

Ατομική Διπλωματική Εργασία

**ΑΝΑΛΥΣΗ ΣΥΝΑΙΣΘΗΜΑΤΟΣ ΣΕ ΚΟΙΝΩΝΙΚΑ ΔΙΚΤΥΑ ΩΣ ΜΕΘΟΔΟΣ
ΠΡΟΒΛΕΨΗΣ ΤΟΥ ΧΡΗΜΑΤΙΣΤΗΡΙΟΥ**

Αντώνης Πουλλής

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ



ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Ιανουάριος 2021

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

**Ανάλυση Συναισθήματος σε Κοινωνικά Δίκτυα ως Μέθοδος Πρόβλεψης του
Χρηματιστήριου**

Αντώνης Πουλλής

Επιβλέπων Καθηγητής

Γιώργος Πάλλης

Η Ατομική Διπλωματική Εργασία υποβλήθηκε προς μερική εκπλήρωση των
απαιτήσεων απόκτησης του πτυχίου Πληροφορικής του Τμήματος Πληροφορικής του
Πανεπιστημίου Κύπρου

Ιανουάριος 2021

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά τον επιβλέπων καθηγητή μου κύριο Γιώργο Πάλλη, για την ευκαιρία που μου έδωσε να γνωρίσω ένα τόσο ενδιαφέρον τομέα της επιστήμης της Πληροφορικής και για την ανεκτίμητη βοήθεια του κατά την διεξαγωγή της μελέτης και την ολοκλήρωση της εργασίας αυτής.

Ευχαριστώ επίσης θερμά τον ερευνητή του Τμήματός Πληροφορικής κύριο Δημοσθένη Στεφανίδη, για την καθοδήγηση του και την άμεση ανταπόκριση σε παντός είδους ζητήματα.

Περίληψη

Τα μέσα κοινωνικής δικτύωσης αποτελούν αναπόσπαστο κομμάτι της σύγχρονης πραγματικότητας. Ζώντας στην εποχή όπου τα κοινωνικά δίκτυα προβάλλονται ως το βασικότερο μέσο κοινωνικής διάδρασης, προκύπτει αβίαστα αυξημένη παρακίνηση από την επιστημονική κοινότητα για συλλογή, ανάλυση και αξιοποίηση τεραστίου όγκου δεδομένων, οριοθετώντας νέες προοπτικές προς την εξέλιξη και ανάδυση πτυχών σε πληθώρα τομέων του σύγχρονου κόσμου.

Η παρούσα εργασία εστιάζεται συνδυαστικά στα Μέσα Κοινωνικής Δικτύωσης και στον οικονομικό τομέα της Χρηματιστηριακής Αγοράς. Πιο συγκεκριμένα, πραγματοποιήθηκε συλλογή δεδομένων από το Μέσο Κοινωνικής Δικτύωσης Twitter για αντιπροσωπευτικό δείγμα 33 εταιρειών από 11 διαφορετικούς επιχειρηματικούς τομείς της Οικονομίας και συλλογή των αντίστοιχων Χρηματιστηριακών τιμών τους, για το ίδιο χρονικό διάστημα. Η επιλογή των εν λόγω εταιρειών βασίστηκε στην εξέταση της επιρροής της πανδημίας του κορωνοϊού στις Χρηματιστηριακές μετοχές τους, κατά την περίοδο του πρώτου κύματος μετάδοσης της, αξιοποιώντας την ανάλυση συναισθήματος, από τα συλλεχθέντα Tweets, ως πρωταρχικό μέσο για εξαγωγή γνώσης. Τα Tweets, έτυχαν προ επεξεργασίας και της ανάλογης διαχείρισης για την παραγωγή αποτελεσμάτων που αφορούν την ανάλυση συναισθήματος και τον υπολογισμό μετρικών στο πεδίο ενδιαφέροντος. Στη συνέχεια, τα αποτελέσματα αναλύθηκαν και αξιολογήθηκαν για εξεύρεση πληροφοριών, σχετικών όχι μόνο με τη συσχέτιση της ευρύτερης κοινωνικής διάθεσης συναρτήσει της κίνησης των μετοχών, αλλά και με την υποβοήθηση της επενδυτικής κοινότητας, για μελλοντικές επιτυχημένες αναζητήσεις και δημοσιεύσεις στην πλατφόρμα Twitter.

Επιπλέον στόχο της παρούσας ατομικής διπλωματικής εργασίας, αποτέλεσε η ανάπτυξη ενός αξιόπιστου μοντέλου πρόβλεψης της Χρηματιστηριακής συμπεριφοράς, με βάση την ανάλυση συναισθήματος αναρτήσεων, από μέσα

κοινωνικής δικτύωσης. Σε αυτό το πλαίσιο, αξιοποιήθηκε σύνολο δεδομένων που περιέχει τίτλους ειδήσεων από τη κοινότητα δημόσιας συζητήσεως Reddit και συλλέχθηκαν οι αντίστοιχες Χρηματιστηριακές τιμές της ίδιας χρονικής περιόδου, για τη πρόβλεψη της διακύμανσης της τιμής του Χρηματιστηριακού Δείκτη Dow Jones. Έπειτα από την απαραίτητη προετοιμασία των δεδομένων και τη πραγματοποίηση ανάλυσης συναισθήματος στο σύνολο των αναρτήσεων, δημιουργήθηκε ένα συγχωνευμένο ενιαίο σύνολο δεδομένων με χαρακτηριστικά, τα οποία χρησιμοποιήθηκαν για την εκπαίδευση 6 διαφορετικών αλγορίθμων μηχανικής μάθησης, προκειμένου να προβλεφθεί η τάση της τιμής κλεισίματος του Δείκτη Dow Jones για κάθε μέρα, σε άγνωστα δεδομένα για τα μοντέλα πρόβλεψης. Για την αξιολόγηση των αποτελεσμάτων από τις προβλέψεις των αλγορίθμων μηχανικής μάθησης, επιλέχθηκαν οι μετρικές της βαθμολογίας F1 και της ορθότητας (accuracy), με σκοπό τη σύγκριση της αποδοτικότητας τους, για την επιλογή του τελικού μοντέλου, ώστε να καταστεί δυνατή η υιοθέτηση και η πρακτική χρησιμοποίηση της θεώρησης περί πρόβλεψης της κίνησης των Χρηματιστηριακών Δεικτών, με την αξιοποίηση της ανάλυσης συναισθήματος σε κοινωνικά δίκτυα και της μηχανικής μάθησης.

Περιεχόμενα

Κεφάλαιο 1 Εισαγωγή	1
1.1. Υποκίνηση Εργασίας	1
1.2. Στόχοι της Εργασίας	2
1.3. Εισαγωγή στα μοντέλα πρόβλεψης των Μέσων Κοινωνικής Δικτύωσης	3
1.4. Περίγραμμα Εργασίας	4
Κεφάλαιο 2 Παρόμοιες Μελέτες	7
2.1. Παρόμοιες μελέτες	7
Κεφάλαιο 3 Θεωρητικό Υπόβαθρο	12
3.1. Θεωρητικό Υπόβαθρο Twitter.....	13
3.1.1. Tweet/ ReTweet/Mention(@)	13
3.1.2. Hashtag (#)	14
3.2. Θεωρητικό Υπόβαθρο Reddit/ Reddit WorldNews Channel	14
3.3. MongoDB.....	15
3.4. Γλώσσα προγραμματισμού Python	16
3.4.1. NLTK.....	18
3.4.2. VADER	18
3.5. Εισαγωγή στο Χρηματιστήριο	19
3.5.1. Μετοχή	20
3.5.2. Πρόβλεψη της Χρηματιστηριακής Αγοράς.....	20
3.5.3. The New York Stock Exchange (NYSE)	20
3.5.4. NASQAD	20
3.5.5. Dow Jones Industrial Average	21
Κεφάλαιο 4 Περιγραφή Μεθοδολογίας	22
4.1. Μεθοδολογία Αξιοποίησης Αναρτήσεων από το Twitter	23
4.1.1. Συλλογή Δεδομένων από το Twitter	23
4.1.2. Προ-επεξεργασία Κειμένου από το Twitter	28
4.1.3. Ανάλυση Συναισθήματος	30
4.1.3.1. Κατηγοριοποίηση αναρτήσεων με βάση το συναίσθημα.....	32

4.1.3.2. Δημοφιλή Hashtags	32
4.1.3.3. Συχνότητα εμφάνισης λέξεων	33
4.1.3.4. Υποκειμενικότητα και Πολικότητα	34
4.2. Μεθοδολογία Ανάπτυξης Μοντέλου Πρόβλεψης	34
4.2.1. Περιγραφή Συνόλου Δεδομένων από το Reddit	34
4.2.2. Προ-επεξεργασία συνόλου δεδομένων	36
4.2.3. Επεξεργασία Συνόλου Δεδομένων	37
4.2.4. Συγχώνευση Συνόλων Δεδομένων.....	38
4.2.5. Μετατροπή Συνόλου Δεδομένων σε Πρόβλημα Μηχανικής Μάθησης...	39
4.2.6. Μηχανική Μάθηση	40
4.2.7. Βελτιστοποίηση αλγορίθμων μηχανικής μάθησης.....	41
4.3. Συλλογή Δεδομένων για τιμές της Χρηματιστηριακής Αγοράς	44
Κεφάλαιο 5 Ανάλυση Συναισθήματος και Covid-19	47
5.1. Covid-19 και Οικονομία	47
5.2. Επιλογή προς εξέταση εταιρειών	48
5.2.1. Αερογραμμές	50
5.2.2. Τουρισμός	50
5.2.3. Ενοικιάσεις Αυτοκινήτων	51
5.2.4. Υπηρεσίες Τηλεδιασκέψεων.....	51
5.2.5. Γραμμές Κρουαζιέρας	51
5.2.6. Υπηρεσίες Streaming.....	52
5.2.7. Ηλεκτρονικά καταστήματα/Αλυσίδες Υπεραγορών	52
5.2.8. Πλατφόρμες Κοινωνικής Δικτύωσης.....	53
5.2.9. Υγεία	53
5.2.10. Τεχνολογία	54
5.2.11. Εστίαση	54
Κεφάλαιο 6 Αξιολόγηση Αποτελεσμάτων	55
6.1. Πειράματα και Αποτελέσματα για Tweets	55
6.1.1. Κατηγοριοποίηση Αναρτήσεων με βάση το Συναίσθημα.....	56
6.1.2. Υποκειμενικότητα.....	64
6.1.3. Συχνότητα Εμφάνισης Λέξεων	65

6.1.4. Ανάλυση αποτελεσμάτων για Hashtags.....	70
6.1.5. Ανάλυση συσχετίσεων	76
6.2. Πειράματα και Αποτελέσματα για Μοντέλο Πρόβλεψης	77
6.2.1. Επιλογή μετρικών.....	77
6.2.2. Ανάλυση αποτελεσμάτων	79
Κεφάλαιο 7 Διαδικτυακές εφαρμογές.....	89
7.1. Αρχική σελίδα διαδικτυακής εφαρμογής	89
7.2. Διαδικτυακές εφαρμογές Streamlit.....	91
7.2.1. Διαδικτυακή εφαρμογή για απεικόνιση Χρηματιστηριακών Δεδομένων	92
7.2.2. Διαδικτυακή εφαρμογή για Ανάλυση Συναισθήματος.....	95
7.2.2.1. Ανάλυση συναισθήματος με δεδομένα εισόδου χρήστη	96
7.2.2.2. Ανάλυση συναισθήματος σε σύνολα δεδομένων	97
Κεφάλαιο 8 Περιορισμοί	102
8.1. Περιορισμοί κατά τη συλλογή δεδομένων	102
8.2. Περιορισμοί κατά την Προ-επεξεργασία Κειμένου	104
8.3. Περιορισμοί κατά την ανάλυση συναισθήματος.....	104
8.4. Περιορισμοί κατά την μετατροπή του συνόλου δεδομένων σε πρόβλημα μηχανικής μάθησης.....	105
8.5. Περιορισμοί αποτελεσμάτων με την χρησιμοποίηση GridSearchCV	106
Κεφάλαιο 9 Συμπεράσματα.....	107
9.1. Συμπεράσματα	107
9.2. Εισηγήσεις	109
Βιβλιογραφία	111
Παράρτημα Α.....	A-1
Παράρτημα Β.....	B-1

Κεφάλαιο 1

Εισαγωγή

1.1.	Υποκίνηση Εργασίας	1
1.2.	Στόχοι της Εργασίας	2
1.3.	Εισαγωγή στα μοντέλα πρόβλεψης των Μέσων Κοινωνικής Δικτύωσης	3
1.4.	Περίγραμμα Εργασίας	4

1.1. Υποκίνηση Εργασίας

Τεράστιες ποσότητες πληροφοριών δημοσιεύονται σε καθημερινή βάση από χρήστες στα μέσα κοινωνικής δικτύωσης. Μεταξύ αυτών, υπάρχουν πολύτιμες πληροφορίες για σχεδόν κάθε θέμα που απασχολεί τους χρήστες. Αν και κάθε ανάρτηση μπορεί να μην φαίνεται εξαιρετικά πολύτιμη, υποστηρίζεται ότι οι συγκεντρώσεις μεγάλου όγκου δεδομένων σε κείμενα από αναρτήσεις και διαδικτυακές δημοσιεύσεις, είναι σε θέση να παρέχουν σημαντικές πληροφορίες σχετικά με τη δημόσια διάθεση και το συναίσθημα [9].

Οι εφαρμογές των πληροφοριών αυτών, κατόπιν επεξεργασίας, μπορούν να αποτελέσουν ωφέλιμο αρωγό στον οικονομικό τομέα, καθώς η εξακρίβωση συναισθήματος του κοινού, μέσω της ανάκτησης διαδικτυακών πληροφοριών, μπορεί να αποτελέσει πολύτιμο εργαλείο για την πραγματοποίηση στρατηγικών συναλλαγών. Εφαρμογές τέτοιου τύπου δεν εμπίπτουν ούτε στο πλαίσιο της θεμελιώδους ανάλυσης, ούτε σε αυτό της τεχνικής ανάλυσης, που αποτελούν τις δύο κύριες μεθόδους ανάλυσης του Χρηματιστηρίου, αλλά αποτελούν ξεχωριστό πεδίο ανάλυσης το οποίο προκύπτει μετά από την εξόρυξη δεδομένων από τα κοινωνικά δίκτυα.

Είναι δεδομένο ότι για τη δημιουργία πολύπλευρων και εμπειριστατωμένων πραγματικών συσχετισμών σε τιμές μετοχών, συνεπάγεται μελέτη πολλαπλών

επιπλέον παραγόντων. Ωστόσο, η παρούσα εργασία στοχεύει στην εξέταση της συσχέτισης μεταξύ του συναισθήματος και της τιμής των μετοχών, απλώς ως ενδιαφέρουσα εφαρμογή ανάλυσης συναισθημάτων για την ενδεχόμενη ανάπτυξη ενός αξιόπιστου μοντέλου προβλέψεων βασιζόμενο στις συσχετίσεις αυτές. Αρκετοί είναι επίσης οι παράγοντες οι οποίοι εμπλέκονται στην πρόβλεψη της κίνησης των τιμών των μετοχών, όμως δεν χρήζει καμίας αμφιβολίας ότι η μελέτη της ευρύτερης κοινωνικής διάθεσης αποτελεί έναν από αυτούς.

1.2. Στόχοι της Εργασίας

Από ανέκαθεν, η πρόβλεψη της Χρηματιστηριακής Αγοράς αποτελούσε ένα ενδιαφέρον θέμα μεταξύ των ερευνητών και επενδυτών. Με την πάροδο του χρόνου μπήκε στο μικροσκόπιο και η εξαιρετικά ιδιαίτερη πρόκληση του συνδυασμού των δεδομένων της Αγοράς με το κοινό συναίσθημα για την ενίσχυση και ενδυνάμωση του εγχειρήματος των προβλέψεων της κίνησης της.

Στόχο της παρούσας ατομικής διπλωματικής εργασίας, αποτελεί η μελέτη της πρόκλησης του συνδυασμού της ανάλυσης συναισθήματος κειμένων από τα μέσα κοινωνικής δικτύωσης και το διαδίκτυο, με τη Χρηματιστηριακή συμπεριφορά συγκεκριμένων δεικτών και εταιρειών, συγκεντρώνοντας γνώση για να καταστεί δυνατή η ανάπτυξη ενός συστήματος πρόβλεψης¹.

Λαμβάνοντας υπόψιν τις πρόσφατες εξελίξεις σε Παγκόσμιο επίπεδο με την έξαρση της πανδημίας του κωρονοϊού Covid-19, στην παρούσα εργασία προστέθηκε ο επιπλέον στόχος της συλλογής αναρτήσεων από το μέσο κοινωνικής δικτύωσης Twitter, σχετικές με μετοχές εταιρειών από πληθώρα τομέων της οικονομίας που ενδεχομένως να επηρεάστηκαν από την πανδημία, με σκοπό την εξέταση κατά πόσο η κίνηση των μετοχών τους μπορεί να είναι ανιχνεύσιμη, μέσω της ανάλυσης συναισθήματος.

¹ Αποποίηση Συγγραφέα: Η παρούσα εργασία δεν αποτελεί χρηματοοικονομική ή επενδυτική συμβουλή και δεν είναι εγγυημένο ότι είναι σε θέση να παρέχει ορθά αποτελέσματα για μελλοντικές προβλέψεις. Ως εκ τούτου, οι αναγνώστες καλούνται να αναλογιστούν το ενδεχόμενο ρίσκο και να μην αξιοποιήσουν το περιεχόμενο της, ως πρωταρχική πηγή επενδυτικών πληροφοριών.

1.3. Εισαγωγή στα μοντέλα πρόβλεψης των Μέσων Κοινωνικής Δικτύωσης

Οι τιμές του Χρηματιστηρίου είναι εξαιρετικά απρόβλεπτες και ασταθείς. Ως εκ τούτου, δεν υπάρχουν σταθερά πρότυπα στα δεδομένα που να επιτρέπουν την μοντελοποίηση των τιμών των μετοχών με υψηλή ακρίβεια, για μεγάλα χρονικά διαστήματα. Ο οικονομολόγος καθηγητής του Πανεπιστημίου του Princeton, Burton Malkiel, στο βιβλίο του "A Random Walk Down Wall Street", το 1996, υποστήριξε ότι: « ...εάν η αγορά είναι πραγματικά αποτελεσματική και η τιμή μιας μετοχής αντανακλά όλους τους παράγοντες που υπάρχουν, τότε αμέσως μετά τη δημοσιοποίηση της, μια μαϊμού με δεμένα τα μάτια που ρίχνει βελάκια σε μια λίστα εφημερίδων, θα είναι σε θέση να επενδύσει τόσο καλά όσο οποιοσδήποτε επαγγελματίας επενδυτής» [20].

Ωστόσο, η πρόβλεψη τιμών στη Χρηματιστηριακή Αγορά δεν αποτελεί στοχαστική ή τυχαία διαδικασία και με την αξιοποίηση της μηχανικής μάθησης είναι δυνατή η μοντελοποίηση των δεδομένων, με τρόπο τέτοιο ώστε οι προβλέψεις που πραγματοποιούνται τουλάχιστον να συσχετίζονται με την πραγματική τους συμπεριφορά. Επομένως, δεν είναι αναγκαία η ακριβής πρόβλεψη στις μελλοντικές τιμές των μετοχών, αλλά οι κινήσεις των τιμών τους, δηλαδή, εάν πρόκειται να αυξηθούν ή να μειωθούν στο εγγύς μέλλον.

Οι προβλέψεις οι οποίες προκύπτουν από την ανάλυση πληροφοριών, που προέρχονται από τα μέσα κοινωνικής δικτύωσης, βασίζονται κατά κύρια ομολογία στις απόψεις και πεποιθήσεις των χρηστών [28]. Οι πληροφορίες αυτές συγκρίνονται αργότερα με πραγματικά δεδομένα και γεγονότα για να εξακριβωθεί η ακρίβεια των προβλέψεων. Ένα από τα πιο συνηθισμένα λάθη στον τομέα των επενδύσεων, αποτελεί η τυφλή αντιγραφή ή μίμηση των επενδυτικών δραστηριοτήτων άλλων επενδυτών. Ακόμα και στο μακρινό 2008, έρευνα έδειξε ότι ένας στους τέσσερις Αμερικανούς επενδυτές βασιζόταν στα μέσα κοινωνικής δικτύωσης για να λάβει επενδυτικές συμβουλές [7]. Αναλογιζόμενοι λοιπόν ότι το Χρηματιστήριο κυβερνάται από τους επενδυτές του, η διάθεση και το συναίσθημα που επικρατεί μεταξύ τους

επηρεάζει την καθημερινή του λειτουργία και αποτελεί εργαλείο για την πρόβλεψη του.



Σχήμα 1.1: Παράδειγμα Μοντέλου Πρόβλεψης Χρηματιστηριακών Δεικτών

1.4. Περίγραμμα Εργασίας

Η παρούσα διπλωματική εργασία αποτελείται συνολικά από εννέα κεφάλαια. Στη συνέχεια, δίδεται σύντομη περιγραφή του περιεχομένου κάθε κεφαλαίου. Σημειώνεται ότι των κεφαλαίων προηγούνται η Περίληψη της εργασίας και τα Περιεχόμενα.

Κεφάλαιο 1 – Εισαγωγή

Στο Κεφάλαιο 1, γίνεται μια σύντομη εισαγωγή στη θεματολογία της εργασίας και σε στοιχεία καίριας σημασίας που υποκίνησαν στη μελέτη και την υλοποίηση της. Επιπρόσθετα, παρατίθενται οι κύριοι στόχοι της. Ακολουθεί επιγραμματική εισαγωγή στο θέμα των μοντέλων πρόβλεψης των Μέσων Κοινωνικής Δικτύωσης για να αξιοποιηθεί αυτή η γνώση στην κατανόηση του προτεινόμενου μοντέλου της εργασίας. Το κεφάλαιο ολοκληρώνει το παρόν Περίγραμμα Εργασίας.

Κεφάλαιο 2 – Παρόμοιες Μελέτες

Στο Κεφάλαιο 2, γίνεται αναφορά σε παρόμοιες μελέτες, την συνεισφορά τους στο πεδίο ενδιαφέροντος και τη συσχέτιση τους με την παρούσα εργασία.

Κεφάλαιο 3 – Θεωρητικό Υπόβαθρο

Στο Κεφάλαιο 3, παρουσιάζεται θεωρητικό υπόβαθρο σε εργαλεία και τεχνολογίες που αξιοποιήθηκαν για τους σκοπούς της παρούσας μελέτης. Επίσης, συμπεριλαμβάνονται βασικές έννοιες και επεξηγήσεις που αφορούν την Χρηματιστηριακή Αγορά σε επίπεδο τέτοιο, ώστε να καταστεί δυνατή η κατανόηση της μεθοδολογίας και των αποτελεσμάτων της εργασίας.

Κεφάλαιο 4 – Μεθοδολογία

Στο Κεφάλαιο 4, περιγράφονται αναλυτικά οι δύο βασικές ξεχωριστές μεθοδολογίες που ακολουθήθηκαν στην παρούσα εργασία.

Στο πρώτο μέρος περιγράφεται αναλυτικά η διαδικασία υλοποίησης ενός συστήματος που επικεντρώνεται στη συλλογή, προ-επεξεργασία και ανάλυση συναισθήματος αναρτήσεων από το κοινωνικό δίκτυο Twitter για συγκεκριμένες εταιρείες και για καθορισμένο χρονικό διάστημα.

Στο δεύτερο μέρος, παρουσιάζονται αναλυτικά τα βήματα ανάπτυξης μοντέλου πρόβλεψης, το οποίο αξιοποιεί συνδυαστικά δεδομένα από το Χρηματιστήριο και πληροφορίες που προκύπτουν μετά την ανάλυση συναισθήματος σε προκαθορισμένο σύνολο δεδομένων. Τα δεδομένα εισόδου χρησιμοποιούνται για την εκπαίδευση αλγορίθμων μηχανικής μάθησης, με σκοπό την πρόβλεψη σε αόρατα για το σύστημα δεδομένα και τη σύγκριση της απόδοσης αλγορίθμων μηχανικής μάθησης, με σκοπό την επιλογή του καλύτερου δυνατού τελικού μοντέλου.

Κεφάλαιο 5 – Ανάλυση Συναισθήματος και Covid-19

Στο Κεφάλαιο 5, γίνεται αναφορά, στη συλλογή αναρτήσεων από το Twitter, για επιχειρήσεις επηρεασμένες από την πανδημία του κορωνοϊού και στην ανάλυση των δεδομένων τους για εξαγωγή γνώσης. Αρχικά, γίνεται σύντομη εισαγωγή στο κίνητρο και στους λόγους συλλογής δεδομένων. Στη συνέχεια, παρουσιάζεται αιτιολόγηση, για κάθε μία από τις επιλογές των τομέων της οικονομίας και των συγκεκριμένων εταιρειών, που τελικά εξετάστηκαν.

Κεφάλαιο 6 – Αξιολόγηση Αποτελεσμάτων

Στο Κεφάλαιο 6, παρουσιάζονται τα αποτελέσματα που παράχθηκαν για τις δύο ξεχωριστές μεθοδολογίες που εφαρμοστήκαν, όπως περιεγράφηκαν στο Κεφάλαιο 4. Τα αποτελέσματα και οι πληροφορίες που εξάχθηκαν, αξιολογούνται για την επισκόπηση της ποιότητας της μελέτης και την εξακρίβωση του βαθμού επίτευξης των αρχικών στόχων που τέθηκαν.

Κεφάλαιο 7 – Διαδικτυακές Εφαρμογές

Στο Κεφάλαιο 7, γίνεται αναφορά στις διαδικτυακές εφαρμογές που υλοποιήθηκαν στο πλαίσιο της διπλωματικής εργασίας και περιέχουν βασικές λειτουργίες του συστήματος. Συγκεκριμένα, αναφέρονται εν συντομία οι δυνατότητες τους, παρατίθενται στιγμιότυπα από τις διεπαφές τους και αναφέρεται η μεθοδολογία ανάπτυξης τους από τα αρχικά στάδια υλοποίησης μέχρις ότου φτάσουν στις συσκευές των τελικών χρηστών.

Κεφάλαιο 8 – Περιορισμοί

Στο Κεφάλαιο 8, αναφέρονται οι περιορισμοί στην έρευνα και υλοποίηση, καθ' όλη τη διάρκεια διεκπεραίωσης της εργασίας. Πιο συγκεκριμένα το κεφάλαιο επικεντρώνεται, σε τεχνικά και θεωρητικά ζητήματα του κάθε σταδίου της μελέτης και παρουσιάζεται αιτιολόγηση για το αναπόφευκτο του κάθε περιορισμού.

Κεφάλαιο 9 – Συμπεράσματα

Στο Κεφάλαιο 9, παρουσιάζονται τα συμπεράσματα που έχουν εξαχθεί μετά την εξαγωγή πληροφορίας από την ανάλυση συναισθήματος των Tweets και την υλοποίηση του μοντέλου πρόβλεψης. Επίσης, παρατίθενται εισηγήσεις για επεκτάσεις και περαιτέρω έρευνα σε θέματα που καταπιάνεται η παρούσα εργασία, με στόχο την εις βάθος μελέτη του πεδίου ενδιαφέροντος, την εξαγωγή εγκυρότερων αποτελεσμάτων και τη μελλοντική πρόοδο στον τομέα που ασχολείται με την πρόβλεψη της Χρηματιστηριακής Αγοράς, μέσω της ανάλυσης συναισθήματος σε μέσα κοινωνικής δικτύωσης.

Κεφάλαιο 2

Παρόμοιες Μελέτες

2.1. Παρόμοιες μελέτες	7
------------------------------	---

2.1. Παρόμοιες μελέτες

Στο παρόν κεφάλαιο, παρουσιάζονται επιγραμματικά ενδιαφέρουσες μελέτες με επίκεντρο την ανάπτυξη μοντέλων πρόβλεψης του Χρηματιστηρίου, χρησιμοποιώντας δεδομένα, όπως αναρτήσεις από το μέσο κοινωνικής δικτύωσης Twitter, ειδήσεις και Χρηματιστηριακές τιμές μετοχών, για καθορισμένη χρονική περίοδο συλλογής δεδομένων.

Αρχική μελέτη βαρόμετρο για τη μελλοντική πρόοδο γύρω από τη θεματολογία αποτελεί αυτή των Bollen κ.ά. [2]. Η έρευνα ασχολήθηκε με την εξερεύνηση πιθανών μοτίβων συσχέτισης μεταξύ της ευρύτερης κοινωνικής διάθεσης με τις μεταβολές στο Χρηματιστηριακό δείκτη Dow Jones (DJIA). Για τη μελέτη αξιοποιήθηκε το Profile of Mood States (POMS) για την ανάλυση συναισθήματος σε Tweets δημοσιευμένα στο δεύτερο μισό του 2008. Η μελέτη αυτή αποτελεί ορόσημο, καθώς ήταν η πρώτη στην εν λόγω θεματολογία και οι μετέπειτα μελέτες στηρίχθηκαν στα αποτελέσματα της.

Μεταγενέστερη μελέτη των Zhang, κ.ά.[36] ασχολήθηκε με την πρόβλεψη των δεικτών Dow Jones (DJIA), NASDAQ, Standard & Poor's (S&P 500) και VIX, βασιζόμενη στην ανάλυση συναισθήματος αναρτήσεων του Twitter. Παρόμοια μελέτη αποτελεί επίσης, το μοντέλο πρόβλεψης των Rao και Srivastava[26] που μεταξύ άλλων αποτελεσμάτων εντόπισαν συσχέτιση της τάξης του 88% μεταξύ της κοινωνική διάθεση από το Twitter και των τιμών των μετοχών, για δείγμα 13 τεχνολογικών εταιρειών και τους δείκτες Nasdaq 100 και Dow Jones.

Την ίδια χρονική περίοδο, οι Mittal και Goel[23] χρησιμοποίησαν δεδομένα από το Twitter για την ευρύτερη κοινωνική διάθεση και τις τιμές του Χρηματιστηριακού δείκτη Dow Jones (DJIA) για την πρόβλεψη των μελλοντικών του διακυμάνσεων. Για την υλοποίηση του συστήματος τους, αξιοποίησαν το Self-Organizing Fuzzy Neural Networks (SOFNN) και πέτυχαν ποσοστό ορθότητας στις προβλέψεις, της τάξης του 75.56%.

Επίσης, η μελέτη του Ding κ.ά.[8] ασχολήθηκε με την πρόβλεψη των κινήσεων του Χρηματιστηριακού Δείκτη Standard and Poor's, βασιζόμενη στην ανάλυση δεδομένων από χρονοσειρές και στην ανάλυση συναισθήματος σε δεδομένα από το Twitter για περίοδο δύο χρόνων (Ιανουάριο 2008 – Απρίλιο 2010). Το σύστημα εκπαιδευσε τους αλγορίθμους, SVM, Logistic Regression και νευρωνικά δίκτυα, ενώ η ανάλυση συναισθήματος πραγματοποιήθηκε, μέσω του Naïve Bayes Classifier της NLTK. Μεταξύ άλλων, τα αποτελέσματα της εν λόγω μελέτης έδειξαν ότι η προσθήκη του συναισθήματος από το Twitter βελτιώνει την απόδοση του μοντέλου, αλλά η βελτίωση ήταν στατιστικά σημαντική μόνο με την χρήση του SVM, αξιοποιώντας δεδομένα 10 ή 30 ημερών πριν την πρόβλεψη.

Ακόμη, μια πιο πρόσφατη αξιολογή μελέτη αποτελεί αυτή των Caliñgo κ.ά.[6], η οποία ασχολήθηκε σε τοπικό επίπεδο με τη πρόβλεψη του κύριου δείκτη στο Φιλιππινέζικο Χρηματιστήριο, μετά από τη συλλογή δεδομένων από το Twitter με 3 διαφορετικές μεθοδολογίες, που ήταν: αναζήτηση αναρτήσεων με βάση τη τοποθεσία των χρηστών, τη παρουσία λέξεων κλειδιών και συγκεκριμένων hashtags στο κείμενο των αναρτήσεων και με βάση Tweets από τοπικά ειδησεογραφικά μέσα. Μετά το τέλος της έρευνας αποδείχθηκε ότι η πρώτη και τρίτη μέθοδος μπορούν να καταστούν βοηθητικές για την πρόβλεψη των διακυμάνσεων στη τιμή του εν λόγω δείκτη.

Μια ενδιαφέρουσα μεθοδολογία για την πρόβλεψη του Χρηματιστηρίου, προτάθηκε από τον Renato Bruni[5], ο οποίος αξιοποιώντας συνεχόμενα Χρηματιστηριακά δεδομένα (τιμή ανοίγματος, τιμή κλεισίματος, μέγιστη τιμή, ελάχιστη τιμή και δείκτες απόδοσης μετοχών που αξιοποιούνται για τεχνική ανάλυση) των Χρηματιστηριακών δεικτών Standard and Poor's και Financial Times Stock Exchange Milano Indice di

Borsa (FTSE MIB), από τις 20 Απριλίου 2010 μέχρι τις 12 Ιουλίου 2016, ασχολήθηκε με την πρόβλεψη αγοραπωλησίας δεικτών μετοχών εντός της ίδιας μέρας ως πρόβλημα δυαδικής ταξινόμησης. Πιο συγκεκριμένα, η μεθοδολογία, μελέτησε το πρόβλημα δυαδικής ταξινόμησης του κατά πόσο η επόμενη χρονική περίοδος είναι ευνοϊκή για ημερήσιες συναλλαγές ή όχι.

Η πρόσφατη μελέτη του ερευνητή Yang Li[19], είχε ως σκοπό την παρουσίαση της φιλοσοφίας σχεδιασμού ενός συνδυαστικού μοντέλου βαθιάς μάθησης, το οποίο συνδυάζει δύο επαναλαμβανόμενα νευρωνικά δίκτυα (RNNs), ακολουθούμενα από ένα πλήρως συνδεδεμένο νευρωνικό δίκτυο (FCNN), για να αποδείξει ότι, τα μοντέλα που συνδυάζουν τεχνολογίες βαθιάς μάθησης μπορούν πραγματικά να προβλέψουν τις μελλοντικές τάσεις των τιμών των μετοχών, πιο αποτελεσματικά από άλλες παραδοσιακές μεθόδους και ότι είναι σε θέση να βοηθήσουν καλύτερα τους επενδυτές στη λήψη σωστών επενδυτικών αποφάσεων. Για τη μελέτη αξιοποιήθηκαν Χρηματιστηριακές τιμές και ειδήσεις που αφορούν τον Χρηματιστηριακό Δείκτη Standard and Poor's, για διάστημα 121 Χρηματιστηριακών ημερών (04/10/2018 – 05/04/2019). Πραγματοποιήθηκε ανάλυση συναισθήματος στις ειδήσεις με το εργαλείο Vader και για την ανάπτυξη του μοντέλου αξιοποιήθηκαν τα νευρωνικά δίκτυα LSTM και GRU. Τα αποτελέσματα της μελέτης έδειξαν ότι το εν λόγω μοντέλο ξεπερνά σε απόδοση το υφιστάμενο βέλτιστο μοντέλο πρόβλεψης, που χρησιμοποιεί ουσιαστικά το ίδιο σύνολο δεδομένων, αυξάνοντας τα ποσοστά ακρίβειας, ανάκλησης και βαθμολογίας F1, κατά 40%, 50% και 44.78%, αντίστοιχα.

Μελέτη	Συνεισφορά
Bollen (2008)[2]	Μελέτη ορόσημο, μετέπειτα μελέτες στηρίχθηκαν στα αποτελέσματα της
Zhang, X., Fuehres, H., και Gloor (2011)[36]	Ασχολήθηκε με την πρόβλεψη των δεικτών Dow Jones (DJIA), NASDAQ και Standard & Poor's (S&P 500) και VIX
Rao και Srivastava (2012)[26]	Ποσοστό συσχέτισης της τάξης του 88%, μεταξύ του Twitter Sentiment και της κίνησης μετοχών
Mittal και Goel (2012)[23]	Ποσοστό ακρίβειας στις προβλέψεις της τάξης του 75.56%
Ding κ.α (2013)[8]	Η προσθήκη ανάλυσης συναισθήματος σε μοντέλο ανάλυσης χρονοσειρών βελτιώνει την απόδοση του
Caliñgo (2016)[6]	Ασχολήθηκε σε τοπικό επίπεδο με τη πρόβλεψη του κύριου δείκτη στο Φιλιππινέζικο Χρηματιστήριο
Renato Bruni (2017)[5]	Μεθοδολογία, επίλυσης προβλήματος δυαδικής ταξινόμησης για ημερήσιες συναλλαγές στο Χρηματιστήριο
Yang Li (2020)[19]	Απόδειξη ότι οι τεχνολογίες βαθιάς μάθησης αποδίδουν καλύτερα στις προβλέψεις από παραδοσιακές μεθοδολογίες

Πίνακας 2.1: Σύνοψη παρόμοιων μελετών και της συνεισφοράς τους στο πεδίο ενδιαφέροντος

Οι τεχνικές που προτείνονται από τις δημοσιεύσεις που αναφέρθηκαν, παρέχουν μια ενδιαφέρουσα επισκόπηση στη θεματολογία της ανάλυσης συναισθήματος και στο πώς αυτή μπορεί να σχετίζεται με το Χρηματιστήριο. Ωστόσο, τα αποτελέσματα διαφέρουν και ποικίλουν αναλόγως της ακρίβειας του αναλυτή συναισθημάτων, της προ-επεξεργασίας και των φιλτραρισμάτων του κειμένου από τα κοινωνικά δίκτυα, καθώς επίσης επηρεάζονται και από τους αλγορίθμους που αξιοποιήθηκαν, για εύρεση συσχέτισης και για ανάπτυξη των μοντέλων πρόβλεψης. Επιπρόσθετα, δεν υπάρχει καταγεγραμμένη εκτενής αναφορά στη χρηματιστηριακή αγορά και στις λέξεις κλειδιά που χρησιμοποιήθηκαν για συλλογή δεδομένων στις μελέτες, στοιχεία

που διαδραματίζουν ισχυρό ρόλο στον προσδιορισμό της αποτελεσματικότητας των συστημάτων πρόβλεψης.

Στην εργασία αυτή, βασιζόμενοι σε προηγούμενες μελέτες, προσδοκούμε στην ανάπτυξη ενός παρόμοιου συστήματος, αξιοποιώντας κοινές, αλλά και διαφορετικές μεθοδολογίες και εργαλεία, για την επίτευξη του επιθυμητού σκοπού.

Το υψηλό ποσοστό ορθότητας(accuracy) που επιτεύχθηκε από τα περισσότερα μοντέλα πρόβλεψης, καταδεικνύει ότι μπορούν να διενεργηθούν προβλέψεις για τη κίνηση των Δεικτών του Χρηματιστηρίου από την ανάλυση συναισθήματος αναρτήσεων στα μέσα κοινωνικής δικτύωσης.

