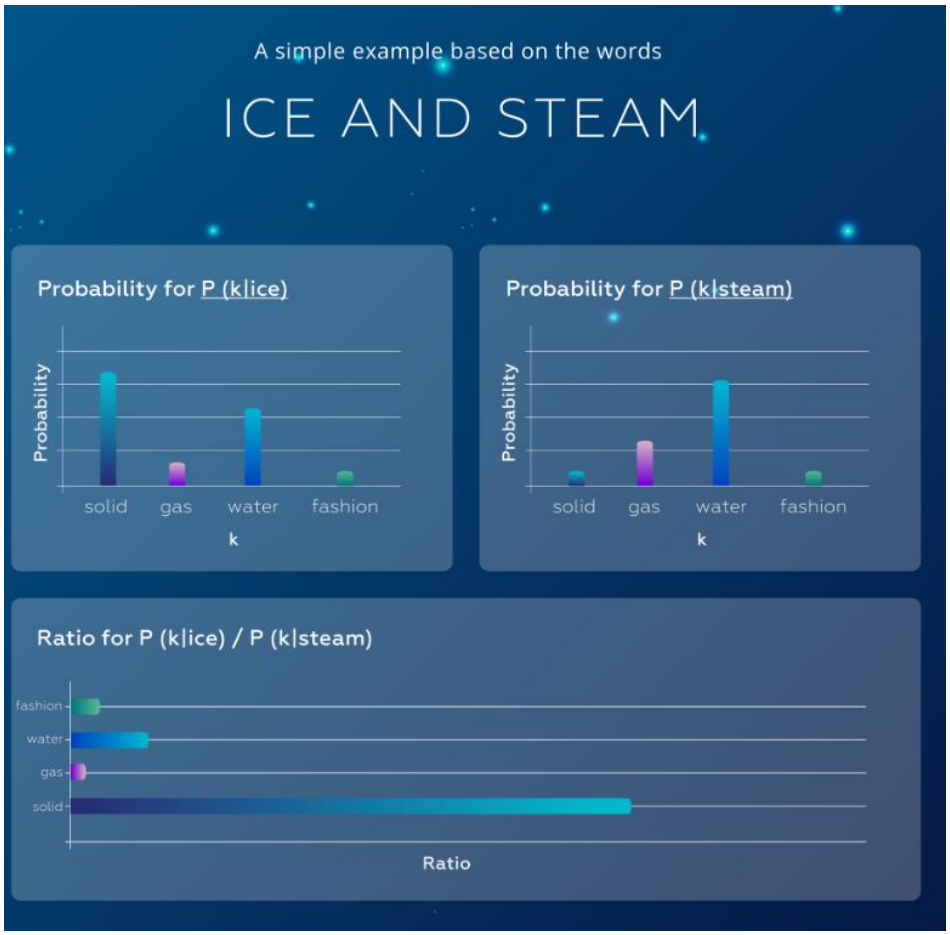


Εικόνα 30: Αναπαράσταση λέξεων ως διανύσματα

Η διανυσματική αναπαράσταση μας δίνει έναν αποδοτικό τρόπο να δημιουργήσουμε πυκνή αναπαράσταση (dense representation) όπου όμοιες σημασιολογικά λέξεις έχουν παρόμοια κρυπτογράφηση (encoding). Λέξεις που είναι συνώνυμες βρίσκονται πιο κοντά από λέξεις που είναι αντώνυμες. Ένα embedding vector είναι ένα πυκνό διάνυσμα από ακέραιους αριθμούς η διάσταση του οποίου καθορίζεται από τον χρήστη και οι τιμές αυτές έχουν προκύψει από εκπαίδευση. Δύο πολύ γνωστές μέθοδοι που χρησιμοποιούν αυτή την τεχνική είναι το GloVe και το word2vec. Ακολουθεί ένα παράδειγμα για το πώς λειτουργεί το GloVe:

3.23.3 Μέθοδος Global Vectors

Έστω P(k|w) είναι η πιθανότητα ότι η λέξη k εμφανίζεται σε κείμενο που ανήκει η λέξη w. Η λέξη ice εμφανίζεται με μεγαλύτερη συχνότητα σε κείμενα που υπάρχει η λέξη solid παρά σε κείμενα που υπάρχει η λέξη gas και ομοίως η λέξη steam εμφανίζεται πιο συχνά σε κείμενα με την λέξη gas παρά με την λέξη solid. Όμως και οι δύο λέξεις μπορούν να εμφανιστούν με παρόμοια συχνότητα σε κείμενα που περιέχουν την λέξη water με την οποία σχετίζονται οι δύο και ομοίως με παρόμοια συχνότητα με την λέξη fashion με την οποία δεν σχετίζεται καμία από τις δύο άμεσα. Με άλλα λόγια η πιθανότητα P(solid | ice) θα είναι σχετικά υψηλότερη από την πιθανότητα P(solid | steam) η οποία θα είναι σχετικά πιο χαμηλή. Άρα ο λόγος P(solid | ice) / P(solid | steam) θα είναι υψηλός. Ομοίως αν πάρουμε την λέξη Gas η οποία σχετίζεται με την λέξη steam και όχι τόσο με την λέξη ice τότε ο λόγος P(gas | ice) / P(gas | steam) θα είναι μικρός. Για λέξεις όπως το water που σχετίζονται και με τις δύο περιμένουμε ο λόγος να είναι κοντά στην μονάδα [54].



Εικόνα 31: Παράδειγμα που βασίζεται σε πιθανότητα εύρεσης λέξεων σε κείμενα

Probability and Ratio	$k = solid$	$k = gas$	$k = water$	$k = fashion$
$P(k\|ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k\|steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$P(k\|ice)/P(k\|steam)$	8.9	8.5×10^{-2}	1.36	0.96

Εικόνα 32: Αποτελέσματα του παραπάνω παραδείγματος

Αναλυτικότερα, το Global Vectors είναι ένα μοντέλο που βασίζεται σε αλγόριθμο μη επιβλεπόμενης μάθησης και έχει ως στόχο να την αντιστοίχιση των λέξεων ενός κειμένου σε διανύσματα ακεραίων αριθμών με αποδοτικό τρόπο, ώστε λέξεις με παρόμοια εννοιολογική σημασία να βρίσκονται πιο κοντά σε σχέση με λέξεις που έχουν διαφορετική σημασία. Το μοντέλο αυτό έχει αναπτυχθεί από μία ομάδα ερευνητών του πανεπιστημίου του Stanford και η εκπαίδευση έχει γίνει πάνω σε τεράστιο όγκο δεδομένων. Το κύριο πλεονέκτημα τους είναι ότι δεν χρειάζονται τον ανθρώπινο παράγοντα όπως οι γράφοι αλλά μπορούν να εξαχθούν μέσα από μεγάλες συλλογές κειμένων όπως είναι το Wikipedia. Ο αλγόριθμος κωδικοποιεί τις λέξεις χρησιμοποιώντας διανυσματικές διαφορές. Παρέχουν διάφορες παραλλαγές του μοντέλου αυτού όπως είναι η χρήση 25, 50, 100 και 300 διαστάσεων των διανυσμάτων βασιζόμενων σε 2, 6, 42, 840 δισεκατομμύρια λέξεις[54].

Word2Vec: Το Word2Vec χρησιμοποιεί ένα Ρηχό Νευρωνικό Δίκτυο (Shallow Neural Network), δηλαδή με μόνο ένα κρυφό επίπεδο (hidden layer). Η τεχνική αυτή εκτελεί το εξής "κόλπο": αρχικά εκπαιδεύουμε ένα νευρωνικό δίκτυο ώστε να μάθει να εκτελεί μία εργασία, όμως αφού το εκπαιδεύσουμε, δεν το χρησιμοποιούμε για τον σκοπό που το εκπαιδεύσαμε. Αυτό το οποίο θα χρησιμοποιήσουμε είναι τα βάρη που θα έχουν σχηματιστεί στο κρυφό επίπεδο του δικτύου! Τα βάρη που θα έχουν σχηματιστεί για κάθε λέξη θα είναι τα τελικά διανύσματα των λέξεων[53].

Οι μέθοδοι word2vec και GloVe παρουσιάζουν αρκετές ομοιότητες όμως εντοπίζονται και κάποιες διαφορές. Πιο συγκεκριμένα, η Word2vec μέθοδος είναι σχεδιασμένη κατά τέτοιο τρόπο ώστε ο υπολογισμός των μεταξύ εμφανίσεων λέξεων να συμβαίνει τοπικά (local). Για παράδειγμα, στην πρόταση 'The cat sat on the mat.' παρατηρούμε ότι η λέξη 'the' θα εμφανιστεί στα συμφραζόμενα της λέξης 'cat' και 'mat όμως δεν προσδίδει κάποιο ιδιαίτερο εννοιολογικό περιεχόμενο καθώς εμφανίζεται σε όλα τα ουσιαστικά της αγγλικής γλώσσας. Ανήκει δηλαδή σε μια κατηγορία λέξεων που ονομάζεται stop words και στην οποία ανήκουν όσες λέξεις γενικά δεν προσθέτουν εννοιολογική πληροφορία όπως τα άρθρα 'a', 'an' και άλλες λέξεις όπως οι 'and', 'but' και 'or'. Το GloVe αντιμετωπίζει το παραπάνω πρόβλημα καθώς είναι σχεδιασμένο κατά τρόπο τέτοιο ώστε να υπολογίζει την συνεμφάνιση (co-occurrence) λέξεων μέσα σε ένα μεγάλο πλήθος κειμένων (corpus) σε καθολικό βαθμό (global). Επίσης, η μέθοδος GloVe δίνει καλά αποτελέσματα και για κείμενα μικρού μήκους με μικρά διανύσματα λέξεων.

Στην εργασία αυτή προτιμήθηκε το GloVe γιατί δίνει καλύτερα αποτελέσματα δεδομένου ότι σαν κείμενο δίνονται μικρές προτάσεις (tweets) από τον διαδικτυακό μέσο twitter. Τέλος, μετά από δοκιμές φάνηκε ότι αυτό με τις 300 διαστάσεις δίνει τα καλύτερα αποτελέσματα και γιαυτό και χρησιμοποιήθηκε σε αυτή την εργασία.

Κεφάλαιο 4

4. Αρχιτεκτονικές Νευρωνικών Δικτύων

Όπως ειπώθηκε και στην εισαγωγή, στην εργασία αυτή έχουμε σαν στόχο να προβλέψουμε μελλοντικές τιμές χρηματιστηριακών μετοχών κάποιων εταιριών. Δοκιμάστηκαν δύο τεχνικές πρόβλεψης. Η μια ήταν να προβλέψουμε την συμπεριφορά των τιμών κάνοντας χρήση μόνο των παρελθοντικών τιμών της κάθε μετοχής και στην άλλη αξιοποιήθηκε επιπλέον και η πληροφορία που αντλήθηκε από σχόλια (tweets) χρηστών στα οποία υπήρχε αναφορά στην μετοχή της εταιρίας της οποίας θέλουμε να προβλέψουμε. Αφού κατασκευάστηκαν αυτές οι τεχνικές πάρθηκαν αποτελέσματα τα οποία συγκρίθηκαν μεταξύ τους με σκοπό να συμπεράνουμε αν με την εισαγωγή πληροφοριών από σχόλια βελτιώθηκε η απόδοση του συστήματος.

4.1 Νευρωνικό Μοντέλο Για Ανάλυση Συναισθήματος

Ο πρώτος στόχος ήταν να αντληθούν πληροφορίες από τα σχόλια των χρηστών. Πιο συγκεκριμένα, σημαντικό ήταν να κατασκευαστεί ένα νευρωνικό δίκτυο το οποίο θα προβαίνει σε ανάλυση συναισθήματος των tweets. Δηλαδή, για κάθε tweet να παράγει ένα αριθμό από το -1 έως το 1 ανάλογα με το πόσο αρνητικό ή θετικό είναι αντίστοιχα. Όπως συμβαίνει στον χώρο της τεχνητής μάθησης και των νευρωνικών δικτύων, κάθε μοντέλο που κατασκευάζεται πρέπει με κάποιο τρόπο να διαπιστωθεί αν είναι αποτελεσματικό και αν επιδέχεται περισσότερη βελτίωση. Για αυτό τον λόγο βρέθηκαν αντίστοιχες εργασίες που έχουν γίνει και συγκρίθηκαν τα αποτελέσματά μας με αυτά που πάρθηκαν σε αυτές τις εργασίες.

Μετά από αναζήτηση διαπιστώσαμε ότι ένα κατάλληλο dataset για το δικό μας πρόβλημα συμπεριλαμβανόταν διαγωνισμό “SemEval-2017 sentiment Analysis in Twitter”[1] . Αυτό το dataset περιέχει σχόλια από δύο διαφορετικές πλατφόρμες. Η μια είναι το microblog messages που αποτελείται από twitter και stockTwits messages και η άλλη είναι news Statements & Headlines. Κάθε ένα περιέχει δύο αρχεία. Ένα train.csv και ένα test.csv. το train.csv αποτελείται από 1700 περίπου σχόλια μαζί με το sentiment score τους. Το περιεχόμενο των tweets και των Headlines αφορά οικονομικά δεδομένα και πρόκειται για κριτικές και σχόλια ανθρώπων για διάφορες εταιρίες-μετοχές. Σκοπός των διαγωνιζόμενων ήταν να κατασκευάσουν ένα μοντέλο το οποίο θα προπονηθεί με βάση το test.csv. έπειτα καλούνται να τρέξουν το μοντέλο αυτό στα σχόλια του train.csv το οποίο δεν περιέχει τα sentiment score τα οποία θα παράξει το μοντέλο τους. Έπειτα στάλθηκαν τα αποτελέσματα στους κριτές και αυτοί που είχαν τις τιμές sentiment score του αρχείου test.csv χρησιμοποιώντας την μετρική συνάρτηση cosine similarity βρήκαν την τελική βαθμολογία της κάθε ομάδας που έλαβε μέρος στον διαγωνισμό.

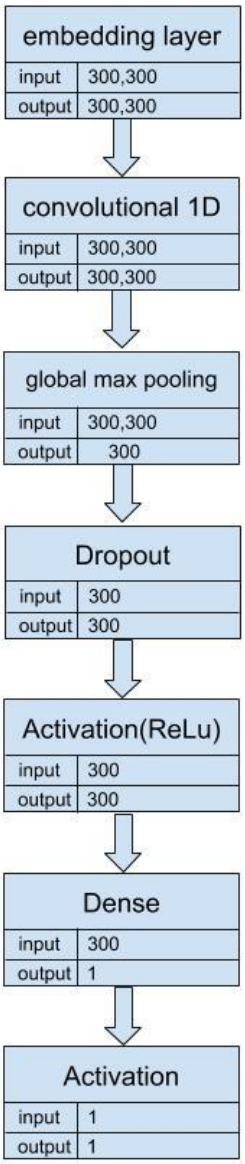
Στην εργασία αυτή κατασκευάστηκαν διάφορα μοντέλα που βασίστηκαν σε μεθόδους των διαγωνιζόμενων και συγκρίθηκαν. Κάποια από τα μοντέλα-νευρωνικά δίκτυα που κατασκευάστηκαν θα αναλυθούν παρακάτω.

[1] http://alt.qcri.org/semeval2017/task5/

.

Αρχικά γίνεται κάποια επεξεργασία των δεδομένων με τα οποία θα προπονηθεί το νευρωνικό. Ξεκινώντας, από κάθε σχόλιο είτε πρόκειται για tweet είτε πρόκειται για news headlines αφαιρέθηκαν τα σημεία στίξης, μετατράπηκαν τα κεφαλαία σε πεζά και διαγράφτηκαν πιθανά tags που μπορεί να είχαν καθώς και σύμβολα όπως είναι τα http, @username. Στην συνέχεια, πραγματοποιήθηκε η διαδικασία του tokenization, δηλαδή χωρίστηκε το κάθε σχόλιο που πρόκειται για μια συμβολοσειρά σε ξεχωριστές λέξεις. Τέλος μέσω του glove σχηματίζονται τα διανύσματα λέξεων τα οποία είναι πλέον σε θέση να αποτελέσουν είσοδο για το νευρωνικό δίκτυο που θα προβεί σε ανάλυση συναισθήματος.

Όλα τα μοντέλα που κατασκευάστηκαν και δοκιμάστηκαν δέχονται ως είσοδο τα word embeddings που έχουν σχηματιστεί, επιλέχθηκε διάσταση ίση με 300. Δοκιμάστηκαν διάφορα μοντέλα που είχαν διαφορές τόσο στην δομή των νευρώνων τους όσο και στις υπερπαραμέτρους τους. Αρχικά, δοκιμάστηκε το παρακάτω μοντέλο που φαίνεται στην εικόνα 34.