Κεφάλαιο 3

Θεωρητικό Υπόβαθρο

3.1.	Θεωρητικό Υπόβαθρο Twitter.....	13
3.1.1.	Tweet/ ReTweet/Mention(@)	13
3.1.2.	Hashtag (#).....	14
3.2.	Θεωρητικό Υπόβαθρο Reddit/ Reddit WorldNews Channel	14
3.3.	MongoDB.....	15
3.4.	Γλώσσα προγραμματισμού Python	16
3.4.1.	NLTK.....	18
3.4.2.	VADER	18
3.5.	Εισαγωγή στο Χρηματιστήριο	19
3.5.1.	Μετοχή	20
3.5.2.	Πρόβλεψη της Χρηματιστηριακής Αγοράς.....	20
3.5.3.	The New York Stock Exchange (NYSE)	20
3.5.4.	NASQAD	20
3.5.5.	Dow Jones Industrial Average.....	21

3.1. Θεωρητικό Υπόβαθρό Twitter

Η διαδικτυακή υπηρεσία Twitter, αποτελεί ένα από τα δημοφιλέστερα μέσα κοινωνικής δικτύωσης στις μέρες μας, με περισσότερους από 300 εκατομμύρια ενεργούς χρήστες. Το Twitter παρέχει τη δυνατότητα στους χρήστες να δημοσιεύουν ανά πάσα στιγμή στους λογαριασμούς τους, σύντομα κείμενα με έκταση μέχρι 280 χαρακτήρες, τα επονομαζόμενα, Tweets[32].

Αξίζει να σημειωθεί, ότι είναι αδύνατη η δημοσίευση Tweet από χρήστες χωρίς λογαριασμό. Παρόλα αυτά, δίδεται ελεύθερη πρόσβαση σε χρήστες χωρίς λογαριασμό, να διαβάσουν όλες τις αναρτήσεις. Η ικανότητα του Twitter να ενημερώνει, να ψυχαγωγεί και να διαδίδει πληροφορίες, είναι αδιαμφισβήτητη.

Καθημερινά δημοσιεύονται στο Twitter πάνω από 300 εκατομμύρια Tweets αναρίθμητων θεματικών και ποικίλων περιεχομένων. Η υπηρεσία του Twitter έχει ισχυροποιηθεί τόσο πολύ τα τελευταία χρόνια που η πλειοψηφία των μεγαλύτερων επιχειρήσεων και διασημότερων προσωπικοτήτων στον κόσμο κατέχουν ενεργούς λογαριασμούς σε αυτήν. Ως λογικό επακόλουθο αυτού, οι αναρτήσεις στο δίκτυο έχουν τη δύναμη να επιφέρουν πολιτικές, κοινωνικές και οικονομικές μεταβολές, καθώς επίσης και διαταράξεις παγκόσμιας εμβέλειας.

3.1.1. Tweet/ ReTweet/Mention(@)

Όπως αναφέρθηκε πιο πάνω, το Tweet είναι ένα σύντομο μήνυμα, έως 280 χαρακτήρων, που δημοσιεύεται στη διαδικτυακή υπηρεσία Twitter. Το μήνυμα αυτό μπορεί να αποτελείται από απλό κείμενο, συνδέσμους άλλων ιστοσελίδων, hashtags, mentions και οπτικοακουστικό υλικό.

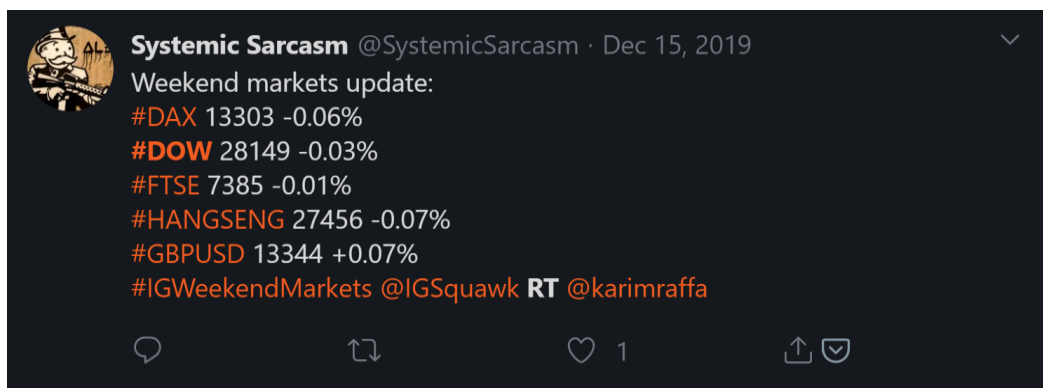
Ένα retweet, είναι η αναδημοσίευση ενός Tweet, είτε από τον χρήστη που το δημοσίευσε αρχικά, είτε από οποιονδήποτε άλλον επιθυμεί και αναγνωρίζεται από την ένδειξη RT ή Retweet, στην αρχή του κειμένου.

Συνήθως, το κείμενο ενός retweet, περιλαμβάνει και τον χαρακτήρα «@» και ακολούθως το όνομα του χρήστη που το δημοσίευσε αρχικά. Το σύμβολο «@» που ακολουθείται πάντοτε από όνομα χρήστη, ονομάζεται mention και χρησιμοποιείται

για να υποδηλώσει πως γίνεται αναφορά σε έναν συγκεκριμένο χρήστη, ενώ αποτελεί και σύνδεσμο, μέσω του οποίου γίνεται μεταφορά στο προφίλ του αναφερόμενου χρήστη.

3.1.2. Hashtag (#)

Ένα από τα πλέον δημοφιλέστερα σύμβολα στα περισσότερα μέσα κοινωνικής δικτύωσης, είναι το hashtag(#). Το hashtag χρησιμοποιείται ως μέσο για διευκόλυνση της αναζήτησης στην θεματολογία ενδιαφέροντος. Η παρουσία του συμβόλου της δίεσης, στη αρχή ή στο τέλος μιας ακολουθίας χαρακτήρων χωρίς κενά, υποδηλώνει ότι κάθε Tweet που περιέχει το ίδιο αναγνωριστικό, κατηγοριοποιείται και σε περίπτωση αναζήτησης του, θα εμφανιστεί στον χρήστη. Οι χρήστες έχουν τη δυνατότητα να δημιουργήσουν νέο hashtag, καθώς επίσης πατώντας ένα hashtag να τους παρουσιαστούν και πάλι όλα τα Tweets που το περιέχουν.



Σχήμα 3.1: Παράδειγμα ανάρτησης (Tweet) στο μέσο κοινωνικής δικτύωσης Twitter

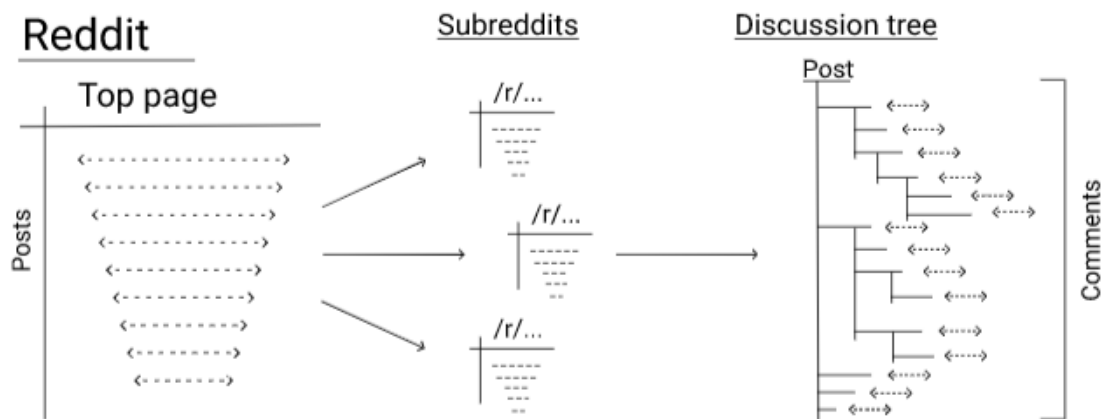
3.2. Θεωρητικό Υπόβαθρό Reddit/ Reddit WorldNews Channel

Το Reddit είναι μια πλατφόρμα συλλογής κοινωνικών ειδήσεων, βαθμολογίας περιεχομένου και ιστοχώρος συζητήσεων που δημιουργήθηκε το 2005. Κατατάσσεται ως η 12^η ιστοσελίδα με τις περισσότερες επισκέψεις στον κόσμο και 7^η στις Η.Π.Α για το 2020, με κατά μέσο όρο 278 εκατομμύρια επισκέπτες τον μήνα και 40 εκατομμύρια αναζητήσεις ανά μέρα[3].

Σημείο εισόδου της πλατφόρμας αποτελεί η αρχική σελίδα του Reddit, η οποία τροφοδοτείται με αναρτήσεις από δευτερεύουσες μονάδες που αποκαλούνται

Subreddits. Τα Subreddits ακολουθούνται από χρήστες και οι αναρτήσεις τους κατατάσσονται σύμφωνα με τις ψήφους και την ηλικία των δημοσιεύσεων. Οι χρήστες είναι σε θέση να μεταβούν στην αρχική σελίδα ενός Subreddit, όπου περιορίζεται η θεματολογία των αναρτήσεων. Κάθε δημοσίευση μπορεί να αξιολογηθεί με θετικούς ή αρνητικούς ψήφους και έχει συνημμένη ενότητα σχολίων. Τα σχόλια είναι δομημένα ως ένα ριζωμένο δέντρο και μπορεί να αποτελούν απάντηση ενός άλλου σχολίου ή ένα ανεξάρτητο σχόλιο[22]. Πιο κάτω απεικονίζεται η σχηματική μορφή της πλατφόρμας Reddit (βλ. Σχήμα 3.2.).

Το Reddit WorldNews Channel(r/worldnews) αποτελεί κοινότητα(subreddit) στο Reddit, όπου αναρτώνται σημαντικές ειδήσεις από τους χρήστες καθημερινά. Η κοινότητα αποτελείται από 25.4 εκατομμύρια μέλη που κατέχουν τη δυνατότητα να αξιολογήσουν τις ειδήσεις, με σκοπό να παρουσιάζονται πρώτες, εκείνες με τις περισσότερους θετικές ψήφους.



Σχήμα 3.2: Σχηματική Δομή της Πλατφόρμας Reddit (Πηγή: “The anatomy of Reddit: An overview of academic research” [22])

3.3. MongoDB

Η MongoDB, είναι μια μη-σχεσιακή βάση δεδομένων (NoSQL database) που χρησιμοποιείται συνήθως για αποθήκευση δεδομένων μεγάλου όγκου. Χρησιμοποιεί κυρίως αρχεία τύπου json και τα δεδομένα αποθηκεύονται ευέλικτα, με πολλαπλές δομές και διαφορετικούς τρόπους. Επομένως, για την αποθήκευση των αναρτήσεων που ανακτήθηκαν στη εργασία, σε αρχικό στάδιο, αξιοποιήθηκε η MongoDB, η οποία διαχειρίζεται αρκετά καλά τα αντικείμενα που επιστρέφονται από το Twitter API.

Κάθε Tweet αποθηκεύεται ως εγγραφή σε μια συλλογή στη βάση δεδομένων. Ωστόσο, δεν αποθηκεύονται όλα τα δεδομένα, πάρα μόνο οι σχετικές πληροφορίες, όπως το κείμενο και η ημερομηνία ανάρτησης του Tweet. Ακόμη ένας λόγος που προτιμήθηκε η βάση δεδομένων MongoDB έναντι άλλων, αποτελεί η απλότητα και αποδοτικότητα στη δημιουργία ερωτημάτων εντός της βάσης, όταν κρίνεται απαραίτητη η ανίχνευση δεδομένων στις συλλογές που εμπεριέχουν Tweets.

3.4. Γλώσσα προγραμματισμού Python

Η γλώσσα προγραμματισμού Python, αποτελεί μια από τις δημοφιλέστερες αντικειμενοστραφής, υψηλού-επιπέδου γλώσσες προγραμματισμού. Εμπεριέχει ενσωματωμένες δομές δεδομένων και χρησιμοποιείται κυρίως για επεξεργασία και ανάλυση μεγάλου όγκου δεδομένων.

Η Python αξιοποιείται ευρέως στην επιστήμη των δεδομένων, καθώς αρκετές από τις βιβλιοθήκες της, προσφέρουν πληθώρα δυνατοτήτων και ευκολιών για τους αναλυτές δεδομένων.

Η “**Numerical Python**”(NumPy)[10] αποτελεί μια από τις πλέον ισχυρότερες και δημοφιλέστερες βιβλιοθήκες της Python. Πρόκειται για μια αριθμητική βιβλιοθήκη ανοιχτού κώδικα, η οποία παρέχει τη δυνατότητα γρήγορης υλοποίησης μαθηματικών υπολογισμών σε πολυδιάστατους πίνακες και δομές δεδομένων. Χρησιμοποιείται ευρέως για εκτέλεση μαθηματικών λειτουργιών, καθώς περιέχει μεγάλο αριθμό μαθηματικών, αλγεβρικών και μετασχηματιστικών συναρτήσεων.

Το **Pandas**[21], είναι ένα εργαλείο χειρισμού δεδομένων υψηλού επιπέδου. Τα αντικείμενα Pandas βασίζονται σε μεγάλο βαθμό σε αντικείμενα NumPy, αφού ουσιαστικά, το Pandas επεκτείνει το NumPy. Η βασική του δομή δεδομένων ονομάζεται DataFrame. Τα DataFrames επιτρέπουν την αποθήκευση και το χειρισμό πινάκων δεδομένων σε σειρές παρατηρήσεων και στήλες μεταβλητών.

Οι παραπάνω βιβλιοθήκες διευκολύνουν και καθιστούν εφικτή την αποδοτική επεξεργασία μεγάλου όγκου δεδομένων, προσφέροντας τη δυνατότητα παράλληλης επεξεργασίας δεδομένων. Είναι για αυτόν το λόγο που αξιοποιήθηκαν στην παρούσα

εργασία, εφόσον ασχολείται με μεγάλο αριθμό Tweets και τίτλων ειδήσεων, τα οποία συλλέχθηκαν, αποθηκεύτηκαν σε βάση δεδομένων ή σε αρχεία τύπου .csv και έτυχαν επεξεργασίας στη συνέχεια κατά την προ επεξεργασία κειμένου, την ανάλυση συναισθήματος και ως επίσης η εξαγόμενη πληροφορία από την ανάλυση επεξεργάστηκε εκ νέου και εκτεταμένα για εξαγωγή γνώσης.

Σε ότι αφορά το στάδιο της προ-επεξεργασίας κειμένου, η Python προσφέρει εξαιρετικές βιβλιοθήκες για μορφοποίηση του περιεχομένου των κειμένων από τα μέσα κοινωνικής δικτύωσης. Αξιοποιήθηκαν βιβλιοθήκες του εργαλείου NLTK (βλ. 3.4.1.), καθώς και η βιβλιοθήκη **re** (Regular Expression)[33]. Συγκεκριμένα, η βιβλιοθήκη **re** χρησιμοποιεί κανονικές εκφράσεις, δηλαδή ειδικές ακολουθίες χαρακτήρων για την εξεύρεση άλλων συμβολοσειρών ή συνόλων συμβολοσειρών, χρησιμοποιώντας μια εξειδικευμένη σύνταξη που εκφράζεται σαν μοτίβο. Το εργαλείο αυτό αξιοποιήθηκε κυρίως για εξαίρεση ανεπιθύμητου κειμένου από το περιεχόμενο των Tweets.

Επιπλέον, η Python προσφέρει βιβλιοθήκες που βοηθούν στην εξαγωγή δεδομένων από το διαδίκτυο. Πιο συγκεκριμένα, η βιβλιοθήκη **Tweepy**[27] παρέχει πρόσβαση στις μεθόδους του Twitter Streaming API. Η κάθε μέθοδος λαμβάνει ορισμένες παραμέτρους και επιστρέφει σχετική απάντηση.

Η απεικόνιση των δεδομένων καθίσταται απαραίτητη για την οπτικογράφηση τους, έτσι ώστε να εξαχθεί πληροφορία από τα συμπεράσματα που θα προκύψουν από την παρατήρηση. Γίνεται λοιπόν αντιληπτό ότι είναι απαραίτητη η αξιοποίηση βιβλιοθηκών που παρέχουν την δυνατότητα κατασκευής γραφικών παραστάσεων, ιστογραμμάτων και άλλων μέσων απεικόνισης δεδομένων. Κάποιες από τις δημοφιλέστερες επιλογές που αξιοποιήθηκαν αποτελούν οι βιβλιοθήκες **Matplotlib**[12], **Wordcloud** και **Seaborn**[34].

3.4.1. NLTK

Το NLTK (Natural Language Toolkit), είναι ένα πακέτο της Python το οποίο υποστηρίζει αλγορίθμους και λειτουργίες για επεξεργασία φυσικής γλώσσας, όπως ανάλυση συναισθήματος, κατηγοριοποίηση θέματος και διάσπαση κειμένου[1].

Το συγκεκριμένο πακέτο επιλέχθηκε για την προ-επεξεργασία κειμένου, ως επίσης και για την ανάλυση συναισθήματος, μέσω του εργαλείου VADER (βλ. 3.4.2. VADER).

Μερικές από τις λειτουργίες της NLTK που αξιοποιήθηκαν στην παρούσα εργασία, αποτελούν οι βιβλιοθήκες με stopwords και punctuation και η λειτουργία του word tokenization.

3.4.2. VADER

Το VADER (Valence Aware Dictionary and Sentiment Reasoner), είναι εργαλείο ανοιχτού κώδικα (MIT License), το οποίο είναι διαθέσιμο, μέσω του εργαλείου NLTK[13]. Συνδυάζει λεξικό και επεξεργασία των χαρακτηριστικών των προτάσεων, μέσω κανόνων, για να καθορίσει το αρνητικό, ουδέτερο ή θετικό τύπο της πολικότητας των συναισθημάτων τους. Το Vader είναι ειδικά διαμορφωμένο και κατάλληλο για ανάλυση συναισθήματος σε κείμενα των μέσων κοινωνικής δικτύωσης. Αυτό συμβαίνει επειδή το Vader δεν προσμετρά μόνο τη βαθμολογία θετικότητας και αρνητικότητας, αλλά επίσης υπολογίζει το πόσο θετικό ή αρνητικό είναι ένα συναίσθημα. Επομένως, σε αντίθεση με άλλα εργαλεία, το Vader παρέχει τη δυνατότητα εκτίμησης ποσοστού του συναισθήματος του κειμένου.

Το Vader αποτελεί πολύ χρήσιμο εργαλείο για ανάλυση κειμένου από το Twitter, καθώς η ανάλυση συναισθήματος που πραγματοποιεί, βασίζεται κυρίως σε συγκεκριμένα στοιχεία που εμφανίζονται συχνά σε αναρτήσεις του εν λόγω κοινωνικού μέσου.

Ένα από τα στοιχεία αυτά είναι τα σημεία στίξης. Για παράδειγμα, η χρήση του θαυμαστικού(!), αυξάνει το μέγεθος της έντασης χωρίς να τροποποιεί τον σημασιολογικό προσανατολισμό. Επομένως, το Tweet "Buy \$SPX tomorrow!" είναι

πιο έντονο από το " Buy \$SPX tomorrow." και μια αύξηση στον αριθμό των θαυμαστικών (!), αυξάνει ανάλογα την βαθμολογία του αναλυτή συναισθήματος.

Επίσης ο αναλυτής συναισθήματος Vader, αποδίδει πολύ καλά όταν αξιολογεί emojis, emoticons, γλώσσα slang και ακρώνυμα (βλ. Πίνακας 3.1).

Παραδείγματα	pos	neg	neu	compound
😊	0.67	0.0	0.33	0.72
😞	0.0	0.71	0.29	-0.34
This stock is a very good deal!!	0.43	0.0	0.57	0.58
This stock is a very good deal!! :-)	0.56	0.0	0.44	0.74
LOL !!!	0.82	0.0	0.19	0.66

Πίνακας 3.1: Πίνακας απεικόνισης παραδειγμάτων αξιολόγησης του συναισθήματος, σε emojis, emoticons και γλώσσα slang από τον αναλυτή συναισθήματος, Vader.

Πλεονεκτήματα του Vader βοηθητικά για την φύση της εργασίας

- ✓ Το εργαλείο Vader λειτουργεί εξαιρετικά καλά σε κείμενο τύπου αναρτήσεων κοινωνικών μέσων.
- ✓ Δεν απαιτεί δεδομένα εκπαίδευσης, αλλά είναι κατασκευασμένο από ένα γενικευμένο, gold standard sentiment lexicon.

Παρόλα τα θετικά στοιχεία του Vader αξιοποιήθηκε και η βιβλιοθήκη ανάλυσης κειμένου **Textblob** της Python για την αξιοποίηση του δείκτη υποκειμενικότητας (subjectivity metric) που προσφέρει, ο οποίος κρίθηκε ότι αποτελεί μια ενδιαφέρουσα προς εξέταση μετρική.

3.5. Εισαγωγή στο Χρηματιστήριο

Ως πηγή των χρηματιστηριακών δεδομένων της μελέτης, χρησιμοποιήθηκε το Χρηματιστήριο των Η.Π.Α, κυρίως για το λόγο ότι σε αυτό ανήκουν τα δύο μεγαλύτερα και δημοφιλέστερα, με διαφορά Χρηματιστήρια στον κόσμο, NYSE και NASDAQ, στα οποία δραστηριοποιούνται οι δημοφιλέστερες και ισχυρότερες εταιρείες και δείκτες μετοχών, με τεράστιο όγκο συναλλαγών καθημερινά[16]. Επιπρόσθετα, για τους

αμερικανικούς δείκτες υπάρχει πληθώρα πληροφοριών στο διαδίκτυο, αναρτήσεις στο κοινωνικό δίκτυο Twitter και μεγάλος αριθμός ειδήσεων καθημερινά.

Πιο κάτω, δίνονται συνοπτικά χρήσιμες πληροφορίες για τα Χρηματιστήρια και τον δείκτη Dow Jones, για τον οποίο χρησιμοποιήθηκε σύνολο δεδομένων.

3.5.1. Μετοχή

Η κυριότητα μιας επιχείρησης υποδιαιρείται, συχνά, σε εκατομμύρια τεμάχια, το καθένα από τα οποία ονομάζεται μετοχή. Όταν κάποιος αγοράζει μετοχές, γίνεται αυτομάτως συνιδιοκτήτης της εταιρείας στην οποία ανήκουν οι τίτλοι που αγόρασε. Από τη στιγμή που «δημιουργηθούν» οι μετοχές, αγοράζονται και πωλούνται μεταξύ των επενδυτών στη «Δευτερογενή αγορά», δηλαδή στο Χρηματιστήριο[11].

3.5.2. Πρόβλεψη της Χρηματιστηριακής Αγοράς

Πρόβλεψη της Χρηματιστηριακής Αγοράς ορίζεται η προσπάθεια της εξεύρεσης των μελλοντικών τιμών της μετοχής μιας εταιρείας ή ενός Χρηματιστηριακού δείκτη. Οι επιτυχημένες προβλέψεις των μελλοντικών τιμών μετοχών, μπορούν να επιφέρουν οικονομικό κέρδος στους επενδυτές.

3.5.3. The New York Stock Exchange (NYSE)



Το Χρηματιστήριο της Νέας Υόρκης αποτελεί, μακράν, το μεγαλύτερο Χρηματιστήριο στον κόσμο, με βάση την κεφαλαιοποίηση (30.1 τρισεκατομμύρια). Βρίσκεται στην 11 Wall Street, Lower Manhattan, New York City, New York.

3.5.4. NASQAD



Το Χρηματιστήριο του NASQAD αποτελεί το δεύτερο μεγαλύτερο Χρηματιστήριο στον κόσμο, με βάση την κεφαλαιοποίηση και είναι το πρώτο σε ότι αφορά τους όγκους συναλλαγών, συγκριτικά με όλα τα υπόλοιπα ηλεκτρονικά Χρηματιστήρια στον κόσμο[24].

3.5.5. Dow Jones Industrial Average

Ένας από τους πλέον αναγνωρισμένους δείκτες της αγοράς στον κόσμο, ο οποίος συγκαταλέγεται στους δείκτες που αναφέρουν οι επιχειρηματικές ειδήσεις, είναι ο δείκτης Dow Jones Industrial Average. Ένας εμβληματικός δείκτης μετοχών στις ΗΠΑ, στον οποίο συμπεριλαμβάνονται σημαντικές παγκόσμιες επωνυμίες, όπως, η Apple, Nike, Microsoft, Disney, Coca-Cola, American Express κλπ.

Ο υπολογισμός του δείκτη βασίζεται στη στάθμιση κατά τιμή μετοχής, δηλαδή οι μετοχές με υψηλότερη τιμή έχουν μεγαλύτερη επίδραση στην απόδοση του δείκτη σε σχέση με τις μετοχές χαμηλότερης αξίας[24].

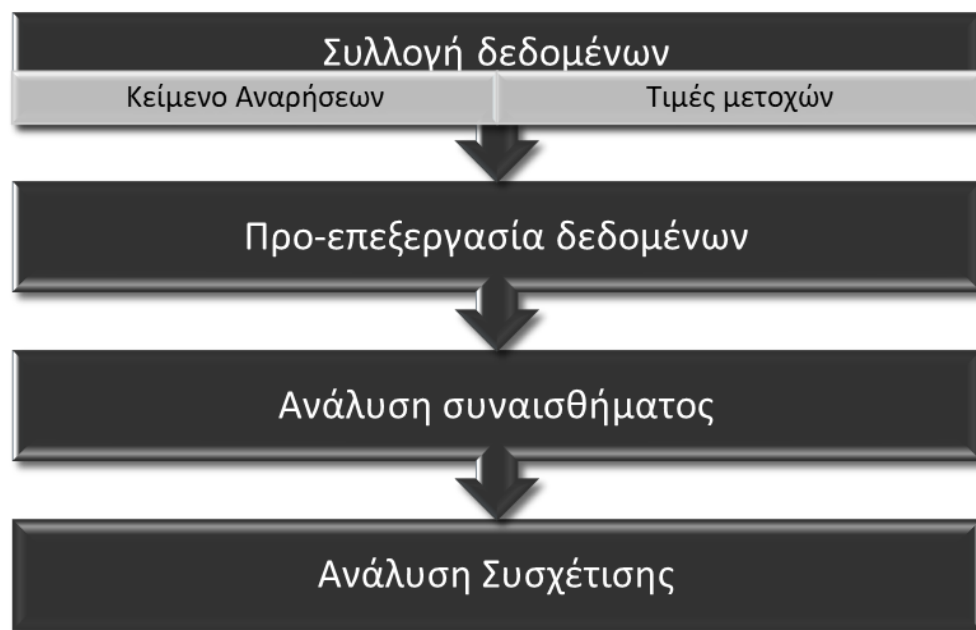
Κεφάλαιο 4

Περιγραφή Μεθοδολογίας

4.1.	Μεθοδολογία Αξιοποίησης Αναρτήσεων από το Twitter	23
4.1.1.	Συλλογή Δεδομένων από το Twitter	23
4.1.2.	Προ-επεξεργασία Κειμένου από το Twitter	28
4.1.3.	Ανάλυση Συναισθήματος	30
4.1.3.1.	Κατηγοριοποίηση αναρτήσεων με βάση το συναίσθημα	32
4.1.3.2.	Δημοφιλή Hashtags	32
4.1.3.3.	Συχνότητα εμφάνισης λέξεων	33
4.1.3.4.	Υποκειμενικότητα και Πολικότητα	34
4.2.	Μεθοδολογία Ανάπτυξης Μοντέλου Πρόβλεψης	34
4.2.1.	Περιγραφή Συνόλου Δεδομένων από το Reddit	34
4.2.2.	Προ-επεξεργασία συνόλου δεδομένων	36
4.2.3.	Επεξεργασία Συνόλου Δεδομένων	37
4.2.4.	Συγχώνευση Συνόλων Δεδομένων	38
4.2.5.	Μετατροπή Συνόλου Δεδομένων σε Πρόβλημα Μηχανικής Μάθησης ...	39
4.2.6.	Μηχανική Μάθηση	40
4.2.7.	Βελτιστοποίηση αλγορίθμων μηχανικής μάθησης	41
4.3.	Συλλογή Δεδομένων για τιμές της Χρηματιστηριακής Αγοράς	44

4.1. Μεθοδολογία Αξιοποίησης Αναρτήσεων από το Twitter

Στο παρόν υποκεφάλαιο, περιγράφονται αναλυτικά τα στάδια για την ανάπτυξη συστήματος, το οποίο συλλέγει αναρτήσεις από το κοινωνικό δίκτυο Twitter, σχετικές με μετοχές εταιρειών του πεδίου ενδιαφέροντος και τις αξιοποιεί έπειτα από την ανάλυση συναισθήματος, για εξαγωγή γνώσης που να συσχετίζεται με την κίνηση των εν λόγω μετοχών στο Χρηματιστήριο.



Σχήμα 4.1: Σύνοψη Μεθοδολογίας Αξιοποίησης Αναρτήσεων από το Twitter

4.1.1. Συλλογή Δεδομένων από το Twitter

Για τη διεξαγωγή της ανάλυσης συναισθήματος καθίσταται απαραίτητη η συλλογή αναρτήσεων από το μέσο κοινωνικής δικτύωσης Twitter. Ως εκ τούτου, αρχικό μέλημα αποτελεί η συλλογή ενός συνόλου Tweets που θα χρησιμεύσουν ως δεδομένα εισόδου για τον αναλυτή συναισθήματος.

Όπως είναι γνωστό στο Twitter δημοσιεύονται καθημερινά εκατομμύρια Tweets και για το λόγο αυτό οι μέθοδοι εξαγωγής δεδομένων πρέπει να είναι όσον το δυνατόν πιο συγκεκριμένοι, ειδικά προσαρμοσμένοι και εύστοχοι, έτσι ώστε να εξασφαλιστεί το υποσύνολο των αναρτήσεων που εμπίπτουν στο πεδίο ενδιαφέροντος του μοντέλου.

Η πλατφόρμα του κοινωνικού μέσου Twitter παρέχει στους προγραμματιστές ένα ισχυρό API για συλλογή Tweets. Υπάρχουν δύο βασικά APIs για συλλογή δεδομένων, τα οποία ονομάζονται Twitter Streaming API και Twitter Search API.

Το Twitter Streaming API επιτρέπει τη πρόσβαση και συλλογή αναρτήσεων σε πραγματικό χρόνο μέσω ερωτημάτων εισαγωγής. Για μια επιτυχημένη πρόσβαση στις αναρτήσεις απαιτείται από τον χρήστη να ζητήσει σύνδεση, σε μια ροή (stream), από τον διακομιστή. Στη συνέχεια, ο διακομιστής ενεργοποιεί μια σύνδεση ροής και οι αναρτήσεις μεταδίδονται σε ροή και εμφανίζονται στον χρήστη. Στην παρούσα διπλωματική εργασία αξιοποιήθηκε αρχικά το Twitter Streaming API, στο δωρεάν του επίπεδο, το οποίο παρουσιάζει περιορισμούς (βλ. 8.1). Είναι για αυτό το λόγο που ακολουθήθηκαν δύο επιπλέον μεθοδολογίες για συλλογή Tweets, έτσι ώστε να υπάρχει η δυνατότητα σύγκρισης και λήψης εγκυρότερων συμπερασμάτων κατά τα στάδια της ανάλυσης αποτελεσμάτων και εξαγωγής πληροφοριών. Συνολικά χρησιμοποιήθηκαν 3 μεθοδολογίες για τη συλλογή δεδομένων, μια που χρησιμοποιεί το Twitter Streaming API και 2 υποστηρίχτηκες που χρησιμοποιούν το Twitter Search API για τη διαμόρφωση μιας πιο ολοκληρωμένης και πολύπλευρης έρευνας για συλλογή Tweets στην υπό μελέτη θεματολογία. Οι δύο τελευταίες μεθοδολογίες αξιοποιήθηκαν για τη συλλογή αναρτήσεων για τις εταιρείες στη λίστα Covid-19 (βλ. Κεφάλαιο 5), και μεταξύ των δύο, η δεύτερη προτιμήθηκε και αξιοποιήθηκε για ανάλυση συναισθήματος και εξαγωγή αποτελεσμάτων.

ΜΕΘΟΔΟΣ Α

Με την αξιοποίηση του Twitter Streaming API συλλέχθηκαν συνολικά 351686 Tweets για τον Χρηματιστηριακό δείκτη Dow Jones και 132,106 για τον Χρηματιστηριακό δείκτη S&P 500, από τις 01/08/2020, μέχρι τις 04/09/2020 (τα οποία αφορούν τους Χρηματιστηριακούς Δείκτες Dow Jones και Standard and Poor). Τα Tweets αποθηκεύτηκαν σε βάση δεδομένων στη MongoDB, σε δύο διαφορετικά Collections, αναλόγως του δείκτη στον οποίον αναφέρονταν. Οι βασικοί κανόνες συλλογής που εφαρμόστηκαν στο Streaming API ήταν αρχικά ένα σύνολο από λέξεις κλειδιά για να επιτευχθεί ανάκτηση αναρτήσεων που αναφέρονται στην προς εξέταση θεματολογία

και επιπρόσθετα το φιλτράρισμα για συλλογή αναρτήσεων μόνο στην αγγλική γλώσσα για να καταστεί δυνατή η προ-επεξεργασία τους και ανάλυση τους για την ανίχνευση συναισθήματος.

ΜΕΘΟΔΟΣ Β

Για τη συλλογή Tweets αξιοποιήθηκε το Twitter Search API με το οποίο παρέχεται η δυνατότητα ανάκτησης των τελευταίων 3000 αναρτήσεων ενός χρήστη. Με τη μέθοδο αυτή ανακτήθηκαν τα 3000 τελευταία Tweets γνωστών λογαριασμών στο Twitter που ασχολούνται με το Χρηματιστήριο και στη συνέχεια, τα αποτελέσματα έτυχαν επεξεργασίας για να κρατηθούν μόνο τα Tweets που αναφέρονται σε εταιρείες συγκεκριμένου ενδιαφέροντος (Covid19). Ανάμεσα στους λογαριασμούς που αξιοποιήθηκαν, εμπεριέχονται ειδησεογραφικές ιστοσελίδες, τηλεοπτικά κανάλια, διάσημες προσωπικότητες στο τομέα των επενδύσεων και γνωστοί λογαριασμοί που δραστηριοποιούνται μόνο στο Twitter και ασχολούνται με ειδήσεις και επενδυτικά ζητήματα. Παρακάτω δίδεται η λίστα των λογαριασμών που αξιοποιήθηκαν και σύντομη περιγραφή για την ιδιότητα τους. Παρόλο που συλλέχθηκαν 3000 αναρτήσεις για 40 λογαριασμούς, δεν εμπεριείχαν όλοι χρήσιμα Tweets για τη δημιουργία του συνόλου δεδομένων και ως εκ τούτου αξιοποιήθηκαν μόνο οι ακόλουθοι (βλ. Πίνακας 4.1).

Λογαριασμός	Ιδιότητα
@Benzinga	Financial news
@jimcramer	Television Personality/Investor
@ianbremmer	Political Scientist
@TheStalwart(Joe Weisenthal)	Executive Editor of news for Bloomberg
@CNBC	Business News Television Channel
@stocktwits	Community of Investors and Traders
@Newsweek	Weekly News Magazine
@washingtonpost	Daily Newspaper
@BreakoutStocks	Stock Performance Statistics Webpage
@bespokeinvest	Financial Markets Research
@WSJmarkets	Breaking news and Financial Markets from The Wall Street Journal.
@Stephanie_Link	Chief Investment Officer and Portfolio Manager at Hightower Advisors
@nytimesbusiness	Financial, tech, media and other business news from The New York Times.
@IBDinvestors	Investor's Business Daily provides leading stock market news and analysis from investors.com
@NorthmanTrader(Sven Henrich)	Market Professional Expert
@Carl_C_Icahn	Businessman and Philanthropist
@Wu_Tang_Finance	Public Figure and Financial Humorists
@BuzzFeed	Internet media, news and entertainment company
@muddywatersre	Freely published investment research
@Schuldensuehner	Market Professional Expert

Πίνακας 4.1: Πίνακας με τους λογαριασμούς που χρησιμοποιήθηκαν για συλλογή Tweets και περιγραφή της ιδιότητας των προσωπικοτήτων ή εταιρειών στις οποίες ανήκουν

Στη συνέχεια τα αποτελέσματα της συλλογής έτυχαν επεξεργασίας, μέσω του λογισμικού λογιστικών φύλλων Microsoft Excel, για να κρατηθούν μόνο τα Tweets που αναφέρονται σε εταιρείες από τη λίστα και να διαμορφωθούν αρχεία τύπου .csv σε μορφή, τέτοια ώστε να επιτρέπουν αποδοτικότερη επεξεργασία από την Python, στα επόμενα βήματα. Τα δεδομένα που συλλέχθηκαν από όλους τους διαφορετικούς λογαριασμούς συνενώθηκαν σε 11 ενιαία αρχεία τύπου csv, όπου το καθένα από αυτά αποτελείται από εγγραφές που αφορούν εταιρείες συγκεκριμένου τομέα της οικονομίας. Οι τομείς της οικονομίας και οι εταιρείες που επιλέχθηκαν παρουσιάζονται και περιγράφονται αναλυτικά στο Κεφάλαιο 5.

ΜΕΘΟΔΟΣ Γ

Το Twitter Search API παρέχει τη δυνατότητα ανάκτησης δεδομένων χρησιμοποιώντας ερωτήματα που αφορούν λέξεις κλειδιά ή hashtags, με τον περιορισμό, η συλλογή να πραγματοποιείται μέχρι 10 ημέρες στο παρελθόν.

Έχοντας υλοποιήσει τη Μέθοδο Β (βλ. Μέθοδος Β), παρουσιάστηκε το ζήτημα ότι λόγω των εξειδικευμένων εταιρειών για τις οποίες συλλέχθηκαν Tweets και δεδομένου ότι οι λογαριασμοί στο Twitter δεν επικεντρώνουν τις αναρτήσεις τους σε συγκεκριμένους δείκτες ή μετοχές, τα τελικά σύνολα δεδομένων που προέκυψε δεν περιείχαν μεγάλη ποσότητα δεδομένων. Ως εκ τούτου, για την επέκταση του συνόλου δεδομένων, αξιοποιήθηκε η δυνατότητα συλλογής δεδομένων για τις πρώτες 10 μέρες του μήνα Σεπτεμβρίου. Η αναζήτηση για τη συλλογή αναρτήσεων της λίστα του COVID-19, πραγματοποιήθηκε μέσω της δημιουργίας ερωτημάτων με τα αναγνωριστικά σύμβολα των εταιρειών για αναζήτηση με λέξεις κλειδιά. Σημειώνεται ότι, τα σύνολα δεδομένων που προέκυψαν από τη Μέθοδο Γ αποθηκεύτηκαν σε αρχεία τύπου csv για εύκολη τοπική ανάγνωση, επεξεργασία και διαχείριση.

Για τις τρεις πιο πάνω αναφερόμενες μεθόδους συλλογής, τα πεδία που συνθέτουν τα σύνολα δεδομένων μετά την συλλογή και την κατηγοριοποίηση τους με βάση την θεματολογία τους, είναι τα εξής:

- Μοναδικό αναγνωριστικό της ανάρτησης (Tweet_ID)
- Αναγνωριστικό σύμβολο εταιρείας στο Χρηματιστήριο
- Κείμενο ανάρτησης
- Όνομα λογαριασμού
- Ημερομηνία δημοσίευσης ανάρτησης
- Αριθμός χρηστών που αποθήκευσαν την ανάρτηση στην βιβλιοθήκη με τις αγαπημένες τους αναρτήσεις.
- Αριθμός χρηστών που αναδημοσίευσαν την ανάρτηση (αριθμός reTweets).
- Τύπος ανάρτησης (Tweet ή ReTweet).
- Αριθμός hashtags στο κείμενο της ανάρτησης.
- Αριθμός mentions στο κείμενο της ανάρτησης.

- Τύπος οπτικοακουστικών μέσων που περιέχονται στην ανάρτηση, εάν υπάρχουν.

Παρόλα αυτά, για διευκόλυνση κατά την ανάλυση συναισθήματος, η τελική μορφή του συνόλου δεδομένων για τη Μέθοδο Α, αποτελείται από το Tweet_ID, το κείμενο του Tweet και την ημερομηνία δημοσίευσης του, ενώ στις Μεθόδους Β και Γ, για ευνόητους λόγους, προστίθεται και το πεδίο του μοναδικού συμβόλου μετοχής, που διευκρινίζει τη μετοχή στην οποία αναφέρεται το κείμενο του Tweet.

Tweet_ID	Date	Text
----------	------	------

Πίνακας 4.2: Μορφή Συνόλου Δεδομένων Μεθόδου Α

Tweet_ID	Date	Text	Symbol
----------	------	------	--------

Πίνακας 4.3: Μορφή Συνόλου Δεδομένων Μεθόδων Β και Γ

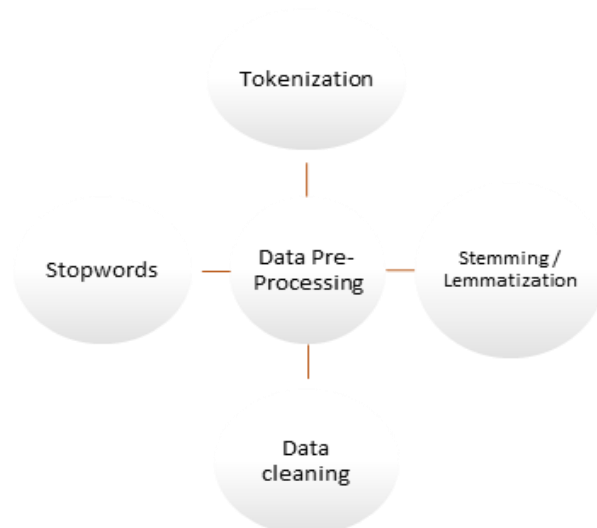
Όπως αναφέρθηκε πιο πάνω για τους σκοπούς της παρούσας εργασίας, στο τελικό σύστημα αξιοποιήθηκε μόνο η Μέθοδος Β, ενώ οι δύο άλλες απορρίφθηκαν, λόγω της μειωμένης ποιότητας των συλλεχθέντων δεδομένων τους σε σχέση με αυτή.

4.1.2. Προ-επεξεργασία Κειμένου από το Twitter

Ως γνωστό, τα δεδομένα που εξάγονται από το Twitter, περιέχουν αχρείαστο κείμενο, το οποίο αποτελεί ανασταλτικό παράγοντα στην ανάλυση συναισθήματος. Εφόσον στην ανάλυση συναισθήματος απαιτείται η απομόνωση του κειμένου, το οποίο εμπεριέχει την επιθυμητή προς ανάλυση πληροφορία, είναι αναγκαίο όπως η επεξεργασία διεξαχθεί απαλλαγμένη από το «θόρυβο» που δημιουργείται από αχρείαστες λέξεις, ώστε η βαθμολογία ανάλυσης συναισθήματος να παρουσιάσει το καλύτερο δυνατό αποτέλεσμα. Αυτό μπορεί να επιτευχθεί κυρίως με τεχνικές οι οποίες προσφέρονται από τη βιβλιοθήκη Natural Language Toolkit (NLTK) της γλώσσας προγραμματισμού Python[18].

Οι τεχνικές που χρησιμοποιήθηκαν για την προεπεξεργασία κειμένου είναι οι εξής:

- Χρήση του tokenization, όπου οι λέξεις διαχωρίζονται με κενά, ώστε να δημιουργείται μια ξεχωριστή λίστα με λέξεις για κάθε διαφορετικό Tweet.
- Εξάλειψη των λέξεων με ουδέτερο νόημα και κανένα συναίσθημα (π.χ. and, also, able, or, etc.), οι οποίες δε διευκολύνουν την ανάλυση συναισθήματος. Η βιβλιοθήκη του εργαλείου NLTK περιέχει λεξικό με stopwords, δηλαδή τέτοιων λέξεων, οι οποίες διαγράφονται, εάν εμπεριέχονται στα προς ανάλυση Tweets.
- Όπως είναι γνωστό στα Tweets συνήθως συμπεριλαμβάνονται χαρακτήρες και σύμβολα, όπως “@”, “#” emojis, και υπερσυνδέσμοι. Τα στοιχεία αυτά αποτελούν επίσης αχρείαστο κείμενο και είναι αναγκαίο να εξαλειφθούν. Το κείμενο που ακολουθεί όμως το σύμβολο της δέσης σε αντίθεση με αυτό του “@” δε διαγράφεται, καθώς είναι πιθανόν να περιέχει πληροφορίες χρήσιμες για την ανάλυση.
- Μια τεχνική που συνήθως ακολουθείται κατά τη διαδικασία της προεπεξεργασίας κειμένου, είναι αυτό του Stemming. Το Stemming είναι η διαδικασία κατά την οποία αφαιρούνται οι καταλήξεις από τις λέξεις (π.χ. "game","gaming","gamed","games") και κρατείται μόνο η ρίζα τους για να διευκολυνθεί το εργαλείο που θα τις χρησιμοποιήσει μετέπειτα στην ανάλυση συναισθήματος.
- Ακόμη μια διαδεδομένη μέθοδος προ-επεξεργασίας κειμένου, είναι αυτή του Lemmatization. Το Lemmatization, είναι παρόμοια τεχνική με το stemming, όμως διαφέρει στο ότι συνδέει λέξεις με παρόμοιο νόημα σε μία λέξη, πραγματοποιώντας ουσιαστικά μορφολογική ανάλυση στο κείμενο (π.χ. -> better : good).



Σχήμα 4.2: Σύνοψη Μεθόδων Προ-επεξεργασίας Κειμένου

Σημειώνεται ότι οι τεχνικές Stemming και Lemmatization απουσιάζουν από το τελικό σύστημα, καθώς κατόπιν αξιολόγησης των αποτελεσμάτων κρίθηκε ότι επηρεάζουν αρνητικά την ποιότητα του κειμένου των αναρτήσεων και κατά συνέπεια την απόδοση του αναλυτή συναισθήματος.

4.1.3. Ανάλυση Συναισθήματος

Βασικό μέρος της εργασίας αποτελεί η ανάλυση των συναισθημάτων στα κείμενα και η εξαγωγή της διάθεσης που επικρατεί είτε αυτή είναι θετική, αρνητική, είτε ουδέτερη. Για την όσον το δυνατό πιο επιτυχημένη ανάλυση χρησιμοποιήθηκαν σχετικά εργαλεία και βιβλιοθήκες με κυριότερο το εργαλείο VADER της βιβλιοθήκης NLTK, το οποίο λαμβάνοντας ως είσοδο το προ-επεξεργασμένο κείμενο από το προηγούμενο στάδιο, παράγει και τα ανάλογα αποτελέσματα ως προς το συναίσθημα.

Με την ανάλυση συναισθήματος επιδιώκεται η εκπαίδευση του συστήματος, έτσι ώστε να εντοπίζει εάν ένα κείμενο είναι “bullish” ή “bearish”, δηλαδή αν αφορά αύξηση ή μείωση τιμών μετοχών. Όπως είναι άλλωστε λογικό, κείμενα που αναφέρονται σε αύξηση τιμής μετοχών χαρακτηρίζονται από θετική διάθεση, καθώς προμηνύουν οικονομικό όφελος. Στον αντίποδα, Tweets που αναφέρονται σε μείωση τιμής μετοχών συσχετίζονται άμεσα με αρνητικό συναίσθημα.

Αποφασίστηκε η κατηγοριοποίηση του κειμένου σε “θετικό” , “αρνητικό” και “ουδέτερο”. Σε αρχικό στάδιο της έρευνας για τον κατατοπισμό της μεθοδολογίας υπήρξε αβεβαιότητα για την ενδεχόμενη συμπερίληψη και της “ουδέτερης” ετικέτας συναισθημάτων, καθώς ο αναλυτής συναισθημάτων Vader που τελικά αξιοποιήθηκε παρέχει αυτή τη δυνατότητα. Αν και ορισμένες παλαιότερες μελέτες[4], αναφέρουν ότι η “ουδέτερη” ετικέτα πιθανώς να επιφέρει μείωση στην ακρίβεια των αποτελεσμάτων, αυτό δεν έχει αποδειχθεί επίσημα. Επιπλέον, ο αριθμός των κειμένων που κατηγοριοποιήθηκαν ως ουδέτερα, ήταν αρκετά μεγάλος και ο ενδεχόμενος αποκλεισμός τους θα αλλοίωνε σημαντικά την εικόνα του συνόλου δεδομένων και θα παραβίαζε την εγκυρότητα της ανάλυσης των αποτελεσμάτων κατά την αξιολόγηση, κάτι που οδήγησε στη συμπερίληψη και των 3 βασικών ετικετών.

Text	Sentiment
Stocks may anticipate economic surge as S&P nears record high.	Bullish/Positive
Here with a comparison chart of \$AAL and \$UAL	Neutral
In case you missed it: All three major US equity benchmarks posted their worst week since Oct2008, and their biggest declines yet of the coronavirus crisis. Dow crashed -17%, S& P 500 plunged -15%, Nasdaq sank -13%.	Bearish/Negative
Dow Jones soars more than 1,600 points or almost 8% as growth rate of new coronavirus cases appears to slow.	Bullish/Positive
Wall St suffered worst trading day since March w/Dow Jones down 6.9%, Boeing plunged 16.4%, Dow Inc almost 10% lower.	Bearish/Negative

Πίνακας 4.4: Παραδείγματα κατηγοριοποίησης συναισθήματος σε Tweets

Έπειτα από την ανάλυση συναισθήματος πραγματοποιήθηκε ανάλυση και απεικόνιση των αποτελεσμάτων για την εξαγωγή γνώσης και την εξέταση ύπαρξης συσχετίσεων μεταξύ των δεδομένων που συλλέγηκαν και των πληροφοριών που παράχθηκαν. Η ανάλυση δεδομένων για εξαγωγή γνώσης, πραγματοποιήθηκε ξεχωριστά στα 11

διαφορετικά σύνολα δεδομένων τομέων της οικονομίας που δημιουργήθηκαν (βλ. Μέθοδος Β) και τα οποία περιέχουν Tweets που αφορούν 33 διαφορετικές εταιρείες. Η διαδικασία ανάλυσης που περιγράφεται στο παρόν κεφάλαιο, εφαρμόστηκε για κάθε σύνολο δεδομένων και ενδεικτικά παρουσιάζονται πίνακες, γραφήματα και στιγμιότυπα για ορισμένα από αυτά, στο στάδιο της αξιολόγησης αποτελεσμάτων.

Τα δεδομένα τα οποία μελετήθηκαν αποθηκεύτηκαν σε dataframes για ευκολότερη διαχείριση τους, όπου υπολογίστηκαν ενδιαφέρον μετρικές και δημιουργήθηκαν γραφικές παραστάσεις αξιοποιώντας τα δεδομένα συνδυαστικά με σχετικές βιβλιοθήκες της Python. Μετά το τέλος της επεξεργασίας των δεδομένων, προστέθηκαν νέα πεδία στα dataframes με τις υπολογισμένες μετρικές για κάθε διαφορετική ανάρτηση.

Τελευταίο στάδιο της διαδικασίας υλοποίησης αποτέλεσε η ανάλυση της συσχέτισης ανάμεσα στο συναίσθημα και στα αντίστοιχα δεδομένα που συλλέχθηκαν από το Χρηματιστήριο, έτσι ώστε να καταστεί δυνατή η αξιολόγηση ως προς το κατά πόσο μπορούν να ληφθούν σημαντικές πληροφορίες για τις μεταβολές των μετοχών.

4.1.3.1. Κατηγοριοποίηση αναρτήσεων με βάση το συναίσθημα

Μετά την ανάλυση συναισθήματος και την κατηγοριοποίηση των αναρτήσεων ανάλογα με το συναίσθημα τους, μετρήθηκε ο αριθμός των Tweets που ανήκουν σε κάθε κατηγορία. Οι αριθμοί που καταμετρήθηκαν έτυχαν επεξεργασίας για να εξαχθεί η μετρική του ποσοστού που καταλαμβάνουν στο σύνολο των Tweets, ενώ δημιουργήθηκαν ενδεικτικές γραφικές παραστάσεις που απεικονίζουν συγκριτικά τον αριθμό των θετικών, αρνητικών και ουδέτερων αναρτήσεων, σύμφωνα με τον αναλυτή συναισθήματος.

4.1.3.2. Δημοφιλή Hashtags

Ενδιαφέρον στατιστικό για τους επενδυτές αποτελεί η πληροφορία για τα δημοφιλέστερα Hashtags, τα οποία χρησιμοποιούνται πολύ συχνά σε αναρτήσεις που αφορούν συγκεκριμένες εταιρείες ή και γενικότερα τον Χρηματιστηριακό κόσμο. Τα αποτελέσματα της εν λόγω μετρικής, είναι σε θέση να υποδείξουν στον χρήστη

hashtags, μέσω των οποίων μπορούν να πραγματοποιηθούν αναζητήσεις με πολλαπλά αποτελέσματα και που αφορούν αναρτήσεις στην κοινότητα των επενδυτών του Twitter.

Για τον υπολογισμό του εν λόγω στατιστικού, πραγματοποιήθηκε εκ νέου η διαδικασία της προ-επεξεργασίας κειμένου, με μόνη διαφορά τη διατήρηση των hashtags στο κείμενο των Tweets, προκειμένου να καταμετρηθούν και να υπολογιστεί η συχνότητα εμφάνισης τους.

Κρίθηκε ενδιαφέρον, μετά την ανάλυση συναισθήματος, να καταμετρηθεί η συχνότητα εμφάνισης των Hashtags για τα κατηγοριοποιημένα θετικά, αρνητικά και ουδέτερα Tweets ξεχωριστά, έτσι ώστε να παραχθεί περισσότερη προς αξιολόγηση πληροφορία.

Για την όσο το δυνατόν καλύτερη αναπαράσταση των αποτελεσμάτων δημιουργήθηκαν γραφικές παραστάσεις με τα δημοφιλέστερα Hashtags για κάθε κατηγορία συναισθήματος συναρτήσει της συχνότητας εμφάνισης τους στο κείμενο των Tweets.

4.1.3.3. Συχνότητα εμφάνισης λέξεων

Εξίσου ενδιαφέρον στατιστικό αποτελεί η απεικόνιση των λέξεων που χρησιμοποιήθηκαν πιο συχνά στο κείμενο του συνόλου των Tweets. Για την αναπαράσταση των αποτελεσμάτων αξιοποιήθηκε η βιβλιοθήκη WordCloud της Python[35], η οποία παρουσιάζει τις συχνά χρησιμοποιημένες λέξεις σε γράφημα και η διαφορά στη συχνότητα χρησιμοποίησης των λέξεων, είναι ανάλογη του μεγέθους της γραμματοσειράς, έτσι ώστε λέξεις οι οποίες είναι πιο δημοφιλής να καταλαμβάνουν μεγαλύτερο χώρο στο γράφημα. Απαραίτητη προϋπόθεση για την ορθή εξαγωγή αποτελεσμάτων στην εν λόγω διεργασία, αποτελεί η αφαίρεση των stopwords, κατά το στάδιο της προ-επεξεργασίας κειμένου, καθώς αυτές χρησιμοποιούνται πολύ συχνά σε κείμενα και η ενδεχόμενη παραμονή τους στο κείμενο, συνεπάγεται την συμπερίληψη τους σε μεγάλο μέρος των αποτελεσμάτων, κάτι που δεν είναι επιθυμητό.

4.1.3.4. Υποκειμενικότητα και Πολικότητα

Η παρούσα εργασία επικεντρώνεται κατά κύριο λόγο στην πολικότητα των αναρτήσεων στα μέσα κοινωνική δικτύωσης. Είναι όμως ευρέως γνωστό, ότι οι χρήστες δεν εκφράζουν πάντοτε τις απόψεις τους με αντικειμενικότητα και ως εκ τούτου η βαθμολογία της πολικότητας, χρήζει αμφισβήτησης. Για το λόγο αυτό, καταμετρήθηκε η βαθμολογία της υποκειμενικότητας των αναρτήσεων κατά το στάδιο της ανάλυσης συναισθήματος.

Παρόλο που η συγκεκριμένη μετρική δεν παρέχει από μόνη της ενδιαφέρουσες πληροφορίες, η σχέση της με την πολικότητα της κάθε ανάρτησης, αποτελεί χρήσιμη προς αξιολόγηση πληροφορία.

Λαμβάνοντας υπόψιν τα πιο πάνω δημιουργήθηκε απεικόνιση της μετρικής της Υποκειμενικότητας συναρτήσει της Πολικότητας, με γράφημα τύπου διασποράς, όπου κάθε συντεταγμένη αποτελεί ανάρτηση, για την οποία απεικονίζεται συσχετιστικά η βαθμολογία πολικότητας και αντικειμενικότητας.

4.2. Μεθοδολογία Ανάπτυξης Μοντέλου Πρόβλεψης

Στο παρόν υποκεφάλαιο, περιγράφεται αναλυτικά η μεθοδολογία ανάπτυξης ενός μοντέλου πρόβλεψης της Χρηματιστηριακής συμπεριφοράς, του οποίου η υλοποίηση βασίζεται σε αλγορίθμους μηχανικής μάθησης και αξιοποιεί ως χαρακτηριστικά εκπαίδευσης του, Χρηματιστηριακές τιμές και τη γνώση από την ανάλυση συναισθήματος σε αναρτήσεις τίτλων ειδήσεων από την πλατφόρμα δημόσιας συζητήσεως Reddit, που αφορούν τον Χρηματιστηριακό Δείκτη Dow Jones.

4.2.1. Περιγραφή Συνόλου Δεδομένων από το Reddit

Για την εφαρμογή και αξιολόγηση του μοντέλου πρόβλεψης αξιοποιήθηκε το σύνολο δεδομένων “Daily News for Stock Market Prediction” το οποίο είναι διαθέσιμο μέσω του Kaggle[17]. Το εν λόγω σύνολο δεδομένων περιέχει αναρτήσεις τίτλων ειδήσεων από την κοινότητα(subreddit) στο Reddit, WorldNews Channel(r/worldnews), στις οποίες εμπεριέχονται θέματα που ενδεχομένως να επηρέασαν την κίνηση του Χρηματιστηριακού Δείκτη Dow Jones από το 2008 μέχρι το 2016 για κάθε

χρηματιστηριακή μέρα. Για κάθε διαφορετική μέρα συλλέχθηκαν οι 25 κυριότερες ειδήσεις, έτσι όπως διαμορφώθηκαν έπειτα από τις αξιολογήσεις των χρηστών του Reddit.

Το σύνολο δεδομένων εμπεριέχει συνολικά 27 στήλες και η μορφή του παρουσιάζεται πιο κάτω.

Date	Label	Top1	...	Top 25
------	-------	------	-----	--------

Πίνακας 4.5: Δομή Συνόλου Δεδομένων με τίτλους ειδήσεων από το Reddit

Date: Η ημερομηνία δημοσίευσης των τίτλων ειδήσεων στο Reddit Worldnews Channel.

Label: '1' για άνοδο της τιμής του δείκτη Dow Jones και '0' για πτώση σε σχέση με την τιμή κλεισίματος της προηγούμενης μέρας.

Top 1 – Top 25: Οι 25 καλύτεροι τίτλοι ειδήσεων της ίδιας μέρας, σύμφωνα με τις αξιολογήσεις των χρηστών, που ενδεχομένως να αφορούν τον Χρηματιστηριακό δείκτη Dow Jones.

Για την ίδια χρονική περίοδο πραγματοποιήθηκε συλλογή ιστορικών Χρηματιστηριακών δεδομένων, μέσω του Yahoo Finance (βλ. 4.3).

Το σύνολο δεδομένων εμπεριέχει 7 στήλες συνολικά και η μορφή του παρουσιάζεται πιο κάτω.

Date	Open	Close	High	Low	Volume	Adj. Close
------	------	-------	------	-----	--------	------------

Πίνακας 4.6: Δομή Συνόλου Δεδομένων από το Χρηματιστήριο

Date: Η ημερομηνία για την οποία συλλέχθηκαν τα δεδομένα.

Open: Η τιμή ανοίγματος του Χρηματιστηριακού Δείκτη Dow Jones, δηλαδή η πρώτη τιμή που καταγράφηκε κατά την έναρξη της Χρηματιστηριακής μέρας.

Close: Η τιμή κλεισίματος του Χρηματιστηριακού Δείκτη Dow Jones, δηλαδή η τελευταία τιμή που καταγράφηκε κατά το κλείσιμο της Χρηματιστηριακής μέρας.

High: Η υψηλότερη τιμή του Χρηματιστηριακού Δείκτη Dow Jones που καταγράφηκε κατά την διάρκεια της Χρηματιστηριακής μέρας.

Low: Η χαμηλότερη τιμή που καταγράφηκε για τον Χρηματιστηριακό Δείκτη Dow Jones κατά την διάρκεια της Χρηματιστηριακής μέρας.

Volume: Ο όγκος των συναλλαγών (αγοραπωλησιών) του Χρηματιστηριακού Δείκτη Dow Jones για το σύνολο της Χρηματιστηριακής μέρας.

Adjusted Closing Price: Η προσαρμοσμένη τιμή κλεισίματος αποτελεί ανάλυση που χρησιμοποιεί κατά βάση την τιμή κλεισίματος, αλλά λαμβάνει υπόψη παράγοντες όπως μερίσματα, διασπάσεις μετοχών και νέες προσφορές μετοχών για τον προσδιορισμό της συνολικής αξίας[15].

4.2.2. Προ-επεξεργασία συνόλου δεδομένων

Για τη διαχείριση των δεδομένων κατά την υλοποίηση του μοντέλου αξιοποιήθηκε η βιβλιοθήκη Pandas της γλώσσας προγραμματισμού Python και η αναπαράσταση των δεδομένων παρουσιάζεται σε μορφή dataframe.

Κύριο στόχο της μετέπειτα επεξεργασίας του συνόλου δεδομένων για την εξαγωγή πληροφοριών, αποτελεί η εξαγωγή πολύτιμων χαρακτηριστικών για την εκπαίδευση του μοντέλου μηχανικής μάθησης. Απαραίτητη προϋπόθεση για την επίτευξη του εν λόγω στόχου, είναι η προσαρμογή του συνόλου δεδομένων, έτσι ώστε οι πληροφορίες που αφορούν την ίδια Χρηματιστηριακή μέρα να βρίσκονται στο ίδιο πεδίο. Επομένως, οι 25 τίτλοι ειδήσεων για κάθε μέρα, συνενώνονται σε ένα νέο πεδίο του συνόλου δεδομένων, το οποίο ονομάζεται “Combined News”.

Πριν από την επεξεργασία των δεδομένων πραγματοποιείται προ-επεξεργασία του κειμένου στο πεδίο “Combined News” και μετά το τέλος των διεργασιών καθαρισμού κειμένου, το επεξεργασμένο κείμενο αντικαθιστά το αρχικό. Όσον αφορά τους τίτλους ειδήσεων, η προ-επεξεργασία τους αποτελεί λιγότερο πολύπλοκη διαδικασία από αυτήν των Tweets, καθώς είναι αρκετή η αφαίρεση περιττών σημείων στίξης και η εξάλειψη ενδεχόμενων stopwords.

Date	Label	Top1	...	Top 25	Combined News
------	-------	------	-----	--------	---------------

Πίνακας 4.7: Δομή Συνόλου Δεδομένων μετά την Προ-επεξεργασία Κειμένου

Μετά τη δημιουργία του πεδίου “Combined News”, δεν είναι πλέον απαραίτητη η ύπαρξη των 25 πεδίων με τους τίτλους ειδήσεων, καθώς αν παραμείνουν θα υπάρχει

επανάληψη χαρακτηριστικών, στοιχείο που πιθανώς να επηρεάσει αρνητικά την αποδοτικότητα του προς υλοποίηση μοντέλου. Για το λόγο αυτό τα 25 πεδία με τους τίτλους ειδήσεων, διαγράφονται.

Date	Label	Combined News
------	-------	---------------

Πίνακας 4.8: Δομή Συνόλου Δεδομένων μετά την διαγραφή περιττών πεδίων

4.2.3. Επεξεργασία Συνόλου Δεδομένων

Κύριος στόχος της επεξεργασίας του συνόλου δεδομένων, ορίζεται η εξαγωγή χαρακτηριστικών και μετρικών από το κείμενο με σκοπό την αξιοποίηση τους στο μοντέλο πρόβλεψης.

Αρχικά, πραγματοποιείται ανάλυση συναισθήματος, μέσω του εργαλείου Vader Sentiment Analyzer σε όλες τις εγγραφές του πεδίου Combined News και δημιουργούνται νέα πεδία με τις ακόλουθες βαθμολογίες:

- Compound Score
- Positive Score
- Negative Score
- Neutral Score

Επιπρόσθετα, δημιουργείται νέο πεδίο που παρουσιάζει την μετρική της υποκειμενικότητας η οποία προσμετράται, μέσω του εργαλείου TextBlob. Τα χαρακτηριστικά, ολοκληρώνει το πεδίο 'Overall Sentiment', που υπολογίζεται μετά την επεξεργασία του Compound Score και παρουσιάζει '1', όταν το συναίσθημα που εξαγάγεται είναι θετικό ή ουδέτερο και '0' όταν το συναίσθημα είναι αρνητικό. Η παρουσία του πεδίου Combined News, πλέον δεν είναι απαραίτητη, καθώς εξάχθηκαν οι επιθυμητές πληροφορίες και ως εκ τούτου διαγράφεται από το ανανεωμένο σύνολο δεδομένων.

Date	Compound	Positive	Negative	Neutral	Subjectivity	Overall Sentiment
------	----------	----------	----------	---------	--------------	-------------------

Πίνακας 4.9: Δομή Συνόλου Δεδομένων μετά την Ανάλυση Συναισθήματος

Σε ότι αφορά την επεξεργασία του συνόλου δεδομένων με τα Χρηματιστηριακά δεδομένα, καθίσταται αναγκαίο να απορριφθούν και να αντικατασταθούν από νέα πεδία δεδομένων, τα οποία να περιέχουν τις χθεσινές τιμές τους, προκειμένου τα χαρακτηριστικά εισόδου του τελικού μοντέλου μηχανικής μάθησης, να προσομοιώνουν πραγματικές συνθήκες. Η παραπάνω αναγκαιότητα, προκύπτει από το γεγονός ότι είναι αδύνατον να γνωρίζονται οι Χρηματιστηριακές τιμές για μια συγκεκριμένη μέρα, πριν το τέλος της. Ως εκ τούτου, για κάθε ημερομηνία παρατίθενται τα αντίστοιχα δεδομένα της ημέρας που προηγήθηκε, τα οποία αποτελούν την πιο κοντινή χρονικά γνώση, που μπορεί να αξιοποιηθεί για την πρόβλεψη της τάσης του Χρηματιστηρίου για την ημέρα που διανύεται.

4.2.4. Συγχώνευση Συνόλων Δεδομένων

Η συγχώνευση των δύο παραχθέντων συνόλων δεδομένων (Χρηματιστηριακών τιμών και Ανάλυσης Συναισθήματος) πραγματοποιείται με βάση την ημερομηνία και κατά αυτόν τον τρόπο, δημιουργείται ένα ενιαίο σύνολο δεδομένων με αριθμητικά στατιστικά και μετρικές από την επεξεργασία κειμένου (numerical and textual analysis), με στόχο την υλοποίηση ενός προβλήματος μηχανικής μάθησης με χαρακτηριστικά του πεδίου ενδιαφέροντος της εργασίας.

Σαν επιπρόσθετο στάδιο της προετοιμασίας του τελικού συνόλου δεδομένων, εξετάστηκε η ύπαρξη συσχέτισης μεταξύ των χαρακτηριστικών, προκειμένου να εντοπιστεί ενδεχόμενο υποσύνολο τους, για το οποίο οι τιμές συσχέτισης ξεπερνούσαν προκαθορισμένο όριο (threshold). Εντοπίστηκαν ζεύγη χαρακτηριστικών από τα Χρηματιστηριακά δεδομένα, τα οποία πρόσφεραν παρόμοια πληροφορία στο μοντέλο και ξεπερνούσαν το προκαθορισμένο όριο συσχέτισης, το οποίο ήταν της τάξης του 0.85, σύμφωνα με τον συντελεστή συσχέτισης γινομένου ροπής Pearson. Από τα ζεύγη χαρακτηριστικών (colinear features) που εντοπίστηκαν κρατήθηκε μόνο το ένα εκ των δύο χαρακτηριστικών για το τελικό σύνολο δεδομένων. Συγκεκριμένα, από το ζεύγος τιμών “High” και “Low”, κρατήθηκε η τιμή “Low”, ενώ από το ζεύγος “Close” και “Adj. Close”, κρατήθηκε η τιμή “Adj. Close”. Από το τελικό σύνολο δεδομένων απουσιάζουν επίσης τα πεδία της υποκειμενικότητας και του όγκου

συναλλαγών, καθώς προτιμήθηκε η επικέντρωση σε χαρακτηριστικά που αφορούν μόνο Χρηματιστηριακές τιμές και πολικότητα συναισθημάτων.

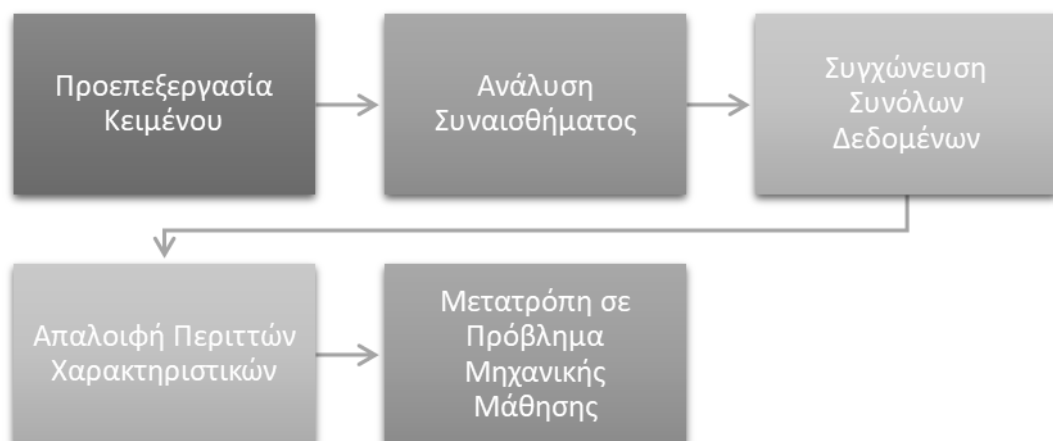
Compound	Positive	Negative	Neutral	Overall Sentiment	Open	Low	Adj. Close
----------	----------	----------	---------	-------------------	------	-----	------------

Πίνακας 4.10: Δομή του Συγχωνευμένου Συνόλου Δεδομένων

4.2.5. Μετατροπή Συνόλου Δεδομένων σε Πρόβλημα Μηχανικής Μάθησης

Για την μετατροπή των δεδομένων σε πρόβλημα μηχανικής μάθησης χρησιμοποιείται η ετικέτα Label του αρχικού συνόλου δεδομένων. Οι ενδείξεις, '1' για την άνοδο και '0' για την κάθοδο της τιμής κλεισίματος μετοχής, ορίστηκαν ως ο στόχος του μοντέλου πρόβλεψης, ενώ τα πεδία του συγχωνευμένου σύνολου δεδομένων που παράχθηκε, αποτελούν τα χαρακτηριστικά εισόδου με τα οποία το μοντέλο εκπαιδεύεται, προκειμένου να διενεργήσει προβλέψεις. Επομένως, με τον τρόπο αυτό έχει δημιουργηθεί ένα πρόβλημα δυαδικής ταξινόμησης, το οποίο είναι σε θέση να επιλύσουν πληθώρα αλγορίθμων μηχανικής μάθησης και σκοπό έχει την πρόβλεψη της τάσης μιας Χρηματιστηριακής μετοχής, με βάση την μεταβολή της τιμής κλεισίματος της, για κάθε μέρα.

Στο παρακάτω σχήμα (βλ. Σχήμα 4.3) συνοψίζεται η διαδικασία προετοιμασίας του μοντέλου, μέχρις ότου μετατραπεί σε πρόβλημα Μηχανικής Μάθησης.



Σχήμα 4.3: Σύνοψη Διαδικασίας Προετοιμασίας Μοντέλου

4.2.6. Μηχανική Μάθηση

Μετά τη διεξαγωγή της προ-επεξεργασίας στα δεδομένα, την προετοιμασία τους για εξαγωγή πληροφοριών και τη δημιουργία ενός επιλύσιμου προβλήματος μηχανικής μάθησης, πραγματοποιείται εκπαίδευση και δοκιμή συγκεκριμένων αλγορίθμων μηχανικής μάθησης, για την πραγματοποίηση προβλέψεων.

Η εκπαίδευση και δοκιμή του συνόλου δεδομένων εφαρμόστηκε χρησιμοποιώντας το TimeSeriesSplit από το model_selection της βιβλιοθήκης sklearn, το οποίο αποτελεί παραλλαγή του k-fold cross validation, προκειμένου να εκτιμηθεί η ικανότητα κάθε μοντέλου σε άγνωστα, για αυτό δεδομένα. Στην εν λόγω παραλλαγή, στον k-ιοστό διαχωρισμό, επιστρέφονται τα πρώτα k-folds σαν εκπαιδευτικό σύνολο(training set) και το (k+1) fold σαν σύνολο δοκιμών. Επιπλέον, σε αντίθεση με τις κλασσικές μεθόδους cross-validation τα διαδοχικά εκπαιδευτικά σύνολα αποτελούν υπερσύνολα των εκπαιδευτικών συνόλων που προηγήθηκαν. Το TimeSeriesSplit, προτιμήθηκε έναντι του k-fold validation, καθώς επιθυμείται η εκπαίδευση των αλγορίθμων με χαρακτηριστικά εισόδου που να ανήκουν στο παρελθόν για να προβλεφθεί το μέλλον και ως εκ τούτου η μεθοδολογία διασπάσεων του k-fold validation δεν ενδείκνυται.

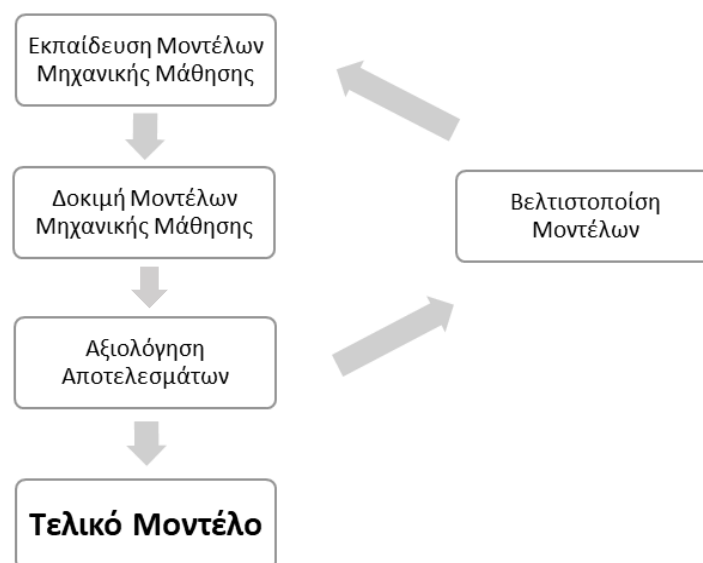
Η υλοποίηση αλγορίθμων μηχανικής μάθησης δεν εμπίπτει στο πεδίο ενδιαφέροντος της παρούσας εργασίας. Ωστόσο, είναι αναγκαία η κατανόηση του τρόπου λειτουργίας των αλγορίθμων που επιλέχθηκαν, καθώς επίσης και η μελέτη του ρόλου των παραμέτρων τους και της επίδρασης αυτών στις προβλέψεις, για την επίτευξη των καλύτερων δυνατών αποτελεσμάτων.

Οι αλγόριθμοι οι οποίοι επιλέχθηκαν έχουν την ιδιότητα να λειτουργούν αποδοτικά και με ακρίβεια σε προβλήματα επιτηρούμενης μάθησης που ασχολούνται με την δυαδική ταξινόμηση. Στον αντίποδα, η φύση του προβλήματος οδήγησε στην απόρριψη αλγορίθμων μηχανικής μάθησης που δεν προβλέπονται για προβλήματα τέτοιου είδους. Οι εργασίες μηχανικής μάθησης ταξινομούνται σε τρεις μεγάλες κατηγορίες: την επιτηρούμενη μάθηση, την μη επιτηρούμενη μάθηση και την ενισχυτική μάθηση, ενώ μεταξύ της επιτηρούμενης και της μη επιτηρούμενης μάθησης, είναι η ημι-επιτηρούμενη μάθηση. Στην παρούσα εργασία χρησιμοποιείται

επιτηρούμενη μάθηση, δηλαδή το είδος μάθησης στην οποία η ακρίβεια των μοντέλων της, συσχετίζεται σε μεγάλο βαθμό με την είσοδο εκ της οποίας το μοντέλο τροφοδοτείται. Οι αλγόριθμοι μηχανικής μάθησης που επιλέχθηκαν είναι οι εξής:

- Logistic Regression
- Linear Discriminant Analysis
- K-Nearest Neighbours
- Random Forest Classifier
- XGBoost Classifier
- Decision Tree Classifier

Πιο κάτω συνοψίζεται η διαδικασία επιλογής του βέλτιστου μοντέλου μηχανικής μάθησης, έτσι όπως προέκυψε, μέσω της επαναληπτικής προσέγγισης που παρουσιάζεται στο σχήμα (βλ. Σχήμα 4.4).



Σχήμα 4.4: Διαδικασία Ανάπτυξης Τελικού Μοντέλου Μηχανικής Μάθησης

4.2.7. Βελτιστοποίηση αλγορίθμων μηχανικής μάθησης

Προκείμενου να επιτευχθούν τα καλύτερα δυνατά αποτελέσματα στις προβλέψεις των μοντέλων από τους αλγορίθμους μηχανικής μάθησης, προβάλλεται ως επιτακτική ανάγκη η κατανόηση των παραμέτρων τους.

Έπειτα από την κατανόηση των παραμέτρων, πραγματοποιήθηκε προσπάθεια χειροκίνητου πειραματισμού με συγκεκριμένες παραμέτρους και αξιολόγηση στη μεταβολή των τιμών των μετρικών. Παρόλα αυτά, η μη αυτοματοποιημένη διεκπεραίωση της διεργασίας αυτής δεν αποτελεί ιδανική επιλογή, καθώς οι παράμετροι του κάθε μοντέλου εξαρτώνται από τη φύση των δεδομένων και τα αποτελέσματα που προκύπτουν είναι περισσότερο εμπειρικά παρά βέλτιστα.

Επομένως, εφαρμόστηκε αυτοματοποιημένη διαδικασία επιλογής συνόλου βέλτιστων παραμέτρων για την εκμάθηση αλγορίθμων πρόβλεψης, το επονομαζόμενο tuning των hyper parameters (ή Optimization) των αλγορίθμων.

Αξιοποιήθηκε λοιπόν η συνάρτηση GridSearchCV[29] από τη βιβλιοθήκη model selection του sklearn[25], η οποία παρέχει τη δυνατότητα εφαρμογής ρύθμισης σε παραμέτρους των αλγορίθμων.

Η διεργασία εφαρμόστηκε στις παραμέτρους όλων των αλγορίθμων που χρησιμοποιήθηκαν. Για τον καθένα από τους αλγορίθμους μηχανικής μάθησης, δημιουργήθηκε ξεχωριστό πλέγμα παραμέτρων με διαφορετικές δοκιμαστικές τιμές για την κάθε παράμετρο.