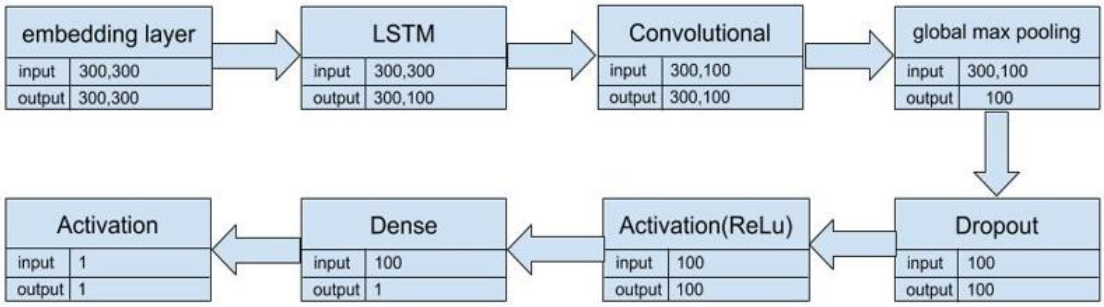


Εικόνα 33: Μοντέλο (1) για την ανάλυση συναισθήματος

Όπως βλέπουμε το πρώτο στρώμα που αποτελεί και την είσοδο του νευρωνικού είναι το embedding Layer, ακολουθεί το convolutional διάστασης 1 το οποίο ως είσοδο δέχεται την έξοδο του embedding layer. Στην συνέχεια ακολουθεί το Global Max Pooling ώστε να κρατήσει το διάνυσμα εκείνο το οποίο έχει το καλύτερο αποτέλεσμα και να αφαιρέσει τα υπόλοιπα. Στην συνέχεια ακολουθεί ο νευρώνας Dropout(0.2) με πιθανότητα p=0.2 ο οποίος είναι σημαντικός ώστε να αποφεύγεται το φαινόμενο του overfitting κατά το στάδιο της εκπαίδευσης. Έπειτα, ακολουθεί νευρώνας ενεργοποίησης με συνάρτηση ενεργοποίησης την ReLU ώστε να μειώσει την διάσταση σε 1 και τέλος το στρώμα dense και activation με συνάρτηση linear δίνουν την τελική έξοδο. Το νευρωνικό αυτό εκπαιδεύτηκε σε batch size=16 και με 50 εποχές και για συνάρτηση κόστους χρησιμοποιήθηκε η mean square error (MSE)

Αφού δοκιμάστηκε στα δεδομένα του test.csv έδωσε σαν αποτέλεσμα cosine similarity=0.67.

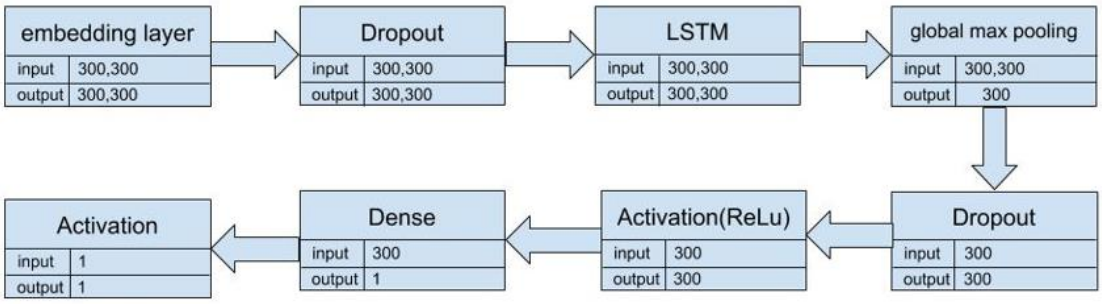
Στην συνέχεια, δοκιμάστηκε ένα παρόμοιο μοντέλο με την σημαντική διαφορά ότι προστέθηκε και ένα επίπεδο LSTM.



Εικόνα 34: Νευρωνικό μοντέλο (2) για την ανάλυση συναισθήματος με convolutional layer Και LSTM

Με την προσθήκη του στρώματος LSTM βελτιώθηκε το αποτέλεσμα σε cosine similarity = 0.70

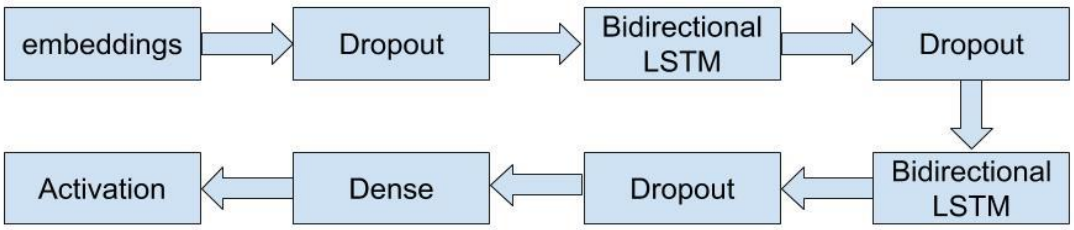
Έπειτα, το καλύτερο αποτέλεσμα δόθηκε από το ακόλουθο μοντέλο το οποίο είναι παρόμοιο με το προηγούμενο με την διαφορά ότι αφαιρέθηκε το στρώμα convolutional 1D και προστέθηκε ένα στρώμα dropout μετά την είσοδο embeddings



Εικόνα 35: Το βέλτιστο μοντέλο ανάλυσης συναισθήματος

Αυτό το μοντέλο προπονήθηκε με batch size=16 και σε 60 epoch και έδωσε ως αποτέλεσμα cosine similarity = 0.74 ! τα καλύτερα αποτελέσματα του διαγωνισμού της semeval κυμαίνονταν κοντά στο 0.75 με την νικήτρια ομάδα να πετυχαίνει σκορ 0.778.

Ένα νευρωνικό δίκτυο που κατάφερε να παράξει cosine similarity περισσότερο από 0.74 ήταν το παρακάτω:



Εικόνα 36: Μοντέλο για ανάλυση συναισθήματος με αμφίδρομα LSTM

Πιο συγκεκριμένα έφτασε να παράξει cosine similarity = 0.80, όμως δεν προτιμήθηκε γιατί είχε ένα σημαντικό μειονέκτημα. Αυτό ήταν το γεγονός ότι σχεδόν για όλα τα tweets έδινε τιμή πρόβλεψης πολύ κοντά στο 0 χωρίς να παίρνει «ρίσκα» δίνοντας πιο ακραίες τιμές μεταξύ του -1 και του 1. Ως εκ τούτου δεν πρόσφερε σημαντική πληροφορία για το πόσο θετικό ή αρνητικό ήταν το κάθε σχόλιο και γιαυτό τον λόγο δεν προτιμήθηκε αυτό το μοντέλο.

Δοκιμάστηκαν και άλλες παραλλαγές του παραπάνω μοντέλου και άλλα είδη νευρωνικών αλλά δεν καταφέραμε να παράξουμε αποτέλεσμα καλύτερο από το 0.74 που να δίνει αποτελεσματική

πληροφορία και σε σύγκριση με το σκορ της νικήτριας ομάδας τα σκορ που παράχθηκαν από εμάς ήταν αρκετά κοντά.

4.2 Ανάλυση του dataset για τα tweets

Αφού συγκρίθηκαν τα σκορ των μοντέλων μας με αυτά του διαγωνισμού και διαπιστώθηκε ότι είναι πολύ παρόμοια, χρησιμοποιήθηκε το καλύτερο για να μετρήσουμε sentiment score στα tweets που αξιοποιήθηκαν.

Η πρώτη προσπάθεια ήταν να συλλέξουμε tweets για κάποιες εταιρίες για κάποια χρονική στιγμή, αλλά αντιμετωπίσαμε διάφορα προβλήματα λόγω της πολιτικής του Twitter API το οποίο δεν επέτρεπε να κατεβάσουμε μεγάλο αριθμό Tweets μέσα σε μικρό χρονικό διάστημα αφού έχει θέσει ένα άνω όριο. Για αυτό τον λόγο κρίθηκε αναγκαία η αναζήτηση στο διαδίκτυο ώστε να βρεθεί dataset το οποίο να περιέχει σχόλια-tweets από χρήστες για διάφορες γνωστές εταιρίες. Αυτό το dataset βρέθηκε και ήταν το NASDAQ[2] του οποίου η περιγραφή δίνεται παρακάτω.

Το dataset που χρησιμοποιήθηκε για την πρόβλεψη των χρηματιστηριακών τιμών είναι το NASDAQ το οποίο περιέχει πληροφορίες για 105 γνωστές εταιρίες. Αναλυτικότερα, για κάθε μια περιέχει ένα csv αρχείο το οποίο έχει tweets με hashtag το όνομα της εταιρίας για μια συγκεκριμένη χρονική περίοδο. Στην επόμενη εικόνα φαίνεται η μορφή του csv αρχείου για την google

Tweet Id	Date	Hour	User Name	Nickname	Tweet content	F
7,43E+17	14/6/2016	6:35	Jim Stevenson	JamesOliverTrad	$GOOG short at 718.02 #trade ideas #investing #stocks #	
7,43E+17	14/6/2016	1:13	Sunny G	ABraveBull	Okay, iMessage revival is finally here. $AAPL $GOOG https://t.co/BgACf2uf3Q	
7,42E+17	13/6/2016	18:08	Mark Ludlow	MarkLudlow	Smarter Apple Maps. Traffic. Food w/ Reviews. Competing with $GOOG Maps and Yelp. https://t.co/OP5Bfyjw3y	
7,42E+17	12/6/2016	0:10	Cornholio	RChang6	@ThomasRice16 the new 6.4" phone with $GOOG tech looks pretty sweet	
7,4E+17	7/6/2016	2:48	Anthony Wang	AnthonyWanger	Google aims to kill passwords by the end of this year https://t.co/XYhmeQqBOw #cybersecurity $GOOG	

Εικόνα 37: δομή του dataset NASDAQ

Όπου τα πεδία που κρίθηκαν σημαντικά ήταν η ημερομηνία Date, hour και το κείμενο (tweet content) ενώ τα πεδία "user Name" και "Nickname" αφαιρέθηκαν αφού δεν δίνουν κάποια αξιοποιήσιμη πληροφορία. Και για τις 105 εταιρίες περιέχει tweets διάρκειας περίπου 2 μηνών (4-4-1016 έως 15-6-2016). Ανάλογα με την εταιρία διαφέρει και ο αριθμός των tweets που περιέχει για αυτό το διάστημα. Στην apple για παράδειγμα υπήρχαν 166632 tweets, μερικά από τα οποία ήταν retweets (RT), ενώ στην google υπήρχαν 37643 tweets.

Προκειμένου να είναι σε κατάλληλη μορφή ώστε να τροφοδοτηθούν στο νευρωνικό σύστημα το οποίο θα προβεί σε ανάλυση συναισθήματος, πρέπει αυτά τα δεδομένα του dataset να υποστούν κάποιο είδος επεξεργασίας. Αρχικά, αφαιρέθηκαν όλα τα σημεία στίξης και όλοι οι χαρακτήρες έγιναν πεζοί. Επίσης κρίθηκε σημαντικό να αφαιρεθούν όλα τα tags καθώς και τα http:// και οι αναφορές σε άλλους users των tweets. Έπειτα έγινε το tokenization προκειμένου να είναι σε θέση να σχηματιστούν τα διανύσματα λέξεων τα οποία θα αποτελέσουν την είσοδο του μοντέλου. Μετά την επεξεργασία τα δεδομένα που φαίνονται στην παραπάνω εικόνα έχουν την μορφή που δίνεται στο παρακάτω σχήμα:

Tweet Id	Date	Text
7 42606E+17	14/6/2016 6:35	GOOG short at trade ideas investing stocks
7 42525E+17	14/6/2016 1:13	Okay iMessage revival is finally here AAPL GOOG
7 42419E+17	13/6/2016 18:08	Smarter Apple Maps Traffic Food Reviews Competing with GOOG Maps and Yelp
7 41785E+17	12/6/2016 0:10	ThomasRice the new phone with GOOG tech looks pretty sweet
7 40012E+17	7/6/2016 2:48	Facebook and Google are the King Kong and Godzilla of digital advertising and everyone els Read more GOOG

Εικόνα 38: Τα δεδομένα μετά την επεξεργασία

Πλέον τα δεδομένα κειμένου είναι έτοιμα να αποτελέσουν είσοδο στο μοντέλο που πραγματοποιεί ανάλυση συναισθήματος. Στο παραπάνω παράδειγμα αφού τα σχόλια περάσουν από το νευρωνικό μοντέλο θα παραχθούν τα παρακάτω αποτελέσματα:

[2] http://followthehashtag.com/datasets/nasdaq-100-companies-free-twitter-dataset/

Tweet Id	Date	Text	prediction
7 42606E+17	14/6/2016 6:35	GOOG short at trade ideas investing stocks	-0.32927915
7 42525E+17	14/6/2016 1:13	Okay iMessage revival is finally here AAPL GOOG	0.03999792
7 42419E+17	13/6/2016 18:08	Smarter Apple Maps Traffic Food Reviews Competing with GOOG Maps and Yelp	-0.120685905
7 41785E+17	12/6/2016 0:10	ThomasRice the new phone with GOOG tech looks pretty sweet	0.3023079
7 40012E+17	7/6/2016 2:48	Facebook and Google are the King Kong and Godzilla of digital advertising and everyone else Read more GOOG	0.24815816

Εικόνα 39: Τα αποτελέσματα που προέκυψαν από το μοντέλο της ανάλυσης συναισθήματος

Όπως φαίνεται οι προβλέψεις δεν απέχουν πολύ από την πραγματικότητα. Με μπλε χρώμα είναι η πρόβλεψή των σχόλιων τα οποία κρίθηκαν ως ουδέτερα, με κόκκινα εκείνα τα οποία προβλέφθηκαν αρνητικά και με πράσινο αυτά που θεωρήθηκαν ότι το περιεχόμενό τους ήταν θετικό. Σε αυτό το σημείο είναι σημαντικό να τονιστεί ότι δυστυχώς υπάρχουν πολλά tweets των οποίο το περιεχόμενο δεν δίνει αξιοποιήσιμη πληροφορία και μάλλον επηρεάζει αρνητικά το τελικό αποτέλεσμα. Για παράδειγμα υπάρχουν tweets που έχουν ως περιεχόμενο για παράδειγμα "AAPL http://" ή μπορεί να περιέχουν την λέξη AAPl και να ακολουθεί κάποια εικόνα η οποία προφανώς δεν μπορεί να αξιοποιηθεί από το μοντέλο μας. Επιπρόσθετα υπάρχουν πολλά retweets RT δηλαδή σχόλια που έχουν γραφτεί από κάποιους χρήστες και έχουν αναρτηθεί και από άλλους. Αυτά όμως δεν αφαιρέθηκαν με την λογική ότι αυτά είναι πιθανό να έχουν ιδιαίτερη σημασία στο ευρύ κοινό αφού αναρτώνται από διαφορετικούς χρήστες.

Είναι σημαντικό σε αυτό το σημείο να τονιστεί ότι το dataset NASDAQ περιέχει μόνο τα tweets και όχι κάποιο σκορ ανάλυσης συναισθήματος. Οπότε, τα αποτελέσματα που παράχθηκαν από το μοντέλο της ανάλυσης συναισθήματος δεν μπορούν να επιβεβαιωθούν όλα για την ορθότητά τους αφού το πλήθος τους είναι υπερβολικά μεγάλο. Όμως το μοντέλο που χρησιμοποιήθηκε έδωσε αρκετά ικανοποιητικά αποτελέσματα κατά την εκπαίδευσή του και κατά την σύγκριση του με τα αποτελέσματα του διαγωνισμού semeval, οπότε υποθέτουμε ότι και τα αποτελέσματα predictions που έχει παράξει στο NASDAQ θα είναι αν όχι όλα τα περισσότερα ικανοποιητικά.