Όπως αναφέρθηκε προηγουμένως, εφαρμόστηκε με μια μέθοδο TimeSeriesSplit 9 μερών (9 folds) και οι καλύτερες παράμετροι βασίστηκαν στο καλύτερα αποτελέσματα ορθότητας (accuracy), ακρίβειας, ανάκλησης και βαθμολογίας F1, σύμφωνα με το εργαλείο GridSearchCV.

Πιο κάτω παρατίθενται τα πλέγματα παραμέτρων κάθε μοντέλου και οι δοκιμαστικές τιμές τους.

Random Forest Classifier	
Παράμετρος	Τιμές
n_estimators	100,200,300,500,600,700,800
class_weight	balanced,balanced_subsample
Criterion	gini,entropy
Max_depth	None, 1, 3, 5, 10
Min_samples_split	5,10
Min_samples_leaf	5, 10

Πίνακας 4.11: Πλέγμα Παραμέτρων αλγορίθμου Random Forest Classifier

K-Nearest Neighbors	
Παράμετρος	Τιμές
n_neighbors	1,5,10,20,25,30,40,50,60,70,100
algorithm	auto,ball_tree,kd_tree,brute
weights	uniform,distance

Πίνακας 4.12: Πλέγμα Παραμέτρων αλγορίθμου K-Nearest Neighbors

XGBoost Classifier	
Παράμετρος	Τιμές
n_estimators	200,300,400,500, 1000
max_depth	3,4,5,6,7,8,9,10

Πίνακας 4.13: Πλέγμα Παραμέτρων αλγορίθμου XGBoost Classifier

Logistic Regression	
Παράμετρος	Τιμές
Solver	newton-cg, lbfgs, liblinear, sag, saga
penalty	l1,l2
C	100, 10, 1.0, 0.1, 0.01

Πίνακας 4.14: Πλέγμα Παραμέτρων αλγορίθμου Logistic Regression

Decision Tree	
Παράμετρος	Τιμές
class_weight	balanced
criterion	giny, entropy
splitter	best, random

Πίνακας 4.15: Πλέγμα Παραμέτρων αλγορίθμου Decision Tree

Linear Discriminant Analysis	
Παράμετρος	Τιμές
solver	svd, lsqr, eigen

Πίνακας 4.16: Πλέγμα Παραμέτρων αλγορίθμου Linear Discriminant Analysis

4.3. Συλλογή Δεδομένων για τιμές της Χρηματιστηριακής Αγοράς

Για τη συλλογή δεδομένων από τη Χρηματιστηριακή Αγορά αξιοποιήθηκε το Yahoo! Finance API για την Γλώσσα Προγραμματισμού Python, που αποτελεί το πιο διαδεδομένο εργαλείο για αυτό το σκοπό.

Το Yahoo! Finance API αποτελεί ένα σύνολο βιβλιοθηκών, APIs και μεθόδων για την λήψη ιστορικών δεδομένων, αλλά και δεδομένων πραγματικού χρόνου για τεράστια ποικιλία χρηματοπιστωτικών αγορών και προϊόντων.

Συλλέχθηκαν δεδομένα και ανακτήθηκαν οι πιο κάτω τιμές για κάθε μέρα, τόσο για τον Χρηματιστηριακό Δείκτη Dow Jones, όσο και για κάθε μια από τις εταιρείες του πεδίου ενδιαφέροντος της μελέτης (βλ. Κεφάλαιο 5).

Τιμή Ανοίγματος: Η τιμή ανοίγματος μετοχής, δηλαδή η πρώτη τιμή που καταγράφηκε κατά την έναρξη της Χρηματιστηριακής μέρας.

Τιμή κλεισίματος: Η τιμή κλεισίματος μετοχής, δηλαδή η τελευταία τιμή που καταγράφηκε κατά το κλείσιμο της Χρηματιστηριακής μέρας.

Μέγιστη τιμή: Η υψηλότερη τιμή κάποιας μετοχής που καταγράφηκε κατά την διάρκεια της Χρηματιστηριακής μέρας.

Ελάχιστη τιμή: Η χαμηλότερη τιμή κάποιας μετοχής κατά την διάρκεια της Χρηματιστηριακής μέρας.

Όγκος Συναλλαγών: Ο όγκος των συναλλαγών (αγοραπωλησιών) κάποιας μετοχής για το σύνολο της Χρηματιστηριακής μέρας.

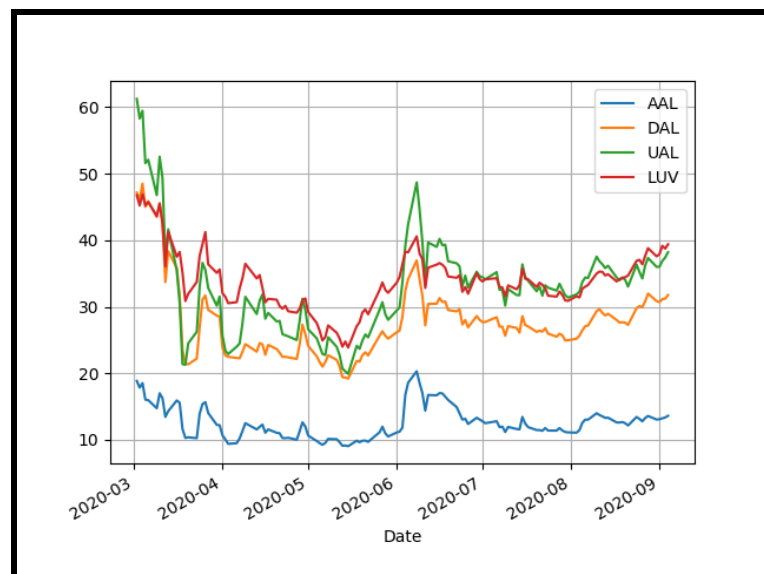
Προσαρμοσμένη Τιμή Κλεισίματος: Η προσαρμοσμένη τιμή κλεισίματος αποτελεί ανάλυση που χρησιμοποιεί κατά βάση την τιμή κλεισίματος, αλλά λαμβάνει υπόψη παράγοντες όπως μερίσματα, διασπάσεις μετοχών και νέες προσφορές μετοχών για τον προσδιορισμό της συνολικής αξίας(31).

Η περίοδος για την οποία συλλέχθηκαν δεδομένα είναι από την 01/03/2020, ημερομηνία του παλαιότερου Tweet που συλλέχθηκε από τους λογαριασμούς, μέχρι

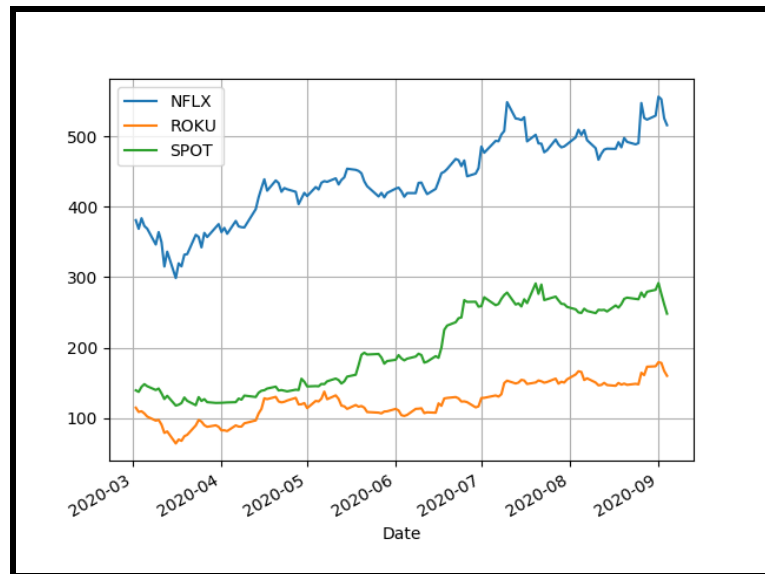
τις 05/09/2020. Για τον δείκτη Dow Jones, πραγματοποιήθηκε συλλογή δεδομένων από 08/06/2008 μέχρι 01/07/2016, περίοδο στην οποία αναφέρονται οι αξιολογημένοι τίτλοι ειδήσεων από το Reddit.

Λαμβάνοντας υπόψιν ότι η αγορά του Χρηματιστηρίου δεν βρίσκεται σε λειτουργία τα Σαββατοκύριακα και τις δημόσιες αργίες, μειώνεται αυτόματα η κατανόηση και η εύκολη διαχείριση τους κατά την επεξεργασία. Για αντιμετώπιση του προβλήματος αυτού λαμβάνονται υπόψιν από το σύστημα μόνο οι μέρες που πραγματοποιούνται συναλλαγές και το Χρηματιστήριο είναι ανοιχτό.

Επιπρόσθετα, μετά από επεξεργασία και αξιοποίηση βοηθητικών βιβλιοθηκών της Ργθηon, πραγματοποιήθηκε απεικόνιση των δεδομένων για ευκολότερη ανάλυση και κατανόηση τους. Πιο κάτω παρατίθεται δείγμα γραφικών παραστάσεων.



Σχήμα 4.5: Γραφική Παράσταση που απεικονίζει τις τιμές κλεισίματος των μετοχών των Αμερικάνων Εταιρειών Αερογραμμών American Airlines, Delta Airlines, United Airlines και Southwest Airlines



Σχήμα 4.6: Γραφική Παράσταση που απεικονίζει τις τιμές κλεισίματος των μετοχών των Αμερικάνων Εταιρειών που παρέχουν υπηρεσίες Streaming: Netflix, Roku και Spotify.

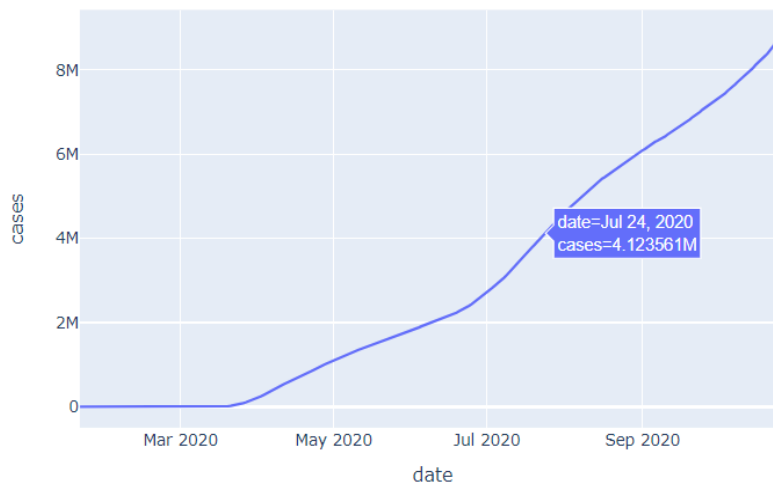
Κεφάλαιο 5

Ανάλυση Συναισθήματος και Covid-19

5.1.	Covid-19 και Οικονομία	47
5.2.	Επιλογή προς εξέταση εταιρειών	48
5.2.1.	Αερογραμμές	50
5.2.2.	Τουρισμός	50
5.2.3.	Ενοικιάσεις Αυτοκινήτων	51
5.2.4.	Υπηρεσίες Τηλεδιασκέψεων.....	51
5.2.5.	Γραμμές Κρουαζιέρας	51
5.2.6.	Υπηρεσίες Streaming.....	52
5.2.7.	Ηλεκτρονικά καταστήματα/Αλυσίδες Υπεραγορών	52
5.2.8.	Πλατφόρμες Κοινωνικής Δικτύωσης.....	53
5.2.9.	Υγεία	53
5.2.10.	Τεχνολογία	54
5.2.11.	Εστίαση	54

5.1. Covid-19 και Οικονομία

Στη δύσκολη χρονική περίοδο που διανύει η ανθρωπότητα με την παρουσία της πανδημίας του κορωνοϊού στις ζωές μας, παρατηρούνται σημαντικές μεταβολές στους τομείς των επιχειρήσεων και του ευρύτερου οικονομικού τομέα. Στην παρούσα διπλωματική εργασία, δραττόμενοι των εξελίξεων της επικαιρότητας εξετάστηκε κατά πόσο το μοντέλο ανάλυσης συναισθήματος που αναπτύχθηκε, μπορεί να παράγει αξιόπιστες πληροφορίες και συμπεράσματα για τη Χρηματιστηριακή συμπεριφορά εν μέσω της πανδημίας.



Σχήμα 5.1: Γραφική απεικόνιση των κρουσμάτων κορωνοϊού στις Η.Π.Α για την περίοδο συλλογής δεδομένων. Η γραφική παράσταση παράχθηκε με την αξιοποίηση της βιβλιοθήκης *plotly* της γλώσσας προγραμματισμού *Python*

5.2. Επιλογή προς εξέταση εταιρειών

Είναι γνωστό ότι, ο κάθε επιχειρηματικός τομέας της οικονομίας επηρεάστηκε διαφορετικά από τη πανδημία. Άλλωστε δεν υπήρξαν μονάχα πληγέντες, αλλά και κερδισμένες επιχειρήσεις, των οποίων οι υπηρεσίες κρίθηκαν απαραίτητες στο πλαίσιο της νέας πραγματικότητας που επέφερε ο κατ' οίκον περιορισμός και τα νέα μέτρα των κυβερνήσεων για αντιμετώπιση του Covid-19.

Δημιουργήθηκε λοιπόν μια λίστα με αντιπροσωπευτικό δείγμα εταιρειών από διάφορους τομείς της οικονομίας για να συλλεχθούν σχετικά δεδομένα από το Twitter και να πραγματοποιηθεί ανάλυση συναισθήματος.

Η μικρή ποσότητα διαθέσιμων δεδομένων για συγκεκριμένες εταιρείες οδήγησε στην αναπροσαρμογή των επιλογών, έτσι ώστε να ανακτηθεί ποσότητα δεδομένων που θα προσφέρει ικανοποιητική απόδοση κατά την επεξεργασία για την εξαγωγή χρήσιμων πληροφοριών.

Παρακάτω δίδεται η λίστα με τους επιχειρηματικούς τομείς, τις εταιρείες των οποίων οι μετοχές εξετάστηκαν από το μοντέλο για τον κάθε τομέα και τα σύμβολα αυτών στο Χρηματιστήριο. Στη συνέχεια, παρουσιάζονται οι τομείς και εταιρείες που συμπεριλήφθηκαν με σύντομη αιτιολόγηση για την κάθε επιλογή.

Επιχειρηματικοί Τομείς	Εταιρείες	Σύμβολα
Τεχνολογία	Apple Inc.	\$AAPL
	Google Group.	\$GOOG
Αερογραμμές	American Airlines	\$AAL
	Delta	\$DAL
	Southwest	\$LUV
	United	\$UAL
Τουρισμός	Booking	\$BKNG
	TripAdvisor	\$TRIP
	Hilton	\$HLT
	Hilton Grand Vacations	\$HGV
	Expedia	\$EXPE
	Marriott	\$MAR
Ενοικιάσεις Αυτοκινήτων	Hertz Global Holdings Inc	\$HTZ
Υπηρεσίες Τηλεδιασκέψεων	Zoom	\$ZM
	Microsoft	\$MSFT
Γραμμές Κρουαζιέρας	Royal Caribbean	\$RCL
	Carnival	\$CCL
Υπηρεσίες Streaming	Netflix	\$NFLX
	Roku	\$ROKU
	Spotify	\$SPOT
Ηλεκτρονικά καταστήματα/Αλυσίδες Υπεραγορών	Target	\$TGT
	Walmart	\$WMT
	Amazon	\$AMZN
	Ebay	\$EBAY
Πλατφόρμες Κοινωνικής δικτύωσης	Facebook	\$FB
	Snapchat	\$SNAP
	Twitter	\$TWTR
Υγεία	Regeneron Pharmaceuticals	\$REGN
	Gilead Sciences	\$GILD
	Inovio Pharmaceuticals	\$INO
	Moderna	\$MRNA
	Novavax	\$NVAX
Εστίαση	Starbucks Corp.	\$SBUX
	McDonald's Corp.	\$MCD

Πίνακας 5.1: Επιχειρηματικοί τομείς και εταιρείες των οποίων οι μετοχές εξετάστηκαν

5.2.1. Αερογραμμές

Η πανδημία και ο κάτοικόν περιορισμός αποτελέσαν τεράστια πληγή για τις αεροπορικές εταιρείες, καθώς οι μετακινήσεις από και προς ξένες χώρες ελαχιστοποιήθηκαν και σε κάποιες περιπτώσεις αεροδρόμια παρέμειναν κλειστά για αξιοσημείωτα χρονικά διαστήματα. Ακόμη και σε περιόδους που επανήλθε εν μέρη η κανονικότητα, υπήρξε και ακόμη υπάρχει επιφύλαξη μεταξύ των πολιτών στις μετακινήσεις με αεροπλάνο, λόγω της πολυκοσμίας σε αεροπλάνα και αεροδρόμια, καθώς και λόγω του αβέβαιου των συνθήκων της πανδημίας σε χώρες του εξωτερικού. Επιλέχθηκαν προς εξέταση, τέσσερις από τις μεγαλύτερες εταιρείες αερογραμμών των Η.Π.Α.

- American Airlines
- Delta Airlines
- Southwest Airlines
- United Airlines

5.2.2. Τουρισμός

Όπως ήταν άλλωστε αναμενόμενο, ο τομέας του Εγχώριου και Διεθνή Παγκόσμιου Τουρισμού πλήγηκε σε τεράστιο βαθμό από τις επιπτώσεις της πανδημίας. Ο εγκλεισμός, τα μέτρα των κυβερνήσεων και το υψηλό επίπεδο ρίσκου στις διακινήσεις, οδήγησαν μεγάλη μερίδα του πληθυσμού να ακυρώσει ή να μη σχεδιάσει καν διακοπές για την καλοκαιρινή περίοδο, που άλλοτε αποτελούσε το κύριο εισόδημα της βιομηχανίας του Τουρισμού. Επιλέχθηκαν ενδεικτικά: η δημοφιλής εταιρεία κρατήσεων χώρων διαμονής Booking.com, το διαδικτυακό ταξιδιωτικό γραφείο και μηχανή μετα-αναζήτησης Expedia, ο πιο διαδεδομένος ταξιδιωτικός οδηγός στον κόσμο TripAdvisor, καθώς επίσης και οι εταιρείες με ξενοδοχειακές μονάδες και εγκαταστάσεις σε όλο τον κόσμο: Hilton, Hilton Grand Holdings Inc, Marriott και Marriot Vacations. Ωστόσο, λόγω της πολύ χαμηλής ζήτησης για διακίνηση των μετοχών τους, δεν έγινε δυνατή η συλλογή μεγάλου αριθμού Tweets για όλες τις προ-αναφερθείσες εταιρείες.

5.2.3. Ενοικιάσεις Αυτοκινήτων

Συμπεριλήφθηκε στη λίστα, ο γνωστός κολοσσός στο χώρο των ενοικιάσεων αυτοκινήτων, Hertz Global Holdings Inc, καθώς λόγω των μειωμένων μετακινήσεων ο τομέας στον οποίο δραστηριοποιείται η επιχείρηση, υπήρξε από τους πλέον χαμένους τη επηρεαζόμενη χρονική περίοδο. Επιπρόσθετα, για την εν λόγω εταιρεία είναι γνωστό ότι πριν την έναρξη της πανδημίας, κήρυξε πτώχευση, στοιχείο που κρίθηκε επίσης ενδιαφέρον για την μελέτη της.

5.2.4. Υπηρεσίες Τηλεδιάσκεψων

Επιτακτική ανάγκη κατά την περίοδο της πανδημίας αποτέλεσαν οι υπηρεσίες τηλεδιάσκεψης. Οι ανάγκες για τηλεργασία και εξ αποστάσεως εκπαίδευση οδήγησαν εκατομμύρια ανθρώπους και εκατοντάδες χιλιάδες επιχειρήσεις να στραφούν στη λύση της διαδικτυακής τηλεδιάσκεψης για τη διεκπεραίωση των έργων τους, με αποτέλεσμα οι εταιρίες που παρέχουν τις εν λόγω υπηρεσίες να γνωρίσουν εξωφρενική ζήτηση και συγκροτούν ίσως τον πιο κερδισμένο οικονομικό τομέα της πανδημίας.

Επιλέχθηκαν, οι εταιρείες Zoom και η Microsoft για το εργαλείο της Microsoft Teams. Παρόλο, που η εταιρεία Microsoft εμπίπτει σε πληθώρα τεχνολογικών τομέων λόγω της ποικιλομορφίας των προϊόντων και υπηρεσιών που παρέχει, περιλήφθηκε στις υπό εξέταση εταιρείες, καθώς κατέχει μια από τις πλέον πολυσυζητημένες και ενδιαφέρουσες μετοχές του Χρηματιστηρίου του NASDAQ και θεωρείται ένας από τους 5 μεγαλύτερους τεχνολογικούς κολοσσούς στον πλανήτη.

5.2.5. Γραμμές Κρουαζιέρας

Παρόμοια με τις αεροπορικές εταιρείες, οι γραμμές κρουαζιέρας πλήγηκαν οικονομικώς από την πανδημία. Επιλέχθηκαν δύο από τις δημοφιλέστερες Αμερικανικές εταιρίες με Γραμμές Κρουαζιέρων:

- Royal Caribbean
- Carnival

Ωστόσο, δεν συλλέχθηκε ικανοποιητική ποσότητα δεδομένων από το Twitter, καθώς οι εν λόγω μετοχές δεν αποτελούν διαδομένη επιλογή για τους επενδυτές.

5.2.6. Υπηρεσίες Streaming

Η καθημερινότητα της καραντίνας και του εγκλεισμού έφεραν τον παγκόσμιο πληθυσμό αντιμέτωπο με μια πρωτόγνωρη πραγματικότητα με απρόσμενα άπλετο χρόνο στο σπίτι. Είναι πολλοί αυτοί που δραττόμενοι της “ευκαιρίας” ή έστω για να σπαταλήσουν ευχάριστα τον επιβεβλημένο ελεύθερο τους χρόνο, διέφυγαν στις υπηρεσίες Streaming για να παρακολουθήσουν τις αγαπημένες τους σειρές, ταινίες, ντοκιμαντέρ, ακόμη και για να ακούσουν περισσότερη μουσική της αρεσκείας τους. Στη λίστα λοιπόν, συμπεριλήφθηκε η μεγαλύτερη streaming υπηρεσία μουσικής Spotify και δύο από τις κυρίαρχες δυνάμεις στον χώρο των υπηρεσιών Streaming, οι Netflix και Roku.

5.2.7. Ηλεκτρονικά καταστήματα/Αλυσίδες Υπεραγορών

Η δυσκολία στις μετακινήσεις, το κλείσιμο καταστημάτων και η αδιάλειπτη και συνεχόμενη ανάγκη των καταναλωτών για προϊόντα, οδήγησαν στη στροφή προς τις διαδικτυακές αγορές όσο ποτέ άλλοτε.

Κυρίαρχες επιχειρήσεις στον τομέα των διαδικτυακών πωλήσεων είναι οι Amazon και Ebay, με την πρώτη να αποτελεί παγκόσμιο φαινόμενο στον κόσμο των επιχειρήσεων κυρίως την τελευταία δεκαετία και να κατέχει μια από τις δημοφιλέστερες επιλογές των επενδυτών στο Χρηματιστήριο του Nasdaq, όντας μια εκ των “Big Five” Τεχνολογικών Εταιρειών.

Ακόμη και γνωστές αλυσίδες υπεραγορών, όπως οι Walmart και Target που στηρίζουν την πλειονότητα των κερδών τους σε πωλήσεις με φυσική παρουσία στα καταστήματα τους, αναπόφευκτα επένδυσαν στο διαδικτυακό εμπόριο, το οποίο αποτέλεσε την βασική πηγή εσόδων τους εν μέσω της πανδημίας.

Αποτελεί λοιπόν ενδιαφέρον έργο η εξέταση της συμπεριφοράς των τεσσάρων προαναφερθέντων μετοχών για αυτήν τη χρονική περίοδο.

5.2.8. Πλατφόρμες Κοινωνικής Δικτύωσης

Η σημαντικότητα και η επιρροή των μέσων κοινωνικής δικτύωσης είναι αδιαμφισβήτητη στις μέρες μας. Όταν λοιπόν, σε αυτήν την ήδη ακατάπαυστη δυναμική ανάπτυξης των μεσών, προστίθεται η ύπαρξη περισσότερου ελεύθερου χρόνου από την πλειοψηφία του παγκόσμιου πληθυσμού, αυτονόητα προκύπτει ότι οι πλατφόρμες κοινωνικής δικτύωσης μπορούν εύκολα να χαρακτηριστούν από τους “κερδισμένους” της πανδημίας. Στη λίστα συμπεριλήφθηκε ο Αμερικανικός όμιλος κοινωνικών μέσων Facebook, επίσης μέλος των “Big Five”, ο οποίος κατέχει μεταξύ άλλων τις πλατφόρμες κοινωνικής δικτύωσης Instagram, Messenger, WhatsApp και φυσικά την ομώνυμη πλατφόρμα Facebook. Επιπρόσθετα, στη λίστα συμπεριλήφθηκαν οι γνωστές εταιρείες με πλατφόρμες κοινωνικής δικτύωσης Snapchat και Twitter.

5.2.9. Υγεία

Ο τομέας της υγείας βρίσκεται στο προσκήνιο περισσότερο από ποτέ, καθώς είναι καθημερινές και ραγδαίες οι εξελίξεις για την πορεία της πανδημίας και είναι οι επαγγελματίες υγείας οι σύγχρονοι ήρωες του 2020. Είναι συνεχόμενη λοιπόν και η προσπάθεια από την παγκόσμια επιστημονική ιατρική κοινότητα για εξεύρεση εμβολίων ή φαρμάκων ικανών να καταστήσουν δυνατή την ίαση του Covid-19.

Γνωστές ιδιωτικές εταιρείες πραγματοποιούν υπό έρευνα θεραπείες και κλινικές δοκιμές. Σε κάθε νέο βήμα προς τη θεραπεία του κορωνοϊού είναι τεράστιο το παγκόσμιο ενδιαφέρον και οι επενδυτές δεν σπαταλούν την ευκαιρία να επενδύσουν στην εταιρεία που πιστεύουν πως θα κερδίσει τη μάχη για έλεγχο της πανδημίας.

Όσο η μάχη εξελίσσεται σε μαραθώνιο, η παγκόσμια προσοχή παραμένει σε υψηλά επίπεδα και αποτελεί εξαιρετικά ενδιαφέρον έργο, η μελέτη πέντε εκ των εταιρειών που ασχολούνται με το εν λόγω εγχείρημα και βρίσκονται στο επίκεντρο της επικαιρότητας. Πρόκειται για τις:

- Regeneron Pharmaceuticals
- Gilead Sciences

- Inovio Pharmaceuticals
- Moderna
- Novavax

5.2.10. Τεχνολογία

Ήδη συμπεριλήφθηκε πληθώρα εταιρειών που πρωτοπορούν στον Τεχνολογικό τομέα. Συμπληρώνοντας την πεντάδα των “Big 5” Τεχνολογικών κολοσσών, συλλέχθηκαν δεδομένα για την ίσως πιο επιτυχημένη επιχείρηση Παγκοσμίως, η οποία κατόρθωσε να γίνει η πρώτη Αμερικανική Εταιρεία με κοστολόγηση άνω των 2 τρισεκατομμυρίων, την Apple. Ως επίσης, η λίστα περιλαμβάνει την πολυεθνική εταιρεία τεχνολογίας Google, η οποία πρωταγωνίστησε και συνεχίζει να πρωταγωνιστεί την τελευταία εικοσαετία με την συνεχή της πρόοδο.

5.2.11. Εστίαση

Ο κατ’ οίκον περιορισμός επηρέασε σε σημαντικό βαθμό τη λειτουργία των επιχειρήσεων στον τομέα της εστίασης, καθώς αναγκάστηκαν να παραμείνουν κλειστά και να διακόψουν τη λειτουργία τους για σημαντικά χρονικά διαστήματα. Εξετάστηκαν λοιπόν οι δύο δημοφιλέστερες μετοχές που αφορούν την εστίαση, οι οποίες ανήκουν στην μεγαλύτερη αλυσίδα εστιατορίων ταχείας εστίασης στον κόσμο, την McDonald’s και την πολυεθνική εταιρεία καφέ Starbucks.

Κεφάλαιο 6

Αξιολόγηση Αποτελεσμάτων

6.1.	Πειράματα και Αποτελέσματα για Tweets	55
6.1.1.	Κατηγοριοποίηση Αναρτήσεων με βάση το Συναίσθημα.....	56
6.1.2.	Υποκειμενικότητα.....	64
6.1.3.	Συχνότητα Εμφάνισης Λέξεων.....	65
6.1.4.	Ανάλυση αποτελεσμάτων για Hashtags.....	70
6.1.5.	Ανάλυση συσχετίσεων	76
6.2.	Πειράματα και Αποτελέσματα για Μοντέλο Πρόβλεψης	77
6.2.1.	Επιλογή μετρικών.....	77
6.2.2.	Ανάλυση αποτελεσμάτων	79

6.1. Πειράματα και Αποτελέσματα για Tweets

Στο παρόν υποκεφάλαιο παρουσιάζονται αναλυτικά τα αποτελέσματα από την ανάλυση και αξιολόγηση της γνώσης που εξάχθηκε, μετά την ανάλυση συναισθήματος και τον υπολογισμό μετρικών, στο σύνολο δεδομένων που συλλέγηκε και αφορά εταιρείες από διαφορετικούς επιχειρηματικούς τομείς της οικονομίας, κατά την περίοδο του πρώτου κύματος εξάπλωσης της πανδημίας του Covid-19. Η αξιολόγηση χωρίζεται σε μικρότερα υποκεφάλαια, το καθένα από τα οποία αναφέρεται σε ανάλυση αποτελεσμάτων διαφορετικής φύσης και για τα οποία παρουσιάζεται ενδεικτικό υποσύνολο πινάκων και γραφικών παραστάσεων, για περαιτέρω κατανόηση της εξαγόμενης γνώσης.

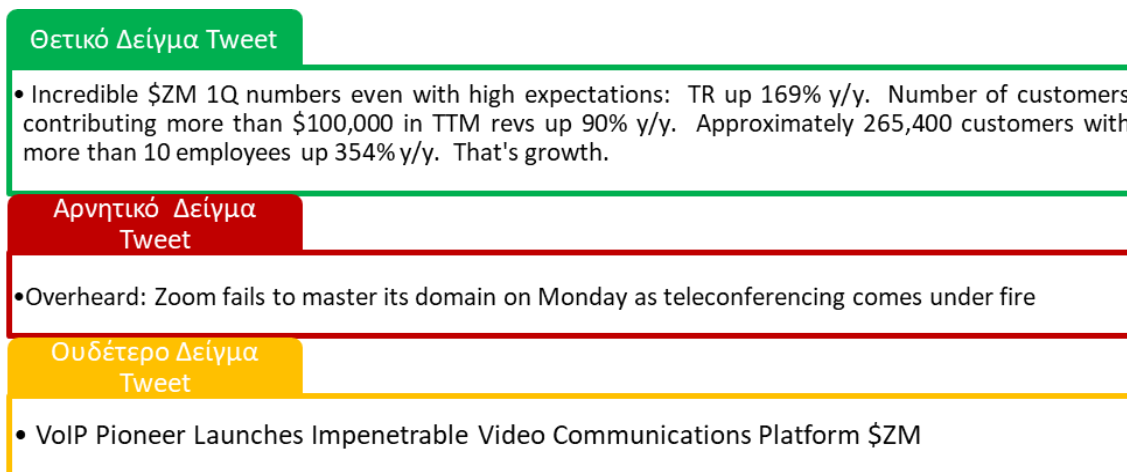
6.1.1. Κατηγοριοποίηση Αναρτήσεων με βάση το Συναίσθημα

Στο παρόν υποκεφάλαιο παρουσιάζεται η ανάλυση των αποτελεσμάτων της κατηγοριοποίησης των Tweets, με βάση το συναίσθημα, για τους 11 διαφορετικούς οικονομικούς τομείς για τους οποίους συλλέχθηκαν δεδομένα (βλ. Κεφάλαιο 5).

Τομέας Τηλεδιασκέψεων

Παρατηρώντας τα ποσοστά των κατηγοριοποιημένων αναρτήσεων στο σύνολο δεδομένων που αφορά εταιρείες οι οποίες δραστηριοποιούνται στον τομέα των τηλεδιασκέψεων, γίνεται αντιληπτό ότι η πλειοψηφία των Tweets κατηγοριοποιήθηκαν από τον αναλυτή συναισθήματος ως θετικά. Συγκεκριμένα, 44.6% των Tweets κατηγοριοποιήθηκαν ως θετικά, 35.5% ως ουδέτερα, ενώ το 15.5% ως αρνητικά.

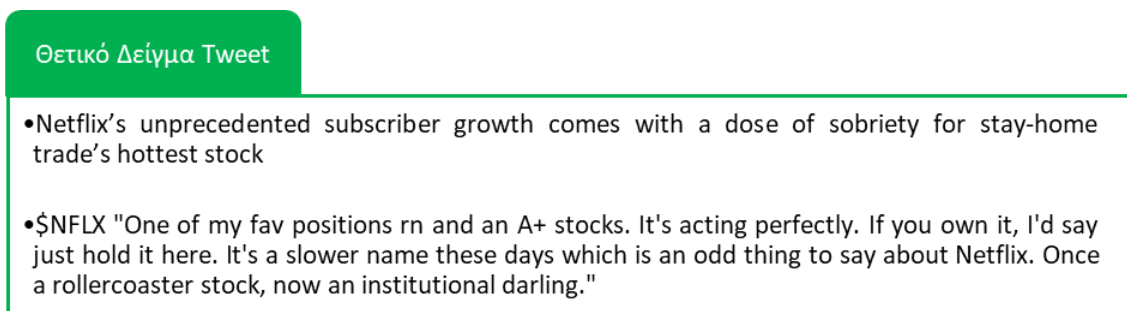
Η επικράτηση του θετικού συναισθήματος στο συγκεκριμένο σύνολο δεδομένων αποτελεί αποτέλεσμα το οποίο συνάδει με την επενδυτική διάθεση της περιόδου, λόγω της εξωφρενικής άνοδου των τιμών των μετοχών, τόσο της εταιρείας Microsoft, όσο κυρίως της εταιρείας Zoom. Επομένως, ο ενθουσιασμός των επενδυτών, οι θετικές εντυπώσεις και ειδήσεις για την πορεία των μετοχών, αναμενόμενα μεταφέρθηκαν στα μέσα κοινωνικής δικτύωσης και ορθά εντοπίστηκαν από τον αναλυτή συναισθήματος. Τα αρνητικά Tweets που ορθά κατηγοριοποιήθηκαν από το σύστημα, αναφέρονται κυρίως σε ημερήσιες πτώσεις των τιμών των μετοχών ή αρνητικές ειδήσεις, για μικρά χρονικά διαστήματα τα οποία εκφράστηκαν από τους χρήστες του Twitter.



Σχήμα 6.1: Απεικόνιση δείγματος θετικά, αρνητικά και ουδέτερα κατηγοριοποιημένων Tweets που αφορούν την εταιρεία Zoom (Θετικό Tweet από Stephani Link στις 02/06/2020, Αρνητικό Tweet από WSJ Markets στις 02/08/2020, Ουδέτερο Tweet από Benzinga στις 11/04/2020)

Τομέας Υπηρεσιών Streaming

Παρόμοια εικόνα καταγράφηκε από την ανάλυση συναισθήματος στο σύνολο δεδομένων που αφορά εταιρείες οι οποίες παρέχουν υπηρεσίες Streaming. Συγκεκριμένα, 49.4% των Tweets κατηγοριοποιήθηκαν ως θετικά, 37.3% ως ουδέτερα, ενώ το 13.3% ως αρνητικά. Πρόκειται για ακόμη ένα λογικό αποτέλεσμα που αντικατοπτρίζει την αντίθεση στη διάθεση γύρω από την θεματολογία των υπηρεσιών Streaming σε σχέση με την ευρύτερη αρνητική παγκόσμια κοινωνική διάθεση λόγω της πανδημίας κατά την περίοδο του εγκλεισμού.

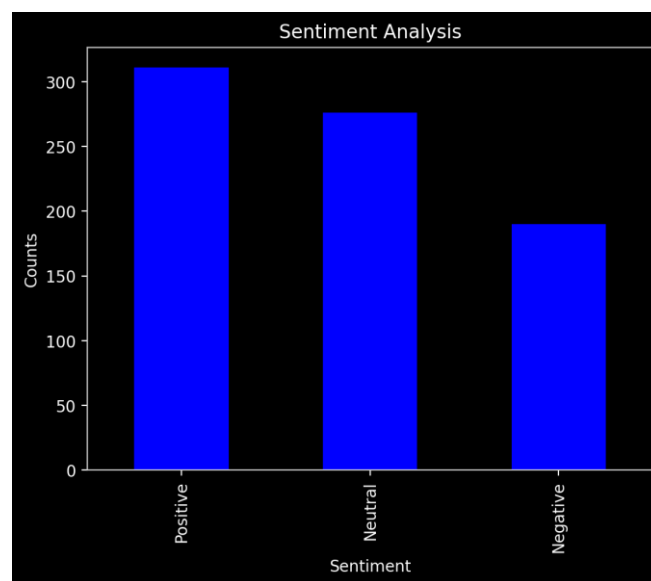


Σχήμα 6.2: Απεικόνιση δείγματος θετικά κατηγοριοποιημένων Tweets που αφορούν την εταιρεία Netflix (Tweet από WSJ Markets στις 21/04/2020, Tweet από IBDinvestors στις 06/05/2020)

Τεχνολογικός Τομέας

Οι μεγάλες τεχνολογικές εταιρείες, κατά την περίοδο της έξαρσης του κορωνοϊού, όχι μόνο δεν επηρεάστηκαν αρνητικά, αλλά σημείωσαν πρωτόγνωρη οικονομική πρόοδο. Αξιολογώντας τα αποτελέσματα από την ανάλυση συναισθήματος στο

σύνολο δεδομένων που αφορά τους διεθνείς κολοσσούς Google και Apple, το θετικό κλίμα που συνόδευε τα οικονομικά επιτεύγματα των εταιρειών, εντοπίστηκε στις αναρτήσεις στο Twitter και μεταφράστηκε σε ένα ποσοστό της τάξης του 40% για τις κατηγοριοποιημένες με θετική πολικότητα. Ακολουθούν οι κατηγοριοποιήσεις για τα ουδέτερα και αρνητικά Tweets με ποσοστά 35.5% και 24.5% αντίστοιχα. Το μεγάλο ποσοστό Tweets που κατηγοριοποιήθηκαν ως ουδέτερα οφείλεται κατά κύρια ομολογία στο μεγάλο αριθμό αναρτήσεων ειδησεογραφικού χαρακτήρα που αναφέρουν πληροφοριακά τις τιμές των μετοχών για την συγκεκριμένη μέρα.



Σχήμα 6.3: Γράφημα απεικόνισης της ποσότητας των κατηγοριοποιημένων αναρτήσεων για εταιρείες Apple και Google

Τομέας Αερογραμμών

Αξιολογώντας το σύνολο δεδομένων που αφορά Αμερικανικές εταιρείες αερογραμμών, παρατηρείται ότι σχεδόν κατά το ήμισυ, τα Tweets είναι θετικά. Συγκεκριμένα 46% των Tweets κατηγοριοποιήθηκαν ως θετικά, 30.6% ως ουδέτερα και 23.4% ως αρνητικά. Τα αποτελέσματα αυτά κρίνονται λογικά, καθώς συσχετίζονται άμεσα με τη θετική γενική άνοδο των τιμών των μετοχών για τις εταιρείες United Airlines, Delta Airlines, American Airlines και Southwest Airlines για την περίοδο συλλογής δεδομένων. Το πρώτο Tweet το οποίο συλλέχθηκε, από το χρονολόγιο των λογαριασμών που αξιοποιήθηκαν, δημοσιεύτηκε στις 24/04/2020. Γίνεται εύκολα αντιληπτό λοιπόν, ότι η περίοδος της εξωφρενικής πτώσης των

μετοχών αυτών, έλαβε τέλος πολύ πριν τις 24/04/2020 όπου και η διακύμανση τους ξεκίνησε σταθερή και σταδιακή, μικρή άνοδο.

Θετικό Δείγμα Tweet

- Major Airlines Reach Deal With Treasury Over Participation In Bailout Program \$AAL \$DAL \$UAL \$LUV \$SAVE \$JBLU <https://t.co/FYjWhiwUMr>
- @Lebeaucarnews Nice news from \$LUV.. the airline i would buy!
- One airline stock is 'simply the best' once air travel demand normalizes, trader says \$JETS \$DAL \$LUV \$UAL \$AAL \$ALK \$JBLU (via @TradingNation) <https://t.co/9pESp2Pzrm>

Σχήμα 6.4: Απεικόνιση δείγματος θετικά κατηγοριοποιημένων Tweets που αφορούν αερογραμμές (Tweet από Benzinga στις 15/04/2020, Tweet από Jim Cramer στις 19/08/2020, Tweet από CNBC στις 25/08/2020)

Τομέας Ενοικίασης Αυτοκινήτων

Η έναρξη της συλλογής των δεδομένων ήταν η 1η Μαρτίου του 2020, με σκοπό την επικέντρωση στο πρώτο κύμα της πανδημίας του Covid-19. Της πανδημίας προηγήθηκε η πτώχευση της εταιρείας Hertz τον Φεβρουάριο, που επέφερε αναμενόμενα καταστροφικές συνέπειες για την τιμή της μετοχής της. Ως εκ τούτου, κατά την περίοδο συλλογής δεδομένων η μετοχή της Hertz ήδη κυμαινόταν σε χαμηλά επίπεδα λόγω της πτώχευσης όπου και παραμένει μέχρι σήμερα. Η αρνητική διάθεση, αποτυπώθηκε στα αποτελέσματα με την πλειοψηφία των Tweets να κατηγοριοποιούνται αρνητικά από τον αναλυτή συναισθήματος. Συγκεκριμένα, 52.9%, των αναρτήσεων χαρακτηρίστηκαν ως αρνητικές.

Το μικρό δείγμα των αναρτήσεων που συλλέχθηκαν για την συγκεκριμένη εταιρεία οδήγησαν στην ενίσχυση της σημαντικότητας κάποιων γεγονότων. Πιο συγκεκριμένα, κατά την περίοδο Ιουνίου, η γενικότερη άνοδος του Χρηματιστηρίου (Bullish Period) επηρέασε θετικά και την εν λόγω μετοχή που σημείωσε σημειωτέα, αλλά όχι σημαντική αύξηση για ορισμένες μέρες, μέχρι να επανέλθει στα ίδια χαμηλά επίπεδα. Ως εκ τούτου, συλλέχθηκαν δεδομένα από αυτή την περίοδο και ορθά εξάχθηκε η θετική πολικότητα των Tweets από το σύστημα για το 25.5% τους.

Συνοψίζοντας, η εταιρεία Hertz επηρεάστηκε τόσο πολύ από την δική της κατάρρευση που δεν υπήρξε περιθώριο για περαιτέρω μείωση, λόγω της πανδημίας και η διάθεση

γύρω από την μετοχή της ήταν πολύ αποθαρρυντική ως προς την αγορά της, κάτι που αποτυπώνεται στα ποσοστά κατηγοριοποίησης, που εξάχθηκαν από το σύστημα.

Ουδέτερο Δείγμα Tweet										
• Hertz	Stock	Jump	115%,	Erases	Post-Bankruptcy	Losses	As	Travel	Revives	\$HTZ
• This	market	is	so	bullish	even	bankrupt	companies	are	rallying.	\$HTZ

Σχήμα 6.5: Απεικόνιση δείγματος ουδέτερα κατηγοριοποιημένων Tweets που αφορούν την εταιρεία Hertz (Tweet από Benzinga στις 09/06/2020, Tweet από Sven Henrich στις 05/06/2020)

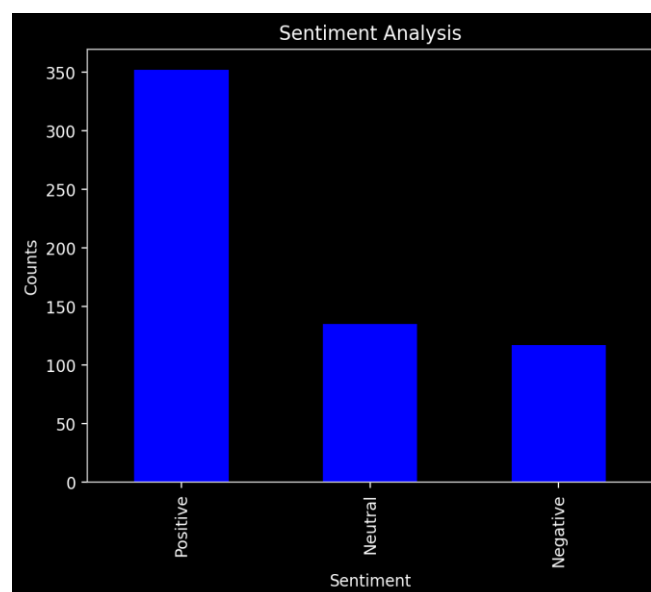
Τομείς Τουρισμού και Γραμμών Κρουαζιέρας

Σε ότι αφορά τα σύνολα δεδομένων που αφορούν τον τουρισμό και τις εταιρείες με κρουαζιερόπλοια, από το στάδιο συλλογής δεδομένων, διαπιστώθηκε ότι λόγω του μικρού όγκου συναλλαγών τους δεν υπάρχουν αρκετά δεδομένα και σπάνια αναρτώνται Tweets από λογαριασμούς για αυτές. Ωστόσο, ο μικρός αριθμός αναρτήσεων έκανε δυνατή τη χειροκίνητη αξιολόγηση των αποτελεσμάτων και εντοπίστηκαν ενδιαφέροντα στοιχεία όχι μόνο για την ποιότητα των αναρτήσεων που συλλέχθηκαν, αλλά και για την απόδοση του συστήματος κατά την ανάλυση τους κάτι που επέτρεψε την αξιολόγηση από μια διαφορετική σκοπιά. Συγκεκριμένα, έπειτα από χειροκίνητη αξιολόγηση διαπιστώθηκε ότι το σύστημα κατηγοριοποίησε ορθά 92% των αναρτήσεων, έχοντας απώλειες σε αναρτήσεις που κατηγοριοποιήθηκαν ως ουδέτερες, ενώ κρίθηκαν αρνητικές, κάτι που αποτελεί λογικό φαινόμενο, ειδικότερα όταν το σύνολο των λέξεων στα Tweets είναι μικρό.

Αξιολογώντας επίσης το μικρό σύνολο δεδομένων που αφορά τις εταιρείες με κρουαζιερόπλοια σε ότι αφορά την ποιότητα του, διαπιστώθηκε ότι το υψηλό ποσοστό των ουδέτερων Tweets που ανέρχεται στο 60% σύμφωνα με το σύστημα, εντοπίστηκε ορθά, καθώς οι πλείστες αναρτήσεις που μιλούν για τις συγκεκριμένες εταιρείες, είτε αναφέρουν το όνομα της επιχείρησης και παραθέτουν απλά ποσοστά αύξησης ή μείωσης της τιμής των μετοχών τους, είτε τις αναφέρουν σε λίστες με πολλές άλλες μετοχές που περιέχουν συνδέσμους που παραπέμπουν σε άλλες ιστοσελίδες.

Τομέας Ηλεκτρονικών Καταστημάτων

Όπως ήταν άλλωστε αναμενόμενο η συντριπτική πλειοψηφία των αναρτήσεων για το σύνολο δεδομένων των ηλεκτρονικών καταστημάτων κατηγοριοποιήθηκε με θετική πολικότητα. Η σημαντική άνοδος των τιμών των μετοχών για τις εταιρείες του τομέα αυτού, αντικατοπτρίζεται στα αποτελέσματα από την ανάλυση συναισθήματος με ένα ποσοστό της τάξης του 58.3% για τα κατηγοριοποιημένα θετικά Tweets. Δεν αποτελεί τυχαίο γεγονός ότι η περίοδος συλλογής δεδομένων συμπίπτει με την περίοδο που οι περισσότερες χώρες πραγματοποίησαν lockdown και οι διαδικτυακές αγορές αποτελούσαν τον μοναδικό τρόπο αγορών ανά το παγκόσμιο. Επομένως, λόγω της άμεσης επιρροής των χρηστών, από την δική τους πραγματικότητα για τις υπηρεσίες των ηλεκτρονικών καταστημάτων, έτειναν να αναρτούν θετικά Tweets για την εν λόγω θεματολογία. Πληροφοριακά αναφέρεται ότι το 19.4% των αναρτήσεων στο Twitter κατηγοριοποιήθηκαν ως αρνητικά και 22.3% ως ουδέτερα.



Σχήμα 6.6: Γράφημα απεικόνισης της ποσότητας των κατηγοριοποιημένων αναρτήσεων για εταιρείες με ηλεκτρονικά καταστήματα

Τομείς Μέσων Κοινωνικής Δικτύωσης, Υγείας και Εστίασης

Σε αρκετά από τα σύνολα δεδομένων, μεγάλο ποσοστό αναρτήσεων χαρακτηρίστηκε ως ουδέτερο από το σύστημα, στοιχείο το οποίο δεν αποτελεί τυχαίο γεγονός.

Παρακάτω παρατίθενται πιθανοί λόγοι για τους οποίους μπορεί να παρατηρηθεί το φαινόμενο αυτό.

- Tweets πληροφοριακού χαρακτήρα είτε που αναφέρουν ειδήσεις, είτε μεταβολές στις τιμές των μετοχών, όπου ορθά κατηγοριοποιούνται ως ουδέτερα, καθώς δεν εμπεριέχουν συναίσθημα.
- Μικρά σε έκταση κειμένου Tweets, όπου συνήθως παραπέμπουν σε συγκεκριμένες σελίδες ή περιέχουν οπτικοακουστικό υλικό.
- Μικρά σε έκταση κειμένου Tweets που παραθέτουν σύντομα μια άποψη. Παρόλα αυτά, έπειτα από τη διαδικασία προ-επεξεργασίας κειμένου και την αφαίρεση των stopwords, το κείμενο είναι πολύ σύντομο για να καταστεί δυνατή η εξεύρεση συναισθήματος από τον αναλυτή Vader.

Ωστόσο, το γεγονός ότι μεγάλο μέρος των Tweets είναι ουδέτερα δεν επηρεάζει την διαδικασία εξαγωγής συμπερασμάτων, καθώς η διαφορά στα ποσοστά των υπολοίπων κατηγοριοποιημένων αναρτήσεων, σε συνδυασμό με την παρατήρηση του συνόλου δεδομένων ποιοτικά, μπορεί να παρέχει ενδιαφέρουσες πληροφορίες. Χαρακτηριστικό παράδειγμα αποτελεί το σύνολο δεδομένων που αφορά μέσα κοινωνικής δικτύωσης, όπου λόγω της ύπαρξης αρκετά μεγάλης ποσότητας Tweets που αναφέρονται συλλογικά σε εταιρείες και παραθέτουν ουδέτερες πληροφορίες, το 36.8% των Tweets κατηγοριοποιήθηκαν ως ουδέτερα. Παρόλα αυτά, τα θετικά Tweets καταλαμβάνουν ένα εξίσου σημειωτέο ποσοστό της τάξης του 33.9%, δίνοντας έτσι την αίσθηση ότι το σύνολο των αποτελεσμάτων τείνει περισσότερο προς την θετική πολικότητα, που συσχετίζεται με την άνοδο των τιμών των μετοχών των εταιρειών Facebook, Snapchat και Twitter κατά την περίοδο συλλογής δεδομένων.

Παρόμοια κατάσταση παρατηρήθηκε κατά την ανάλυση των αποτελεσμάτων από το σύνολο δεδομένων που αφορά την υγεία. Στο συγκεκριμένο σύνολο δεδομένων το 47.5% των Tweets κατηγοριοποιήθηκαν ως ουδέτερα, κυρίως για το λόγο ότι πολλά από τα Tweets ανέφεραν πληροφοριακά σύντομες ειδήσεις για την πρόοδο των κλινικών δοκιμών ή την πορεία των ερευνών σχετικά με το εμβόλιο κατά του Covid-19. Ωστόσο, η μεγάλη διαφορά που παρατηρείται από την κατηγοριοποίηση των 37% των αναρτήσεων ως θετικές και το 15.5% ως αρνητικές, επιτρέπουν την εξαγωγή του

συμπεράσματος ότι πρόκειται για μια θετική χρονική περίοδο για τις εταιρείες του εν λόγω τομέα.

Θετικό Δείγμα Tweet

- Moderna Rallies On Positive Phase 1 Coronavirus Vaccine Data, Aims For Phase 3 Trial In July \$MRNA <https://t.co/V6YhMB3Zcg>
- These 6 Coronavirus Vaccine Candidates Are The Likeliest To Succeed, Says Morgan Stanley \$MRNA \$SNY \$GSK \$JNJ \$PFE \$BNTX <https://t.co/vil4Awmj5E>

Ουδέτερο Δείγμα Tweet

- Gilead Reports Q1 Earnings Beat, Anticipates Short-Term Impact From Coronavirus \$GILD <https://t.co/JXPSZVJJvK>

Σχήμα 6.7: Απεικόνιση δείγματος ουδέτερα και θετικά κατηγοριοποιημένων Tweets που αφορούν τον τομέα της Υγείας (Θετικό Tweet από Bezinga στις 18/05/2020, Θετικό Tweet από από Bezinga στις 18/05/2020, Ουδέτερο Tweet από Benzinga, 30/04/2020)

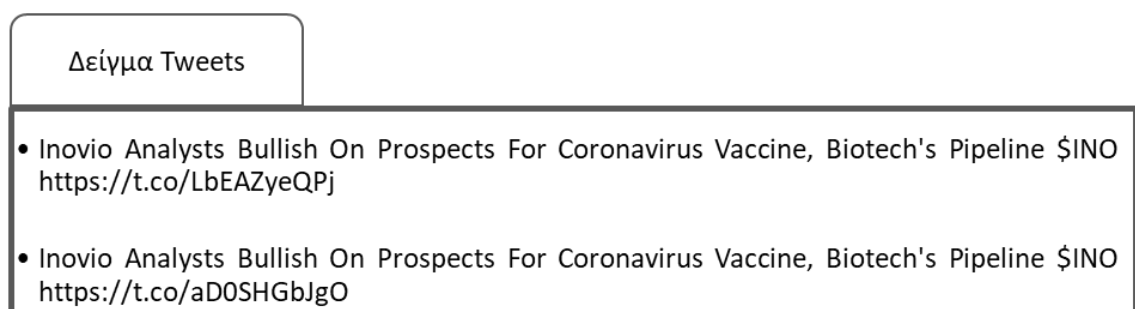
Αντίθετα με τις δύο πιο πάνω περιπτώσεις, η γενική εικόνα η οποία αφήνει η αξιολόγηση της ανάλυσης συναισθήματος στο σύνολο δεδομένων που αφορά τους χώρους εστίασης, αναμενόμενα παραπέμπει σε αρνητική πολικότητα. Συγκεκριμένα, το 46.2% των Tweets κατηγοριοποιήθηκαν ως ουδέτερα, το 32.1% ως αρνητικά και το 21.8% ως θετικά, στοιχείο που επιβεβαιώνει την αρνητική κοινωνική διάθεση γύρω από το κλείσιμο των χώρων εστίασης και συσχετίζεται με τη συνεχή πτώση των τιμών των μετοχών για τις εταιρείες Starbucks και Mcdonalds, λόγω της έξαρσης της πανδημίας. Παρόμοια, αλλά αντίθετα με τις ηλεκτρονικές αγορές, λόγω της αρνητικής επίδρασης στον τρόπο ζωής του πληθυσμού από το κλείσιμο των χώρων εστίασης, έτειναν να δημοσιεύονται αρνητικές σε συναισθήμα αναρτήσεις γύρω από τη θεματολογία, κατά την περίοδο συλλογής δεδομένων.

Αρνητικό Δείγμα Tweet

- Starbucks sales are still getting crushed by coronavirus
- McDonald's, Apple Further Delay Reopening Due To Coronavirus Cases Spike

Σχήμα 6.8: Απεικόνιση δείγματος αρνητικά κατηγοριοποιημένων Tweets που αφορούν τον τομέα της Εστίασης (Tweet από Stocktwits στις 29/07/2020, Tweet από Bezinga στις 02/07/2020)

Μια επιπρόσθετη παρατήρηση για το περιεχόμενο των αναρτήσεων που συλλέχθηκαν, αποτελεί η αναδημοσίευση τους από τους λογαριασμούς σε διαφορετικές ημερομηνίες ή ακόμη και την ίδια ημέρα, με την συμπερίληψη διαφορετικού οπτικοακουστικού υλικού ή υπερσυνδέσμων στο κείμενο τους. Οι αναρτήσεις αυτές κρατήθηκαν, καθώς εκτός από το γεγονός ότι αποτελούν διαφορετικές αναρτήσεις από τις πρωτότυπες, υποδηλώνουν τη διατήρηση της ίδιας κατάστασης και κατά συνέπεια υποδηλώνουν την αμετάβλητη κοινωνική διάθεση, γύρω από τη θεματολογία που αφορά τα εκάστοτε προς ανάλυση δεδομένα.



Σχήμα 6.9: Παράδειγμα αναρτήσεων με το ίδιο ακριβώς κείμενο για την εταιρεία Inovio, που παραπέμπουν σε διαφορετικούς συνδέσμους (Tweet από Bezinga στις 29/04/2020, Tweet από Bezinga στις 12/05/2020)

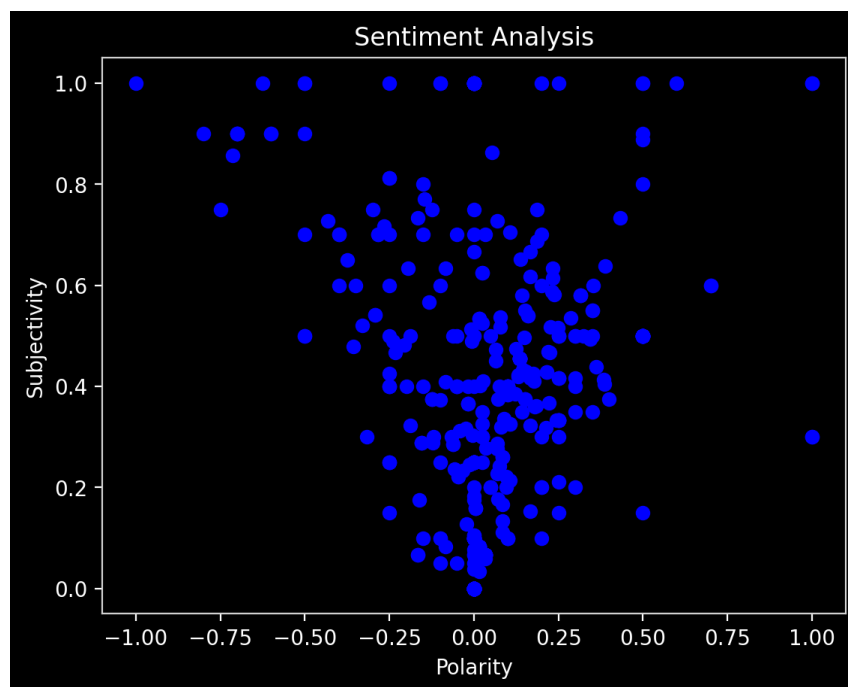
6.1.2. Υποκειμενικότητα

Παρατηρώντας τις γραφικές παραστάσεις που απεικονίζουν την ανάλυση συναισθήματος συναρτήσει της υποκειμενικότητας, αποκτάται μια γενική εικόνα της συσχέτισης των δύο μετρικών στα σύνολα δεδομένων. Η μορφή σχεδόν όλων των γραφικών παραστάσεων που παράχθηκαν είναι η επιθυμητή, καθώς η πλειοψηφία των σημείων βρίσκεται στο κάτω μέρος των γραφημάτων μέχρι το μέσο τους, στοιχείο που υποδεικνύει ότι η υποκειμενικότητα των αναρτήσεων κυμαίνεται σε χαμηλά επίπεδα.

Σε όλες τις γραφικές παραστάσεις ανεξαρτήτως, συναντώνται outliers με υψηλή υποκειμενικότητα η οποία συνήθως δεν συνοδεύεται και με κάποιο ξεκάθαρο συναίσθημα, καθώς το συναίσθημα των outliers μπορεί να είναι είτε θετικό, είτε αρνητικό, αλλά σπανιότερα ουδέτερο. Στον αντίποδα, οι αναρτήσεις με πολύ χαμηλή βαθμολογία υποκειμενικότητας στις πλείστες των περιπτώσεων έχουν ουδέτερο

συναίσθημα, στοιχείο το οποίο είναι αναμενόμενο, καθώς Tweets που δεν έχουν συναίσθημα, όπως για παράδειγμα ειδήσεις, παραθέτουν αντικειμενικά πληροφορίες, χωρίς ενδείξεις αυξημένης υποκειμενικότητας.

Η συγκεκριμένη γραφική παράσταση αξιοποιείται κυρίως ως μέθοδος αξιολόγησης του συνόλου δεδομένων, όπου σε περίπτωση μεγαλύτερης ποσότητας outliers πιθανότατα κάποιο σύνολο δεδομένων θα μπορούσε να απορριφθεί, καθώς υψηλά επίπεδα υποκειμενικότητας είναι ανεπιθύμητα και υποβαθμίζουν την εγκυρότητα του συνόλου δεδομένων. Το φαινόμενο αυτό δεν παρατηρήθηκε στα σύνολα δεδομένων της παρούσας εργασίας, κυρίως λόγω της συλλογής Tweets από γνωστούς λογαριασμούς στο επενδυτικό κόσμο που παραθέτουν περισσότερο γεγονότα παρά υποκειμενικές απόψεις.



Σχήμα 6.10: Γραφική Παράσταση απεικόνισης της υποκειμενικότητας συναρτήσει της πολικότητας των Tweets για το σύνολο δεδομένων των μέσων κοινωνικής δικτύωσης. Στο πάνω μέρος παρατηρείται η ύπαρξη outliers, ενώ οι πλείστες αναρτήσεις έχουν υποκειμενικότητα με βαθμολόγηση μικρότερη του 0.5, κάτι το οποίο είναι επιθυμητό

6.1.3. Συχνότητα Εμφάνισης Λέξεων

Πιο κάτω περιγράφονται αναλυτικά αποτελέσματα για ορισμένα από τα σύνολα δεδομένων τα οποία εξετάστηκαν για να υποστηριχθεί η εγκυρότητα του εντοπισμού γεγονότων από το σύστημα, καθώς επίσης και να δοθεί μια εμπεριστατωμένη

προσέγγιση για το πως το εν λόγω στατιστικό μπορεί να υποβοηθήσει τους επενδυτές κατά την αναζήτηση επενδυτικών συμβουλών.

Γενικότερα, στα αποτελέσματα όλων των συνόλων δεδομένων, διαπιστώθηκε η μεγάλη συχνότητα εμφάνισης ορολογιών οι οποίες χρησιμοποιούνται στο Χρηματιστήριο, στοιχείο που είναι λογικό και επιβεβαιώνει ότι τα συλλεγμένα Tweets εμπύπτουν στο πεδίο ενδιαφέροντος της μελέτης. Τέτοιες λέξεις είναι: “stocks”, “market”, “buy”, “sell”, “rise”.

Μελετώντας την συχνότητα εμφάνισης λέξεων για το σύνολο δεδομένων του Τεχνολογικού τομέα, παρατηρείται, μέσω και του γραφήματος, ότι εντοπίστηκαν πολλά σημαντικά γεγονότα της περιόδου για την οποία συλλέχθηκαν αναρτήσεις για τις δύο εταιρείες Google και Apple.

Αρχίζοντας από τα γεγονότα που εντοπίστηκαν και ενδιαφέρουν από κοινού τις δύο επιχειρήσεις, τόσο η λέξη “earnings” που εμφανίστηκε αρκετές φορές, όσο και η λέξη “report” σε μικρότερο βαθμό την περίοδο Ιουλίου, παρουσιάστηκε συχνά λόγω των τότε επερχόμενων δημοσιεύσεων των τετραμηνιαίων εκθέσεων κερδών (Earnings Reports) των δύο εταιρειών στις 30 Ιουλίου 2020. Είναι άλλωστε γνωστό στις τάξεις των επενδυτών ότι οι μέρες που προηγούνται και ακολουθούν της δημοσίευσης των εκθέσεων κερδών αποτελούν κρίσιμη περίοδο μεταβολής των τιμών των μετοχών, πόσο μάλλον όταν αναμενόταν οι εταιρείες Google και Apple να παρουσιάσουν θετικά αποτελέσματα.

Επιπρόσθετα, η συχνή παρουσία των λέξεων “coronavirus” και “pandemic” αποτελεί αυτονόητη ένδειξη της περιόδου που διανύεται με την έξαρση της πανδημίας. Συνεχίζοντας, η συχνή καταγραφή της λέξης “Fortnite” υποδεικνύει τον εντοπισμό ενός ακόμη σημαντικού γεγονότος που τυχαίνει να αφορά και τις δύο εταιρείες. Πρόκειται για τη μήνυση που κατέθεσε η εταιρεία ανάπτυξης παιχνιδιών Epic Games στις 14 Αυγούστου 2020, λόγω της απαγόρευσης του δημοφιλούς παιχνιδιού της Fortnite από τα διαδικτυακά καταστήματα εφαρμογών Google Play της Google και App Store (λέξη η οποία εμφανίστηκε επίσης συχνά) της Apple.

Επιπρόσθετα, η συχνή εμφάνιση ονομάτων άλλων εταιρειών στο σύνολο δεδομένων, δεν αποτελεί τυχαίο γεγονός, εφόσον οι εταιρείες Google και Apple συγκαταλέγονται μεταξύ των εξέχουσων αμερικανικών τεχνολογικών εταιρειών, οι οποίες συνθέτουν ευρέως διαδεδομένα ακρωνύμια στον επενδυτικό κόσμο του Αμερικάνικου Χρηματιστηρίου, που αντιπροσωπεύουν ομάδες συνεχώς αυξανόμενων στην πάροδο του χρόνου τεχνολογικών μετοχών υψηλής αξίας. Τα δημοφιλέστερα ακρωνύμια είναι τα ακόλουθα:

- FANG: Facebook/Apple/Netflix/Google.
- FAANG: Facebook/Apple/Amazon/Netflix/Google.
- FAAMG: Facebook/Apple/Amazon/Microsoft/Google.
- FANGMAN: Facebook, Apple, Netflix, Google, Microsoft, Amazon, Nvidia
- FANTASIA: Facebook/Apple/Netflix/Tesla/Alphabet/Salesforce/Intel/Amazon.

Ως εκ τούτου, είναι σύνηθες πολλά Tweets να αναφέρονται στις πιο πάνω εταιρείες συλλογικά παραθέτοντας τα ποσοστά μεταβολής των μετοχών για μια συγκεκριμένη ημέρα ή ένα γενικό σχόλιο που τις αφορά, με κατά κύρια ομολογία ουδέτερο συναίσθημα. Επομένως, η συχνή εμφάνιση των λέξεων “Amazon”, “Tesla”, “tech”, “Netflix”, “Microsoft”, “Companies” στα σύνολα δεδομένων που περιέχουν εταιρείες από τα ακρώνυμα που αναφέρθηκαν, αποτελεί λογικό και σύνηθες φαινόμενο.

Δείγμα Tweets

- Wed - CEOs of Amazon, Apple, Facebook, Google: There is lots of competition. Our dominance is overstated.
- OUCH! Combined Mkt Cap of #FANGMAN (#Facebook, #Apple, #Netflix, #Google, #Microsoft, #Amazon, #Nvidia) drops \$400bn today, equal to GDP of #Norway.
- Tech is eating the world. The combined market cap of FANGMAN (Facebook, Apple, Netflix, Google, Microsoft, Amazon, Nvidia) has hit fresh ATH at \$8.4tn, almost equal to the combined GDP of the world's #3 #Japan and #4 #Germany.
- The combined market cap of Alphabet \$GOOGL, Apple \$AAPL, Amazon \$AMZN, Facebook \$FB, and Microsoft \$MSFT hit \$5.73 trillion today, a new all-time high.

Σχήμα 6.11: Δείγμα Tweets που στο κείμενο τους, γίνεται αναφορά σε περισσότερες από μία εταιρείες (Tweet από NYT Business στις 31/07/2020, Tweet από Holger Zschaepitz στις 03/09/2020, Tweet από Holger Zschaepitz στις 02/09/2020, Tweet από Bespoke στις 05/06/2020)

“wants sell” παρουσιάζουν ξεκάθαρα την ευρύτερη επενδυτική διάθεση και προτρέπουν έντονα στην πώληση της εν λόγω μετοχής.

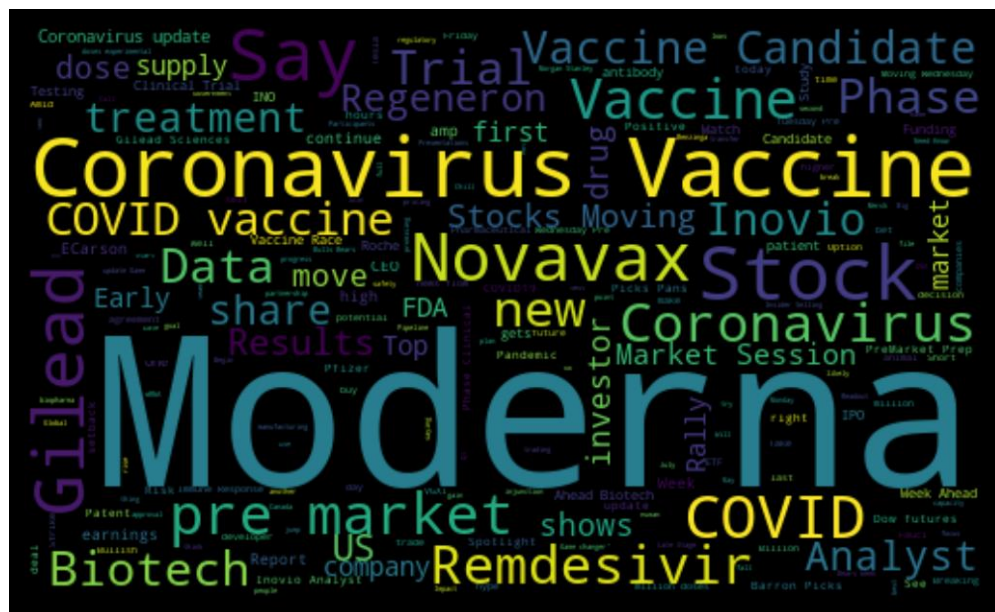


Σχήμα 6.13: Γράφημα τύπου WordCloud από τα κείμενα των αναρτήσεων για την εταιρεία Hertz

Στο σύνολο δεδομένων που αφορά την υγεία, αναμενόμενα διακρίνονται τα ονόματα των εταιρειών για τις οποίες συλλέχθηκαν δεδομένα, ενώ η λέξη “Moderna” παρουσιάζεται να έχει την μεγαλύτερη συχνότητα, καθώς τα περισσότερα Tweets τα οποία συλλέχθηκαν την αφορούν, λόγω και της μεγαλύτερης εμπλοκής της στην επικαιρότητα σε καθημερινή βάση σε σχέση με τις υπόλοιπες εταιρείες.

Επιβεβαιώνεται για ακόμα μια φορά ότι το γράφημα της συχνότητας εμφάνισης λέξεων δημιουργεί μια ικανοποιητική επισκόπηση της περιόδου με σημαντικές λέξεις και φράσεις που ενημερώνουν επαρκώς για την κατάσταση. Αρχικά, όπως ήταν λογικό, παρατηρείται σε μεγάλο βαθμό η αναφορά στον Covid 19, με λέξεις όπως: “Coronavirus” και “COVID”. Επιπρόσθετα, κυριαρχούν στο γράφημα οι αναφορές στο εμβόλιο κατά του Covid19, το οποίο αποτελεί την κύρια δραστηριότητα των εν λόγω εταιρειών την περίοδο συλλογής δεδομένων. Εμφανίζονται, με υψηλή συχνότητα, φράσεις όπως: “Coronovirus Vaccine”, “COVID vaccine”, “Vaccine” και “Vaccine Candidates” που αντικατοπτρίζουν την πραγματικότητα για την ενασχόληση των υποψήφιων, για εύρεση φαρμάκου κατά του κορωνοϊού, εταιρειών. Η λέξη “treatment” παραπέμπει στην θεραπεία του ιού, ενώ οι κλινικές δοκιμές της

Επιπλέον, εντοπίστηκε από το σύστημα η συχνή χρησιμοποίηση στα Tweets, της λέξης “Remdesivir”, δηλαδή της ρεμτεσιβίρης που είναι ένα αντιικό φάρμακο ευρέος φάσματος και αναπτύχθηκε από τη βιοφαρμακευτική εταιρεία Gilead Sciences. Τέλος, η ύπαρξη της λέξης “FDA” δεν αποτελεί τυχαίο γεγονός, εφόσον το φάρμακο κατά του Covid19 πρέπει να εγκριθεί από το Food and Drug Administration ή αλλιώς FDA.



6.1.4. Ανάλυση αποτελεσμάτων για Hashtags

70

προσμετράται η συχνότητα εμφάνισης των hashtags για κάθε σύνολο δεδομένων, σύμφωνα με την κατηγοριοποίηση των Tweets, δηλαδή αναλόγως της πολικότητά τους. Ως γενική παρατήρηση, διαπιστώθηκε ότι το σύνολο των hashtags αντικατοπτρίζουν συχνά στοιχεία όπως την περίοδο, γεγονότα της επικαιρότητας, λέξεις κλειδιά στο πεδίο ενδιαφέροντος, ειδησεογραφικά hashtags ή hashtags προσωπικοτήτων και ονόματα διαφόρων εταιρειών.

Αρχίζοντας από στοιχεία τα οποία εντοπίστηκαν και στην αξιολόγηση προηγούμενων μετρικών, καταγράφηκε συχνή χρησιμοποίηση hashtags με ονόματα άλλων εταιρειών από εκείνες οι οποίες εξετάζονται στο σύνολο δεδομένων για τις τεχνολογικές εταιρείες μετοχών υψηλής αξίας που ανήκουν σε ακρώνυμα, αλληλοσυσχετίζονται και εμπεριέχονται κυρίως σε πληροφοριακά Tweets με ουδέτερο συναίσθημα. Επίσης τα ονόματα των εταιρειών που εξετάζονται συνήθως αποτελούν τα πιο δημοφιλή hashtags, αποδεικνύοντας ότι η αναζήτηση στο Twitter με το όνομα της εταιρείας που ο χρήστης επιθυμεί είναι πάντοτε εύστοχη στρατηγική.

Παρατηρήθηκε συχνή αξιοποίηση hashtags με την ρίζα της λέξης Invest, όπως: “investing”, “Investing”, “investor”, “investors”, InvestingStrategies”, “#investingwithIBD”, στοιχείο που επιδεικνύει ότι hashtags με την συγκεκριμένη ρίζα λέξεων χρησιμοποιούνται ευρέως σε Tweets που αφορούν το Χρηματιστήριο. Επιπλέον, τα hashtags #stock και #stocks που ήταν λογικό να ανιχνευθούν, βρέθηκαν με επίσης υψηλή συχνότητα στα πλείστα σύνολα δεδομένων. Πιο κάτω παρατίθεται λίστα με βοηθητικά hashtags για αναζήτηση Χρηματοοικονομικών αναρτήσεων, σύμφωνα με τα αποτελέσματα από την ανάλυση των συνόλων δεδομένων.

#stock #stocks #investing #investor #investors #WJSWhatsNow #investingwithIBD #Trading #InvestingStrategies #Investing #Business #premarket #market
--

Σχήμα 6.15: Λίστα με βοηθητικά hashtags για αναζήτηση Χρηματοοικονομικών αναρτήσεων

Αξίζει να αναφερθεί ότι εντοπίστηκαν σε μεγάλο βαθμό και σε όλα τα σύνολα δεδομένων, hashtags τα οποία παραπέμπουν έμμεσα και άμεσα στην πανδημία, όπως φαίνεται στο πιο κάτω σχήμα (βλ. Σχήμα 6.16). Σημειωτέο είναι ότι παρόλο που η

παρουσία των λέξεων αυτών ήταν συλλογική και στις 3 λίστες ετικετών πολικότητας (θετικά, αρνητικά, ουδέτερα) για κάθε σύνολο δεδομένων, παρουσιάστηκαν πιο συχνά στις πρώτες θέσεις χρησιμοποίησης των κατηγοριοποιημένων αρνητικά Tweets, κάτι το οποίο μοιάζει αναμενόμενο και μπορεί να θεωρηθεί θετικό για την αποδοτικότητα του συστήματος. Χαρακτηριστικό παράδειγμα αποτελεί το γεγονός ότι στα σύνολα δεδομένων της Τεχνολογίας και υπηρεσιών Streaming και Τηλεδιασκέψεων, που ήταν τα μεγαλύτερα σε ποσότητα δεδομένων εμφανίζεται στην πρώτη θέση της λίστας για τα αρνητικά Tweets, το “#coronavirus”. Τέλος, όπως ήταν άλλωστε λογικό για το σύνολο δεδομένων που αφορά την Υγεία, εμφανίζεται πρώτο σε συχνότητα το hashtag “COVID19”, ακόμη και στη λίστα των πιο δημοφιλών Tweets με θετική πολικότητα.