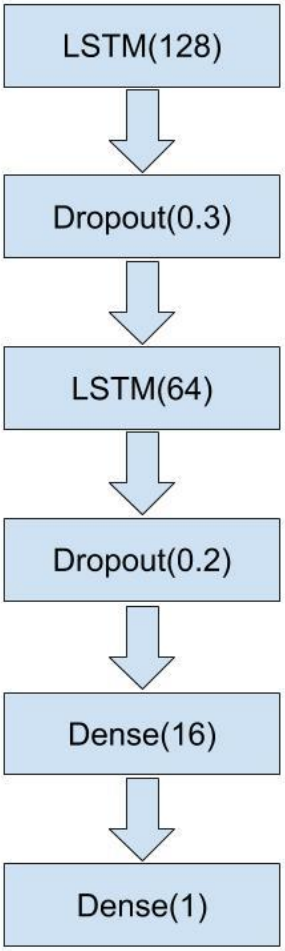
4.3 Νευρωνικό δίκτυο για πρόβλεψη χρηματιστηριακών τιμών

Το επόμενο βήμα της εργασίας ήταν να προβλέψουμε την συμπεριφορά των χρηματιστηριακών τιμών με βάση μόνο τις τιμές του παρελθόντος. Παρακάτω θα επεξηγηθούν κάποιες χρηματιστηριακές τιμές που χρησιμοποιήθηκαν σε αυτή την εργασία

- **Μετοχή (stock)** είναι ένα από τα ίσα μερίδια, στα οποία διαιρείται το κεφάλαιο μιας εταιρίας. Η μετοχή, ως αξιόγραφο, ενσωματώνει τα δικαιώματα του μετόχου που πηγάζουν από τη συμμετοχή του στην εταιρία. Τα δικαιώματα αυτά, είναι ανάλογα του αριθμού μετοχών που κατέχει ο μέτοχος
- **Open price:** είναι η τιμή που έχει μια μετοχή όταν ανοίγει το χρηματιστήριο σε μια μέρα συναλλαγών. Για παράδειγμα το χρηματιστήριο της νέας Υόρκης (NYSE) ανοίγει τις καθημερινές ακριβώς στις 9:30 π.μ. Η τιμή της πρώτης συναλλαγής είναι η «τιμή ανοίγματος» ή αλλιώς open price. Ομοίως για τις μετοχές κάποιας εταιρίας.
- **Close price:** πρόκειται για την τιμή που έχει μια μετοχή όταν κλείνει το χρηματιστήριο για εκείνη την ημέρα συναλλαγών. Δηλαδή είναι η τιμή της τελευταίας συναλλαγής που πραγματοποιείται για εκείνη την ημέρα. Επίσης η τιμή κλεισίματος θεωρείται η ακριβέστερη αποτίμηση μιας μετοχής μέχρι να συνεχιστεί την επόμενη μέρα.
- **High price:** είναι η μέγιστη τιμή που φτάνει μια μετοχή κατά την διάρκεια μιας ημέρας συναλλαγών.
- **Low price:** είναι η ελάχιστη τιμή που φτάνει μια μετοχή κατά την διάρκεια μια ημέρας συναλλαγών.
- **Volume:** πρόκειται για το πλήθος των συναλλαγών μιας μετοχής που πραγματοποιούνται κατά την διάρκεια μια ημέρας συναλλαγών[55].

Στόχος σε αυτό το βήμα είναι η πρόβλεψη των μελλοντικών τιμών του close μιας μετοχής αξιοποιώντας τις ιστορικές τιμές της. Όπως έγινε και στο μοντέλο της ανάλυσης συναισθήματος έτσι και εδώ έπρεπε να κατασκευαστεί ένα νευρωνικό και μετά να συγκριθεί η απόδοσή του σε σχέση με

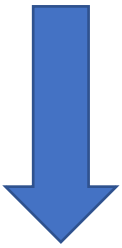
άλλες ερευνητικές εργασίες στα ίδια dataset που χρησιμοποίησαν. Πιο συγκεκριμένα το μοντέλο που κατασκευάστηκε βασίστηκε στο paper " Predicting Stock Prices Using LSTM" που γράφτηκε από τους Murtaza Roondiwala, Harshal Patel, Shraddha Varma[30] οι οποίοι χρησιμοποιώντας νευρωνικά δίκτυα LSTM προσπαθούν να προβλέψουν την συμπεριφορά του close στο άμεσο μέλλον. Ακολουθώντας εν μέρει την μεθοδολογία τους χρησιμοποιήσαμε το παρακάτω νευρωνικό δίκτυο που απεικονίζεται στην εικόνα 41.



Εικόνα 40: Νευρωνικό δίκτυο (1) για την πρόβλεψη χρηματιστηριακών τιμών

Για να αξιολογηθεί αυτό το μοντέλο χρησιμοποιήθηκε ένα dataset που περιείχε χρηματιστηριακές τιμές του χρηματιστηρίου της NIFTY. Αναλυτικότερα, είχε τις τιμές close, open, high, low, volume για 6 χρόνια (1462 ημέρες), από 3/11/2011 έως 30/12/2016. Πρώτού τροφοδοτηθεί το νευρωνικό σύστημα που φαίνεται παραπάνω έπρεπε να γίνει μια επεξεργασία στα τα δεδομένα. τα δεδομένα δεν είναι κανονικοποιημένα και το εύρος για κάθε χρηματιστηριακή τιμή ποικίλλει, ιδιαίτερα η τιμή Volume. Η κανονικοποίηση των δεδομένων βοηθάει το νευρωνικό σύστημα να συγκλίνει δηλαδή να βρίσκει πιο εύκολα τα τοπικά και ολικά ελάχιστα. Οπότε σημαντικό ήταν να γίνει scaler transform (MinMaxScalere from Sci-kit Learn) δηλαδή να μετατραπούν όλα τα δεδομένα της εισόδου σε αριθμούς δεκαδικούς ανάμεσα στο -1 και το 1. Παρακάτω φαίνεται αυτή η ενέργεια.

Date	Open	High	Low	Close1
2011-01-03	6177.450195	6178.549805	6147.200195	6157.600098
2011-01-04	6172.750000	6181.049805	6124.399902	6146.350098
2011-01-05	6141.350098	6141.350098	6062.350098	6079.799805
2011-01-06	6107.000000	6116.149902	6022.299805	6048.250000
2011-01-07	6030.899902	6051.200195	5883.600098	5904.600098



Date	Open	High	Low	Close1
2011-01-03	0.346478	0.345948	0.367752	0.362395
2011-01-04	0.345430	0.346504	0.362564	0.359868
2011-01-05	0.338431	0.337674	0.348444	0.344920
2011-01-06	0.330774	0.332069	0.339330	0.337833
2011-01-07	0.313810	0.317623	0.307767	0.305567

Εικόνα 41: λειτουργία της κανονικοποίησης

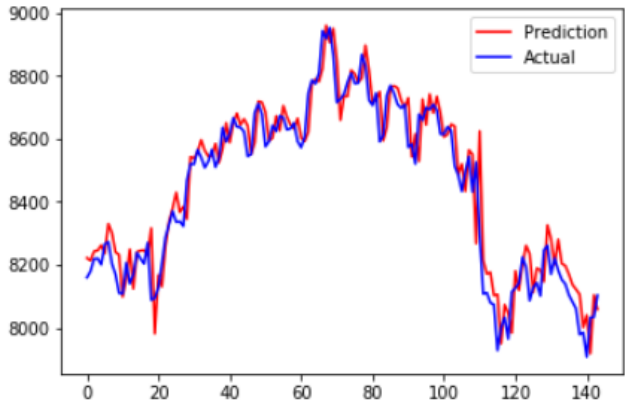
Έπειτα αναγκαίο ήταν να χωριστούν τα δεδομένα σε δεδομένα που θα αξιοποιηθούν για το training και σε δεδομένα που θα χρησιμοποιηθούν για τον έλεγχο (test). επιλέχθηκε το 90% των συνολικών δεδομένων να χρησιμοποιηθούν για την εκπαίδευση (training) και το 10% για τον έλεγχο του μοντέλου. Στην συνέχεια, έπρεπε να οριστεί και το διάστημα των ημερών του παρελθόντος που θα αξιοποιούνταν για την πρόβλεψη των μελλοντικών τιμών. Αυτό ορίζεται ως (window) και δόθηκε η τιμή ίση με 22. Δηλαδή το μοντέλο θα «κοιτάει» τις τιμές των προηγούμενων 22 ημερών για να προβλέψει την τιμή της επόμενης.

Για να γίνει κατανοητό, έστω ότι για απλότητα επιλέγεται window=3, δηλαδή θα αξιοποιηθούν οι 3 προηγούμενες μέρες για να προβλεφθεί η επόμενη όπως φαίνεται στο επόμενο σχήμα.

	Open	High	Low	Close	Volume
0	0.6277	0.6362	0.6201	0.6201	2575579
1	0.6201	0.6201	0.6122	0.6201	1764749
2	0.6201	0.6201	0.6037	0.6122	2194010
3	0.6122	0.6122	0.5798	0.5957	3255244
4	0.5957	0.5957	0.5716	0.5957	3696430
5	0.5957	0.6037	0.5878	0.5957	2778285
6	0.5957	0.6037	0.5957	0.5957	2337096

Εικόνα 42: Παράδειγμα πρόβλεψης με window=3

Αφού εκπαιδεύτηκε το νευρωνικό που φαίνεται στην παραπάνω εικόνα χρησιμοποιώντας ως συνάρτηση κόστους την mean square error (MSE), batch size = 50 και για 500 epoch και με optimizer τον “adam” προέκυψαν οι παρακάτω προβλέψεις.



Εικόνα 43: Πρόβλεψη τιμής κλεισίματος του NIFTY από το νευρωνικό (1). Η μπλε καμπύλη είναι της πραγματικής τιμής και η κόκκινη είναι η καμπύλη της πρόβλεψης

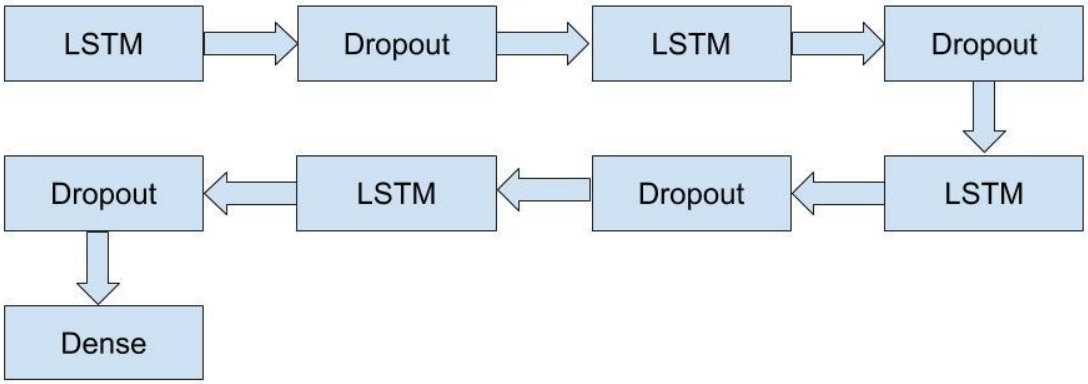
Τα αποτελέσματα του χρηματιστηρίου NIFTY που κατάφεραν στην ερευνητική εργασία σε σχέση με αυτά που καταφέραμε εμείς φαίνονται στον παρακάτω πίνακα

	RMSE του μοντέλου μας	RMSE του μοντέλου στο Paper
TRAIN	0.015485	0.01332
TEST	0.016297	0.01236

Βλέπουμε ότι τα αποτελέσματα μας είναι αρκετά παρόμοια. Σε αυτό το σημείο κληθήκαμε να αντιμετωπίσουμε ένα πρόβλημα. Όπως αναφέρθηκε και νωρίτερα το dataset που χρησιμοποιήθηκε για τα tweets-σχόλια των χρηστών είχε διάρκεια μόλις δύο μηνών ! ενώ το παραπάνω μοντέλο αξιοποίησε δεδομένα $3^{ων}$ χρόνων ! Είναι προφανές ότι το νευρωνικό σύστημα προκειμένου να παράξει ορθά αποτελέσματα χρειάζεται αρκετά περισσότερα δεδομένα εισόδου και όχι μόνο 70 τιμές, όσες δηλαδή είναι και οι ημέρες των οποίων είναι διαθέσιμα τα Tweets. Αυτό το πρόβλημα αντιμετωπίστηκε βρίσκοντας χρηματιστηριακές τιμές κάθε ώρας για τις μέρες που ήταν ανοικτά τα χρηματιστήρια. Δηλαδή, για το διάστημα (1/4/2016 έως 15/6/2016) αντί να χωριστούν τα δεδομένα σε τιμές ανά μέρα χωρίστηκαν ανά ώρα. Τα χρηματιστήρια ξεκινούν την λειτουργία τους στις 9:30 το πρωί και λειτουργούν μέχρι τις 15:00 το απόγευμα. Άρα βρέθηκαν οι intraday τιμές, 7 στο σύνολο για κάθε μέρα στο διάστημα των 2 μηνών που εξετάζουμε. Έτσι αυξήθηκαν τα δεδομένα σε περίπου 400 δεδομένα εισόδου αριθμός ο οποίος είναι αρκετά πιο ικανοποιητικός σε σχέση με τα 60 δεδομένα. Ομοίως και σε αυτή την περίπτωση ορίστηκε ως window = 22 intraday τιμές.

Για αυτά τα δεδομένα που χρησιμοποιήθηκε το νευρωνικό δίκτυο που περιεγράφηκε παραπάνω κάνοντας κάποιες αλλαγές σε τιμές υπερπαραμέτρων προκειμένου να προκύψουν καλύτερα αποτελέσματα τα οποία είναι συγκεντρωμένα στην επόμενη ενότητα.

Στην συνέχεια, κατασκευάστηκε ένα άλλο νευρωνικό δίκτυο με κάποιες αλλαγές προκειμένου να παραχθούν περισσότερα αποτελέσματα και να συγκριθούν οι αποδόσεις των δύο νευρωνικών μοντέλων. Σε αυτό χρησιμοποιήθηκαν τέσσερα LSTM με ενδιάμεσους νευρώνες Dropout και έναν νευρώνα Dense στο τέλος ώστε να παράξει το τελικό αποτέλεσμα-πρόβλεψη. Στο επόμενο σχήμα δίνεται η μορφή του μοντέλου



Εικόνα 44: νευρωνικό δίκτυο (2) για την πρόβλεψη χρηματιστηριακής τιμής κλεισίματος

Όπως και στο προηγούμενο μοντέλο έτσι και σε αυτό ως συνάρτηση κόστους χρησιμοποιήθηκε η MSE με batch size 15, validation split=0.1 και 250 epoch.

Στο τελικό στάδιο την εργασίας χρησιμοποιήθηκαν τα δύο μοντέλα που προβαίνουν πρόβλεψη μελλοντικών χρηματιστηριακών τιμών και απλά προστέθηκε στην είσοδο τους επιπλέον και τα sentiment score τα οποία παράχθηκαν από το μοντέλο του sentiment analysis. Αναλυτικότερα, από το dataset NASDAQ που περιέχει τα tweets-σχόλια χωρίστηκαν ανά ώρες εντός του διαστήματος 8:30 έως 15:00 και για κάθε ώρα υπολογίστηκε ο Μέσος Όρος των sentiment scores των tweets και προστέθηκαν δίπλα στις τιμές close, open, high, low, volume, sentiment score, ώστε να δοθούν ως είσοδο στο νευρωνικό. Ο σκοπός ήταν αν αποφανθεί αν όντως θα βελτιωθούν τα αποτελέσματα που παράγει το σύστημα αξιοποιώντας επιπλέον και πληροφορίες κειμένων και σχόλιων των ανθρώπων.

Κεφάλαιο 5

5. Αποτελέσματα

Χρησιμοποιώντας τα νευρωνικά δίκτυα που παρουσιάστηκαν στην προηγούμενη ενότητα παράχθηκαν κάποια αποτελέσματα πρόβλεψης χρηματιστηριακών τιμών κάποιων μεγάλων και γνωστών εταιριών-μετοχών. Η μεθοδολογία που ακολουθήθηκε προκειμένου να καταλήξουμε στα αποτελέσματα για την κάθε εταιρία-μετοχή ήταν η ακόλουθη. Αρχικά από το dataset NASDAQ αποσυμπιέστηκε το αρχείο εκείνο που περιείχε τα tweets των χρηστών για την περίοδο των 2 περίπου μηνών. Έπειτα έγινε η προ-επεξεργασία του κειμένου δηλαδή αφαίρεση σημείων στίξης, όλοι οι χαρακτήρες πεζοί, tokenization και σχηματισμός της διανυσματικής αναπαράστασης λέξεων ώστε να είναι έτοιμο να τροφοδοτηθεί στο νευρωνικό δίκτυο το οποίο πραγματοποιεί την ανάλυση συναισθήματος. Είναι σημαντικό να επισημανθεί ότι το μοντέλο αυτό έχει ήδη εκπαιδευτεί και δεν χρειάζεται να το εκπαιδεύουμε ξανά κάθε φορά που θέλουμε να παράξουμε sentiment score για τα tweets. Έτσι λοιπόν για κάθε tweet παράγει ένα sentiment score από το -1 έως το 1 αναλόγως το πόσο θετικό ή αρνητικό είναι.