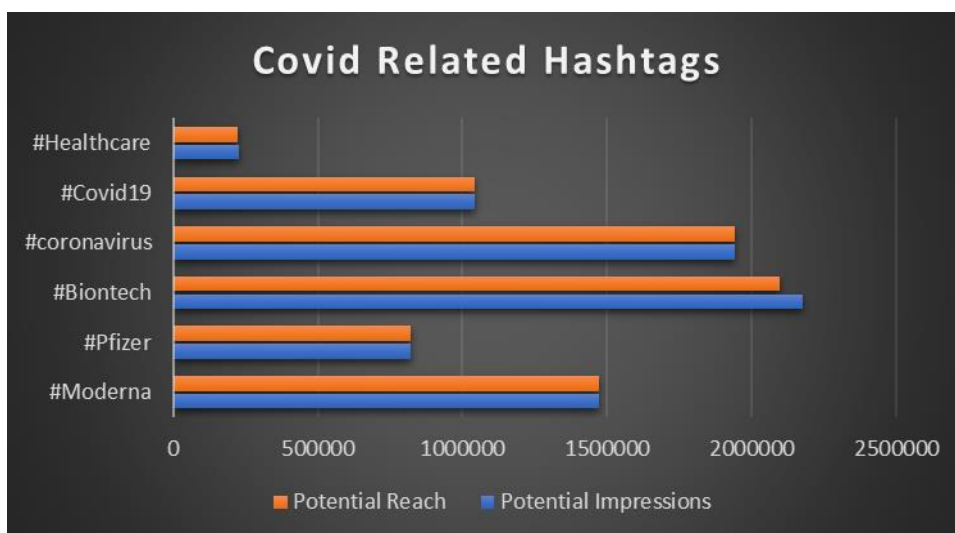
#coronavirus
#Coronavirus
#Covid19
#COVID19
#Healthcare
#Moderna

Σχήμα 6.16: Λίστα με βοηθητικά hashtags για αναζήτηση αναρτήσεων σε σχέση με τον κορωνοϊό

Είναι αδιαμφισβήτητο ότι η επιλογή ενός δημοφιλούς hashtag κρίνεται αναγκαία τόσο για την επιτυχή αναζήτηση αναρτήσεων στο πεδίο ενδιαφέροντος που επιθυμεί ο χρήστης, όσο και για την συγγραφή και δημοσιοποίηση μιας επιτυχημένης ανάρτησης. Με τον όρο «επιτυχημένη» χαρακτηρίζεται η ανάρτηση με πολλές προβολές και αντιδράσεις κάτι που μπορεί να επιτευχθεί με την προσθήκη hashtags που συγκεντρώνουν υψηλό ενδιαφέρον. Μετρικές αξιολόγησης ενός επιτυχημένου hashtag είναι τα Impressions και τα Reaches των αναρτήσεων που τα εμπεριέχουν στο κείμενο τους. Ως εκ τούτου, συλλέχθηκαν τα εν λόγω στατιστικά(βλ. Παράρτημα Α) για υποσύνολο των hashtags που βρέθηκαν από το σύστημα και είχαν υψηλή συχνότητα στα σύνολα δεδομένων της μελέτης, ενώ συμπεριλήφθηκαν ορισμένα επιπλέον συγκεκριμένα hashtags για σκοπούς συγκρίσεως με βάση την επικαιρότητα και εύλογα ερωτήματα στο πεδίο ενδιαφέροντος. Ένα Impression δηλώνει ότι μια ανάρτηση εμφανίστηκε στο χρονολόγιο του λογαριασμού ενός χρήστη, ενώ το Reach

δηλώνει το σύνολο των λογαριασμών στους οποίους ένα Tweet/hashtag/twitter campaign έφτασε.

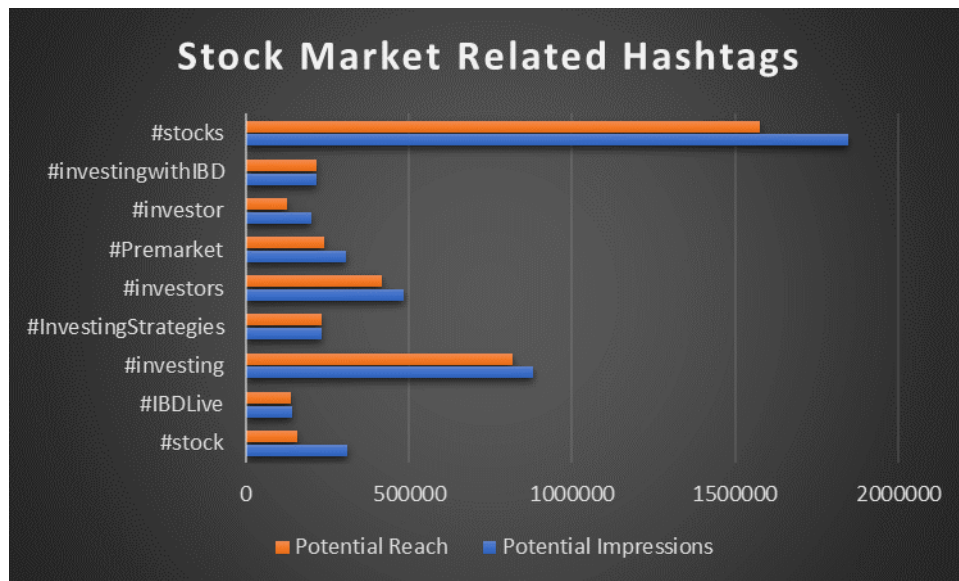
Η συλλογή των δεδομένων πραγματοποιήθηκε με την βοήθεια προηγμένου εργαλείου παρακολούθησης του Twitter που βασίζεται σε τεχνητή νοημοσύνη και ανήκει στην εταιρεία TrackMyHashtag[30]. Σημειώνεται ότι η ανάλυση που παρουσιάζεται είναι για την ημερομηνία 10/12/2020 και τα εν λόγω αποτελέσματα μεταβάλλονται συνεχώς, καθώς δίνεται έμφαση στην ανάλυση των τελευταίων 100 Tweets της συγκεκριμένης μέρας. Παρόλα αυτά καταδεικνύεται ότι μπορούν να ληφθούν ενδιαφέρον πληροφορίες που αντικατοπτρίζουν την περίοδο στην οποία υλοποιείται η εν λόγω εργασία.



Σχήμα 6.17: Γραφική παράσταση που παρουσιάζει συγκριτικά τα ενδεχόμενα impression και reach σε αναρτήσεις που αξιοποιούν hashtags σχετικά με τον κωρονοϊό

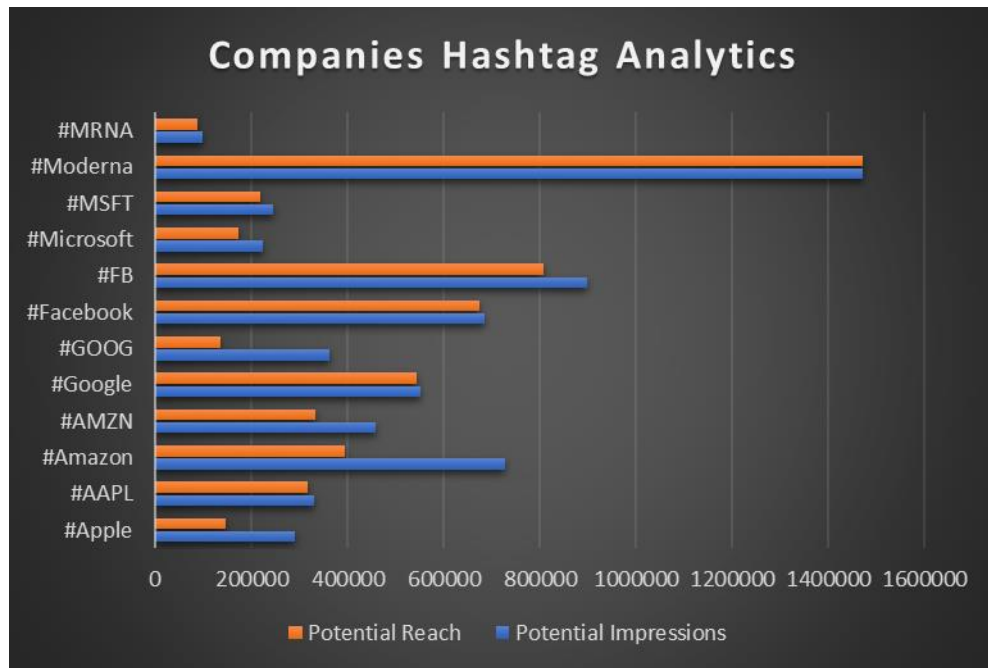
Για τη σύγκριση hashtags σχετικά με την πανδημία, συμπεριλήφθηκαν στη λίστα hashtags με τα ονόματα των εταιρειών Pfizer και Biontech, οι οποίες βρίσκονται κοντά στην ανάπτυξη εμβολίου κατά του Covid19 και θα ήταν ενδιαφέρον να παρατηρηθούν οι μετρικές και για αυτές. Αρχικά, παρατηρείται ότι το #coronavirus είναι πιο επιτυχημένο από το Covid19. Επίσης, το #Healthcare αναμενόμενα καταλαμβάνει την τελευταία θέση στις συγκρίσεις, ενώ λόγω της επικαιρότητας η εταιρεία Biontech έχει τις υψηλότερες τιμές σε πιθανά impressions και reach. Σημειώνεται, ότι για τα

#Covid19 και #coronavirus, περιλαμβάνονται στις μετρήσεις και οι διάφορες παραλλαγές τους με κεφαλαίους και πεζούς χαρακτήρες.



Σχήμα 6.18: Γραφική παράσταση που παρουσιάζει συγκριτικά τα ενδεχόμενα impression και reach σε αναρτήσεις που αξιοποιούν hashtags σχετικά με το Χρηματιστήριο

Στη σύγκριση hashtags σχετικά με ορολογίες που αξιοποιούνται σε Tweets του Χρηματιστηρίου, παρατηρήθηκε ότι η χρησιμοποίηση του #stocks συγκεντρώνει πολύ μεγαλύτερο ενδιαφέρον από τα υπόλοιπα hashtags που συλλέχθηκαν. Δεν είναι άλλωστε τυχαίο ότι #stocks και #investing συγκαταλέγονται στη λίστα με τα πιο δημοφιλή hashtags για το Χρηματιστήριο από αρκετές έρευνες δημοσιογραφικών ιστοσελίδων. Τονίζεται ότι στη λίστα δεν συμπεριλήφθηκε το #WSJWhatsNow, το οποίο είναι με διαφορά το πιο επιτυχημένο hashtag που εντοπίστηκε με 74526956 Potential Impressions και 19084367 Potential Reach, εφόσον οι μεγάλοι αριθμοί που καταγράφηκαν το κατατάσσουν ως αποκλειστικό (outlier) και η συμπερίληψη του θα αλλοίωνε κατά πολύ την συγκριτική φύση της γραφικής παράστασης.



Σχήμα 6.19: Γραφική παράσταση που παρουσιάζει συγκριτικά τα ενδεχόμενα *impression* και *reach* σε αναρτήσεις που αξιοποιούν *hashtags* με ονόματα γνωστών εταιρειών

Αξιολογήθηκαν επίσης *hashtags* με ονόματα εταιρειών. Στη λίστα προστέθηκαν τα σύμβολα των εταιρειών στο Χρηματιστήριο για να παρατηρηθεί η απόδοση τους σε σχέση με τα *hashtags* που εντόπισε το σύστημα και αναφέρουν ονομαστικά τις εταιρείες. Παρατηρήθηκε περισσότερο πιθανό *reach* για τα ονομαστικά *hashtags*, με εξαίρεση αυτό της εταιρείας Apple. Πιθανώς το αποτέλεσμα αυτό να οφείλεται στο ότι Tweets ειδησεογραφικού χαρακτήρα και μη επενδυτών, συνήθως χρησιμοποιούν ονομαστικά *hashtags*, ενώ τα *hashtags* συμβόλων αξιοποιούνται μόνο μεταξύ των επενδυτών. Επίσης, σύμφωνα με την εμπειρία που λήφθηκε από τα σύνολα δεδομένων με τα οποία καταπιάστηκε η μελέτη, τα σύμβολα των μετοχών συνήθως εμφανίζονται στο κείμενο των Tweets έχοντας την μορφή: “\$Σύμβολο”, παρά σε μορφή *hashtag*. Λαμβάνοντας υπόψιν τα πιο πάνω, οι χρήστες του Twitter προτρέπονται να χρησιμοποιούν *hashtags* με το όνομα των εταιρειών παρά με το σύμβολο, τόσο για τις αναζητήσεις τους, όσο και για την ανάρτηση νέου Tweet.

Τέλος, λόγω της επικαιρότητας, το *hashtag* της εταιρείας Moderna σημειώνει πολύ μεγαλύτερο ενδιαφέρον συγκριτικά με αυτά των 5 μεγαλύτερων τεχνολογικών εταιρειών στον επενδυτικό κόσμο. Ωστόσο, όπως αναφέρθηκε και πιο πάνω, αυτό αποτελεί συμπέρασμα για την περίοδο μελέτης της εργασίας και τα αποτελέσματα

μεταβάλλονται συνεχώς, ανάλογα με την επικαιρότητα και τις “τάσεις” στα κοινωνικά δίκτυα τη χρονική περίοδο αξιολόγησης.

6.1.5. Ανάλυση συσχετίσεων

Οι τιμές από το σύνολο δεδομένων του Χρηματιστήριου και τα αποτελέσματα από την ανάλυση συναισθήματος συγκρίθηκαν για διαπίστωση συσχέτισης με στόχο την πρόβλεψη της κάθε Χρηματιστηριακής μέρας. Για την εξεύρεση συσχετίσεων αξιοποιήθηκε ο συντελεστής συσχέτισης γινομένου ροπής Pearson, με τη βοήθεια της σχετικής βιβλιοθήκης της Pandas που τον υλοποιεί. Η ποσότητα δεδομένων τα οποία δεν ήταν συνεχόμενα και είχαν μεταβλητή ποσότητα για κάθε μέρα, συντέλεσαν έπειτα από την εξέταση των p-values, στο να παρατηρηθεί ότι δεν υπήρξαν στατιστικά σημαντικές συσχετίσεις ($< .05$) μεταξύ του συναισθήματος και της κίνησης των μετοχών.

Επομένως, παρόλο που η ευρύτερη κοινωνική διάθεση και τα σημαντικά γεγονότα μιας περιόδου μπορούν να εντοπιστούν από στατιστικά που παράγονται από την ανάλυση συναισθήματος, η πολικότητα δεν είναι σε θέση να υποδείξει επενδυτικές συμβουλές για την κάθε χρηματιστηριακή μέρα. Συμπεραίνεται λοιπόν, ότι η ανάλυση συναισθήματος πρέπει να διαδραματίζει περισσότερο συμβουλευτικό και όχι πρωταγωνιστικό ρόλο, τουλάχιστον όταν τα σύνολα δεδομένων δεν περιέχουν συνεχόμενα και σταθερά σε ποσότητα δεδομένα για την περίοδο ενδιαφέροντος.

6.2. Πειράματα και Αποτελέσματα για Μοντέλο Πρόβλεψης

Στο παρόν υποκεφάλαιο παρουσιάζονται τα αποτελέσματα, τα οποία προέκυψαν μετά την ολοκλήρωση των εκτελέσεων του προβλήματος μηχανικής μάθησης από τους αλγορίθμους που επιλέχθηκαν για τη μελέτη, με χαρακτηριστικά εισόδου από το συγχωνευμένο σύνολο δεδομένων, που παράχθηκε έπειτα από την επεξεργασία του συνόλου δεδομένων Daily Stock News και των δεδομένων, που συλλέχθηκαν από το Χρηματιστήριο για την ίδια χρονική περίοδο.

6.2.1. Επιλογή μετρικών

Η επιλογή των κατάλληλων μετρικών αποτελεί εξαιρετικά σημαντικό στοιχείο για την ορθή αξιολόγηση των προβλέψεων από τα μοντέλα, καθώς επηρεάζουν άμεσα τα αποτελέσματα που εξάγονται. Η **ορθότητα** (accuracy) αποτελεί μια από τις δημοφιλέστερες μετρικές αξιολόγησης μοντέλων πρόβλεψης και ως εκ τούτου συμπεριλήφθηκε σαν μια εκ των βασικών μετρικών αξιολόγησης. Συγκεκριμένα, ορθότητα ορίζεται ο λόγος του αριθμού των σωστών προβλέψεων του μοντέλου με τον συνολικό αριθμό δειγμάτων εισόδου. Η εν λόγω μετρική, θεωρείται αρκετά αξιόπιστη όταν τα δεδομένα βρίσκονται καλά διανεμημένα στις κλάσεις. Ωστόσο, η ορθότητα από μόνη της δεν είναι αρκετή για την επιλογή του αποδοτικότερου μοντέλου, καθώς δεν μεταφέρει την γνώση των λανθασμένων περιπτώσεων ταξινόμησης.

Για την προσμέτρηση των περιπτώσεων αυτών εισάχθηκαν οι ακόλουθες μετρικές:

- **Ακρίβεια** (Precision): Φανερώνει το ποσοστό των δειγμάτων που ταξινομήθηκαν ορθά από όλα τα θετικά δείγματα. Η μετρική αυτή επικεντρώνεται στα ψευδή θετικά (False Positives), δηλαδή στα δείγματα που ταξινομούνται λανθασμένα ως ορθά.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

Εξίσωση 6.1: Εξίσωση Ακρίβειας

- **Ανάκληση (Recall):** Φανερώνει το ποσοστό των δειγμάτων που ταξινομήθηκαν ορθά από όλα τα πραγματικά θετικά δείγματα. Η μετρική αυτή επικεντρώνεται στα ψευδή αρνητικά (False Positives), δηλαδή στα δείγματα που ο αλγόριθμος απέτυχε να ταξινομήσει σωστά.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

Εξίσωση 6.2: Εξίσωση Ανάκλησης

- **F1 Score:** Μετρική που συνδυάζει τις ακρίβεια και ανάκληση υπολογίζοντας τον αρμονικό μέσο.

$$F1\ Score = 2 * \left(\frac{Precision * Recall}{Precision + Recall} \right)$$

Εξίσωση 6.1: Εξίσωση F1 Score

Η ανάκληση και ακρίβεια μπορούν να εκφραστούν σε σχέση με τον αριθμό των αληθώς θετικών (True Positives), αληθώς αρνητικών (True Negative), ψευδώς θετικών (False Positive), και ψευδώς αρνητικών (False Negatives), τα οποία για τους σκοπούς της παρούσας εργασίας ερμηνεύονται ως εξής:

Αληθώς Θετικά: Το μοντέλο προέβλεψε ορθά την άνοδο ('1') της τιμής κλεισίματος μετοχής σε σχέση με την χθεσινή.

Αληθώς Αρνητικά: Το μοντέλο προέβλεψε ορθά τη μείωση ('0') της τιμής κλεισίματος μετοχής σε σχέση με την χθεσινή.

Ψευδώς Θετικά: Το μοντέλο προέβλεψε λανθασμένα την άνοδο ('1') της τιμής κλεισίματος μετοχής σε σχέση με την χθεσινή.

Ψευδώς Αρνητικά: Το μοντέλο προέβλεψε λανθασμένα τη μείωση ('0') της τιμής κλεισίματος μετοχής σε σχέση με την χθεσινή.

6.2.2. Ανάλυση αποτελεσμάτων

Για την αξιολόγηση του ρόλου που διαδραματίζει η ποσότητα των δεδομένων, παράχθηκαν αποτελέσματα, τόσο για ολόκληρο το σύνολο δεδομένων, όσο και για υποσύνολο αυτού, το οποίο περιορίζεται στα τελευταία δύο χρόνια των συλλεχθέντων δεδομένων, δηλαδή από το 2014 μέχρι το 2016. Με τις δύο περιπτώσεις συνόλου δεδομένων για τα οποία παρουσιάζονται αποτελέσματα, εξετάζεται επίσης και η αποδοτικότητα των μοντέλων όταν το σύνολο δεδομένων με το οποίο εκπαιδεύονται περιορίζεται σε μικρότερο χρονικό πλαίσιο.

Για τη βελτιστοποίηση της απόδοσης των μοντέλων, χρησιμοποιήθηκε η τεχνική της κλιμάκωσης των χαρακτηριστικών (feature scaling). Από την θεωρία, γνωρίζεται ότι η μέθοδος κλιμάκωσης των χαρακτηριστικών είναι απαραίτητη για βελτιστοποίηση της απόδοσης του αλγόριθμου μηχανικής μάθησης K-Nearest Neighbours, καθώς αλγόριθμοι, οι οποίοι χρησιμοποιούν αποστάσεις ή ομοιότητες μεταξύ των δειγμάτων δεδομένων, όπως ο K-NN, είναι ευαίσθητοι σε μετασχηματισμούς των χαρακτηριστικών. Για τους περισσότερους από τους επιλεχθέντες αλγορίθμους, η τεχνική της κλιμάκωσης, δεν ενδείκνυται πάντοτε, καθώς παραμένουν αναλλοίωτοι στην κλιμάκωση, όμως εξετάστηκε η μεταβολή των αποτελεσμάτων τους με την επίδραση της εν λόγω τεχνικής. Για όλους τους αλγορίθμους αξιοποιήθηκε η τεχνική της ρύθμισης των υπερπαραμέτρων τους, η οποία υποβοήθησε στη βελτιστοποίηση των τελικών τιμών των μετρικών αξιολόγησης.

Παρακάτω παρουσιάζονται οι τιμές των μετρικών πριν και μετά την προσπάθεια βελτιστοποίησης των αποτελεσμάτων με τις τεχνικές της κλιμάκωσης των χαρακτηριστικών για ορισμένους αλγορίθμους και της ρύθμισης των υπερπαραμέτρων για όλα τα αναπτυχθέντα μοντέλα μηχανικής μάθησης. Τα εν λόγω στάδια και η σειρά εκτέλεσης τους διαδραμάτισαν σημαντικό ρόλο για την επίτευξη επιτυχημένων προβλέψεων με υψηλά επίπεδα ορθότητας.

Linear Discriminant Analysis				
	Precision	Recall	F1 Score	Accuracy
Βασικό Μοντέλο	0.60	0.61	0.60	0.61
Βελτιστοποίηση	0.60	0.61	0.60	0.61

Πίνακας 6.1: Πίνακας απεικόνισης της μεταβολή των μετρικών μετά την εφαρμογή των τεχνικών βελτιστοποίησης για τον αλγόριθμο Linear Discriminant Analysis

Logistic Regression				
	Precision	Recall	F1 Score	Accuracy
Βασικό Μοντέλο	0.60	0.60	0.59	0.60
Βελτιστοποίηση	0.61	0.61	0.60	0.61

Πίνακας 6.2: Πίνακας απεικόνισης της μεταβολή των μετρικών μετά την εφαρμογή των τεχνικών βελτιστοποίησης για τον αλγόριθμο Logistic Regression

Decision Tree Classifier				
	Precision	Recall	F1 Score	Accuracy
Βασικό Μοντέλο	0.52	0.52	0.52	0.52
Βελτιστοποίηση	0.54	0.54	0.54	0.54

Πίνακας 6.3: Πίνακας απεικόνισης της μεταβολή των μετρικών μετά την εφαρμογή των τεχνικών βελτιστοποίησης για τον αλγόριθμο Decision Tree Classifier

Random Forest Classifier				
	Precision	Recall	F1 Score	Accuracy
Βασικό Μοντέλο	0.52	0.52	0.52	0.52
Βελτιστοποίηση	0.55	0.55	0.55	0.55

Πίνακας 6.4: Πίνακας απεικόνισης της μεταβολή των μετρικών μετά την εφαρμογή των τεχνικών βελτιστοποίησης για τον αλγόριθμο Random Forest Classifier

XGBoost Classifier				
	Precision	Recall	F1 Score	Accuracy
Βασικό Μοντέλο	0.52	0.54	0.52	0.54
Βελτιστοποίηση	0.54	0.55	0.55	0.55

Πίνακας 6.5: Πίνακας απεικόνισης της μεταβολή των μετρικών μετά την εφαρμογή των τεχνικών βελτιστοποίησης για τον αλγόριθμο XGBoost Classifier

K-Nearest Neighbours				
	Precision	Recall	F1 Score	Accuracy
Βασικό Μοντέλο	0.53	0.52	0.52	0.52
Βελτιστοποίηση	0.51	0.52	0.52	0.52

Πίνακας 6.6: Πίνακας απεικόνισης της μεταβολή των μετρικών μετά την εφαρμογή των τεχνικών βελτιστοποίησης για τον αλγόριθμο K-Nearest Neighbours

Πιο κάτω, παρουσιάζονται πίνακες με τις τιμές των μετρικών της ορθότητας, της ακρίβειας, της ανάκλησης και της βαθμολογίας του F1, για κάθε αλγόριθμο μηχανικής μάθησης που αξιοποιήθηκε και για τις δύο περιπτώσεις συνόλων δεδομένων. Επιπλέον, παρουσιάζονται γραφήματα που απεικονίζουν τα αποτελέσματα των μετρικών για κάθε αλγόριθμο, υποβοηθώντας με αυτόν τον τρόπο, την κατανόηση των αποτελεσμάτων. Σημειώνεται ότι, όλες οι τιμές από τις μετρικές που παρουσιάζονται αποτελούν τους σταθμισμένους μέσους όρους από τις τιμές όλων των συλλεχθέντων αποτελεσμάτων.

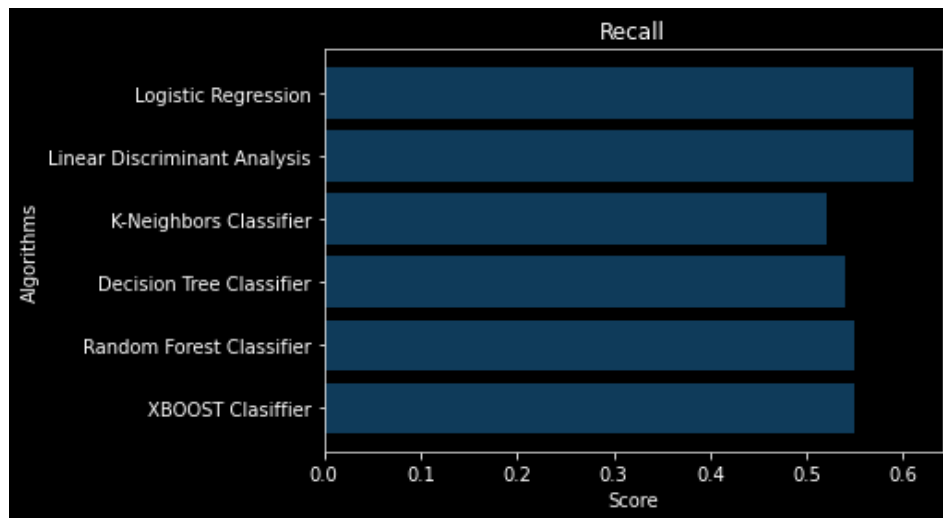
Επιπλέον, εφόσον είναι επιθυμητή η αποφυγή της ανισορροπίας μεταξύ των τιμών της ανάκλησης και της ακρίβειας, ως εκ τούτου είναι άξια παρατήρησης η βαθμολογία της μετρικής F1 για κάθε μοντέλο αλγορίθμου μηχανικής μάθησης το οποίο υλοποιήθηκε.

Αλγόριθμος	Precision	Recall	F1 Score	Accuracy
Linear Discriminant Analysis	0.60	0.61	0.60	0.61
Logistic Regression	0.61	0.61	0.60	0.61
K-Nearest Neighbors	0.51	0.52	0.52	0.52
Decision Tree Classifier	0.54	0.54	0.54	0.54
Random Forest	0.55	0.55	0.55	0.55
XGBoost Classifier	0.54	0.55	0.55	0.55

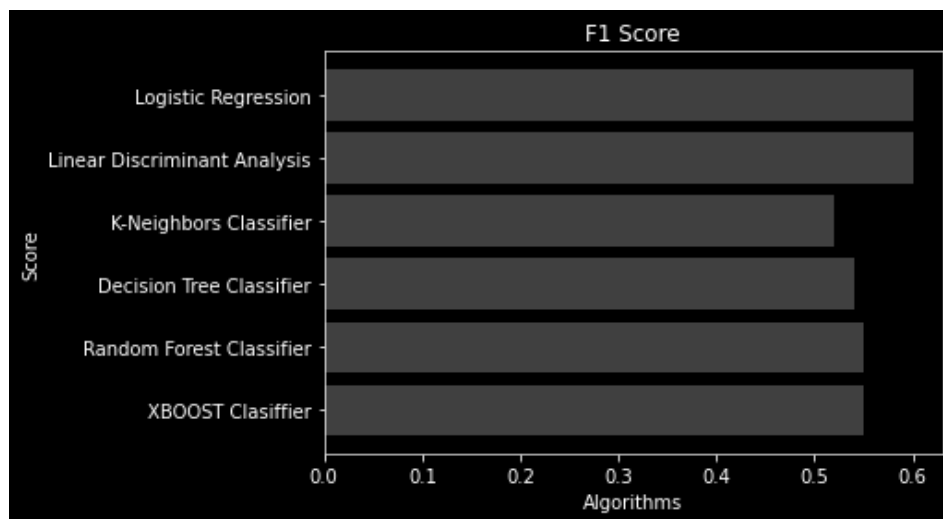
Πίνακας 6.7: Πίνακας απεικόνισης των αποτελεσμάτων των μετρικών για όλα τα μοντέλα μηχανικής μάθησης που αναπτύχθηκαν

Παρατηρώντας τον πίνακα αποτελεσμάτων γίνεται αντιληπτό ότι ο αλγόριθμος Logistic Regression, είχε την καλύτερη απόδοση για τις προβλέψεις σε ολόκληρο το σύνολο δεδομένων, με ποσοστό ορθότητας 61% και βαθμολογίας F1 της τάξης του 60%. Εξίσου καλή απόδοση είχε ο αλγόριθμός Linear Discriminant Analysis με ποσοστό ορθότητας επίσης, 61%, με ελαφρώς μειωμένο ποσοστό ακρίβειας, 60%.

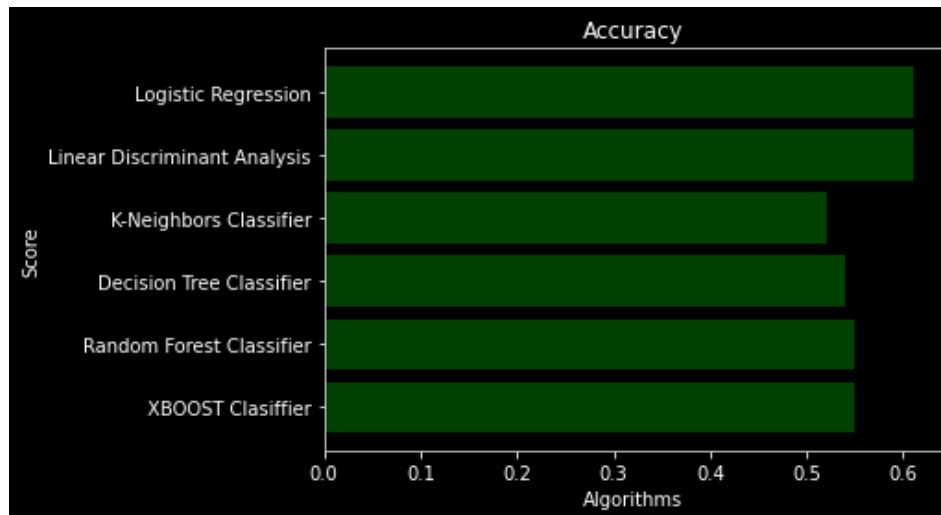
Ακολουθεί γραφική απεικόνιση, για την σύγκριση των βαθμολογιών των μετρικών κάθε αλγορίθμου σε ολόκληρο το σύνολο δεδομένων.



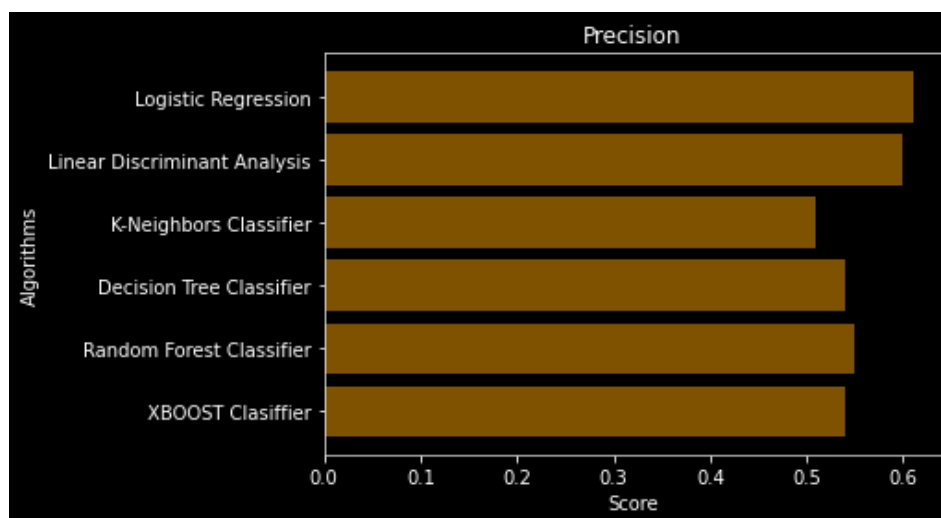
Σχήμα 6.20: Γραφική παράσταση σύγκρισης της βαθμολογίας ανάκλησης των αλγορίθμων



Σχήμα 6.21: Γραφική παράσταση σύγκρισης της βαθμολογίας F1 των αλγορίθμων



Σχήμα 6.22: Γραφική παράσταση σύγκρισης της βαθμολογίας ορθότητας των αλγορίθμων



Σχήμα 6.23: Γραφική παράσταση σύγκρισης της βαθμολογίας ακρίβειας των αλγορίθμων

Πιο κάτω παρουσιάζονται τα αποτελέσματα των μετρικών της ακρίβειας, της ανάκλησης, της ορθότητας και της βαθμολογίας F1 για το μικρότερο σύνολο δεδομένων, έτσι όπως προέκυψαν μετά την βελτιστοποίηση της απόδοσης των αλγορίθμων, μέσω των τεχνικών που αξιοποιήθηκαν.

Algorithm	Precision	Recall	F1 Score	Accuracy
Linear Discriminant Analysis	0.64	0.64	0.64	0.64
Logistic Regression	0.66	0.66	0.66	0.66
K-Nearest Neighbors	0.55	0.54	0.53	0.54
Decision Tree Classifier	0.56	0.56	0.56	0.56
Random Forest	0.56	0.56	0.56	0.56
XGBoost Classifier	0.59	0.59	0.59	0.59

Πίνακας 6.8: Πίνακας απεικόνισης των αποτελεσμάτων των μετρικών σε όλα τα μοντέλα μηχανικής μάθησης που αναπτύχθηκαν, για το μικρότερο σύνολο δεδομένων

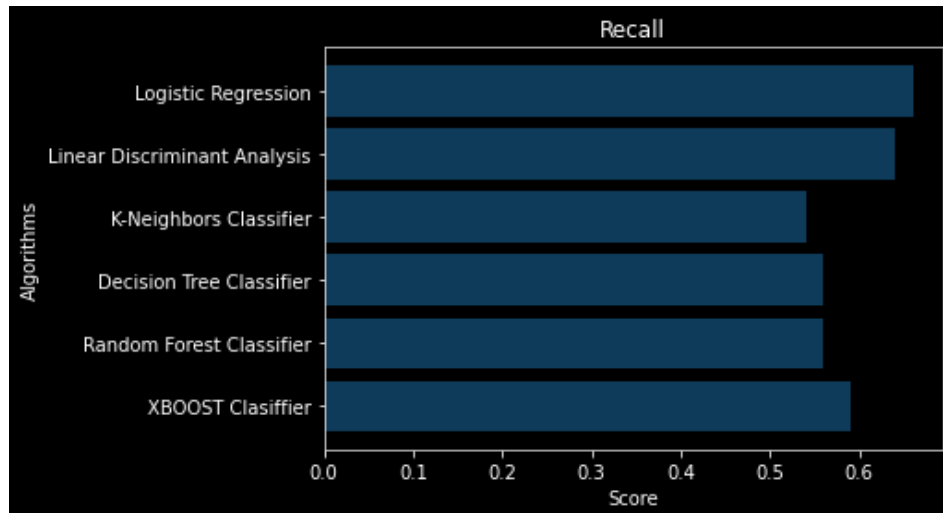
Όπως παρατηρείται και στα αποτελέσματα του μεγάλου συνόλου δεδομένων, οι ίδιοι δύο αλγόριθμοι είχαν την καλύτερη απόδοση στις προβλέψεις τους και για το μικρότερο σύνολο δεδομένων, με τον αλγόριθμο Logistic Regression να σημειώνει την καλύτερη απόδοση με ποσοστά ορθότητας και βαθμολογίας F1, 66%, και τον αλγόριθμο Linear Discriminant Analysis να ακολουθεί με ορθότητα και F1 που ανέρχεται στο ποσοστό του 64%.

Η αρκετά καλή απόδοση του αλγορίθμου Logistic Regression και για τα δύο σύνολα δεδομένων, ήταν αναμενόμενη, καθώς ο εν λόγω αλγόριθμος λειτουργεί καλύτερα σε προβλήματα δυαδικής ταξινόμησης, παρόμοια δηλαδή με αυτό, με το οποίο ασχολείται η παρούσα εργασία.

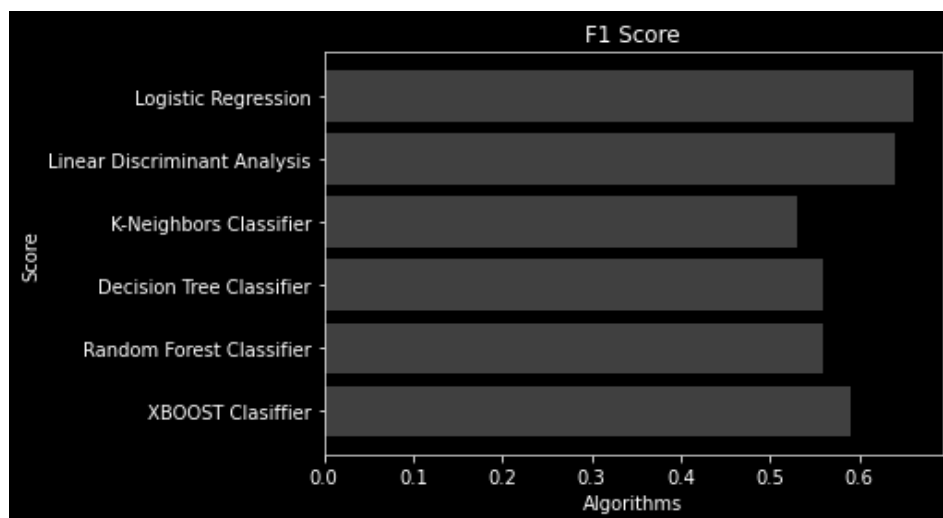
Σημειώνεται επίσης ότι, οι εκτελέσεις επαναλήφθηκαν αρκετές φορές και η καλύτερη βαθμολογία F1 επιτεύχθηκε από τον αλγόριθμο Logistic Regression, για το μικρότερο σύνολο των δεδομένων, με ποσοστό ορθότητας 72%.

Σαν γενικό σχόλιο, για τους υπόλοιπους αλγορίθμους παρατηρείται ότι η απόδοση τους, βελτιώθηκε ελάχιστα συγκριτικά με την απόδοση τους για το μεγαλύτερο σύνολο δεδομένων. Ωστόσο, παρατηρείται, ότι ο αλγόριθμός με την χειρότερη απόδοση ήταν ο KNN, ο οποίος διατήρησε την απόδοση του για το μικρότερο σύνολο δεδομένων, σημειώνοντας βαθμολογία F1, 53% και δεν λειτούργησε ικανοποιητικά στο πρόβλημα δυαδικής ταξινόμησης της παρούσας μελέτης.

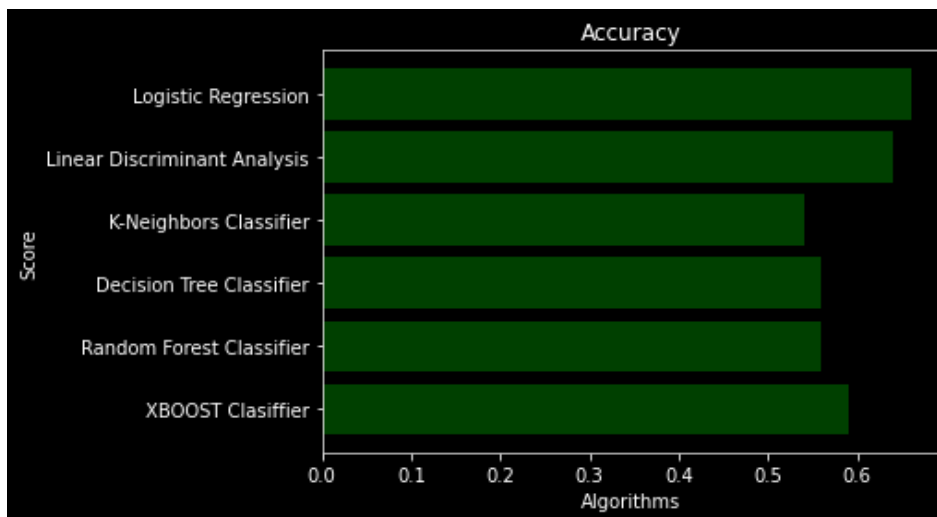
Ακολουθεί γραφική απεικόνιση, για την σύγκριση των βαθμολογιών των μετρικών κάθε αλγορίθμου, στο υποσύνολο του συνόλου δεδομένων.