Στο επόμενο βήμα, θα εκτελεστούν τα επόμενα δύο νευρωνικά συστήματα τα οποία θα πραγματοποιήσουν και την πρόβλεψη της μελλοντικής τιμής του close των μετοχών που επιθυμούμε να εξετάσουμε. Πραγματοποιήθηκαν διάφορες εκτελέσεις των μοντέλων αυτών δίνοντας κάθε φορά διαφορετικό συνδυασμό εισόδων από τα δεδομένα μας. Παρακάτω φαίνεται μια εικόνα που περιέχει την μορφή των δεδομένων που θα χρησιμοποιηθούν για την πρόβλεψη.

DATE	OPEN	HIGH	CLOSE	LOW	VOL	SENTIMENT
2016.04.01 09:30:00	738.6	742.4	741.77	737	739397	0.002987505500000001
2016.04.01 10:00:00	741.77	745.92	745.5	740.61	743888	0.16166392514285713
2016.04.01 11:00:00	745.55	748.98	747.2	744.54	746937	0.220657753
2016.04.01 12:00:00	747.44	747.92	747.59	744.01	746114	0.1690107558
2016.04.01 13:00:00	747.1	748.73	747.25	746.83	747.8	0.17310998261538463
2016.04.01 14:00:00	747.66	749	748.34	746.75	747788	0.13911896790862066
2016.04.01 15:00:00	748.41	750.34	749.91	747.8	749157	0.16621910582954544
2016.04.04 09:30:00	750.06	752.8	748.87	747.05	749737	0.08087288763636365
2016.04.04 10:00:00	749	749.42	744.05	744.05	746241	0.08615547400000001
2016.04.04 11:00:00	744.19	744.74	743.21	742.43	743393	0.1407594
2016.04.04 12:00:00	742.65	746.97	745.57	742.45	745362	0.12941858776470588
2016.04.04 13:00:00	745.44	746.72	744.08	744.08	745402	0.19201470810000001
2016.04.04 14:00:00	744.31	744.39	743.07	742.7	743608	0.13146961740526317
2016.04.04 15:00:00	743.06	745.31	745.31	742.75	744156	0.10326999711538463
2016.04.05 09:30:00	738	742.8	737.93	735.37	738391	0.17289755125

Εικόνα 45: Μορφή των δεδομένων που θα αξιοποιηθούν από τα νευρωνικά μοντέλα

Όπως φαίνεται υπάρχουν 6 τιμές που μπορούν να αξιοποιηθούν για την πρόβλεψη οι οποίες είναι:

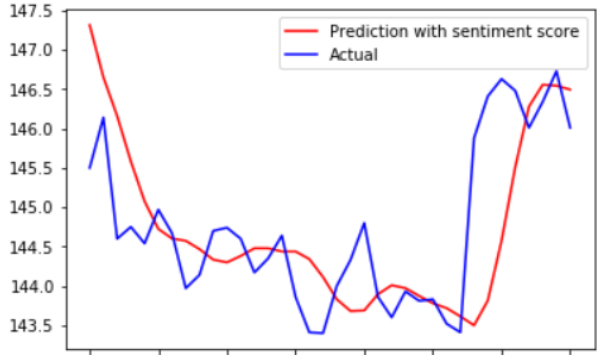
- Open
- High
- Close
- Low
- Volume
- Sentiment score

Όπως θα φανεί και παρακάτω έγιναν κάποιοι συνδυασμοί με σκοπό την επίτευξη καλύτερων αποτελεσμάτων. Το μέτρο σύγκρισης είναι το Root Mean Square Error το οποίο είναι το Mean Square error στην τετραγωνική ρίζα. Όσο πιο μικρή είναι η τιμή τόσο καλύτερη είναι η πρόβλεψη, δηλαδή οι γραφικές παραστάσεις των πραγματικών τιμών και των προβλέψεων δεν αποκλίνουν σημαντικά. Θεωρούμε την παραδοχή ότι όπου αναφέρεται η λέξη «νευρωνικό(1)» εννοούμε ότι είναι το νευρωνικό της εικόνα 41 που αποτελείται από LSTM 16 νευρώνων ενώ όταν αναφέρεται το

«νευρωνικό(2) εννοούμε ότι είναι το νευρωνικό της εικόνας 45 το οποίο αποτελείται από LSTM 96 νευρώνων.

5.1 Αποτελέσματα για την εταιρία Apple

Αρχικά δίνοντας στο νευρωνικό σύστημα (1) μόνο τις παρελθοντικές τιμές της τιμής close και της ανάλυσης συναισθήματος και αξιοποιώντας αυτές τις πληροφορίες για διάστημα 22 (window=22) προηγούμενων τιμών έδωσε τα ακόλουθα αποτελέσματα:

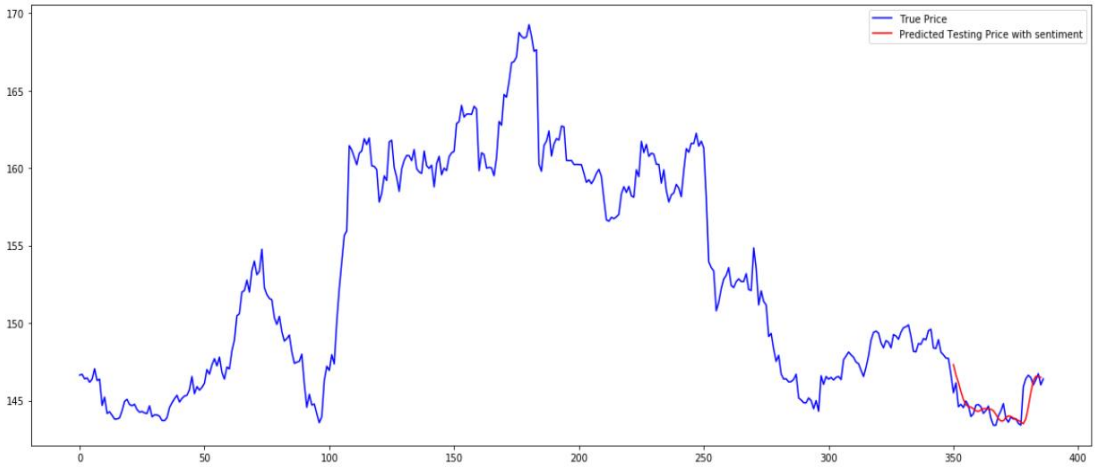


Εικόνα 46: Πρόβλεψη μετοχής AAPL του νευρωνικού (1), χρησιμοποιώντας close και sentiment score

Στο διπλανό σχήμα η μπλε γραμμή αντιστοιχεί στις πραγματικές τιμές της χρηματιστηριακής τιμής close ενώ η κόκκινη στην πρόβλεψη που πραγματοποίησε το νευρωνικό σύστημα (1)

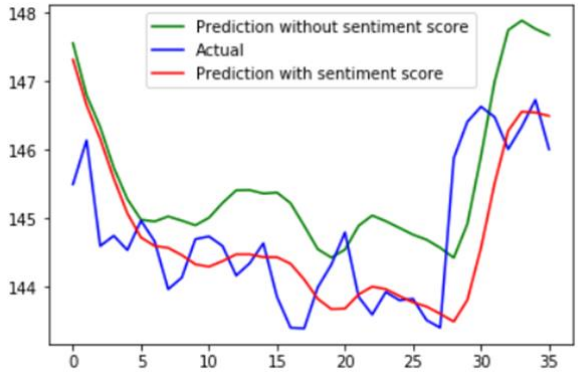
Το root mean square error (RMSE) ήταν 0.024526 και όπως βλέπουμε τα αποτελέσματα είναι αρκετά ικανοποιητικά.

Παρακάτω δίνεται η εικόνα της τιμής close της μετοχής καθώς και η πρόβλεψη της πρόβλεψης σε όλο το διάστημα που αξιοποιήθηκε



Εικόνα 47: Τιμή κλεισίματος της μετοχή AAPL σε όλο το διάστημα που αξιοποιήθηκε για την πρόβλεψη

Στην συνέχεια, το ίδιο νευρωνικό (1) τροφοδοτήθηκε μόνο από τις παρελθοντικές τιμές του close χωρίς να δοθούν οι τιμές του sentiment analysis. Τα αποτελέσματα που προέκυψαν φαίνονται παρακάτω

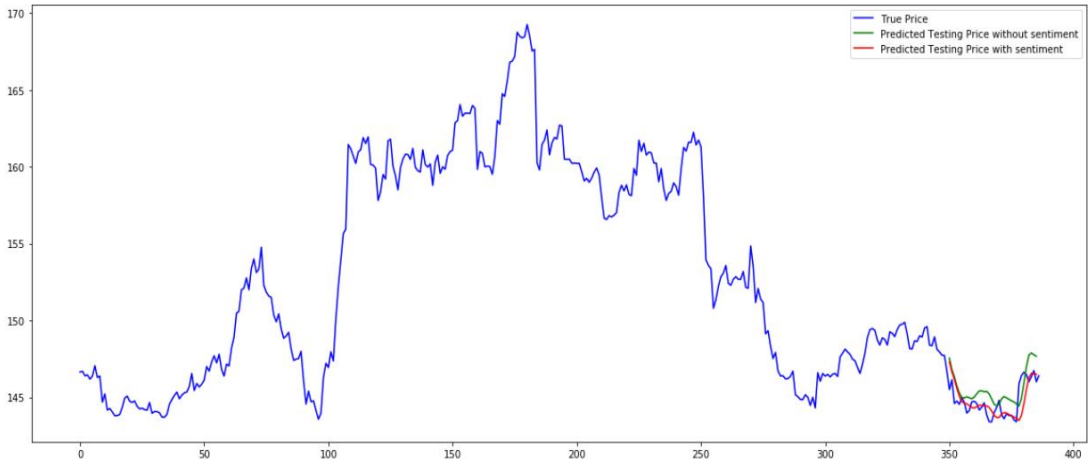


Εικόνα 48: Πρόβλεψη μετοχής AAPL του νευρωνικού (1) χρησιμοποιώντας α) close και sentiment β) close

Σε αυτή την περίπτωση με μπλε χρώμα είναι ομοίως η πραγματική συμπεριφορά της τιμής close ενώ με κόκκινη είναι η πρόβλεψη χρησιμοποιώντας τις τιμές close και sentiment analysis ενώ η πράσινη δείχνει τις προβλέψεις αξιοποιώντας μόνο τις παρελθοντικές τιμές της τιμής close.

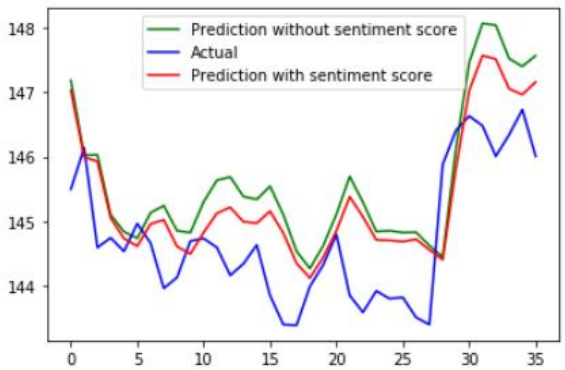
Το RMSE της πράσινης 0.050174 δηλαδή μεγαλύτερο από αυτό που πετύχαμε όταν δόθηκαν ως είσοδο και οι τιμές του sentiment analysis. Αυτό φανερώνεται και στο σχήμα αφού εύκολα κανείς παρατηρεί ότι η κόκκινη γραμμή βρίσκεται πιο κοντά στην μπλε σε σύγκριση με την πράσινη.

Ομοίως στο επόμενο σχήμα δίνεται η πρόβλεψη και όλο το διάστημα που αξιοποιήθηκε



Εικόνα 49: Τιμή κλεισίματος της μετοχής AAPL του νευρωνικού (1) σε όλο το διάστημα που αξιοποιήθηκε για την πρόβλεψη α) με sentiment scores β) χωρίς sentiment scores

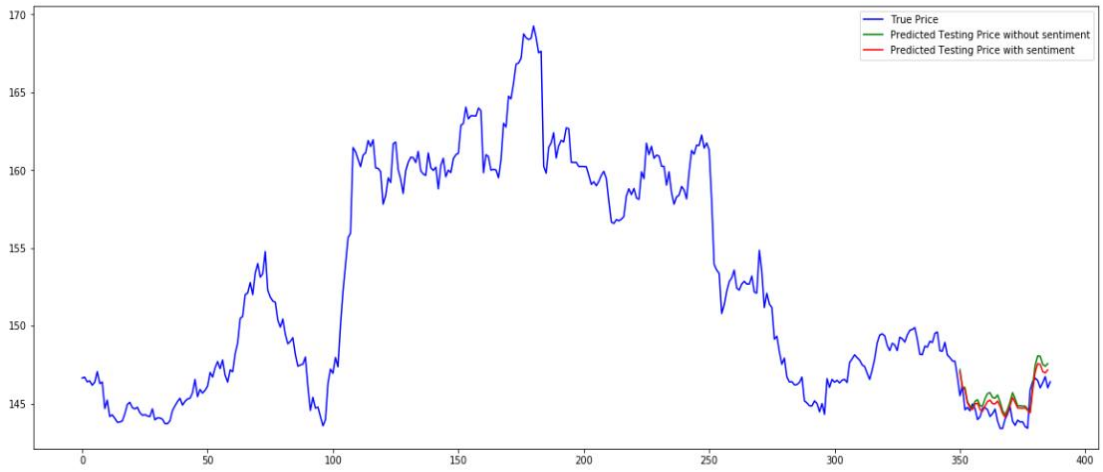
Στην συνέχεια ακολουθήθηκε η ίδια διαδικασία, όμως αυτή την φορά δοκιμάστηκε το μοντέλο (2). Παρακάτω φαίνονται τα συγκριτικά αποτελέσματα ανάλογα με το αν δόθηκαν οι sentiment τιμές ή όχι.



Εικόνα 50: Πρόβλεψη μετοχής AAPL του νευρωνικού (2), χρησιμοποιώντας α)close και sentiment score β) close

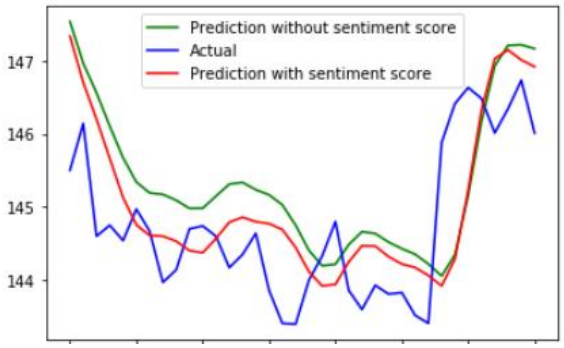
	RMSE
Με sentiment score	0.035162
Χωρίς sentiment score	0.043574

Και πάλι το σύστημα εξήγαγε, έστω και λίγο, καλύτερα αποτελέσματα όταν δόθηκαν και τα sentiment score, το οποίο φαίνεται αν προσέξει κανείς την διπλανή εικόνα.



Εικόνα 51: Τιμή κλεισίματος της μετοχής AAPL του νευρωνικού (2) σε όλο το διάστημα που αξιοποιήθηκε για την πρόβλεψη α) με sentiment scores β) χωρίς sentiment scores

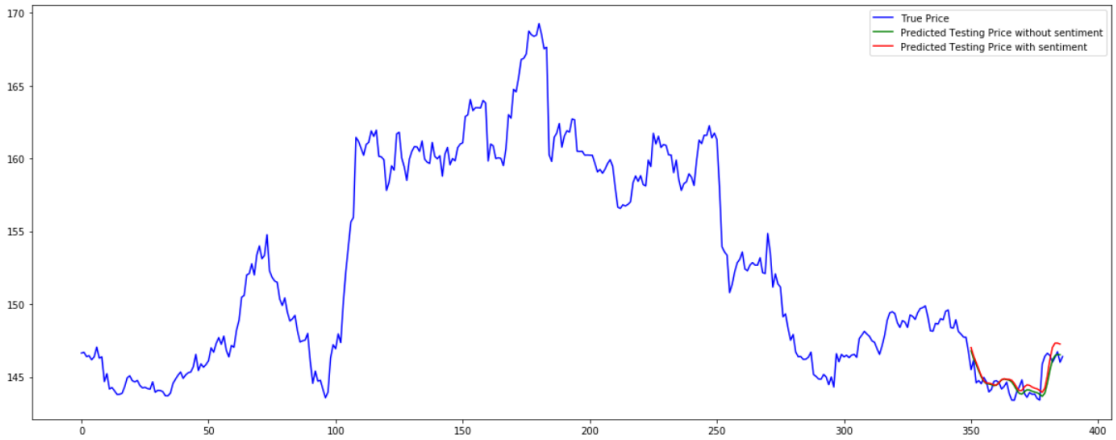
Στην συνέχεια, κρίθηκε σκόπιμο να παραχθούν αποτελέσματα όμως αυτή την φορά δόθηκαν σαν είσοδο στο νευρωνικό (1) όλες οι διαθέσιμες χρηματιστηριακές τιμές δηλαδή close, open, high, low, volume, sentiment score ώστε να συγκριθούν αυτά τα αποτελέσματα με τα προηγούμενα. Παρακάτω φαίνεται η πρόβλεψη όταν δοκιμάστηκε με sentiment score (κόκκινο) και αλλά και χωρίς sentiment score (πράσινο)



Εικόνα 52: Πρόβλεψη μετοχής AAPL του νευρωνικού (1), χρησιμοποιώντας α)όλες τις τιμές και sentiment score β) όλες τις τιμές

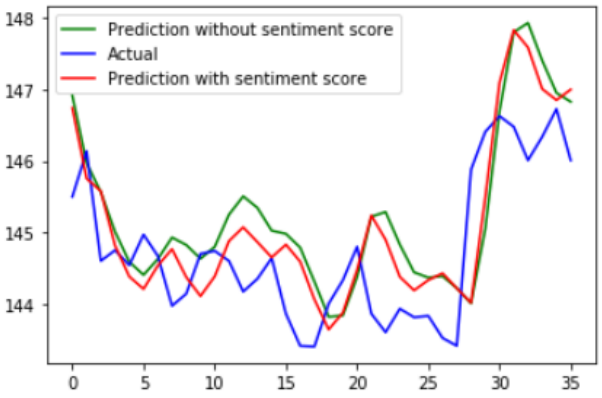
	RMSE
Με sentiment score	0.034513
Χωρίς sentiment score	0.041416

Παρατηρούμε ότι η κόκκινη γραμμή η οποία είναι αυτή που αξιοποίησε και το sentiment score έδωσε προβλέψεις πιο κοντά στις πραγματικές.



Εικόνα 53: Τιμή κλεισίματος της μετοχή AAPL του νευρωνικού (1) σε όλο το διάστημα που αξιοποιήθηκε για την πρόβλεψη α) όλες τις τιμές με sentiment scores β) όλες τις τιμές χωρίς sentiment scores

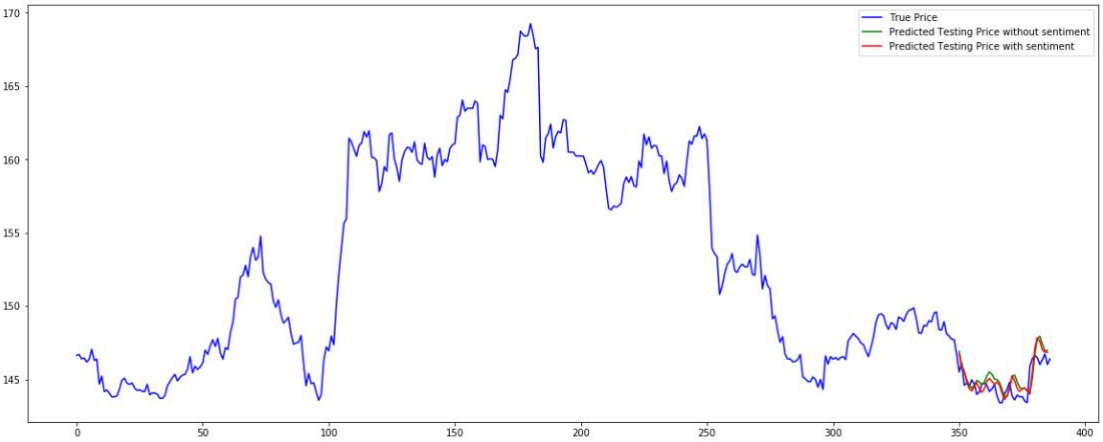
Τροφοδοτώντας το μοντέλο (2) με όλες τις διαθέσιμες τιμές προκύπτουν τα παρακάτω αποτελέσματα:



Εικόνα 54: Πρόβλεψη μετοχής AAPL του νευρωνικού (2), χρησιμοποιώντας α)όλες τις τιμές και sentiment score β) όλες τις τιμές

	RMSE
Με sentiment score	0.032012
Χωρίς sentiment score	0.036795

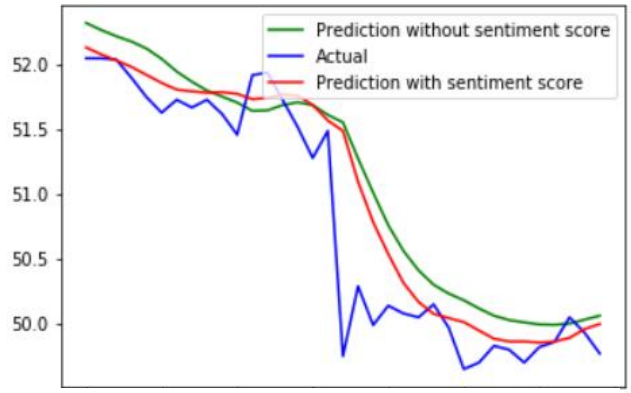
Και σε αυτή την περίπτωση φαίνεται ότι όταν δόθηκε στην είσοδο και τα sentiment score οι προβλέψεις ήταν πιο επιτυχείς.



Εικόνα 55: Τιμή κλεισίματος της μετοχή AAPL του νευρωνικού (2) σε όλο το διάστημα που αξιοποιήθηκε για την πρόβλεψη α)όλες τις τιμές με sentiment scores β)όλες τις τιμές χωρίς sentiment scores

5.2 Αποτελέσματα για την εταιρία Microsoft

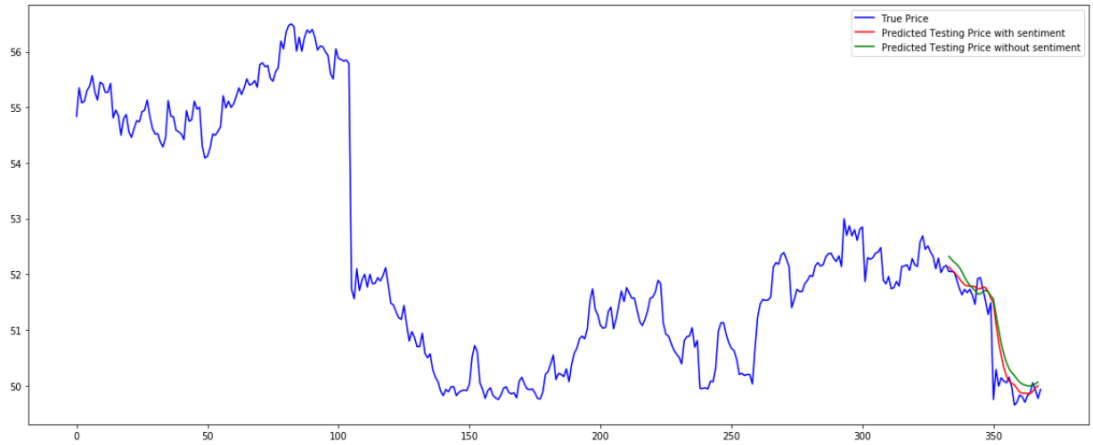
Ομοίως με πριν παράχθηκαν αποτελέσματα και για την γνωστή εταιρία Microsoft (MSFT). Αρχικά χρησιμοποιώντας το μοντέλο (1) και δίνοντας ως είσοδο την τιμή close και sentiment score παράχθηκαν τα παρακάτω αποτελέσματα



Εικόνα 56: Πρόβλεψη μετοχής MSFT του νευρωνικού (1) χρησιμοποιώντας α) close και sentiment β) close

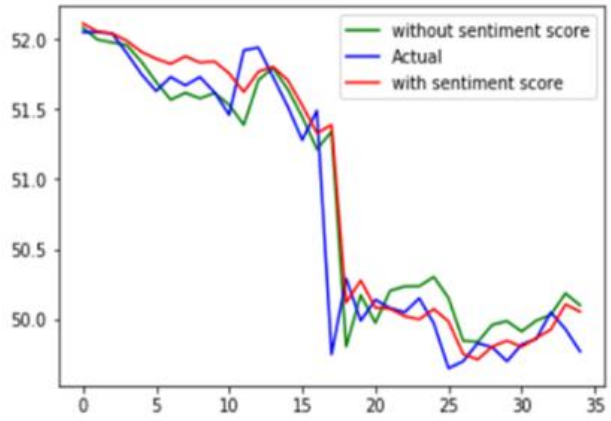
	RMSE
Με sentiment score	0.04157
Χωρίς sentiment score	0.04955

Όπως θα περίμενε κανείς όταν το νευρωνικό (1) αξιοποιεί και τα sentiment score δίνει ακριβέστερα αποτελέσματα σε σχέση με όταν δεν δίνονται. Αυτό φαίνεται και στην διπλανή εικόνα στην οποία η κόκκινη γραμμή βρίσκεται πιο κοντά στις πραγματικές τιμές



Εικόνα 57: Τιμή κλεισίματος της μετοχής MSFT του νευρωνικού (1) σε όλο το διάστημα που αξιοποιήθηκε για την πρόβλεψη α) με sentiment scores β) χωρίς sentiment scores

Δοκιμάζοντας τώρα το νευρωνικό σύστημα (2) πήραμε τα ακόλουθα αποτελέσματα

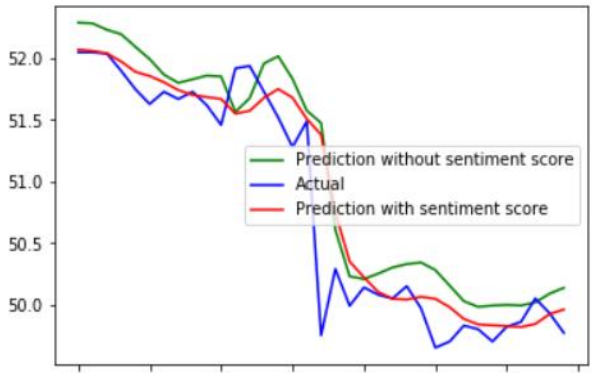


Εικόνα 58: Πρόβλεψη μετοχής MSFT του νευρωνικού (2), χρησιμοποιώντας α)close και sentiment score β) close

	RMSE
με sentiment score	0.03923
Χωρίς sentiment score	0.04320

Για άλλη μια φορά η περίπτωση που δόθηκαν οι πληροφορίες των κειμένων το μοντέλο εξήγαγε καλύτερα αποτελέσματα. Αξίζει επίσης να τονίσουμε ότι το νευρωνικό (2) έδωσε πιο ακριβή προβλέψεις σε σχέση με το (1) ενώ στην περίπτωση της google τα αποτελέσματα των δύο μοντέλων ήταν παρόμοια.

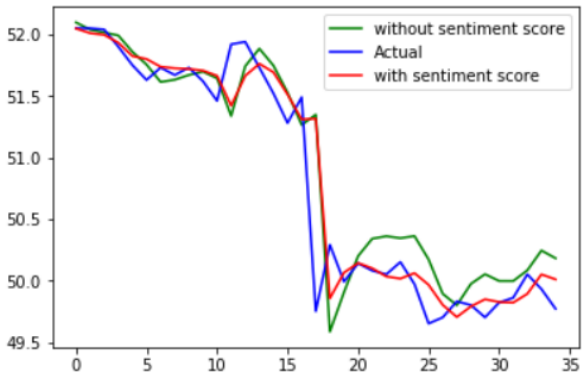
Όταν δόθηκαν ως είσοδοι όλες οι τιμές (close, open, high, low, sentiment score) στο νευρωνικό (1) εξάχθηκαν τα παρακάτω αποτελέσματα



Εικόνα 59: Πρόβλεψη μετοχής MSFT του νευρωνικού (1), χρησιμοποιώντας α) όλες τις τιμές και sentiment score β) όλες τις τιμές

	RMSE
με sentiment score	0.03739
Χωρίς sentiment score	0.04791

Και όταν έγινε χρήση του νευρωνικού (2) τα αποτελέσματα ήταν τα εξής:



	RMSE
Με sentiment score	0.03673
Χωρίς sentiment score	0.04948

Παρατηρούμε ότι το νευρωνικό (2) έδωσε ακριβέστερη πρόβλεψη σε σχέση με το νευρωνικό (1)

Εικόνα 60: Πρόβλεψη μετοχής MSFT του νευρωνικού (2), χρησιμοποιώντας α)όλες τις τιμές και sentiment score β) όλες τις τιμές

5.3 Αποτελέσματα για άλλες γνωστές μετοχές

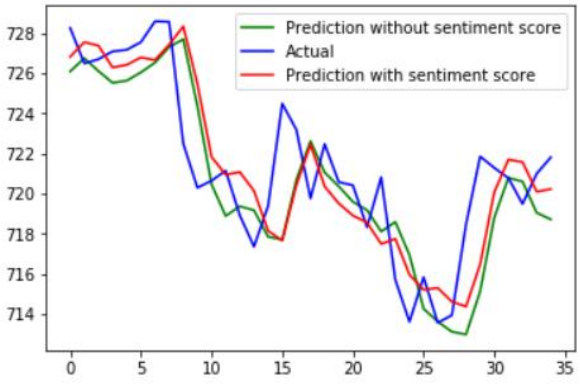
Παρομοίως δόθηκαν ως είσοδοι οι τιμές και για κάποιες επιπλέον γνωστές εταιρίες, οι οποίες υπήρχαν στο dataset NASDAQ ώστε να αντληθούν και πληροφορίες από τα tweets-σχόλια των χρηστών που αφορούσαν τις εταιρίες αυτές. Τα αποτελέσματα συνοπτικά δίνονται στον επόμενο πίνακα:

Μοντέλο/RMSE	AAPL	GOOG	MSFT	FB	NFLX
LSTM(1)	0.02452 0.05017	0.02728 0.02865	0.04157 0.04955	**0.01894** **0.02025**	0.07161 0.06921
LSTM(2)	0.03516 0.04357	0.02721 0.02881	0.03923 0.04320	0.02248 0.02341	0.04998 0.06588

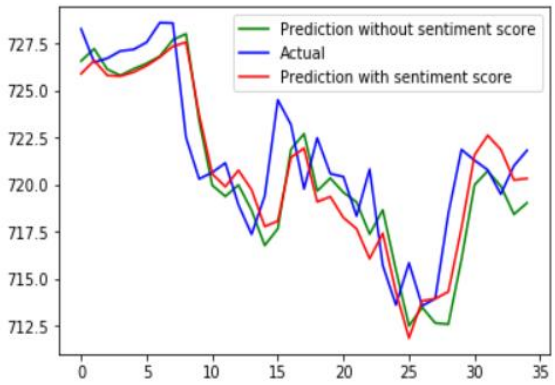
Εικόνα 61: Πίνακας 1: RMSE μετοχών των δύο μοντέλων α)close και sentiment score β)close

Και παρακάτω δίνονται οι γραφικές παραστάσεις των δύο νευρωνικών μοντέλων

Τιμές μέτοχών της εταιρίας Google

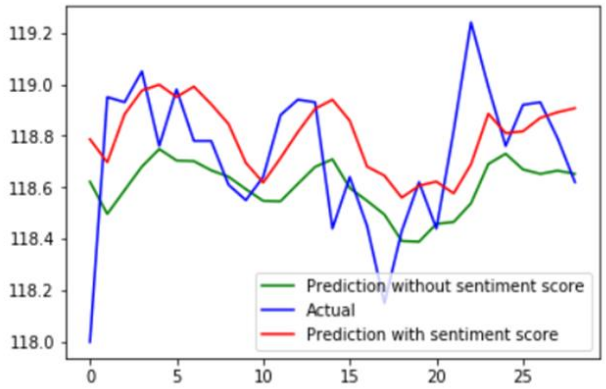


Εικόνα 63: Πρόβλεψη μετοχής GOOG του νευρωνικού (1), χρησιμοποιώντας α) close και sentiment score β) close

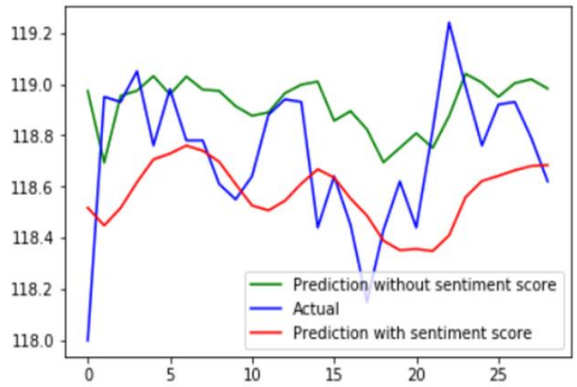


Εικόνα 62: Πρόβλεψη μετοχής GOOG του νευρωνικού (2), χρησιμοποιώντας α) close και sentiment score β) close