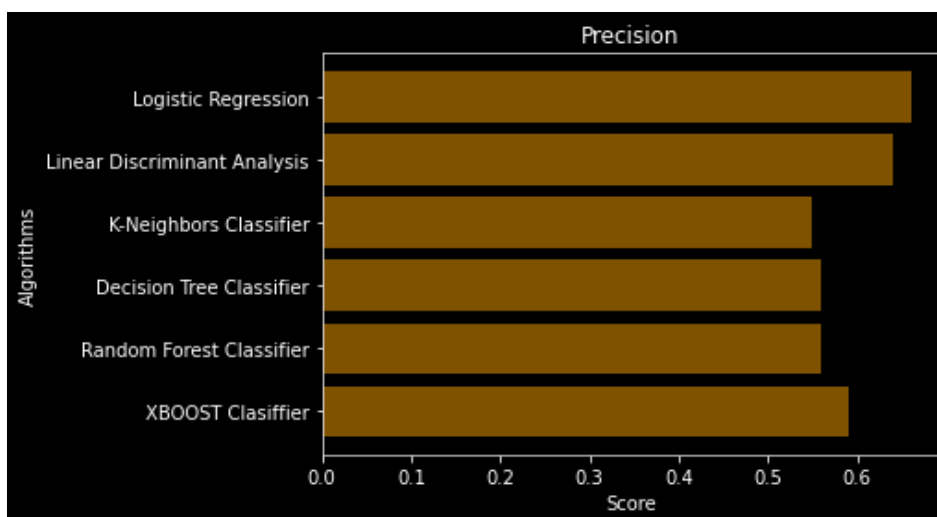
Σχήμα 6.24: Γραφική παράσταση σύγκρισης της βαθμολογίας ανάκλησης των αλγορίθμων στο μικρότερο σύνολο δεδομένων



Σχήμα 6.25: Γραφική παράσταση σύγκρισης της βαθμολογίας F1 των αλγορίθμων στο μικρότερο σύνολο δεδομένων



Σχήμα 6.26: Γραφική παράσταση σύγκρισης της βαθμολογίας ορθότητας των αλγορίθμων στο μικρότερο σύνολο δεδομένων



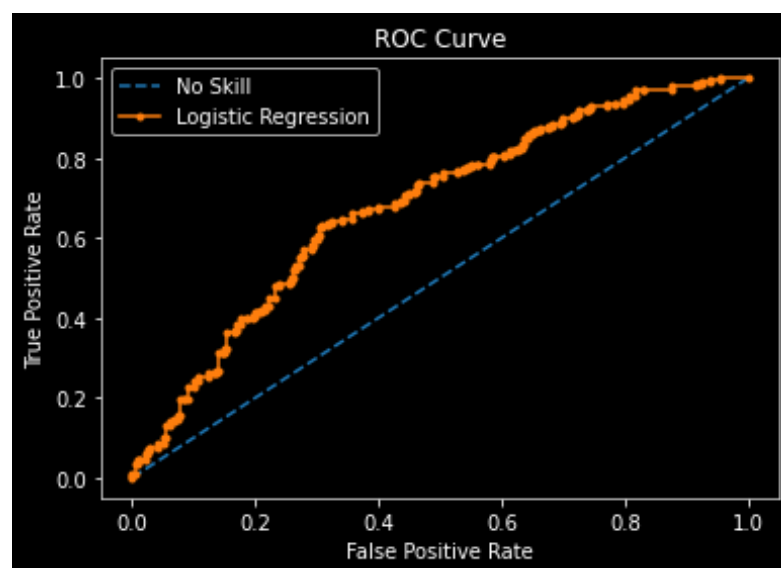
Σχήμα 6.27: Γραφική παράσταση σύγκρισης της βαθμολογίας ακρίβειας των αλγορίθμων στο μικρότερο σύνολο δεδομένων

Αξίζει να σημειωθεί ότι, η παρουσία βαθμολογιών ακρίβειας και ανάκλησης στα μοντέλα πρόβλεψης, οι οποίες είναι πανομοιότυπες ή απέχουν ελάχιστα από τις αντίστοιχες βαθμολογίες της ορθότητας και F1, διασφαλίζει ότι τα παραχθέντα αποτελέσματα είναι αξιόπιστα και λογικά.

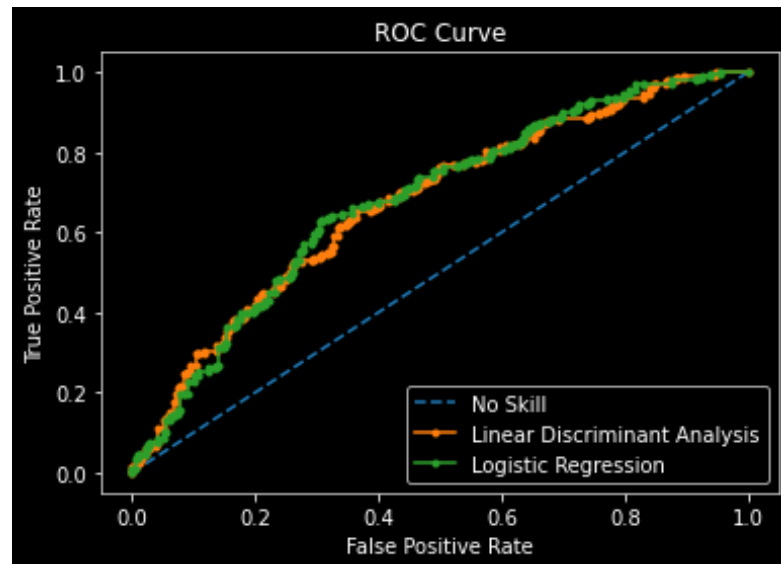
Για την καλύτερη κατανόηση των αποτελεσμάτων από τα μοντέλα μηχανικής μάθησης, κρίθηκε χρήσιμο να υπολογιστεί το κατώφλι ορθότητας των προβλέψεων, προκειμένου να αξιολογηθεί η απόδοση των αλγορίθμων και του συστήματος.

Για τον προσδιορισμό του κατωφλίου των προβλέψεων, υπολογίστηκαν τιμές πρόβλεψης, στο σύνολο των δοκιμών, που βασίζονται στην τυχαιότητα (0 ή 1). Πραγματοποιήθηκε μεγάλος αριθμός εκτελέσεων και υπολογίστηκε ο μέσος όρος της τιμής ορθότητας (accuracy), ο οποίος ανέρχεται σε ποσοστό της τάξης 44.95%. Επομένως, επενδυτές, οι οποίοι για τα διαστήματα δοκιμών πραγματοποιούσαν, επενδύσεις βασιζόμενοι στην τυχαιότητα, μακροπρόθεσμα θα έχαναν λεφτά με πολύ υψηλή πιθανότητα. Στον αντίποδα, με την αξιοποίηση του τελικού μοντέλου του συστήματος, για το ίδιο διάστημα, εάν ένας επενδυτής ακολουθούσε τις επενδυτικές συμβουλές του συστήματος για καθημερινές συναλλαγές, με σταθερό ποσό επένδυσης και ελεγχόμενο ποσοστό απωλειών, κατά πάσα πιθανότητα θα λάμβανε οικονομικό όφελος με το τέλος της κάθε περιόδου.

Για κάθε μοντέλο μηχανικής μάθησης δημιουργήθηκε γράφημα, το οποίο απεικονίζει την καμπύλη ROC (Receiver Operator Characteristic), προκειμένου να αναπαρασταθεί η διαγνωστική ικανότητα του κάθε δυαδικού ταξινομητή (binary classifier). Στο πιο κάτω σχήμα παρουσιάζεται το γράφημα για τα μοντέλα με την καλύτερη απόδοση, σε ολόκληρο το σύνολο δεδομένων. Ταξινομητές, των οποίων η καμπύλη προσεγγίζει την πάνω αριστερή γωνιά, υποδεικνύουν υψηλά επίπεδα αποδοτικότητας. Στον αντίποδα, όσο περισσότερο η καμπύλη ενός ταξινομητή προσεγγίζει την διαγώνιο των 45 μοιρών, τόσο χαμηλότερα είναι και τα επίπεδα ακρίβειας των αποτελεσμάτων του.



Σχήμα 6.28: Η καμπύλη ROC για το μοντέλο Μηχανικής Μάθησης, Logistic Regression



Σχήμα 6.29: Η καμπύλη ROC για τα μοντέλα Μηχανικής Μάθησης, *Logistic Regression Linear* και *Discriminant Analysis*

Κεφάλαιο 7

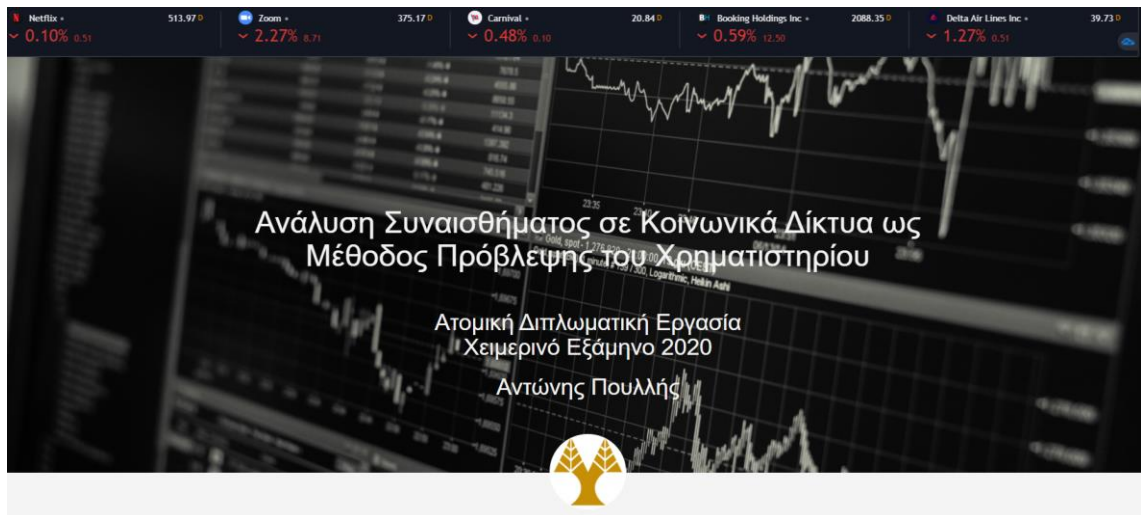
Διαδικτυακές Εφαρμογές

7.1.	Αρχική σελίδα διαδικτυακής εφαρμογής	89
7.2.	Διαδικτυακές εφαρμογές Streamlit.....	91
7.2.1.	Διαδικτυακή εφαρμογή για απεικόνιση Χρηματιστηριακών Δεδομένων	92
7.2.2.	Διαδικτυακή εφαρμογή για Ανάλυση Συναισθήματος.....	95
7.2.2.1.	Ανάλυση συναισθήματος με δεδομένα εισόδου χρήστη	96
7.2.2.2.	Ανάλυση συναισθήματος σε σύνολα δεδομένων	97

7.1. Αρχική σελίδα διαδικτυακής εφαρμογής

Προκειμένου να γίνουν αντιληπτές και να αξιοποιηθούν από χρήστες οι δυνατότητες του συστήματος που προτείνεται στη μελέτη, δημιουργήθηκε διαδικτυακή εφαρμογή η οποία εμπεριέχει πληροφορίες για την Α.Δ.Ε και συνδέσμους που οδηγούν σε εφαρμογές που παρουσιάζουν βασικές λειτουργίες του συστήματος. Οποιοσδήποτε χρήστης μπορεί μέσω της εφαρμογής να αλληλοεπιδράσει για την επίτευξη διαφορετικών στόχων. Σημειώνεται ότι, η διαδικτυακή εφαρμογή, είναι ειδικά σχεδιασμένη για να λειτουργεί και σε κινητές συσκευές. Η αρχική σελίδα της διαδικτυακής εφαρμογής υλοποιήθηκε με την χρήση των τεχνολογιών HTML5, CSS και Javascript, ενώ το deployment της σελίδας πραγματοποιήθηκε, μέσω της δωρεάν πλατφόρμας Infinity Free[14], στον σύνδεσμο <http://antonispoullis.epizy.com/>.

Αξιοποιήθηκαν επίσης widgets για απεικόνιση Χρηματιστηριακών πληροφοριών σε πραγματικό χρόνο από το γνωστό κοινωνικό δίκτυο για επενδυτές, Trading View[31].

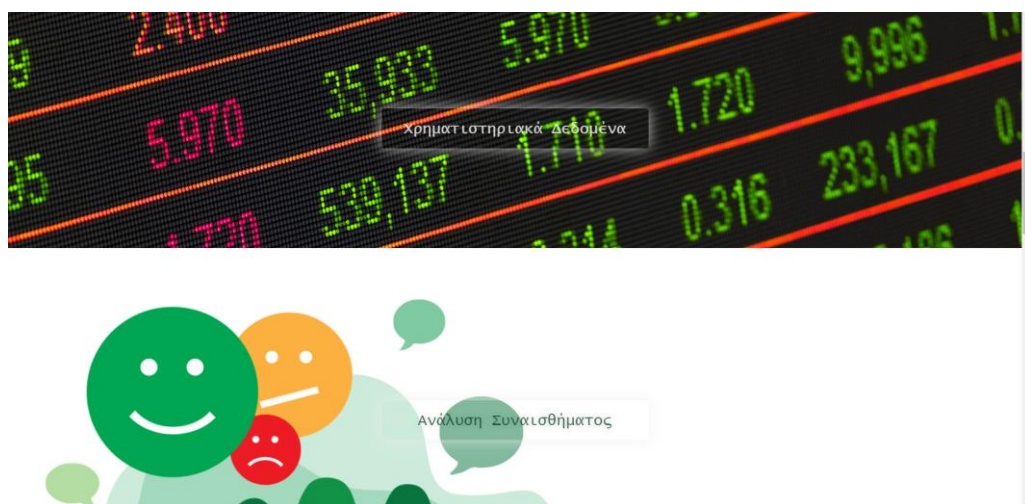


Σχήμα 7.1: Στιγμιότυπο από την αρχική σελίδα της διαδικτυακής εφαρμογής

Δημοφιλή Hashtags

Hashtags Σχετικά με τον Covid-19	Hashtags Σχετικά με το Χρηματιστήριο	Hashtags Σχετικά με Ονόματα Εταιρειών
#coronavirus #Coronavirus #Covid19 #COVID19 #Healthcare #Moderna #Pfizer #Biontech	#stock #stocks #investing #investor #investors #WJSWhatsNow #investingwithIBD #Trading #investing #investingStrategies #Business #market #premarket	#Microsoft #MSFT #Google #GOOG #Apple #AAPL #Amazon #AMZN #Facebook #FB #Moderna #Pfizer #Biontech

Σχήμα 7.2: Στιγμιότυπο απεικόνισης των hashtags από την αρχική σελίδα της διαδικτυακής εφαρμογής



Σχήμα 7.3: Στιγμιότυπο απεικόνισης των συνδέσμων για τις εφαρμογές Streamlit από την αρχική σελίδα της διαδικτυακής εφαρμογής

Αποτελέσματα Μοντέλων Πρόβλεψης

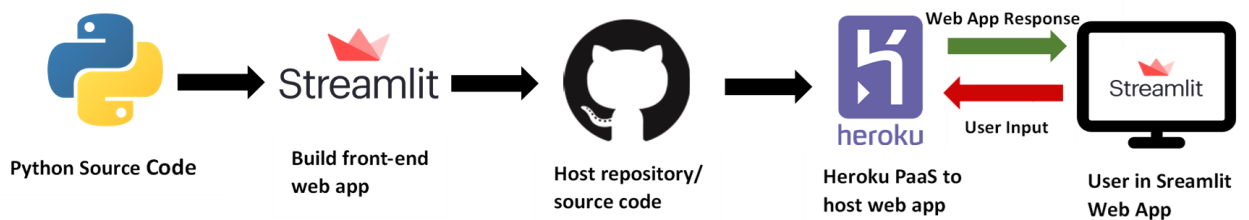
Αλγόριθμος	Precision	Recall	F1 Score	Accuracy
Linear Discriminant Analysis	0.64	0.64	0.64	0.64
Random Forest Classifier	0.56	0.56	0.56	0.56
K-Nearest Neighbours	0.55	0.54	0.53	0.54
XBoost Classifier	0.59	0.59	0.59	0.59
Logistic Regression	0.66	0.66	0.66	0.66
Decision Tree Classifier	0.56	0.56	0.56	0.56

Σχήμα 7.4: Στιγμιότυπο απεικόνισης των αποτελεσμάτων των μοντέλων πρόβλεψης από την αρχική σελίδα της διαδικτυακής εφαρμογής

7.2. Διαδικτυακές εφαρμογές Streamlit

Οι διαδικτυακές εφαρμογές που περιέχουν λειτουργικότητα, υλοποιήθηκαν στη γλώσσα προγραμματισμού Python αξιοποιώντας την βιβλιοθήκη Streamlit για την προσαρμογή και παρουσίαση του κώδικα σε μορφή διαδικτυακής εφαρμογής. Για το deployment των σελίδων και την παροχή δυνατότητας αλληλεπίδρασης του χρήστη με το σύστημα, ο κώδικας εκτελείται διαδικτυακά από τον χρήστη, μέσω της πλατφόρμα εφαρμογών ως υπηρεσία (PaaS) Heroku, ενώ ο κώδικας και τα σύνολα δεδομένων που χρησιμοποιούνται βρίσκονται αποθηκευμένα στη πλατφόρμα Github.

Η διαδικασία ανάπτυξης των διαδικτυακών εφαρμογών, παρουσιάζεται στο Σχήμα 7.5.



Σχήμα 7.5: Διαδικασία Deployment των Διαδικτυακών Εφαρμογών που υλοποιήθηκαν με την βιβλιοθήκη Streamlit

7.2.1. Διαδικτυακή εφαρμογή για απεικόνιση Χρηματιστηριακών Δεδομένων

Τα αριθμητικά δεδομένα των εταιρειών τα οποία συλλέγηκαν (βλ. 4.3) από το Χρηματιστήριο, τα νέα υπολογισμένα στατιστικά και γραφικές παραστάσεις με χρήσιμες πληροφορίες, παρουσιάζονται στους χρήστες, μέσω διαδικτυακής εφαρμογής που βρίσκεται στον σύνδεσμο <https://stocks-visualization.herokuapp.com/>.

Με αυτή την διαδικτυακή εφαρμογή, παρέχεται η δυνατότητα επιλογής ανάμεσα σε 33 εταιρείες για τις οποίες υπάρχουν διαθέσιμες πληροφορίες και στατιστικά, καθώς επίσης και δυνατότητα εισαγωγής εύρους ημερομηνίας για την οποία επιθυμείται η προβολή των αποτελεσμάτων. Μετά την επιλογή των δεδομένων εισόδου εμφανίζονται τα εξής στοιχεία:

- Κεφαλίδα και αντιπροσωπευτική φωτογραφία της προς εξέταση εταιρείας.
- Παρουσίαση συνόλου δεδομένων στο εύρος του προκαθορισμένου χρονικού διαστήματος που επέλεξε ο χρήστης. Το σύνολο δεδομένων περιέχει τα ακόλουθα πεδία:
 - Ημερομηνία
 - Τιμή ανοίγματος μετοχής
 - Τιμή κλεισίματος μετοχής
 - Μέγιστη Τιμή μετοχής
 - Ελάχιστη τιμή μετοχής
 - Όγκος συναλλαγών μετοχής
 - Προσαρμοσμένη τιμή κλεισίματος μετοχής
- Διαδραστική γραφική παράσταση απεικόνισης της προσαρμοσμένης τιμής κλεισίματος μετοχής συναρτήσει του προκαθορισμένου χρονικού διαστήματος που επέλεξε ο χρήστης.
- Διαδραστική γραφική παράσταση που απεικονίζει στον ίδιο άξονα, τις τιμές ανοίγματος, κλεισίματος και τη μέγιστη και ελάχιστη τιμή, συναρτήσει του προκαθορισμένου χρονικού διαστήματος που επέλεξε ο χρήστης. Μπορούν επίσης να κρατηθούν οι καμπύλες που επιθυμεί ο χρήστης.

- Διαδραστική γραφική παράσταση απεικόνισης του όγκου συναλλαγών μετοχής συναρτήσει του προκαθορισμένου χρονικού διαστήματος που επέλεξε ο χρήστης.
- Παρουσίαση των πιο κάτω στατιστικών, για τα πεδία του αρχικού συνόλου δεδομένων:
 - Αριθμός εγγραφών πεδίου
 - Μέσος όρος
 - Τυπική απόκλιση
 - Ελάχιστη τιμή
 - 25%, 50%, 75%
 - Μέγιστη τιμή

Dates

You can insert dates from 2020-03-01 until 2020-09-05.

Start Date

2020-03-01

End Date

2020-09-05

Select Stock Symbol

AAPL

Apple Inc.

NASDAQ:AAPL



Σχήμα 7.6: Στιγμιότυπα πεδίων δεδομένων εισόδου χρήστη

Apple Inc. Dataset

	Open	High	Low	Close	Adj Close	Vol
Mar 2, 2020 2:00 AM	70.5700	75.3600	69.4300	74.7025	74.3667	34139
Mar 3, 2020 2:00 AM	75.9175	76	71.4500	72.3300	72.0049	31947
Mar 4, 2020 2:00 AM	74.1100	75.8500	73.2825	75.6850	75.3448	21917
Mar 5, 2020 2:00 AM	73.8800	74.8875	72.8525	73.2300	72.9009	18757
Mar 6, 2020 2:00 AM	70.5000	72.7050	70.3075	72.2575	71.9327	22617
Mar 9, 2020 2:00 AM	65.9375	69.5225	65.7500	66.5425	66.2434	28674
Mar 10, 2020 2:00 AM	69.2850	71.6100	67.3425	71.3350	71.0144	28529
Mar 11, 2020 2:00 AM	69.3475	70.3050	67.9650	68.8575	68.5480	25559
Mar 12, 2020 2:00 AM	63.9850	67.5000	62	62.0575	61.7786	41847
Mar 13, 2020 2:00 AM	66.2225	69.9800	63.2375	69.4925	69.1802	37073
Mar 16, 2020 2:00 AM	60.4875	64.7700	60	60.5525	60.2803	32242

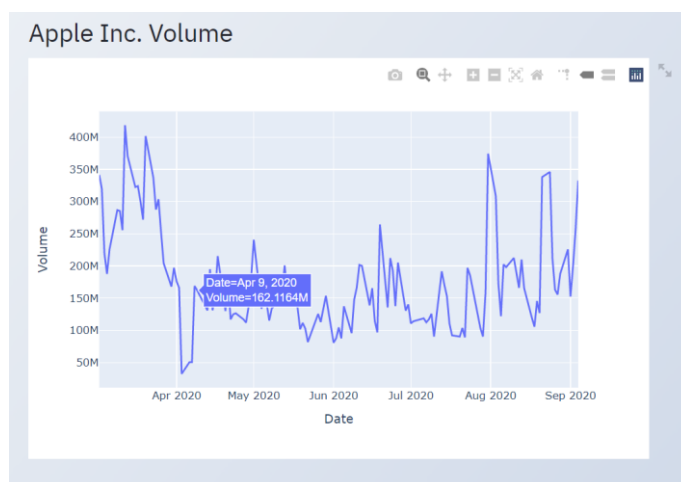
Σχήμα 7.7: Στιγμιότυπο απεικόνισης συνόλου δεδομένων για την εταιρεία Apple Inc.



Σχήμα 7.8: Στιγμιότυπο απεικόνισης γραφήματος τιμής κλεισίματος συναρτήσει του χρόνου για την εταιρεία Apple Inc.



Σχήμα 7.9: Στιγμιότυπο απεικόνισης γραφήματος με τα συλλεγμένα Χρηματιστηριακά δεδομένα για την εταιρεία Apple Inc.



Σχήμα 7.10: Στιγμιότυπο απεικόνισης γραφήματος όγκου συναλλαγών συναρτήσει του χρόνου για την εταιρεία Apple Inc.

Data Statistics

	Open	High	Low	Close	Adj Close	Volume
count	132	132	132	132	132	132
mean	86.1758	87.5676	84.8787	86.3523	86.1651	175,976,327.2727
std	19.5436	19.5443	19.0799	19.3731	19.4633	78,215,687.0756
min	57.0200	57.1250	53.1525	56.0925	55.8404	32470000
25%	70.5525	71.7056	69.5406	71.2781	70.9578	119335100
50%	80.9863	81.4775	80.3850	81.0575	80.9116	158529600
75%	96.4538	97.1744	94.8431	96.2244	96.0512	204476800
max	137.5900	137.9800	130.5300	134.1800	134.1800	418474000

Σχήμα 7.11: Στιγμιότυπο απεικόνισης στατιστικών του συνόλου δεδομένων για την εταιρεία Apple Inc.

7.2.2. Διαδικτυακή εφαρμογή για Ανάλυση Συναισθήματος

Στα πλαίσια της παρούσας μελέτης, υλοποιήθηκε επίσης διαδικτυακή εφαρμογή, μέσω της οποίας δίνεται η δυνατότητα στον χρήστη να αλληλοεπιδράσει και να παρουσιάσει αποτελέσματα αξιοποιώντας το μοντέλο επεξεργασίας κειμένου και ανάλυσης φυσικής γλώσσας της διπλωματικής εργασίας. Η διαδικτυακή εφαρμογή είναι προσβάσιμη, μέσω του συνδέσμου, <https://sentiment-analysis-thesis.herokuapp.com/>.

Η διαδικτυακή εφαρμογή χωρίζεται σε δύο βασικές λειτουργίες:

- Πραγματοποίηση Ανάλυσης Συναισθήματος και παρουσίασης αποτελεσμάτων για δεδομένα εισόδου που εισάγει ο χρήστης.
- Δυνατότητα εκτέλεσης του μοντέλου επεξεργασίας κειμένου και ανάλυσης φυσικής γλώσσας, για διαφορετικά σύνολα δεδομένων.

Select Operation

Options

Analyze a single Tweet

Analyze a single Tweet

Analyze stock related Tweet Datasets

Σχήμα 7.12: Στιγμιότυπο που παρουσιάζει τις επιλογές του χρήστη για εκτέλεση των δύο βασικών λειτουργιών του συστήματος

7.2.2.1. Ανάλυση συναισθήματος με δεδομένα εισόδου χρήστη

Ο χρήστης της διαδικτυακής εφαρμογής ανάλυσης συναισθήματος, έχει την δυνατότητα να εισάγει την ανάρτηση από το Twitter που επιθυμεί στο σχετικό πεδίο και να κάνει υποβολή, έτσι ώστε να πραγματοποιηθεί ανάλυση των δεδομένων εισόδου.

Μετά την εκτέλεση της ανάλυσης, παρουσιάζεται κατάλογος αποτελεσμάτων, ο οποίος περιλαμβάνει τα ακόλουθα:

- Παρουσίαση επεξεργασμένου κειμένου απαλλαγμένου από περιττό θόρυβο που πιθανώς να επηρεάσει αρνητικά τον αναλυτή συναισθήματος.
- Παρουσίαση μηνύματος που αναφέρει την υπολογισμένη πολικότητα του Tweet, δηλαδή αν κατηγοριοποιήθηκε ως θετικό, αρνητικό ή ουδέτερο με βάση το συναίσθημα.
- Παρουσίαση τιμής της μετρικής της υποκειμενικότητας.
- Παρουσίαση των τιμών του αναλυτή συναισθήματος Vader.
- Παρουσίαση γραφήματος που υποδεικνύει την βαθμολογία του αναλυτή συναισθήματος για κάθε λέξη του επεξεργασμένου κειμένου του Tweet.



Σχήμα 7.13: Στιγμιότυπο του περιβάλλοντος αλληλεπίδρασης του χρήστη για τη λειτουργία ανάλυσης συναισθήματος ενός Tweet με κενό πεδίο εισαγωγής δεδομένων



Σχήμα 7.14: Στιγμιότυπο από τα αποτελέσματα για την ανάλυση συναισθήματος της διαδικτυακής εφαρμογής

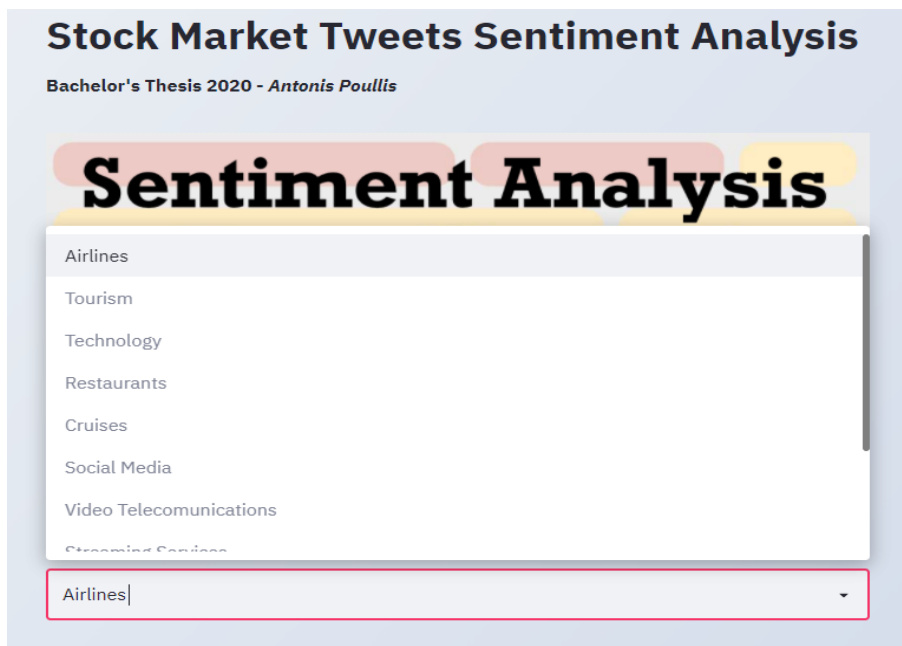
7.2.2.2.Ανάλυση συναισθήματος σε σύνολα δεδομένων

Υλοποιήθηκε επίσης λειτουργία στη διαδικτυακή εφαρμογή, η οποία επιτρέπει στον χρήστη να εκτελέσει το μοντέλο επεξεργασίας κειμένου και ανάλυσης φυσικής γλώσσας για όλα τα διαφορετικά σύνολα δεδομένων με αναρτήσεις του Twitter που συλλέχθηκαν κατά την φάση συλλογής με την Μέθοδο B.

Με την λειτουργία αυτή, παρέχεται η δυνατότητα επιλογής ανάμεσα σε 11 σύνολα δεδομένων που αφορούν διαφορετικούς τομείς της οικονομίας και εμφάνισης των δεδομένων τους. Σημειώνεται ότι κάνοντας hover πάνω σε κείμενο κελιού του οποίου το κείμενο δεν φαίνεται ολόκληρο, το κείμενο θα εμφανιστεί.

Μετά την εκτέλεση της ανάλυσης παρουσιάζεται λίστα αποτελεσμάτων που περιλαμβάνει τα ακόλουθα:

- Παρουσίαση επεξεργασμένου κειμένου απαλλαγμένου από περιττό θόρυβο που πιθανώς να επηρεάσει αρνητικά τον αναλυτή συναισθήματος.
- Παρουσίαση γραφήματος τύπου WordCloud που φανερώνει τη συχνότητα εμφάνισης λέξεων στο σύνολο δεδομένων, απεικονίζοντας λέξεις με τρόπο τέτοιο ώστε το μέγεθος γραμματοσειράς να αντιστοιχεί με τη σημαντικότητα τους.
- Παρουσίαση γραφήματος με μπάρες που απεικονίζει τον αριθμό των αναρτήσεων από κάθε κατηγορία, μετά την ανάλυση.
- Παρουσίαση των ποσοστών, θετικών, αρνητικών και ουδέτερων αναρτήσεων.
- Παρουσίαση γραφήματος με μπάρες που απεικονίζει τα δημοφιλέστερα hashtags των αναρτήσεων που κατηγοριοποιήθηκαν ως θετικές.
- Παρουσίαση γραφήματος με μπάρες που απεικονίζει τα δημοφιλέστερα hashtags των αναρτήσεων που κατηγοριοποιήθηκαν ως αρνητικές.
- Παρουσίαση γραφήματος που απεικονίζει τις τιμές της πολικότητας των Tweets, συναρτήσει της υποκειμενικότητας τους
- Παρουσίαση του συνόλου δεδομένων με την προσθήκη των ακόλουθων υπολογισμένων πεδίων:
 - Τιμές του αναλυτή συναισθήματος Vader.
 - Μετρική της υποκειμενικότητας.
 - Παρουσίαση μηνύματος που αναφέρει την υπολογισμένη πολικότητα του Tweet, δηλαδή αν κατηγοριοποιήθηκε ως θετικό, αρνητικό ή ουδέτερο με βάση το συναίσθημα.



Σχήμα 7.15: Στιγμιότυπο του περιβάλλοντος αλληλεπίδρασης χρήστη για την λειτουργία ανάλυσης συναισθήματος σε σύνολα δεδομένων με Tweets. Παρουσιάζονται ορισμένες από τις διαθέσιμες επιλογές

Social Media Dataset

NASDAQ: FB, NYSE:SNAP, NYSE: TWTR

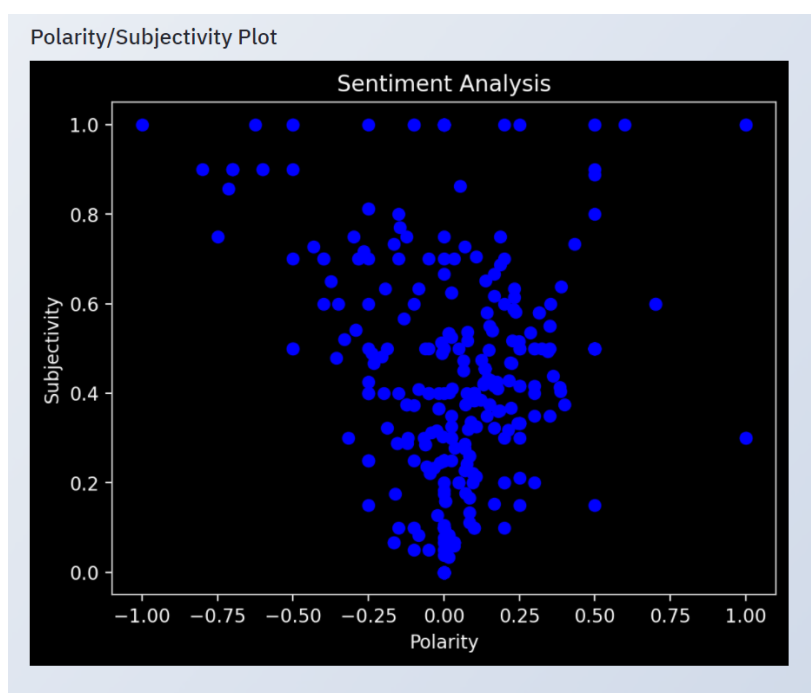
	Symbol	Text	Name	Date	Retweets
83	FB	How Cameo, Facebook An...	Benzinga	03/06/2020	0
84	FB	The combined market ca...	Bespoke	05/06/2020	34
85	FB	Facebook \$FB breakout ...	Top Breakout Stocks	05/06/2020	0
86	FB	#Facebook & the cr...	Holger Zschaepitz	07/06/2020	90
87	FB	RT @GunjanJS : Crazy t...	WSJ Markets	07/06/2020	0
88	FB	RT @IBD_ECarson : Dow ...	Investors.com	07/06/2020	0
89	FB	The Nasdaq composite l...	Investors.com	09/06/2020	1
90	FB	In case you missed it:...	Holger Zschaepitz	10/06/2020	387
91	FB	Facebook \$FB breakout ...	Top Breakout Stocks	10/06/2020	0
92	FB	Today's drop hardly no...	Bespoke	11/06/2020	10
93	FB	Apple and Microsoft ar...	NYT Business	13/06/2020	28

Apply Analysis

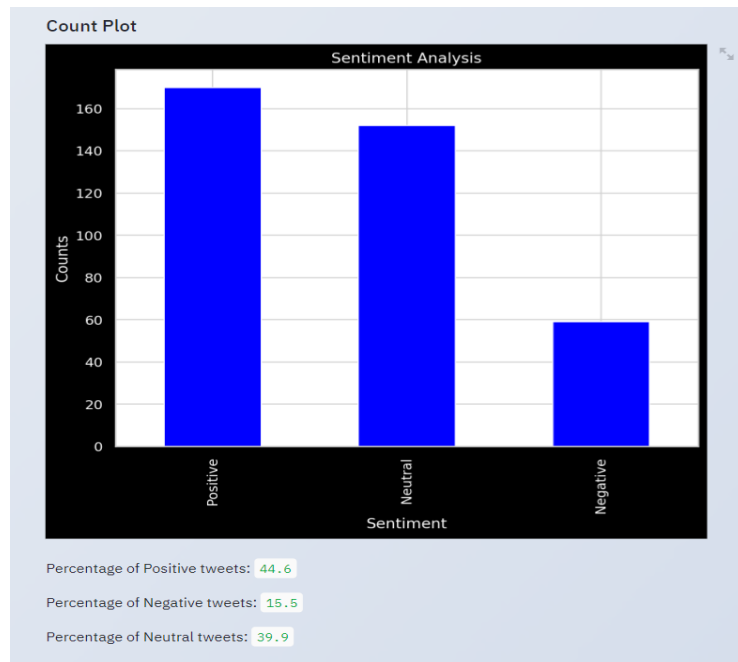
Σχήμα 7.16: Στιγμιότυπο συνοπτικής απεικόνισης του σύνολο δεδομένων που επιλέγεται από τον χρήστη πριν την υποβολή για ανάλυση συναισθήματος



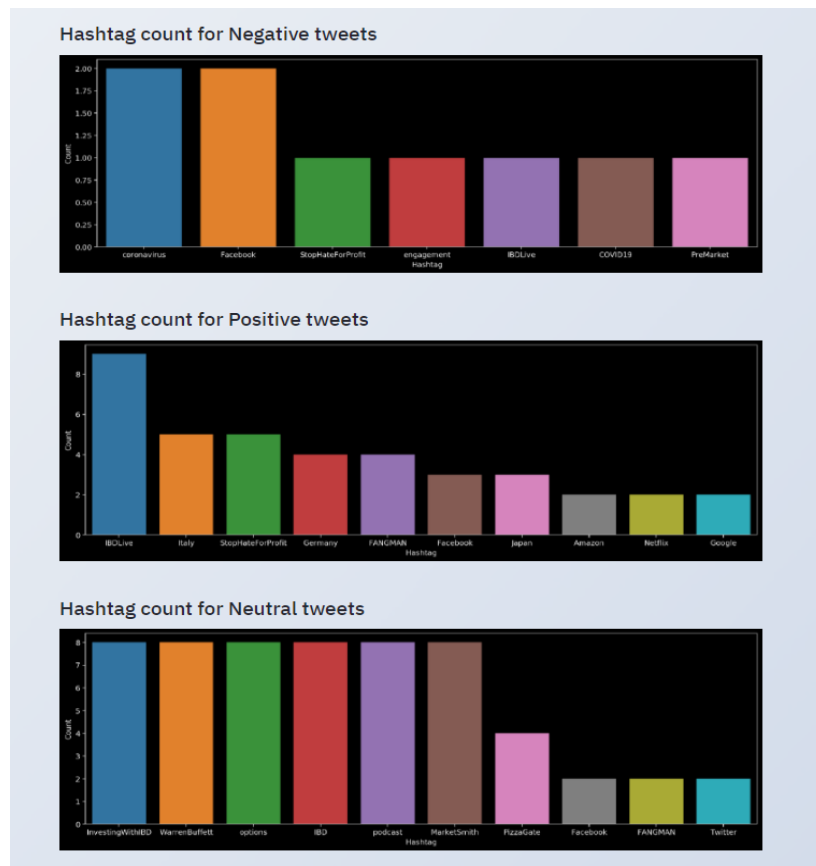
Σχήμα 7.17: Στιγμιότυπο αποτελεσμάτων από τη δημιουργία γραφήματος τύπου Word Cloud από τα εισερχόμενα δεδομένα του συνόλου δεδομένων.