Τιμές μέτοχών της εταιρίας Facebook

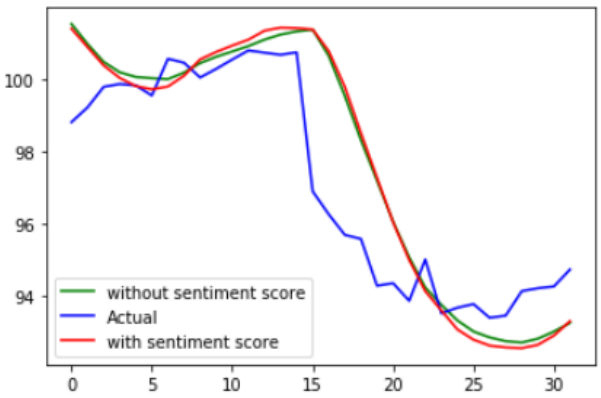


Εικόνα 64: Πρόβλεψη μετοχής FB του νευρωνικού (1), χρησιμοποιώντας α) close και sentiment score β) close

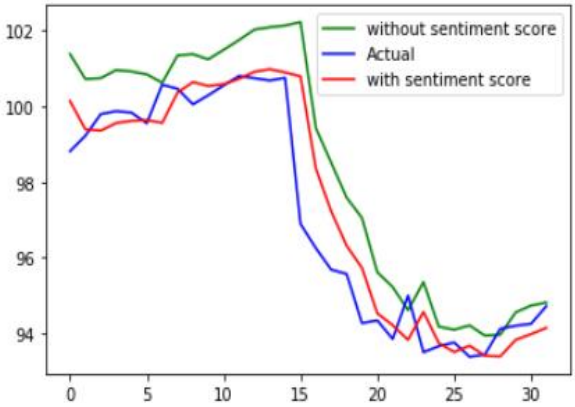


Εικόνα 65 Πρόβλεψη μετοχής FB του νευρωνικού (2), χρησιμοποιώντας α) close και sentiment score β) close

Τιμές μέτοχών της εταιρίας Netflix



Εικόνα 66: Πρόβλεψη μετοχής NFLX του νευρωνικού (1), χρησιμοποιώντας α) close και sentiment score β) close



Εικόνα 67: Πρόβλεψη μετοχής NFLX του νευρωνικού (2), χρησιμοποιώντας α) close και sentiment score β) close

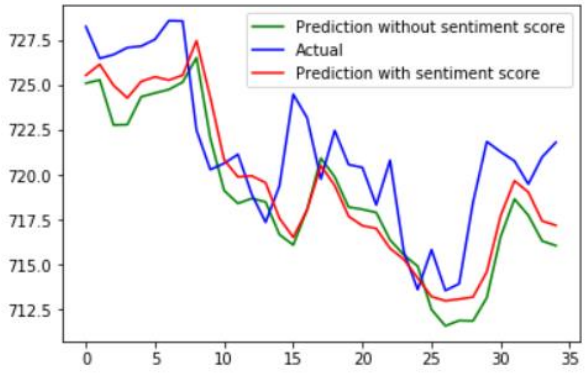
Στον επόμενο πίνακα παρουσιάζονται τα αποτελέσματα των δύο νευρωνικών όταν δόθηκαν ως είσοδοι όλες οι χρηματιστηριακές τιμές (close, open, high, low, volume) με sentiment score ή χωρίς. Με κόκκινο φαίνονται οι τιμές των προβλέψεων όταν δόθηκαν και τα sentiment scores ενώ με πράσινο φαίνονται τα αποτελέσματα όταν αξιοποιήθηκαν μόνο οι χρηματιστηριακές τιμές του παρελθόντος.

Μοντέλο/RMSE	AAPL	GOOG	MSFT	FB	NFLX
LSTM(1)	0.03451 0.04141	0.03736 0.04521	0.03739 0.04791	**0.01850** **0.01984**	0.04999 0.04688
LSTM(2)	0.03201 0.03679	0.03256 0.05453	0.03673 0.04948	0.02660 0.03024	0.03364 0.03593

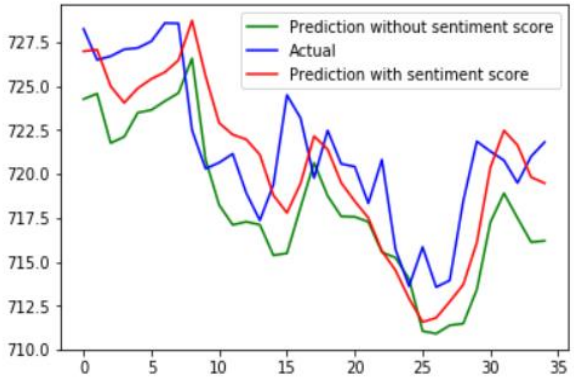
Εικόνα 68: Πίνακας 2: RMSE μετοχών των δύο μοντέλων α)όλες οι τιμές και sentiment score β)όλες οι τιμές

Παρακάτω δίνονται οι γραφικές τους παραστάσεις

Τιμές μέτοχών της εταιρίας Google

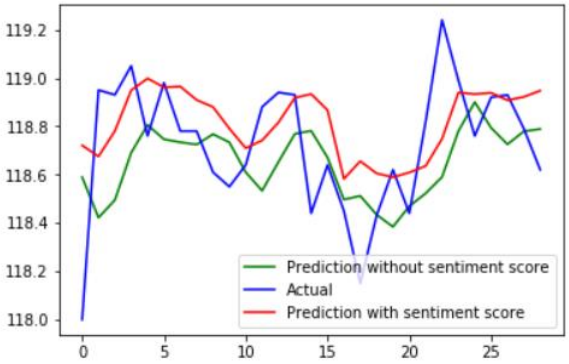


Εικόνα 69: Πρόβλεψη μετοχής GOOG του νευρωνικού (1), χρησιμοποιώντας α) όλες τις τιμές και sentiment score β) όλες τις τιμές

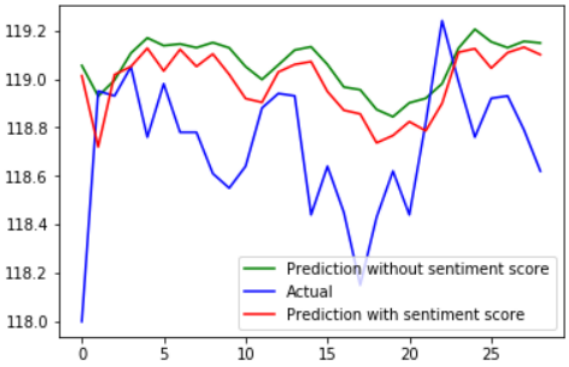


Εικόνα 70: Πρόβλεψη μετοχής GOOG του νευρωνικού (2), χρησιμοποιώντας α) όλες τις τιμές και sentiment score β) όλες τις τιμές

Τιμές μέτοχών της εταιρίας Facebook

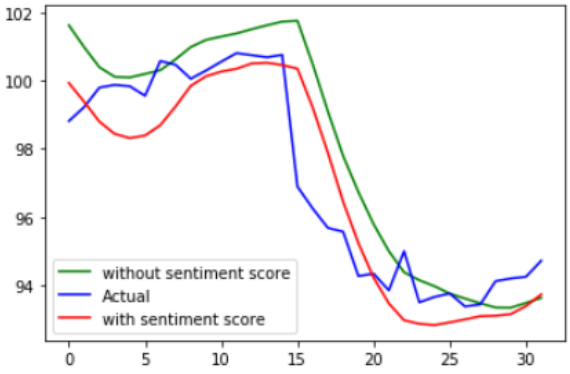


Εικόνα 71: Πρόβλεψη μετοχής FB του νευρωνικού (1), χρησιμοποιώντας α) όλες τις τιμές και sentiment score β) όλες τις τιμές

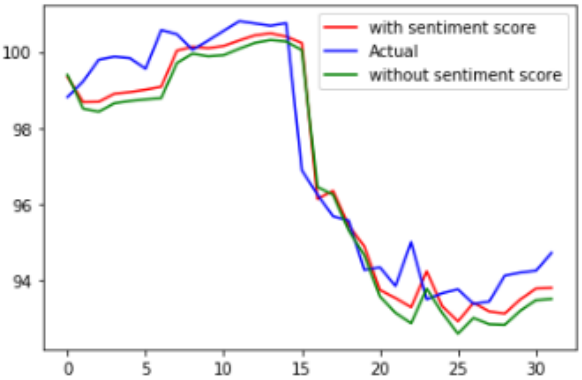


Εικόνα 72: Πρόβλεψη μετοχής MSFT του νευρωνικού (2), χρησιμοποιώντας α) όλες τις τιμές και sentiment score β) όλες τις τιμές

Τιμές μέτοχών της εταιρίας Netflix



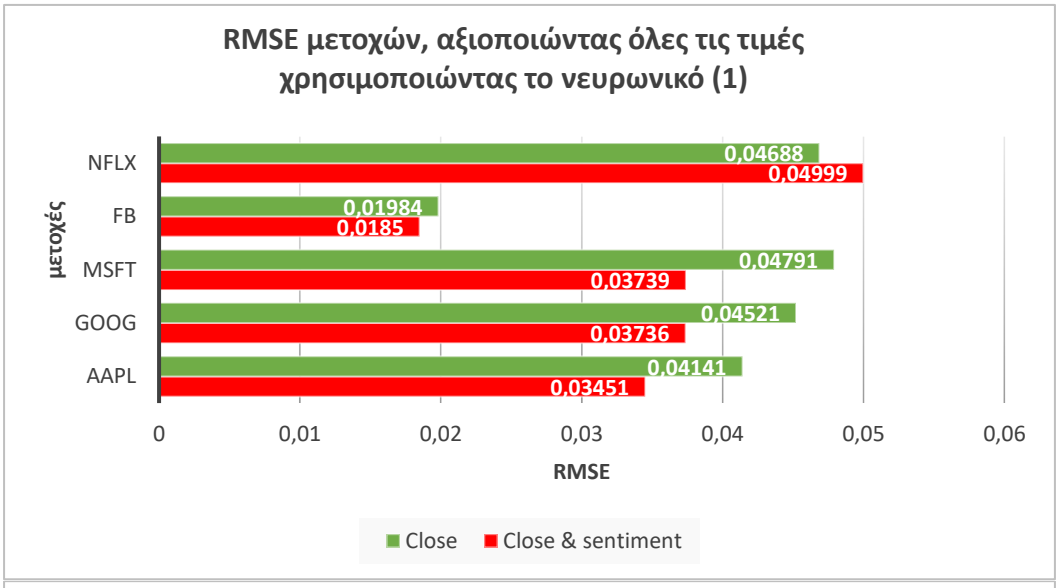
Εικόνα 74: Πρόβλεψη μετοχής NFLX του νευρωνικού (1), χρησιμοποιώντας α) όλες τις τιμές και sentiment score β) όλες τις τιμές

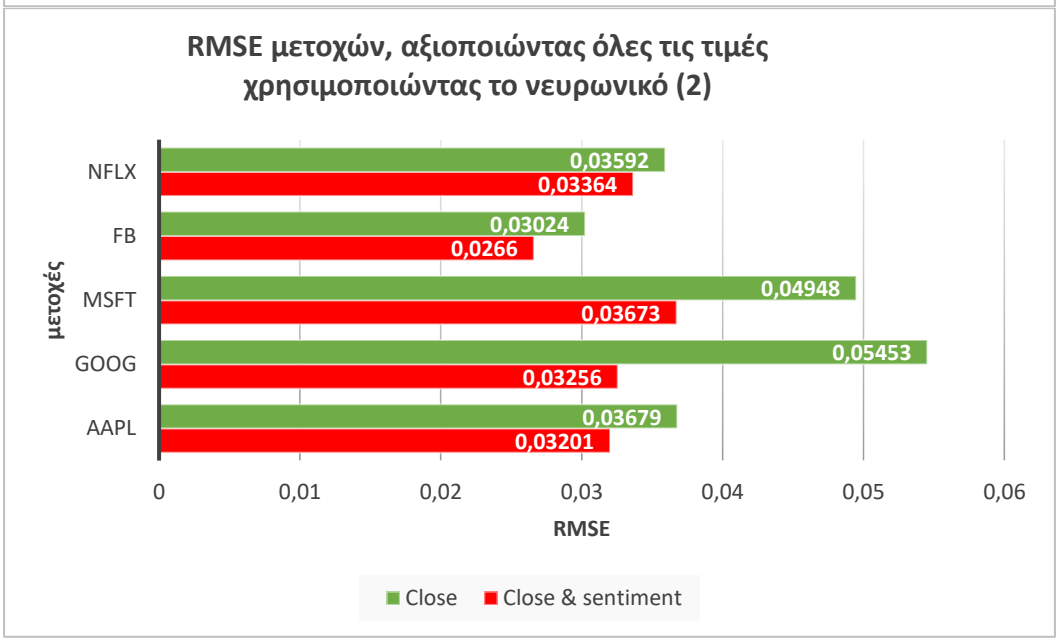


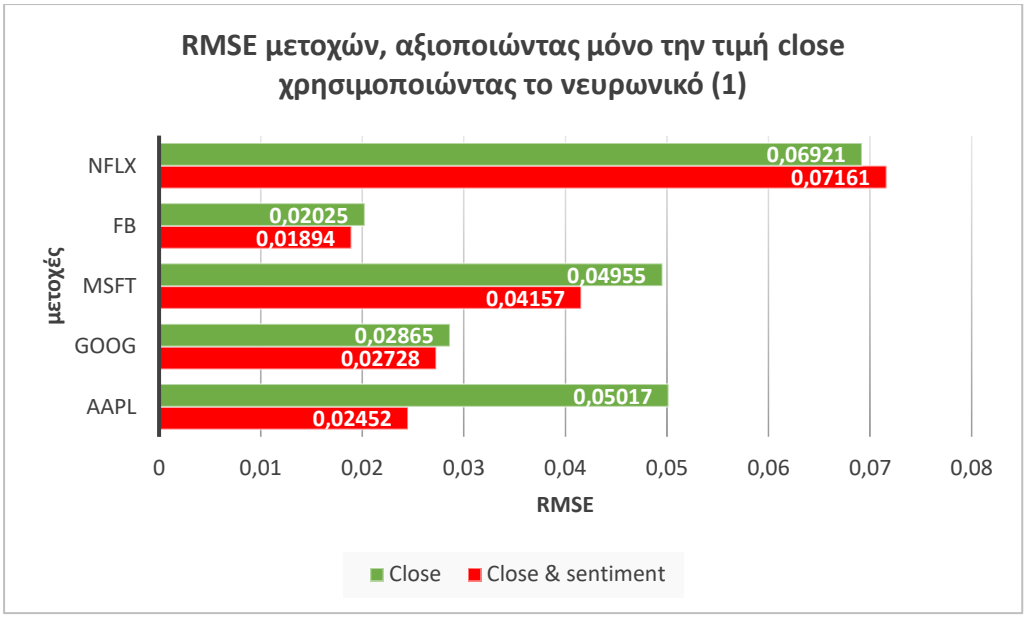
Εικόνα 73: Πρόβλεψη μετοχής NFLX του νευρωνικού (2), χρησιμοποιώντας α) όλες τις τιμές και sentiment score β) όλες τις τιμές

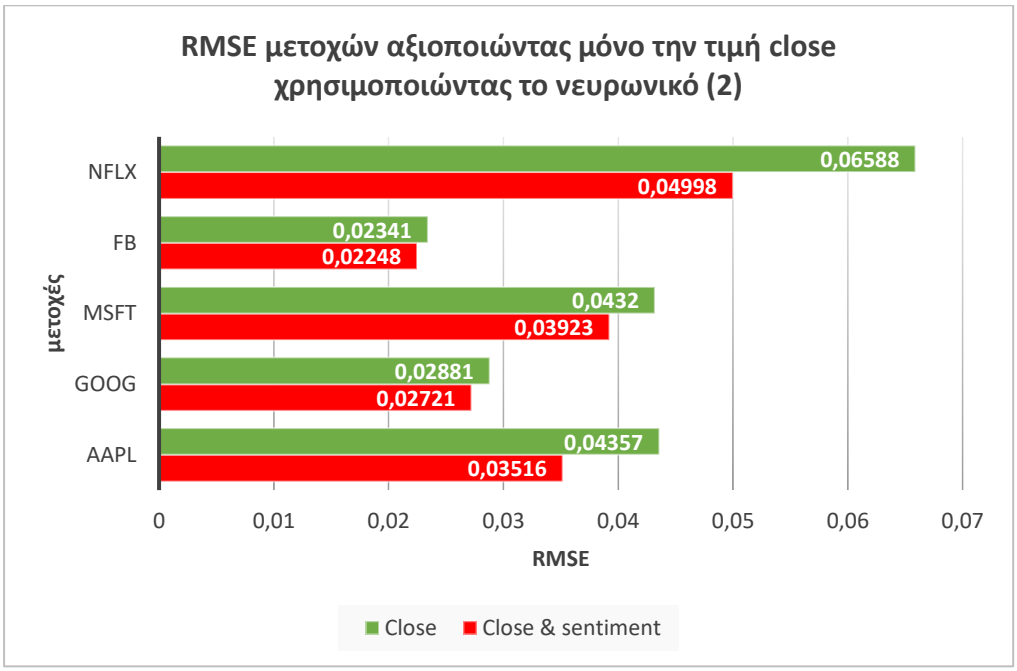
5.4 Συμπεράσματα Μετρήσεων

Στα παρακάτω γραφήματα φαίνονται συγκεντρωτικά τα αποτελέσματα ώστε να διευκολυνθεί η σύγκριση μεταξύ των προβλέψεων που χρησιμοποίησαν μόνο χρηματιστηριακές τιμές και αυτών που αξιοποίησαν και την ανάλυση συναισθήματος.









Από τα αποτελέσματα που φαίνονται στους πίνακες 1 και 2 και από τα παραπάνω γραφήματα παρατηρούμε ότι στις περισσότερες περιπτώσεις που δόθηκαν και τα sentiment scores τα νευρωνικά ανταποκρίθηκαν καλύτερα δίνοντας πιο ακριβή αποτελέσματα σε σύγκριση με τα αποτελέσματα που έδωσαν όταν δεν δόθηκαν τα sentiment scores. Να τονισθεί για να μην προκληθεί σύγχυση ότι στα παραπάνω γραφήματα όσο πιο κοντές είναι οι γραμμές τόσο πιο μικρό είναι το RMSE και άρα τόσο πιο ακριβής είναι η πρόβλεψη. Ως συνέπεια λοιπόν μπορούμε να συμπεράνουμε ότι η ανάλυση συναισθήματος των σχολίων των ανθρώπων για αυτές τις εταιρίες και αυτή την χρονική περίοδο επηρέασε τις μελλοντικές τιμές των μετοχών και εξού και τα μοντέλα πρόβλεψαν καλύτερα την μελλοντική τους συμπεριφορά. Μάλιστα σε μερικές περιπτώσεις βλέπουμε σημαντική βελτίωση των προβλέψεων. Για παράδειγμα, η πρόβλεψη της τιμής close της μετοχής apple όταν δόθηκαν τα sentiment scores βελτιώθηκε κατά 34% ! Άρα μπορούμε να ισχυριστούμε ότι ισχύει η βασική προϋπόθεση και ότι επιτεύχθηκε ο στόχος της εργασίας ο οποίος ήταν να δούμε αν παράγονται πιο ακριβείς προβλέψεις δίνοντας και τα sentiment scores. Υπήρχαν βέβαια και περιπτώσεις στις οποίες δεν υπήρξε τόσο μεγάλη βελτίωση.

Παρατηρώντας τα αποτελέσματα που έδωσε το νευρωνικό μοντέλο (1) για την μετοχή του Netflix βλέπουμε ότι όταν δεν δόθηκαν τα sentiment scores παράχθηκαν, έστω και ελάχιστα, πιο ακριβείς προβλέψεις. Αυτό μπορεί να συμβαίνει διότι το πλήθος των σχολίων που χρησιμοποιήθηκαν, τα οποία ήταν στο dataset NASDAQ, ήταν συγκριτικά μικρότερο σε σχέση με το πλήθος των tweets άλλων μετοχών. Δηλαδή υπήρχαν ώρες στις οποίες δεν αναρτήθηκε κάποιο σχόλιο από κάποιον χρήστη και ως ετούτου το sentiment score για αυτή την ώρα ήταν 0, δηλαδή δεν δόθηκε καμία πληροφορία από τα σχόλια. Επιπρόσθετα όπως ειπώθηκε και κατά την περιγραφή του dataset υπήρχαν αρκετά σχόλια-tweets τα οποία δεν πρόσφεραν κάποια πληροφορία όπως για παράδειγμα αυτά που είχαν μόνο URL ή αποτελούνταν μόνο από εικόνα. Αυτά αποτελούν «θόρυβο» για το νευρωνικό και επηρεάζουν αρνητικά τα αποτελέσματα. Παρακάτω φαίνεται το πλήθος των tweets, ο μέσος αριθμός tweets ανά ώρα, και ο μέγιστος αριθμός για την κάθε μετοχή κατά το χρονικό διάστημα που μελετήθηκε.

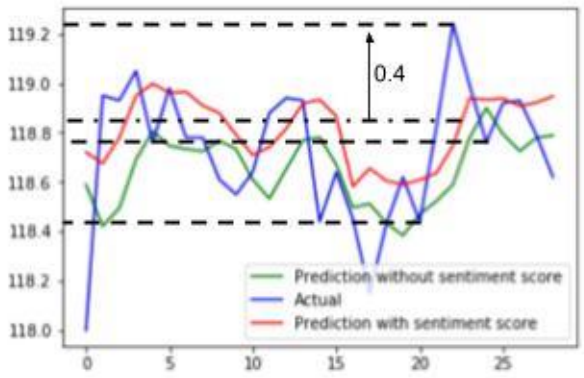
Μετοχή	Πλήθος tweets	Μέσος όρος ανά ώρα	Μέγιστος αριθμός ώρας
AAPL	85204	149,21	1606
FB	23280	50,28	541
GOOG	10785	20,54	134
MSFT	14341	27,06	1556
NFLX	37082	31,64	1459

Εικόνα 75:Στατιστικά των tweets που χρησιμοποιήθηκαν

Όσον αφορά την σύγκριση των δύο νευρωνικών μοντέλων 1 και 2 παρατηρούμε ότι σε κάποιες μετοχές ακριβέστερα αποτελέσματα δόθηκαν από το ένα και σε άλλες από το άλλο. Αξίζει να σημειώσουμε ότι τα δύο μοντέλα όταν δόθηκαν όλες οι χρηματιστηριακές τιμές είχαν πολύ παρόμοια αποτελέσματα κυρίως όταν αξιοποιήθηκαν και τα sentiment scores.

Όπως μπορεί να αποφανθεί από τους δύο παραπάνω πίνακες, τις καλύτερες προβλέψεις τις δίνει η μετοχή Facebook και στα δύο νευρωνικά μοντέλα. Αυτό μπορεί να συμβαίνει γιατί αφενός οι μεταβολές της τιμής του close είναι πιο ομαλές χωρίς πολλά τοπικά μέγιστα και αφετέρου γιατί τα σχόλια-tweets των ανθρώπων παρείχαν περισσότερη αξιόπιστη πληροφορία κάθε ώρα σε σχέση με τα tweets των άλλων μετοχών. Επίσης για αυτό τον λόγο δεν είναι αποτελεσματική η σύγκριση μεταξύ των αποτελεσμάτων διαφορετικών μετοχών γιατί όπως είπαμε κάθε μετοχή περιέχει διαφορετικό πλήθος tweets.

Επιπρόσθετα, σημαντικό είναι να επισημάνουμε ότι οι γραφικές παραστάσεις των προβλέψεων των νευρωνικών είναι σχετικά πιο ομαλές και δεν σχηματίζουν τόσες γωνίες όσες έχουν οι γραφικές παραστάσεις των πραγματικών τιμών. Δηλαδή η διακύμανση της πρόβλεψης είναι μικρότερη από την διακύμανση των πραγματικών τιμών. Αυτό είναι λογικό αφού τα μοντέλα που κατασκευάστηκαν είναι σε θέση να προβλέπουν την συμπεριφορά του άμεσου μέλλοντος και όχι τις ακριβείς χρηματιστηριακές τιμές μιας και αυτό δεν είναι δυνατόν να υλοποιηθεί. Γιαυτό τον λόγο όταν η πραγματική τιμή μεταβεί σε μια μεγάλη αλλαγή-διακύμανση από μια ώρα στην επόμενη τότε τα συστήματα των νευρωνικών που κατασκευάστηκαν δεν θα είναι σε θέση να προβλέψουν αυτή την σημαντική αλλαγή, αλλά θα το προσδιορίσουν πιθανών με μια πιο ομαλή διακύμανση. Για παράδειγμα στην διπλανή εικόνα, η πραγματική μέγιστη τιμή του close είναι 119,21 ενώ η αντίστοιχη τιμή που προβλέπει το νευρωνικό δίκτυο είναι 118,81. Όπως μπορεί αν παρατηρήσει κάποιος η τιμή οι γειτονικές τιμές της 119,21 είναι 118,42 και 118,64 άρα φτάνει σε μέγιστο για μια μόνο χρονική στιγμή και μετά επανέρχεται σε



Εικόνα 76: Διακύμανση τιμών πρόβλεψης και πραγματικών τιμών.

πιο χαμηλές τιμές σχηματίζοντας έτσι μια απότομη μεταβολή στην καμπύλη. Αυτό είναι πρακτικά αδύνατο να προβλεφθεί από το σύστημα πρόβλεψης διότι οι τιμές που προβλέπει συνήθως έχουν πιο ομαλή διακύμανση, πράγμα το οποίο φαίνεται και στην εικόνα.

5.5 Πρόβλεψη αξιοποιώντας μόνο ιστορικές τιμές

Τα νευρωνικά δίκτυα είναι στατιστικές μέθοδοι και γιαυτό απαιτούν μεγάλο όγκο δεδομένων, διαφορετικά υπάρχει ο κίνδυνος να μην λειτουργήσουν σωστά. Γενικά μπορούμε να θεωρήσουμε ότι όταν σε ένα νευρωνικό δίκτυο δοθεί μεγαλύτερο πλήθος δεδομένων training τότε αυτό πιθανόν να παράξει καλύτερα αποτελέσματα και πιο ακριβείς προβλέψεις. Μέχρι τώρα για την πρόβλεψη της χρηματιστηριακής τιμής close αξιοποιήθηκαν 450 περίπου τιμές ανά ώρα(intraday τιμές). αν υπήρχαν περισσότερα σχόλια και μεγαλύτερης περιόδου των ανθρώπων που αφορούν τις εξετάζουσες μετοχές πολύ πιθανό να καταφέρναμε να εξάγουμε πιο ακριβή αποτελέσματα μιας και το training data θα περιείχε περισσότερα δεδομένα. Στη συνέχεια θα δοκιμαστεί να τροφοδοτηθεί το νευρωνικό σύστημα με περισσότερες τιμές δεδομένων για training, χωρίς τα sentiment score τα οποία δεν υπάρχουν διαθέσιμα, και θα συγκριθούν τα αποτελέσματα.

Σε αυτό το σημείο αξίζει να δούμε τα αποτελέσματα που δίνει το νευρωνικό σύστημα (1) αξιοποιώντας μόνο ιστορικές χρηματιστηριακές τιμές αλλά αυτή την φορά για μεγαλύτερες περιόδους. Επίσης δεν θα έχουμε Intraday δεδομένα αλλά κάθε μέρα θα έχει από μια τιμή close, open, high, low, volume όπως φαίνεται στην παρακάτω εικόνα.

date	Open	High	Low	Close	Adj Close	Volume
2013-01-02	358.366760	362.142609	356.937103	360.274597	360.274597	5101500
2013-01-03	361.111481	364.598389	359.014313	360.483826	360.483826	4653700
2013-01-04	363.308228	369.350586	362.481323	367.607117	367.607117	5547600
2013-01-07	366.351837	368.309479	363.925903	366.003143	366.003143	3323800
2013-01-08	366.396667	366.775238	360.862396	365.280823	365.280823	3364700
2013-01-09	364.767761	367.796417	362.939606	367.681824	367.681824	4064500
2013-01-10	370.028046	371.108978	365.380463	369.355560	369.355560	3685000
2013-01-11	369.614594	369.828796	366.775238	368.613342	368.613342	2579900
2013-01-14	367.123932	369.714233	359.826294	360.274597	360.274597	5749200
2013-01-15	358.321930	366.127655	354.720428	361.111481	361.111481	7884700
2013-01-16	359.851196	360.817566	355.502502	356.259644	356.259644	4061900
2013-01-17	357.514954	358.476349	354.182434	354.331879	354.331879	4439500
2013-01-18	353.853668	355.054169	349.355530	350.939606	350.939606	6477700
2013-01-22	351.014313	351.353027	346.461365	350.122650	350.122650	7613100
2013-01-23	366.620819	373.101532	366.521179	369.365540	369.365540	11862400
2013-01-24	369.236023	377.001892	368.872375	375.696808	375.696808	6790600
2013-01-25	373.983215	377.823822	373.724182	375.427795	375.427795	4468400
2013-01-28	374.476379	376.389191	372.548584	373.963287	373.963287	3266300
2013-01-29	371.980713	377.061676	371.876129	375.432800	375.432800	3507200
2013-01-30	375.462677	379.054199	375.049225	375.507507	375.507507	3478900
2013-01-31	373.853699	377.395416	373.724182	376.434021	376.434021	3280500
2013-02-01	377.684357	386.850006	377.634521	386.351868	386.351868	7520100

Εικόνα 77: Μορφή δεδομένων που θα αξιοποιήσει το νευρωνικό (1)

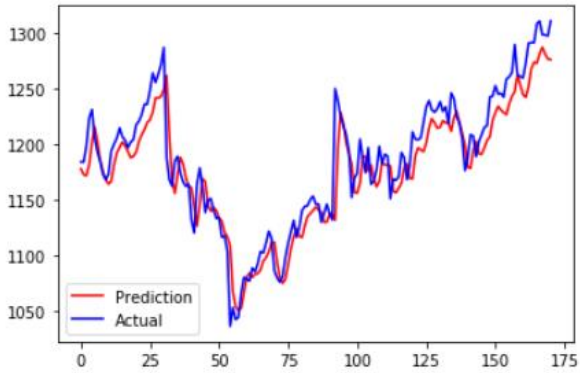
Δόθηκαν οι χρηματιστηριακές τιμές του close των μετοχών (google, apple, msft, fb) για την περίοδο 1/1/2013 έως 15/11/2019 (1700 τιμές). Παρακάτω φαίνονται τα αποτελέσματα που εξήγαγε το νευρωνικό μοντέλο (1) δίνοντας ως υπερπαραμέτρους

- Neurons = 64
- Batch size = 20
- Epoch = 150

Μετοχή/RMSE	Intraday (1/4/2016-15/6/2016)	1/1/2013-15/11/2019
Google	0.02865	0.020686
Apple	0.05017	0.016748
MSFT	0.04157	0.017869
FB	0.01894	0.01697

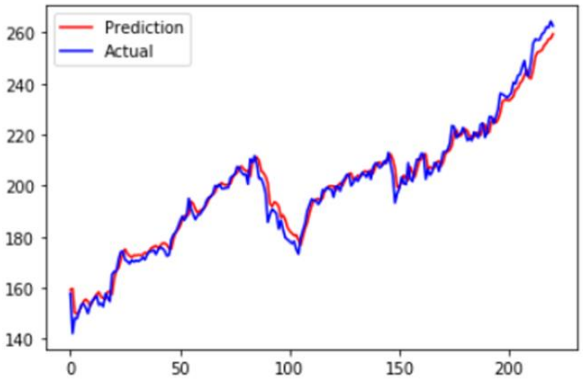
Εικόνα 78: Σύγκριση προβλέψεων μεταξύ δύο περιόδων

Μετοχές Google



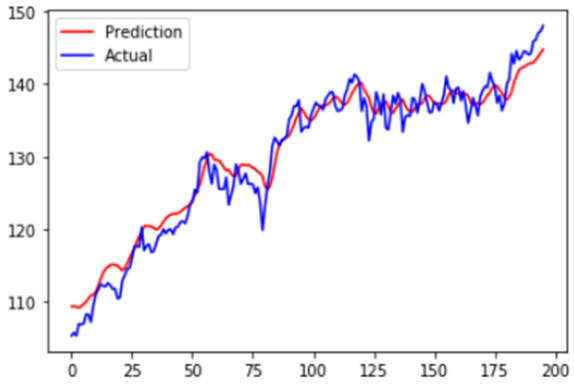
Εικόνα 80: Αποτέλεσμα πρόβλεψης της μετοχής GOOG για διάρκεια 6 χρόνων

Μετοχές Apple



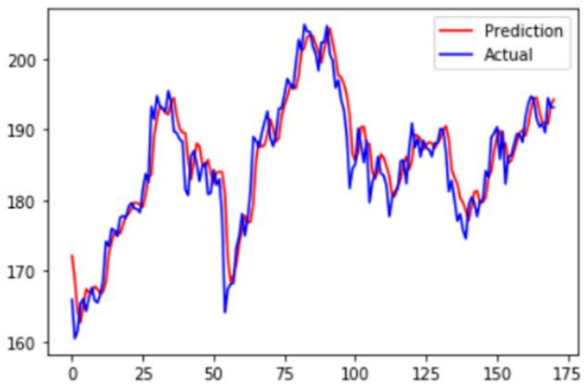
Εικόνα 79: Αποτέλεσμα πρόβλεψης της μετοχής AAPL για διάρκεια 6 χρόνων

Μετοχές Microsoft



Εικόνα 81: Αποτέλεσμα πρόβλεψης της μετοχής MSFT για διάρκεια 6 χρόνων

Μετοχές Facebook



Εικόνα 82: Αποτέλεσμα πρόβλεψης της μετοχής FB για διάρκεια 6 χρόνων

Από τα αποτελέσματα γίνεται φανερό ότι όταν δόθηκαν περισσότερες ιστορικές τιμές, το νευρωνικό μοντέλο ανταποκρίθηκε καλύτερα δίνοντας πιο ακριβή πρόβλεψη. Αυτό ήταν αναμενόμενο αφού γενικά όταν δίνονται περισσότερα δεδομένα σε ένα νευρωνικό δίκτυο είναι πιθανό να παράγει καλύτερα αποτελέσματα. Αυτό φαίνεται και από τις γραφικές τους παραστάσεις όπου φαίνεται ότι η κόκκινη γραμμή η οποία συμβολίζει την πρόβλεψη «ακολουθεί» καλύτερα την πραγματική (μπλε) σε σχέση με τις γραφικές παραστάσεις των intraday στις οποίες τα δεδομένα ήταν πιο λίγα.