Σχήμα 7.18: Στιγμιότυπο απεικόνισης γραφικής παράστασης της πολικότητας των συναισθημάτων συναρτήσει της μετρικής υποκειμενικότητας



Σχήμα 7.19: Στιγμιότυπο απεικόνισης γραφήματος με τα αποτελέσματα της κατηγοριοποίησης των Tweets σύμφωνα με την πολικότητά τους



Σχήμα 7.20: Στιγμιότυπο αποτελεσμάτων για τα δημοφιλέστερα Hashtags από το σύνολο δεδομένων

Κεφάλαιο 8

Περιορισμοί

8.1. Περιορισμοί κατά τη συλλογή δεδομένων	102
8.2. Περιορισμοί κατά την Προ-επεξεργασία Κειμένου	104
8.3. Περιορισμοί κατά την ανάλυση συναισθήματος	104
8.4. Περιορισμοί κατά την μετατροπή του συνόλου δεδομένων σε πρόβλημα μηχανικής μάθησης.....	105
8.5. Περιορισμοί αποτελεσμάτων με την χρησιμοποίηση GridSearchCV	106

8.1. Περιορισμοί κατά τη συλλογή δεδομένων

Η συλλογή Tweets, μέσω του δωρεάν επιπέδου του Twitter Streaming API που χρησιμοποιήθηκε σε αρχικό στάδιο της μελέτης, περιορίζει τη συλλογή ολόκληρου του σώματος της ανάρτησης, σε μόνο ένα μέρος της, πράγμα το οποίο συμβάλλει αρνητικά στην ανάλυση συναισθήματος και καθιστά το σύνολο δεδομένων μη βέλτιστο. Για το λόγο αυτό δεν αξιοποιήθηκαν στο τελικό σύστημα τα δεδομένα, τα οποία συλλέχθηκαν για την ανάλυση συναισθήματος με αυτή τη μέθοδο συλλογής.

Στο πλαίσιο της εργασίας πραγματοποιήθηκε συλλογή δεδομένων σχετικά με Tweets που αφορούν συγκεκριμένες εταιρείες, οι οποίες εναρμονίζονται με τις ανάγκες του συστήματος και το πεδίο ενδιαφέροντος. Ωστόσο, για πολλές από αυτές δεν υπήρχε αξιοσημείωτη ποσότητα δεδομένων κατά την περίοδο συλλογής τους, εφόσον δεν αποτελούν περιζήτητες και διαδομένες επιλογές μεταξύ των επενδυτών.

Επιπρόσθετα, οι επενδυτές τείνουν να δημοσιεύουν Tweets για αρκετές εταιρίες μόνο όταν προβλέπουν κάποια ενδεχόμενη αξιοσημείωτη μεταβολή βραχυπρόθεσμα, με αποτέλεσμα να λαμβάνεται αρκετή πληροφορία για ορισμένες ημέρες και ελάχιστη έως καθόλου πληροφορία για μεγάλα χρονικά διαστήματα.

Η επιλογή λογαριασμών από τους οποίους συλλέχθηκαν Tweets, υποβοήθησε στην ενίσχυση της ποιότητας των αναρτήσεων, οι οποίες κατείχαν τις ιδιότητες, να προέρχονται από αξιόπιστες πηγές και να είναι σχετικές με το πεδίο ενδιαφέροντος. Παρόλα αυτά, ως λογικό επακόλουθο των θετικών στοιχείων, προκύπτουν μειονεκτήματα. Συγκεκριμένα, τα σύνολα δεδομένων που προκύπτουν με την εν λόγω μέθοδο συλλογής είναι μικρότερα σε όγκο δεδομένων και για συγκεκριμένες εταιρίες δεν εντοπίζονται συνεχόμενα δεδομένα, καθώς όσο πιο στοχευμένες είναι οι αναζητήσεις στο Twitter, τόσο μικρότερος και ο αριθμός των αποτελεσμάτων. Επίσης, σημειώνεται ότι αυτή η διαδικασία συλλογής, είναι πιο χρονοβόρα και περίπλοκη κατά τον διαχωρισμό των Tweets σε μικρότερα σύνολα δεδομένων, τα οποία πρέπει να αφορούν συγκεκριμένες εταιρίες, στοιχεία ωστόσο που σε μεγάλα συστήματα είναι εφικτό να αυτοματοποιηθούν.

Όπως αναφέρθηκε στη αξιολόγηση αποτελεσμάτων για τη δημιουργία αξιόπιστων μοντέλων πρόβλεψης, απαιτείται μεγάλη ποσότητα δεδομένων, συνεχόμενα δεδομένα και ιδανικά, ισορροπημένος σε μέγεθος, όγκος δεδομένων για κάθε διαφορετική μέρα, στοιχεία τα οποία δεν χαρακτηρίζουν το σύνολο δεδομένων που συλλέχθηκε και γι' αυτούς τους λόγους η ανάπτυξη αξιόπιστων μοντέλων πρόβλεψης, τα οποία να βασίζονται σε αυτό παρουσιάζει μειονεκτήματα.

Σε αρχικό στάδιο, αξιοποιήθηκε η Μέθοδος Γ για τη συλλογή Tweets, με σκοπό την επέκταση του τελικού συνόλου δεδομένων, λαμβάνοντας μεγάλο αριθμό Tweets για τις τελευταίες 10 μέρες. Εντοπίστηκε όμως, ότι η παρουσία μεγάλου αριθμού δεδομένων για τις τελευταίες 10 μέρες υπερίσχυε σε ποσότητα, του αριθμού των συλλεχθέντων Tweets για όλη την χρονική περίοδο συλλογής, με την Μέθοδο Β. Επομένως, η κατανομή των δεδομένων ήταν μη ισορροπημένη σε ποσότητα για όλη την διάρκεια συλλογής και ως εκ τούτου τα αποτελέσματα με την συμπερίληψη αυτής της μεθόδου, αλλοίωναν το γενικό συναίσθημα της περιόδου, εστιάζοντας στο συναίσθημα που εντοπίστηκε τις τελευταίες 10 μέρες συλλογής δεδομένων.

Επιπρόσθετα, παρατηρήθηκε ότι η ποιότητα των συλλεχθέντων δεδομένων με την Μέθοδο Γ, διέφερε από εκείνην της Μεθόδου Β, καθώς τα Tweets από την Μέθοδο Γ,

προέρχονταν από οποιοδήποτε λογαριασμό, χωρίς τις εγγυήσεις ότι επρόκειτο για αξιόπιστες πηγές πληροφοριών και ότι το περιεχόμενο τους θα ήταν πάντοτε σχετικό με το πεδίο ενδιαφέροντος.

8.2. Περιορισμοί κατά την Προ-επεξεργασία Κειμένου

Η Προ-επεξεργασία Κειμένου αποτελεί ένα από τα σημαντικότερα στάδια σε συστήματα επεξεργασίας φυσικής γλώσσας και η εξαγόμενη πληροφορία τους μπορεί να διαφοροποιήσει σημαντικά το τελικό αποτέλεσμα. Περιορισμό κατά την εφαρμογή της προ-επεξεργασίας στο κείμενο των Tweets, αποτέλεσε η αλλοίωση του επεξεργασμένου κειμένου, που επέφερε η χρησιμοποίηση των τεχνικών Stemming και Lemmitization. Η αρνητική επίδραση στις αναρτήσεις, οδήγησε στην μη χρησιμοποίηση των δύο τεχνικών, κυρίως λόγω της αδυναμίας του αναλυτή συναισθήματος να αναγνωρίσει λέξεις παραγόμενες έπειτα από την επεξεργασία κειμένου τους.

Επομένως, το τελικό σύστημα δεν επωφελήθηκε των θετικών που επιφέρουν οι εν λόγω τεχνικές, καθώς προτιμήθηκε η αποφυγή της πιθανής αρνητικής τους επίδρασης σε μεγάλο μέρος του συνόλου δεδομένων.

8.3. Περιορισμοί κατά την ανάλυση συναισθήματος

Η ανάλυση συναισθήματος με οποιαδήποτε μέθοδο ή εργαλείο και να πραγματοποιείται δεν είναι πάντοτε άρτια και μπορεί να περιέχει αρκετές παγίδες.

Για παράδειγμα, στο κείμενο "Amazon crashed it! 🤖", ο χρήστης ανάρτησε ένα θετικό σχόλιο για την άνοδο της τιμής μετοχής της Amazon, αλλά το γενικό συναίσθημα του αναλυτή συναισθήματος είναι αρνητικό.

Το παράδειγμα αυτό επιβεβαιώνει το γεγονός ότι είναι πάντοτε δύσκολο, ακόμη και για κάποιον άνθρωπο, να ερμηνεύσει ποια ήταν η πραγματική πρόθεση ενός κειμένου, χωρίς κάποιο πρόσθετο πλαίσιο, πόσο μάλλον για τον αναλυτή συναισθήματος.

8.4. Περιορισμοί κατά την μετατροπή του συνόλου δεδομένων σε πρόβλημα μηχανικής μάθησης

Κατά τη μετατροπή του συνόλου δεδομένων σε πρόβλημα μηχανικής μάθησης, χρειάστηκε να ληφθούν υπόψιν συγκεκριμένοι περιορισμοί, έτσι ώστε η μεθοδολογία που ακολουθείται να συμβαδίζει με την πραγματικότητα του κόσμου των επενδύσεων. Συγκεκριμένα, εντοπίστηκε ότι τα Χρηματιστηριακά δεδομένα του συνόλου δεδομένων εμπεριέχουν γνώση, την οποία ο επενδυτής δεν κατέχει στις συγκεκριμένες χρονικές στιγμές. Για παράδειγμα, εάν είναι επιθυμητή η πρόβλεψη της τάσης μιας μετοχής για την ημερομηνία 14/04 και το μοντέλο πρόβλεψης έχει σαν χαρακτηριστικό εκπαίδευσης του την τιμή κλεισίματος της 14/04, το αποτέλεσμα της πρόβλεψης δίδεται εσφαλμένα ως είσοδος, μέσω της τιμής κλεισίματος και θα διενεργηθεί ορθή πρόβλεψη, η οποία βασίζεται σε εσφαλμένη μεθοδολογία. Επομένως, εάν εφαρμοστεί το εσφαλμένο μοντέλο πρόβλεψης σε πραγματικό χρόνο δεν θα είναι σε θέση να έχει δεδομένα εισόδου και θα αποτύχει.

Για το λόγο αυτό, στο μοντέλο πρόβλεψης της μελέτης αξιοποιήθηκαν για κάθε διαφορετική μέρα τα Χρηματιστηριακά δεδομένα της ημέρας που προηγήθηκε. Κατά συνέπεια, γίνεται αντιληπτό ότι η εν λόγω μεθοδολογία απαιτεί συνεχόμενα Χρηματιστηριακά δεδομένα και επομένως συνεχόμενα καθημερινά δεδομένα από τα κοινωνικά δίκτυα, για την ορθή εκπαίδευση των αλγορίθμων μηχανικής μάθησης, στοιχείο που αποτελεί περιορισμό κατά την ανάπτυξη παρόμοιων συστημάτων.

Αξίζει να σημειωθεί, ότι παρατηρήθηκε πως, ορισμένες μελέτες, οι οποίες αξιοποιούν το συγκεκριμένο σύνολο δεδομένων, χρησιμοποιούν εκτός από τους τίτλους ειδήσεων, δεδομένα από το Χρηματιστήριο (π.χ. Low, High, Close, Adj. Close) για την ίδια ημέρα πρόβλεψης, εκπαιδεύοντας το μοντέλο με δεδομένα, τα οποία ανήκουν στο μέλλον και το μοντέλο δε θα γνώριζε υπό κανονικές συνθήκες, με αποτέλεσμα να προβλέπουν το πεδίο Label (στόχο) με εσφαλμένη μεθοδολογία. Ως εκ τούτου, τα υψηλά επίπεδα ακρίβειας των μοντέλων αυτών, είναι πλασματικά και δεν απεικονίζουν την πραγματικότητα.

8.5. Περιορισμοί αποτελεσμάτων με την χρησιμοποίηση GridSearchCV

Πρέπει να ληφθεί υπόψιν ότι η υπόθεση πως οι “καλύτεροι” υπερπαραμέτροι που βρέθηκαν κατά την χρησιμοποίηση του εργαλείου GridSearchCV παράγουν και τα καλύτερα αποτελέσματα για το άγνωστο σύνολο δοκιμής (test dataset), δεν ισχύει, αφού δεν υπάρχει καμία απολύτως διασφάλιση αυτής της θεωρίας.

Η φιλοσοφία πίσω από την επιλογή υπερπαραμέτρων, είναι ότι επιλέγονται με τον καλύτερο δυνατό τρόπο, σύμφωνα με τις περιορισμένες πληροφορίες που έχει στη διάθεσή του το εργαλείο, τη στιγμή της προσαρμογής του μοντέλου. Πρόκειται λοιπόν, για μια περισσότερο λογική παρά εγγυημένα βέλτιστη αποδοτικά προσέγγιση επιλογής παραμέτρων, λόγω της φύσεως της αβεβαιότητας που προκύπτει από την ύπαρξη των αόρατων δεδομένων.

Αξίζει να σημειωθεί ότι εξ ορισμού τα αποτελέσματα που προκύπτουν από το cross validation δεν είναι μεροληπτικά μόνο σε ότι αφορά το σύνολο δεδομένων που χρησιμοποιείται, αλλά είναι μεροληπτικά και στον συγκεκριμένο διαχωρισμό των δεδομένων σε πτυχές εκπαίδευσης και επικύρωσης (training & validation folds). Επομένως, υπάρχει πάντοτε η πιθανότητα, χρησιμοποιώντας διαφορετικό διαχωρισμό κατά το στάδιο του cross validation με τα ίδια δεδομένα, να προκύψουν διαφορετικές «βέλτιστες τιμές» για τις ίδιες υπερπαραμέτρους, ειδικότερα όταν χρησιμοποιούνται ασταθής ταξινομητές (π.χ. Decision Tree).

Συνοψίζοντας, το cross validation είναι πράγματι μια χρήσιμη και ορθολογική προσέγγιση, ωστόσο η αναμονή οποιουδήποτε είδους μαθηματικής διασφάλισης, ότι τα αποτελέσματά που παράγονται για δεδομένα που δεν παρατηρούνται (test dataset) θα είναι βέλτιστα, είναι αβάσιμη.

Κεφάλαιο 9

Συμπεράσματα

9.1. Συμπεράσματα	107
9.2. Εισηγήσεις	109

9.1. Συμπεράσματα

Στην παρούσα διπλωματική εργασία εξετάστηκε η δυνατότητα πρόβλεψης της Χρηματιστηριακής συμπεριφοράς, αξιοποιώντας ως εργαλείο την ανάλυση συναισθήματος σε αναρτήσεις από μέσα κοινωνικής δικτύωσης. Για το σκοπό αυτό, συλλέχθηκαν και αναλύθηκαν δεδομένα από το κοινωνικό μέσο Twitter και αναπτύχθηκε μοντέλο πρόβλεψης της Χρηματιστηριακής Αγοράς, βασισμένο στην ανάλυση συναισθήματος σε κοινωνικά δίκτυα και στη Μηχανική Μάθηση.

Έπειτα από την πολύπλευρη ανάλυση των αποτελεσμάτων της παρούσας εργασίας, μπορεί να εξαχθεί το συμπέρασμα, ότι τα Tweets αποτελούν ωφέλιμο αρωγό ενημέρωσης, καθώς μέσω τους αναρτώνται καθημερινά χρήσιμες ειδήσεις και πληροφορίες για τον Χρηματιστηριακό κόσμο. Ωστόσο, ο καθένας καλείται να ερμηνεύσει με τον δικό του τρόπο, τέτοιες πληροφορίες και να προβεί σε επενδύσεις αναλογιζόμενος την δική του άποψη και αναλαμβάνοντας το δικό του προσωπικό ρίσκο.

Συμπεραίνεται επίσης, ότι η συχνότητα εμφάνισης λέξεων μπορεί να αποτελέσει ένα αρκετά χρήσιμο επενδυτικό βοηθό. Πιο συγκεκριμένα, είναι σε θέση να συνεισφέρει στον εντοπισμό σημαντικών γεγονότων που αφορούν εταιρείες για τις οποίες ενδιαφέρονται οι επενδυτές. Επιπλέον, μέσα από την ανάλυση των αποτελεσμάτων παρατηρήθηκε ότι ένα γράφημα τύπου Word Cloud, μπορεί να προσφέρει μια πολύ χρήσιμη σύνοψη για την περίοδο ενδιαφέροντος μιας μετοχής, εισάγει τον χρήστη

στη θεματολογία της επικαιρότητας και του εμφανίζει λέξεις κλειδιά, προκειμένου να πραγματοποιήσει στοχευμένες αναζητήσεις και να ενημερωθεί περαιτέρω, πριν προβεί σε επενδύσεις.

Αξίζει να σημειωθεί ότι λόγω της χρήσης της Μεθόδου Β που συλλέγει δεδομένα μόνο από συγκεκριμένους λογαριασμούς του πεδίου ενδιαφέροντος και έπειτα από το φιλτράρισμα και την κατηγοριοποίηση των Tweets με βάση τις εταιρείες και τους τομείς που εξετάστηκαν για όλα τα σύνολα δεδομένων, συλλέχθηκαν 100% σχετικά δεδομένα με τη θεματολογία του κάθε συνόλου δεδομένων. Λογικό επακόλουθο αυτού, σε συνδυασμό με την απόρριψη μεγάλου αριθμού συλλεγμένων αναρτήσεων, ήταν τα τελικά σύνολα δεδομένων να μην είναι αρκετά μεγάλα και ικανοποιητικά σε ποσότητα για να πραγματοποιηθούν περαιτέρω αναλύσεις και στατιστικά σημαντικές προβλέψεις για κάθε Χρηματιστηριακή μέρα.

Τα πειράματα που καλύπτονται στην παρούσα εργασία προτείνουν την αξιοποίηση διαφορετικών μοντέλων και την επιλογή των κατάλληλων παραμέτρων τους, έτσι ώστε να επιτευχθεί η υψηλότερη δυνατή βαθμολογία πρόβλεψης του Χρηματιστηρίου.

Ίσως το σημαντικότερο συμπέρασμα που εξάγεται αποτελεί ο καίριος ρόλος που διαδραματίζει το σύνολο των δεδομένων, όχι μόνο ως προς το μέγεθος, αλλά και ως προς την ποιότητα του. Παρατηρήθηκε ότι δεν απαιτείται μόνο μεγάλη ποσότητα δεδομένων, αλλά και συνεχόμενα δεδομένα από έγκυρες πηγές για κάθε μέρα, των οποίων ο θόρυβος επιθυμείται να είναι όσο το δυνατόν μειωμένος. Γίνεται λοιπόν αντιληπτό, ότι ο αντίκτυπος του συνόλου δεδομένων στην πρόβλεψη της τάσης των τιμών του Χρηματιστηρίου, είναι υψηλότερος από εκείνο των μοντέλων μηχανικής μάθησης, κάτι που υποδηλώνεται εν μέρη και από τις διαφορετικές τιμές αποτελεσμάτων για το μικρότερο σύνολο δεδομένων στην παρούσα εργασία.

Η αξιολόγηση των αποτελεσμάτων από τις μετρικές της ανάκλησης, ακρίβειας, ορθότητας και βαθμολογίας F1 για όλους τους αλγόριθμους μηχανικής μάθησης που συγκρίθηκαν, έδειξαν ότι ο αλγόριθμός Logistic Regression είχε την καλύτερη απόδοση με ποσοστό ορθότητας 66% για το μικρότερο σύνολο δεδομένων και 61%

για ολόκληρο το σύνολο δεδομένων, συγκριτικά με τους υπόλοιπους αλγορίθμους μηχανικής μάθησης που εξετάστηκαν στο πλαίσιο της εργασίας.

Τα ποσοστά ορθότητας των μοντέλων πρόβλεψης των μελετών στο πεδίο ενδιαφέροντος, δεν είναι συγκρίσιμα μεταξύ τους, καθώς δεν αναφέρονται στις ίδιες μετοχές, για τα ίδια χρονικά διαστήματα και δεν προβλέπουν την κίνηση των μετοχών πάντοτε από την ίδια σκοπιά (βλ. Κεφάλαιο 2). Παρόλα αυτά, τόσο οι εν λόγω μελέτες, όσο και η παρούσα μελέτη, με την μερική ανάλυση που πραγματοποιήθηκε σε σύνολο δεδομένων με τίτλους ειδήσεων από το κοινωνικό δίκτυο συζητήσεως Reddit, επιβεβαιώνουν, ότι η αξιοποίηση της μηχανικής μάθησης, σε συνδυασμό με την ανάλυση συναισθήματος σε μέσα κοινωνικής δικτύωσης και τις Χρηματιστηριακές τιμές, είναι σε θέση να βοηθήσει τους επενδυτές, αποτελώντας πολύτιμο εργαλείο στην πρόβλεψη της συμπεριφοράς της Χρηματιστηριακής Αγοράς.

9.2. Εισηγήσεις

Κατά τη διάρκεια, της μελέτης του πεδίου ενδιαφέροντος της διπλωματικής εργασίας και της ανάπτυξης του συστήματος, εντοπίστηκαν στοιχεία, τα οποία θεωρήθηκαν ενδιαφέρον να συμπεριληφθούν σε μελλοντικές μελέτες ή να βελτιώσουν υπάρχοντα συστήματα. Σε αυτό το πλαίσιο, αναφέρονται εν συντομία, ορισμένες εισηγήσεις.

Η πρώτη εισήγηση, αφορά την συλλογή σχετικότερων δεδομένων για τη δημιουργία ενός ποιοτικότερου τελικού συνόλου δεδομένων, που θα επιφέρει επιπλέον χρήσιμη γνώση στην επιστημονική κοινότητα. Προτείνεται ουσιαστικά, η αξιοποίηση μέσων κοινωνικής δικτύωσης, τα οποία να επικεντρώνονται μόνο σε επενδυτικά θέματα. Τα κοινωνικά δίκτυα, Public.com και Etoro αποτελούν παραδείγματα πλατφόρμων, μέσω των οποίων οι επενδυτές μπορούν να δημοσιεύσουν νέα, απόψεις και να σχολιάσουν στις δημοσιεύσεις, άλλων χρηστών. Επιπρόσθετα, είναι δυνατή η προβολή των επενδυτικών τακτικών και κερδών των χρηστών, στοιχείο που θα μπορούσε να αξιοποιηθεί για την συλλογή δεδομένων μόνο από επιτυχημένους στις προβλέψεις τους επενδυτές, για αποφυγή θορύβου και λανθασμένων απόψεων από υποσύνολο των χρηστών.

Όπως αναφέρθηκε στους περιορισμούς της μελέτης, είναι πιθανό να προκύψουν εσφαλμένες κατηγοριοποιήσεις από τον αναλυτή συναισθήματος, σε συστήματα που αξιολογούν κείμενα από κοινωνικά δίκτυα σχετικά με το Χρηματιστήριο. Για την άμβλυνση της παρουσίας προβλημάτων τέτοιας φύσης στο πεδίο ενδιαφέροντος, προτείνεται σε μελλοντικά συστήματα, η ενημέρωση του λεξικού του αναλυτή συναισθήματος με σύνολα δεδομένων που εμπεριέχουν ορολογίες και τις ανάλογες βαθμολογίες για την πολικότητα τους, οι οποίες να συμβαδίζουν με τον τρόπο που χρησιμοποιούνται οι εν λόγω φράσεις σε αναρτήσεις σχετικές με το Χρηματιστήριο από τους χρήστες των κοινωνικών δικτύων.

Όσον αφορά τη μεθοδολογία υλοποίησης του μοντέλου πρόβλεψης, λαμβάνοντας υπόψιν ότι οι τιμές των Χρηματιστηριακών μετοχών αποτελούν χρονοσειρές, μπορούν να υλοποιηθούν υβριδικά μοντέλα που συνδυάζουν την μεταβολή της τιμής κλεισίματος μετοχής για συγκεκριμένη χρονική περίοδο ως μια χρονοσειρά και την ανάλυση συναισθήματος σε αναρτήσεις από μέσα κοινωνικής για την ίδια περίοδο με συνεχόμενα δεδομένα. Παραδείγματα προσεγγίσεων που ενδείκνυνται για τέτοιας φύσης υβριδικά μοντέλα αποτελούν, η αρχιτεκτονική νευρωνικών δικτύων LSTM (Long short-term memory) και το μοντέλο ARIMA (Autoregressive integrated moving average). Στο Παράρτημα Β, παρατίθεται σύντομη εισαγωγή γύρω από την θεματολογία των χρονοσειρών, για κατανόηση του βασικού θεωρητικού υπόβαθρου από ενδιαφερόμενους.

Βιβλιογραφία

- [1] Bird S, Klein E, Loper E. Natural language processing with Python: analyzing text with the natural language toolkit. “ O’Reilly Media, Inc.”; 2009.
- [2] Bollen J, Mao H, Zeng X. Twitter mood predicts the stock market. J Comput Sci. 2011 Mar;2(1):1–8.
- [3] BroadbandSearch, The Most Visited Internet Sites of 2020 - BroadbandSearch [Internet]. [cited 2020 Nov 22. Available from: <https://www.broadbandsearch.net/blog/most-visited-popular-websites>
- [4] Bromberg A. Second Try: Sentiment Analysis in Python : Andy Bromberg [Internet]. [cited 2020 Nov 22]. Available from: https://github.com/abromberg/sentiment_analysis_python
- [5] Bruni R. Stock Market Index Data and indicators for Day Trading as a Binary Classification problem. Data Br. 2017 Feb 1;10:569–75.
- [6] Caliñgo AR, Sison AM, Tanguilig III BT. Prediction Model of the Stock Market Index Using Twitter Sentiment Analysis. Int J Inf Technol Comput Sci. 2016 Oct 8;8(10):11–21.
- [7] Chen H, De P, Hu Y, Hwang B-H, Faccio M, Kim S, et al. WISDOM OF CROWDS: THE VALUE OF STOCK OPINIONS TRANSMITTED THROUGH SOCIAL MEDIA [Internet]. [cited 2020 Nov 21]. Available from: <http://ssrn.com/abstract=1807265>
- [8] Ding T. Stock Market Prediction based on Time Series Data and Market Sentiment. 2012;
- [9] Elson SB, Yeung D, Roshan P, Bohandy SR, Nader A. Iranian Public Opinion About Specific Topics in the Aftermath of the 2009 Election. In: Using Social Media to Gauge Iranian Public Opinion and Mood After the 2009 Election. 2012.

- [10] Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. Nature [Internet]. 2020 Sep 17 [cited 2020 Nov 21];585(7825):357–62. Available from: <http://www.nature.com/articles/s41586-020-2649-2>
- [11] Hayes A. Stock Definition [Internet]. [cited 2020 Dec 24]. Available from: <https://www.investopedia.com/terms/s/stock.asp>
- [12] Hunter JD. Matplotlib: A 2D Graphics Environment. Comput Sci Eng [Internet]. 2007 [cited 2020 Nov 21];9(3):90–5. Available from: <http://ieeexplore.ieee.org/document/4160265/>
- [13] Hutto, C.J. & Gilbert E. (15) (PDF) VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text [Internet]. [cited 2020 Nov 21]. Available from: https://www.researchgate.net/publication/275828927_VADER_A_Parsimonious_Rule-based_Model_for_Sentiment_Analysis_of_Social_Media_Text
- [14] InfinityFree, Free and Unlimited Web Hosting with PHP and MySQL - InfinityFree [Internet]. [cited 2020 Dec 4]. Available from: <https://infinityfree.net/>
- [15] Investopedia, Adjusted Closing Price Definition [Internet]. [cited 2020 Dec 7]. Available from: https://www.investopedia.com/terms/a/adjusted_closing_price.asp
- [16] Investopedia, An Introduction to U.S. Stock Market Indexes [Internet]. [cited 2020 Nov 21]. Available from: <https://www.investopedia.com/insights/introduction-to-stock-market-indices/>
- [17] Kaggle, Daily News for Stock Market Prediction | Kaggle [Internet]. [cited 2020 Nov 22]. Available from: <https://www.kaggle.com/aaron7sun/stocknews>
- [18] Kordonis J, Symeonidis S, Arampatzis A. Stock price forecasting via sentiment analysis on Twitter. In: ACM International Conference Proceeding Series. Association for Computing Machinery; 2016.

- [19] Li Y, Pan Y. A Novel Ensemble Deep Learning Model for Stock Prediction Based on Stock Prices and News. arXiv [Internet]. 2020 Jul 23 [cited 2020 Dec 29]; Available from: <http://arxiv.org/abs/2007.12620>
- [20] Malkiel, Burton G, A random walk down Wall Street: including a life-cycle guide to personal investing. Choice Rev Online. 1996;
- [21] McKinney W. Data Structures for Statistical Computing in Python. In: Proceedings of the 9th Python in Science Conference. SciPy; 2010. p. 56–61.
- [22] Medvedev AN, Lambiotte R, Delvenne J-C. The anatomy of Reddit: An overview of academic research. Springer Proc Complex [Internet]. 2018 Oct 25 [cited 2020 Nov 22];183–204. Available from: <http://arxiv.org/abs/1810.10881>
- [23] Mittal A, Goel A. Stock Prediction Using Twitter Sentiment Analysis.
- [24] Nasdaq, Difference Between Dow, Nasdaq, and S&P 500: Major Facts & Opportunities | Nasdaq [Internet]. [cited 2020 Nov 23]. Available from: <https://www.nasdaq.com/articles/difference-between-dow-nasdaq-and-sp-500-major-facts-opportunities-2018-08-03>
- [25] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python . J Mach Learn Res. 2011;12:2825–30.
- [26] Rao T, Srivastava S. Analyzing Stock Market Movements Using Twitter Sentiment Analysis. Proc 2012 Int Conf Adv Soc Networks Anal Min. 2012;
- [27] Roesslein J. Tweepy: Twitter for Python! Available from: <https://github.com/tweepy/tweepy> . 2020;
- [28] Sheng Yu, A Survey of Prediction Using Social Media [Internet]. [cited 2020 Nov 21]. Available from: https://www.researchgate.net/publication/221672859_A_Survey_of_Prediction_Using_Social_Media

- [29] Sklearn, sklearn.model_selection.GridSearchCV — scikit-learn 0.23.2 documentation [Internet]. [cited 2020 Dec 4]. Available from: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html#sklearn.model_selection.GridSearchCV
- [30] TrackMyHashtag, Hashtag Tracking tool for Twitter - TrackMyHashtag [Internet]. [cited 2020 Dec 24]. Available from: <https://www.trackmyhashtag.com/>
- [31] TradingView, Free Stock Charts, Stock Quotes and Trade Ideas — TradingView [Internet]. [cited 2020 Dec 4]. Available from: <https://www.tradingview.com/>
- [32] Twitter, About [Internet]. [cited 2020 Nov 21]. Available from: https://about.twitter.com/en_gb.html
- [33] Van Rossum G. The Python Library Reference, release 3.8.2. Python Software Foundation; 2020.
- [34] Waskom M, the seaborn development team. mwaskom/seaborn [Internet]. Zenodo; 2020. Available from: <https://doi.org/10.5281/zenodo.592845>
- [35] WordCloud, wordcloud.WordCloud — wordcloud 1.8.1 documentation [Internet]. [cited 2020 Dec 24]. Available from: https://amueller.github.io/word_cloud/generated/wordcloud.WordCloud.html
- [36] Zhang X, Fuehres H, Gloor PA. Predicting Stock Market Indicators Through Twitter “I hope it is not as bad as I fear.” *Procedia - Soc Behav Sci*. 2011;26:55–62.

Παράρτημα Α

Παρατίθενται τα στατιστικά που συλλέχθηκαν χρησιμοποιώντας το εργαλείο TrackMyHashtag και αξιοποιήθηκαν για την δημιουργία γραφικών παραστάσεων.

Hashtag	Potential Impressions	Potential Reach
#stock	309637	158488
#WSJWhatsNow	74526956	19084367
#coronavirus	1943631	1943383
#Covid19	1042378	1041716
#IBDLive	139278	139238
#investing	881017	818564
#InvestingStrategies	231221	231221
#investors	482807	414998
#Premarket	307493	239868
#investor	200149	124825
#investingwithIBD	214473	214473
#stocks	1847202	1574378
#Healthcare	227133	224222
#Apple	291277	145252
#AAPL	329948	317209
#Amazon	728694	393647
#AMZN	457768	332335
#Google	551808	544131
#GOOG	363088	135380
#Facebook	684005	674030
#FB	897894	807747
#Microsoft	223998	173698
#MSFT	244174	218065
#Moderna	1472387	1471671
#PFIZER	822331	822248
#BIONTECH	2179368	2096301

Παράρτημα Β

Όπως αναφέρθηκε στο υποκεφάλαιο Μελλοντικές Μελέτες, μια διαφορετική προσέγγιση υλοποίησης μοντέλου πρόβλεψης αποτελεί η μελέτη χρονοσειρών. Πιο κάτω δίδεται μια σύντομη εισαγωγή στη θεματολογία των χρονοσειρών και στους αλγόριθμους που προτείνονται για μελέτη.

Ανάλυση με χρονοσειρές

Μια χρονοσειρά είναι ένα σύνολο παρατηρήσεων, οι οποίες συλλέγονται συνήθως σε ίσα διαστήματα. Στον άξονα Χ παρουσιάζεται ο χρονικός διαχωρισμός σε ίσα διαστήματα και στον άξονα τον Υ το προς εξέταση μέγεθος (magnitude). Κύριος σκοπός της ανάλυσης των δεδομένων μιας χρονοσειράς είναι εξαγωγή χρήσιμων στατιστικών και χαρακτηριστικών.

Τα μοντέλα πρόβλεψης χρονοσειρών είναι σε θέση να προβλέψουν μελλοντικές τιμές βάσει τιμών που παρατηρήθηκαν προηγουμένως.

Η πρόβλεψη χρονοσειρών εφαρμόζεται ευρέως σε μη στάσιμα δεδομένα, δηλαδή τα δεδομένα των οποίων οι στατιστικές ιδιότητες π.χ. η μέση και τυπική απόκλιση, δεν παραμένουν σταθερές στην πορεία του χρόνου, αλλά μεταβάλλονται.

Αυτά τα μη στατικά δεδομένα εισόδου ονομάζονται συνήθως χρονοσειρές. Μερικά παραδείγματα χρονοσειρών είναι οι μεταβαλλόμενες τιμές της θερμοκρασίας, μιας μετοχής ή ενός σπιτιού, δεδομένα τα οποία διαφοροποιούνται συνεχώς με την πάροδο του χρόνου.

Συνοπτικά, μια χρονοσειρά είναι μια ακολουθία παρατηρήσεων που λαμβάνονται διαδοχικά στο χρόνο.

Σημαντικότητα Χρονοσειρών

Μελετώντας τη σημαντικότητα της ανάλυσης με χρονοσειρές αξίζει να αναφερθούν ορισμένα βασικά στοιχεία.

Σημαντικότερο στοιχείο στις χρονοσειρές αποτελεί η επιχειρηματική πρόβλεψη (business forecasting), όπου βασίζεται στην ιδέα ότι το παρελθόν καθορίζει τι θα συμβεί στο μέλλον. Παραδείγματα επιχειρηματικών προβλέψεων αποτελούν, η αξιοποίηση χρονοσειρών από επαγγελματίες αναλυτές και επενδυτές, οι οποίοι προσπαθούν να προβλέψουν τις μελλοντικές τιμές μετοχών στο Χρηματιστήριο και η προσπάθεια από λιανοπωλητές να προβλέψουν τις επερχόμενες πωλήσεις τους.

Ακόμη, οι χρονοσειρές αποτελούν αρωγό για τη κατανόηση των παλαιών συμπεριφορών και δεδομένων και ως επίσης αποτελούν εργαλείο για τον προγραμματισμό του μέλλοντος, αξιολογώντας τα δεδομένα συναρτήσει του χρόνου.

Τέλος, αποτελούν σημαντικό εργαλείο για την αξιολόγηση τρεχόντων επιτευγμάτων που έχουν τεθεί παρελθοντικά.



Συστατικά Ανάλυσης Χρονοσειρών

Βασικό συστατικό των χρονοσειρών αποτελεί η τάση (trend), η οποία ορίζει προς ποια κατεύθυνση (αύξηση ή μείωση) τείνουν να κινηθούν τα δεδομένα για συγκεκριμένη χρονική περίοδο. Επομένως, εάν στην ανάλυση μιας χρονοσειράς παρουσιάζεται μια γενική τάση προς τα πάνω, το μοτίβο καλείται ως ανοδική τάση, ενώ στην αντίθετη περίπτωση ως πτωτική τάση και όταν δεν παρατηρείται κάποια διακύμανση το καλούμε οριζόντια τάση. Μια παρατηρήσιμη τάση, μπορεί να ισχύει για κάποιο χρονικό διάστημα και μετά εξαφανίζεται.

Συστατικό αποτελεί επίσης η εποχικότητα (seasonality), η οποία είναι ένα επαναλαμβανόμενο μοτίβο προς τα πάνω ή προς τα κάτω για μια καθορισμένη χρονική περίοδο. Παράδειγμα, η ανοδική πορεία πωλήσεων, κάθε χρονιά τον μήνα Δεκέμβριο λόγω των Χριστουγέννων, η παρατήρηση του οποίου προέρχεται από την ανάλυση της χρονοσειράς πωλήσεων των προηγούμενων χρόνων.

Ακόμη ένα συστατικό αποτελούν οι παρατυπίες ή θόρυβος (irregularity or noise), που ορίζονται ως απρόσμενες μετακινήσεις μικρής διάρκειας οι οποίες δεν επαναλαμβάνονται. Οι παρατυπίες εμφανίζονται τυχαία και ακανόνιστα στη πάροδο του χρόνου και είναι αδύνατον να προβλεφθούν.

Συστατικό αποτελεί επίσης η κυκλικότητα (cyclic), η οποία ορίζεται ως επαναλαμβανόμενη διακύμανση προς τα πάνω ή προς τα κάτω, χωρίς να παρατηρείται κάποιο σταθερό μοτίβο. Επομένως, είναι δυνατόν να εμφανιστεί οποιαδήποτε στιγμή και είναι αρκετά δύσκολο να προβλεφθεί.