Κεφάλαιο 6

6. Συμπεράσματα

Η παρούσα εργασία ασχολήθηκε με την εφαρμογή μηχανικής μάθησης στον χώρο των χρηματοοικονομικών και πιο συγκεκριμένα του χρηματιστηρίου. Πιο συγκεκριμένα, αντιμετωπίζουμε το θέμα της πρόβλεψης των χρηματιστηριακών τιμών στο άμεσο μέλλον. Αυτό το πρόβλημα προσεγγίσθηκε από διάφορες οπτικές γωνίες. Αρχικά δοκιμάστηκε η πρόβλεψη της συμπεριφοράς των τιμών αξιοποιώντας μόνο τις ιστορικές τιμές ενώ στην συνέχεια επιδιώξαμε να παράξουμε καλύτερα αποτελέσματα χρησιμοποιώντας και πληροφορίες από σχόλια των χρηστών για τις μετοχές που εξετάζουμε. Αναλυτικότερα αυτά τα σχόλια-tweets υπέστησαν ανάλυση συναισθήματος με σκοπό να αποφανθεί πόσο θετικά ή αρνητικά είναι και έπειτα τροφοδοτήθηκαν στο νευρωνικό σύστημα το οποίο αξιοποίησε και αυτή την επιπλέον πληροφορία προκειμένου να προβλέψει την μελλοντική συμπεριφορά.

Δοκιμάστηκαν αρκετά μοντέλα νευρωνικών δικτύων τα οποία χρησιμοποιήθηκαν για την επίτευξη του στόχου τα οποία χωρίστηκαν σε δύο κατηγορίες αναλόγως την λειτουργία που επιτελούν. Αρχικά κατασκευάστηκαν μοντέλα τα οποία είναι υπεύθυνα να πραγματοποιήσουν μια αξιόλογη ανάλυση συναισθήματος το οποίο ήταν και το πρώτο βήμα της εργασίας αυτής. Στην συνέχεια σχηματίστηκαν νευρωνικά δίκτυα των οποίων η δουλεία ήταν να επεξεργάζονται τα δεδομένα εισόδου με σκοπό την πρόβλεψη της χρηματιστηριακής τιμής του κλεισίματος. Από τα διάφορα νευρωνικά δίκτυα που κατασκευάστηκαν επιλέχθηκαν εκείνα τα οποία έδιναν τα καλύτερα αποτελέσματα.

Από την ανάλυση που πραγματοποιήθηκε στις προηγούμενες ενότητες γίνεται φανερό ότι ακριβέστερες προβλέψεις παράχθηκαν όταν το σύστημα αξιοποίησε τόσο τις ιστορικές χρηματιστηριακές τιμές όσο και τις πληροφορίες κειμένων δηλαδή τα sentiment scores. Οπότε έχουμε επιτύχει τον σκοπό αυτής της εργασίας ο οποίος ήταν η σύγκριση και το αν θα παράγονταν καλύτερα αποτελέσματα αν αξιοποιούνταν και οι πληροφορίες από τις απόψεις των ανθρώπων σχετικά με τις εξετάζουσες μετοχές. Τέλος αν υπήρχαν δεδομένα για μεγαλύτερη χρονική περίοδο τότε πολύ πιθανόν το σύστημα να έδινε καλύτερες προβλέψεις.

Κεφάλαιο 7

7. Μελλοντική εργασία

Από τα αποτελέσματα που παράχθηκαν σε αυτή την εργασία μπορούμε να συμπεράνουμε πως δίνοντας και τις κατάλληλες πληροφορίες από τα σχόλια των ανθρώπων για τις εταιρίες-μετοχές παράγονται σχετικά ακριβέστερα αποτελέσματα σε σύγκριση με τις παρελθοντικές τιμές μόνο. Υπήρξαν βέβαια κάποια προβλήματα και δυσκολίες που έπρεπε να αντιμετωπιστούν. Η σημαντικότερη ίσως ήταν το γεγονός ότι αξιοποιήθηκαν κειμενικά δεδομένα για διάστημα μόνο 2 περίπου μηνών. Αυτό έχει ως συνέπεια ότι το νευρωνικό σύστημα ίσως δεν είχε την απαραίτητη πληροφορία-δεδομένα για να προπονηθεί ώστε να προβεί σε μια ακόμα καλύτερη πρόβλεψη της συμπεριφοράς των χρηματιστηριακών τιμών. Αυτό γίνεται φανερό από το γεγονός ότι όταν δόθηκαν μόνο χρηματιστηριακές τιμές για διάρκεια 6 χρόνων τα αποτελέσματα των προβλέψεων ήταν καλύτερα. Και δεδομένου ότι συνήθως όταν δίνονται και τα sentiment scores το σύστημα παράγει καλύτερη πρόβλεψη καταλήγουμε στο συμπέρασμα ότι αν υπήρχαν κειμενικά δεδομένα για μεγαλύτερη χρονική περίοδο πιθανότατα θα είχαν παραχθεί ακόμα ακριβέστερα αποτελέσματα. Άρα μελλοντική έρευνα θα μπορούσε να αποτελέσει η εύρεση και αξιοποίηση περισσότερων σχολίων για μεγαλύτερη χρονική περίοδο από αυτή που χρησιμοποιήθηκε σε αυτή την εργασία.

Τα αποτελέσματα των προβλέψεων της τιμής κλεισίματος θα μπορούσαν να είχαν βελτιωθεί αν γινόταν χρήση και χρηματοοικονομικών νέων ή απόψεων ειδικών σε αυτό τον τομέα. Αντλώντας δηλαδή, πληροφορίες από κείμενα ειδήσεων ή ακόμα και από blogs θα μπορούσαμε να εξορύξουμε περισσότερη πληροφορία την οποία θα εκμεταλλευόταν το νευρωνικό δίκτυο που κατασκευάστηκε στην εργασία αυτή.

Στην συνέχεια, όσον αφορά το μοντέλο της ανάλυσης συναισθήματος θα μπορούσε να βελτιωθεί αξιοποιώντας κάποιες άλλες ιδέες που χρησιμοποίησαν άλλοι ερευνητές. Πιο συγκεκριμένα με την προσθήκη κάποιου λεξικού και κυρίως με χρηματοοικονομικές έννοιες πολύ πιθανόν το μοντέλο που κατασκευάσαμε θα βελτιωνόταν ακόμα περισσότερο και θα έδινε ακριβέστερα sentiment scores. Για παράδειγμα θα μπορούσε να χρησιμοποιηθεί το λεξικό Vader το οποίο για κάθε λέξη περιέχει και μια τιμή για το πόσο θετική ή αρνητική είναι. Δίνοντας και αυτή την τιμή στο νευρωνικό που κατασκευάσαμε πιθανόν να παρήγαγε καλύτερο αποτέλεσμα.

Τέλος, όπως αναφέρθηκε και στην ενότητα related work πλέον υπάρχουν τρόποι και μέθοδοι να αναλύσουμε τα κειμενικά δεδομένα ώστε να παράξουμε περισσότερη πληροφορία. Αναλυτικότερα σε αυτή την εργασία αξιοποιήθηκε μόνο το πόσο θετικό ή αρνητικό είναι το κάθε σχόλιο. Ως μελλοντική έρευνα θα μπορούσε κάποιος να αξιοποιήσει και άλλους «άξονες» συναισθήματος όπως είναι το πόσο ήρεμο, αγχώδης και πόσο σίγουρο είναι το κάθε σχόλιο. Με αυτό τον τρόπο το νευρωνικό μοντέλο που προβλέπει την μελλοντική συμπεριφορά των χρηματιστηριακών τιμών θα τροφοδοτούνταν με περισσότερες τιμές και αρκετά πιθανόν να προέβαινε σε καλύτερες και ακριβέστερες προβλέψεις.

Κεφάλαιο 8

8. Βιβλιογραφία

[1] J. Bollen, H. Mao, and X. Zeng, “Twitter mood predicts the stock market,” *J. Comput. Sci.*, vol. 2, no. 1, pp. 1–8, 2011.

[2] J. Dean, “Distilling the Knowledge in a Neural Network,” pp. 1–9.

[3] J. D. Hamilton, “Time Series Analysis.”

[4] B. G. Malkiel, “Critics,” vol. 17, no. 1, pp. 59–82, 2003.

[5] B. Qian and K. Rasheed, “Stock market prediction with multiple classifiers,” no. November 2006, pp. 25–33, 2007.

[6] M. Mittermayer and G. F. Knolmayer, “Text Mining Systems for Market Response to News : A Survey,” vol. 41, no. 184, 2006.

[7] W. Antweiler and M. Z. Frank, “Is All That Talk Just Noise ? The Information Content of Internet Stock Message Boards,” vol. LIX, no. 3, 2004.

[8] E. Gilbert and K. Karahalios, “Predicting Tie Strength With Social Media,” 2009.

[9] M. Makrehchi, S. Shah, and W. Liao, “Stock Prediction Using Twitter Sentiment Analysis,” *Stanford*, vol. 1, no. June, pp. 337–342, 2009.

[10] J. A. Puppim *et al.*, “Promoting win e win situations in climate change mitigation , local environmental quality and development in Asian cities through co-bene fi ts Development Goals :,” *J. Clean. Prod.*, vol. 58, pp. 1–6, 2013.

[11] J. Li, H. Bu, and J. Wu, “Sentiment-aware stock market prediction: A deep learning method,” *14th Int. Conf. Serv. Syst. Serv. Manag. ICSSSM 2017 - Proc.*, pp. 1–6, 2017.

[12] Y. Lin, H. Guo, and J. Hu, “An SVM-based Approach for Stock Market Trend Prediction,” *2013 Int. Jt. Conf. Neural Networks*, pp. 1–7.

[13] B. Liu, *Sentiment Analysis*. 2012.

[14] T. Nasukawa, “Sentiment Analysis --Capturing favorability using Natural Language Processing--,” no. March, 2015.

[15] C. O. Alm, D. Roth, and R. Sproat, “Emotions from text : machine learning for text-based emotion prediction,” no. October, pp. 579–586, 2005.

[16] M. Lemmon, “Consumer Confidence and Asset Prices : Some Empirical Evidence,” 2006.

[17] K. Cortis, T. Daudert, H. Manuela, M. Zarrouk, S. Handschuh, and B. Davis, “SemEval-2017 Task 5 : Fine-Grained Sentiment Analysis on Financial Microblogs and News,” pp. 519–535, 2017.

[18] I. Symeonidis, A. Aly, and M. A. Mustafa, “SePCAR : A Secure and Privacy-Enhancing Protocol for Car Access Provision SePCAR : A Secure and Privacy-Enhancing Protocol for Car Access Provision,” no. September, 2017.

[19] P. Saleiro and C. Soares, “FEUP at SemEval-2017 Task 5: Predicting Sentiment Polarity and Intensity with Financial Word Embeddings,” no. April, 2017.

[20] A. Seyeditabari, U. N. C. Charlotte, and U. N. C. Charlotte, “Emotion Detection in Text : a Review,” 2009.

[21] M. S. Akhtar, A. Kumar, D. Ghosal, A. Ekbal, and P. Bhattacharyya, “A Multilayer Perceptron based Ensemble Technique for Fine-grained Financial Sentiment Analysis,” pp. 540–546, 2018.

[22] L. Pivovarova and R. Yangarber, “HCS at SemEval-2017 Task 5 : Sentiment Detection in Business News Using Convolutional Neural Networks,” pp. 842–846, 2017.

[23] A. Hedayati, M. Hedayati, and M. Esfandyari, “Stock market index prediction using artificial neural network,” *J. Econ. Financ. Adm. Sci.*, vol. 21, no. 41, pp. 89–93, 2016.

[24] T. Quah and B. Srinivasan, “Improving returns on stock investment through neural network selection,” vol. 17, pp. 295–301, 1999.

[25] P. Hájek, “Municipal credit rating modelling by neural networks,” *Decis. Support Syst.*, vol. 51, no. 1, pp. 108–118, 2011.

[26] E. Guresen, G. Kayakutlu, and T. U. Daim, “Expert Systems with Applications Using artificial neural network models in stock market index prediction,” *Expert Syst. Appl.*, vol. 38, no. 8, pp. 10389–10397, 2011.

[27] A. Chen, M. T. Leung, and H. Daouk, “Application of neural networks to an emerging ÿnancial market : forecasting and trading the Taiwan Stock Index,” vol. 30, pp. 901–923, 2003.

[28] R. J. Kuo, C. H. Chen, and Y. C. Hwang, “An intelligent stock trading decision support system through integration of genetic algorithm based fuzzy neural network and artiÿcial neural network,” vol. 118, pp. 21–45, 2001.

[29] G. S. Atsalakis and K. P. Valavanis, “Expert Systems with Applications Forecasting stock market short-term trends using a neuro-fuzzy based methodology,” *Expert Syst. Appl.*, vol. 36, no. 7, pp. 10696–10707, 2009.

[30] M. Roondiwala, H. Patel, and S. Varma, “Predicting Stock Prices Using LSTM,” no. September, pp. 1–4, 2018.

[31] Κ. Μ. Μάθηση, “ΚΕΦΑΛΑΙΟ 4 –Μηχανική Μάθηση.”

[32] “https://towardsdatascience.com/understanding-neural-networks-19020b758230.”

[33] Chris Chatfield, “The Analysis of Time Series.”

[34] R. S. T. Daniel Peña, George C. Tiao, “A Course in Time Series Analysis,” 2001.

[35] Douglas C. Montgomery, *Introduction to Time Series Analysis and Forecasting*. .

[36] “Πρόβλεψη, Μιχάλης Βαϊδάνης, 2005 http://www.metal.ntua.gr/uploads/3469/447/forecasting.pdf,” pp. 1–18, 2005.

[37] Https://en.wikipedia.org/wiki/Twitter, “Twitter.”

[38] “https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17.”

[39] B. Pang and L. Lee, “Opinion mining and sentiment analysis,” vol. 2, no. 1, 2008.

[40] “https://colah.github.io/posts/2015-08-Understanding-LSTMs/.”

[41] R. Cascade-correlation and N. S. Chunking, “2 PREVIOUS WORK,” vol. 9, no. 8, pp. 1–32, 1997.

[42] “https://machinelearningmastery.com/develop-bidirectional-lstm-sequence-classification-python-keras/.”

[43] “https://medium.com/the-theory-of-everything/understanding-activation-functions-in-neural-networks-9491262884e0.” .

[44] “towardsdatascience.com/coding-deep-learning-for-beginners-linear-regression-part-2-cost-function-49545303d29f.” .

[45] “https://en.wikipedia.org/wiki/Evaluation_function.”

[46] “https://en.wikipedia.org/wiki/Overfitting.”

[47] “https://machinelearningmastery.com/pooling-layers-for-convolutional-neural-networks/.”

[48] “https://keras.io/layers/embeddings/.”

[49] “https://keras.io/layers/core/.”

[50] “https://machinelearningmastery.com/difference-between-a-batch-and-an-epoch/.”

[51] “https://medium.com/syncedreview/iclr-2019-fast-as-adam-good-as-sgd-new-optimizer-has-both-78e37e8f9a34.”

[52] S. Golson, “One-hot state machine design for FPGAs History of one-hot encoding Why use one-hot Example design,” pp. 1–6, 1993.

[53] “Introduction to Word Embedding and Word2Vec.” [Online]. Available: https://towardsdatascience.com/introduction-to-word-embedding-and-word2vec-652d0c2060fa.

[54] “GloVe: Global Vectors for Word Representation.” [Online]. Available: https://nlp.stanford.edu/projects/glove/.

[55] “https://www.investopedia.com/.”